

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

SIDNEY COMANDULLI

**MÉTODO BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA A AVALIAÇÃO E
MELHORIA DA QUALIDADE DO PRODUCT BACKLOG EM AMBIENTES ÁGEIS
EM ESCALA: UMA ABORDAGEM BASEADA EM DESIGN SCIENCE RESEARCH**

CURITIBA

2026

SIDNEY COMANDULLI   

**MÉTODO BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA A AVALIAÇÃO E
MELHORIA DA QUALIDADE DO PRODUCT BACKLOG EM AMBIENTES ÁGEIS
EM ESCALA: UMA ABORDAGEM BASEADA EM DESIGN SCIENCE RESEARCH**

**Evaluation of Product Backlog Quality in Large-Scale Agile Environments:
A Design Science Research and Artificial Intelligence–Based Approach**

Dissertação apresentada como requisito para obtenção do título de Mestre em Engenharia de Produção e Sistemas da Universidade Tecnológica Federal do Paraná (UTFPR). Área de Concentração: Nome da Área.

Orientador:

► Prof. Dr. Sergio Eduardo Gouvea da Costa   

Coorientador:

► Prof. Dr. Edson Pinheiro de Lima   

CURITIBA

2026



4.0 Internacional

Esta licença permite apenas que outros façam download dos trabalhos licenciados e os compartilhem desde que atribuam crédito ao autor, mas sem que possam alterá-los de nenhuma forma ou utilizá-los para fins comerciais. Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.



SIDNEY COMANDULLI

**MÉTODO BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA A AVALIAÇÃO E MELHORIA DA QUALIDADE DO PRODUCT
BACKLOG EM AMBIENTES ÁGEIS EM ESCALA: UMA ABORDAGEM BASEADA EM DESIGN SCIENCE RESEARCH**

Trabalho de pesquisa de mestrado apresentado como requisito para
obtenção do título de Mestre Em Engenharia De Produção E Sistemas
da Universidade Tecnológica Federal do Paraná (UTFPR). Área de
concentração: Gestão Dos Sistemas Produtivos.

Data de aprovação: 15 de Julho de 2026

Dr. Sergio Eduardo Gouvea Da Costa, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Daniel Pacheco Lacerda, Doutorado - Universidade Federal de Santa Catarina (Ufsc)

Dr. Jose Marcelo Almeida Prado Cestari, Doutorado - Universidade Federal do Paraná (Ufpr)

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 15/07/2026.

AGRADECIMENTOS

A Deus, pelo dom da vida, pela saúde, pela proteção ao longo de minha caminhada e por me permitir compartilhar esta trajetória com uma família tão especial e com amigos que tornam meus dias mais significativos.

Ao meu pai, Orlando Comandulli (in memoriam), cuja presença permanece viva em meus valores, em minhas escolhas e em tudo aquilo que procuro construir. À minha mãe, Maria de Lourdes, pelo amor incondicional, pelos ensinamentos e pelo exemplo de força e dedicação que sempre orientaram minha vida. A eles, minha eterna gratidão por terem formado os alicerces de quem sou.

À minha esposa, Márcia Elisa, companheira de uma vida, pelo amor, pela compreensão e pela paciência ao longo desta caminhada. Agradeço por tudo o que construímos e vivemos juntos, por estar sempre ao meu lado em cada desafio e pela força que me transmite diariamente, tornando mais leves os momentos difíceis e mais significativas as conquistas. Obrigado por celebrar comigo cada vitória e, sobretudo, por nossa maior realização: nossos filhos, Gabriel e Leonardo, fontes permanentes de orgulho, inspiração e esperança.

Aos professores doutores Sérgio Gouveia e Edson Pinheiro, da Universidade Tecnológica Federal do Paraná, por terem aceitado o desafio de me orientar, pela confiança depositada em meu trabalho e pela generosidade com que compartilharam seus conhecimentos. Nossa sintonia, a liberdade para a troca de ideias e as reflexões construídas ao longo desta jornada foram fundamentais para o desenvolvimento e o êxito desta pesquisa.

A todos que, de alguma maneira, estiveram presentes nesta caminhada, oferecendo apoio, conhecimento, incentivo ou uma palavra amiga, registro minha mais sincera gratidão. Esta conquista também pertence a vocês.

RESUMO

A gestão da qualidade do *Product Backlog* constitui um desafio recorrente em ambientes ágeis em escala, nos quais múltiplas equipes trabalham simultaneamente no desenvolvimento de produtos complexos. Sob a perspectiva da Engenharia de Produção, o backlog pode ser compreendido como o principal insumo informacional do processo de desenvolvimento de software, desempenhando papel análogo ao da matéria-prima em sistemas produtivos. Quando esse insumo apresenta ambiguidades, inconsistências ou baixa qualidade informacional, o processo de desenvolvimento tende a apresentar retrabalho, aumento da variabilidade e perda de previsibilidade nas entregas. Apesar de sua importância, a avaliação do backlog costuma ocorrer de forma subjetiva, pouco padronizada e com baixa escalabilidade, o que compromete a maturidade dos processos de planejamento e a governança do desenvolvimento de software em ambientes organizacionais complexos.

Diante desse contexto, esta dissertação tem como objetivo propor, implementar e avaliar um método estruturado para a avaliação e melhoria da qualidade do *Product Backlog* apoiado por Inteligência Artificial. A pesquisa adota a abordagem de *Design Science Research* (DSR), envolvendo a identificação do problema, a concepção do artefato, sua implementação e a avaliação em contexto organizacional real.

O método proposto estrutura-se como um processo integrado composto por três componentes complementares: (i) avaliação sistemática da qualidade dos itens do backlog, (ii) apoio à escrita e melhoria contínua das descrições de requisitos, e (iii) monitoramento contínuo do processo de planejamento ágil. A avaliação é fundamentada em quatro dimensões centrais da engenharia de requisitos em ambientes ágeis: descrição da história de usuário, critérios de aceitação, *Definition of Ready* e *Definition of Done*. Para cada dimensão são definidos critérios explícitos e uma escala ordinal de mensuração, permitindo transformar avaliações qualitativas em indicadores estruturados de qualidade.

A operacionalização do método é realizada por meio de um artefato computacional baseado em modelos de *Large Language Models* (LLMs¹), capaz de analisar automaticamente artefatos textuais do backlog, gerar diagnósticos estruturados e fornecer recomendações de melhoria. O método também incorpora ciclos periódicos de avaliação coletiva e mecanismos de gestão à vista, com o objetivo de promover aprendizado organizacional e estimular a evolução contínua da qualidade dos requisitos.

A aplicação do método em um ambiente organizacional que adota metodologias ágeis em escala demonstrou sua viabilidade técnica e operacional. Os resultados indicam que a solução contribui para reduzir a subjetividade das avaliações, ampliar a escalabilidade do processo de análise do backlog e apoiar a identificação sistemática de fragilidades recorrentes nos requisitos. Dessa forma, o artefato proposto contribui para o fortalecimento da governança do planejamento ágil e para o aprimoramento da qualidade informacional que sustenta o desenvolvimento de software em larga escala, reforçando princípios clássicos da Engenharia de Produção relacionados à qualidade na origem, padronização de processos e melhoria contínua.

¹ LLMs - do inglês *Large Language Models* (Modelos de Linguagem de Grande Porte) são modelos de inteligência artificial baseados em redes neurais profundas, construídos tipicamente sobre a arquitetura Transformer (Alawaji; Hakami; Alshemaimri, 2024).

Palavras-chave: Product Backlog; Métodos Ágeis; Design Science Research; Inteligência Artificial; Qualidade de Requisitos.

ABSTRACT

Managing the quality of the *Product Backlog* represents a recurring challenge in scaled agile environments, where multiple teams simultaneously develop complex and continuously evolving products. From a Production Engineering perspective, the backlog can be understood as the primary informational input of the software development process, playing a role analogous to that of raw materials in traditional production systems. When this input contains ambiguities, inconsistencies, or low informational quality, the development process tends to exhibit rework, increased variability, and reduced delivery predictability. Despite its importance, backlog evaluation is often conducted in a subjective manner, with limited standardization and low scalability, which compromises the maturity of planning processes and the governance of software development in complex organizational environments.

In this context, this dissertation aims to propose, implement, and evaluate a structured method for assessing and improving the quality of the *Product Backlog* supported by Artificial Intelligence. The research adopts the *Design Science Research* (DSR) approach, involving problem identification, artifact design, implementation, and evaluation within a real organizational context.

The proposed method is structured as an integrated process composed of three complementary components: (i) systematic evaluation of backlog item quality, (ii) support for writing and continuously improving requirement descriptions, and (iii) continuous monitoring of the agile planning process. The evaluation framework is based on four central dimensions of agile requirements engineering: user story description, acceptance criteria, *Definition of Ready*, and *Definition of Done*. For each dimension, explicit evaluation criteria and an ordinal measurement scale are defined, enabling qualitative assessments to be transformed into structured quality indicators.

The method is operationalized through a computational artifact based on LLM capable of automatically analyzing textual backlog artifacts, generating structured diagnostics, and providing improvement recommendations. The method also incorporates periodic collective evaluation cycles and visual management mechanisms aimed at promoting organizational learning and stimulating the continuous improvement of requirement quality.

The application of the method in an organizational environment adopting scaled agile methodologies demonstrated its technical and operational feasibility. The results indicate that the solution contributes to reducing evaluation subjectivity, increasing the scalability of backlog analysis, and supporting the systematic identification of recurring weaknesses in requirements. Therefore, the proposed artifact contributes to strengthening agile planning governance and improving the informational quality that supports large-scale software development, reinforcing classical Production Engineering principles related to quality at the source, process standardization, and continuous improvement.

Keywords: Product Backlog; Agile Methods; Design Science Research; Artificial Intelligence; Requirements Quality.

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação - Lei de Little	24
Figura 2 – Modelo dos Três Cs das histórias de usuário	28
Figura 3 – Ciclo de trabalho no Scrum	30
Figura 4 – Ciclo conceitual de avaliação do backlog e execução do desenvolvimento ágil	31
Figura 5 – Estrutura simplificada de coordenação em ambientes de agilidade em escala	32
Figura 6 – Uso estratégico de marcadores e legendas - Manter foco nas informações relevantes (não poluir)	35
Figura 7 – Elementos centrais da arquitetura Data Warehouse (DW)/Business intelligence (BI) de Kimball	36
Figura 8 – Fluxo do processo de revisão da literatura com destaque para as etapas automatizadas pelo programa Python.	43
Figura 9 – Fluxograma de seleção de artigos Eixo 1	47
Figura 10 – Evolução anual das publicações de artigos Eixo 1	48
Figura 11 – Nuvem de palavra Eixo 1	48
Figura 12 – Países publicações - Eixo 1	48
Figura 13 – Fluxograma de seleção de artigos Eixo 2	51
Figura 14 – Evolução anual das publicações de artigos Eixo 2	52
Figura 15 – Nuvem de palavra Eixo 2	52
Figura 16 – Países publicações - Eixo 2	52
Figura 17 – Estrutura conceitual da classe de problema	57
Figura 18 – Encadeamento metodológico da pesquisa à luz da Design Science Research.	58
Figura 19 – Ciclo da Design Science Research adotado na pesquisa	62
Figura 20 – Fluxo metodológico da pesquisa à luz da Design Science Research	64
Figura 21 – Modelo conceitual do artefato Backlog Analyzer	71
Figura 22 – Fluxo Iterativo do Ciclo de Melhoria do Backlog.	85
Figura 23 – Diagrama esquemático (simplificado) Backlog Analyzer	86
Figura 24 – Diagrama esquemático (simplificado) Agente	87
Figura 25 – Diagrama esquemático (simplificado) Feedback Analyzer	88
Figura 26 – Diagrama esquemático do Método de	90
Figura 27 – Diagrama esquemático do Backlog Analyzer	92
Figura 28 – Diagrama esquemático do Prompt Booster	96
Figura 29 – Diagrama esquemático do Prompt Booster	97
Figura 30 – Fluxograma resumido do artefato	99
Figura 31 – Fluxograma das atividades de planejamento e utilização do artefato	99
Figura 32 – Análise longitudinal de tempos de processamento	103
Figura 33 – Avaliação inicial da qualidade dos itens de backlog antes da aplicação do método. Data: 25/08/2025, times analisados: 16, times com nota acima da nota 3: 1, média das notas: 0,98	108
Figura 34 – Avaliação inicial da qualidade dos itens de backlog após a aplicação do método. Data: 18/03/2026, times analisados: 15, times com nota acima da nota 3: 11, média das notas: 3,85	109

Figura 35 – Distribuição das avaliações de qualidade do backlog. Início do processo e após implantação.	111
Figura 36 – Evolução das pontuações médias de qualidade dos itens do <i>Product Backlog</i> ao longo do período de aplicação do método	112
Figura 37 – Distribuição das notas semanais do backlog por <i>Program Increment</i> (boxplot)	114
Figura 38 – Evolução semanal da nota média do backlog e volume de cards ao longo dos PI-21, PI-22 e PI-23	116
Figura 39 – Evolução de % <i>Allocation</i> e % <i>Concluded</i> por <i>Program Increment</i>	118
Figura 40 – Comparação das médias das avaliações Likert por <i>sprint</i> (2025 vs. 2026)	120
Figura 41 – Distribuição geral dos comentários por categoria de sentimento	121
Figura 42 – Temas recorrentes nos comentários: comparação entre menções negativas e positivas	122

LISTA DE TABELAS

Tabela 1 – Comparação entre sistemas industriais e sistemas digitais sob a ótica da Engenharia de Produção	26
Tabela 2 – Estratégia de busca utilizada na revisão da literatura	44
Tabela 3 – Principais estudos que discutem desafios associados ao <i>Product Backlog</i>	50
Tabela 4 – Estudos recentes sobre modelos de linguagem na análise de requisitos: Perspectiva Técnica	53
Tabela 5 – Correspondência entre as etapas da Design Science Research e sua aplicação na dissertação	63
Tabela 6 – Mapeamento de Requisitos e Constructos	69
Tabela 7 – Rastreabilidade entre requisitos de projeto e componentes do artefato	79
Tabela 8 – Critérios estruturados, indicadores observáveis e classificação INVEST para avaliação da qualidade do backlog	81
Tabela 9 – Escala de maturidade utilizada para avaliação dos itens do backlog	82
Tabela 10 – Configuração dos hiperparâmetros utilizados no módulo Backlog Analyzer	94
Tabela 11 – Comparação da distribuição das avaliações de qualidade do backlog	111
Tabela 12 – Evolução dos indicadores de qualidade do <i>Product Backlog</i> por <i>Program Increment</i>	113
Tabela 13 – Padrão de queda da nota ao final de cada <i>Program Increment</i> .	113
Tabela 14 – Evolução integrada dos indicadores de qualidade do backlog e eficiência do planejamento por PI	117
Tabela 15 – Médias das avaliações Likert por mês e dimensão (histórias de usuário)	119
Tabela 16 – Síntese dos principais achados da pesquisa	125

SUMÁRIO

1	INTRODUÇÃO	12
1.1	Problema da Pesquisa	16
1.2	Questão de Pesquisa	17
1.3	Objetivo Geral	17
1.3.1	Objetivos Específicos	17
1.4	Contribuições da Pesquisa	18
1.4.1	Contribuições Científicas	19
1.4.2	Contribuições para a Prática Organizacional	19
1.5	Justificativa e Relevância da Pesquisa	20
1.6	Delimitação da pesquisa	21
1.7	Estrutura do trabalho	22
2	FUNDAMENTAÇÃO TEÓRICA	23
2.1	Engenharia de Produção, Gestão de Operações e Fluxo de Trabalho	23
2.2	Métodos Ágeis e Agilidade em Escala	27
2.2.1	Fundamentos da Agilidade	27
2.2.2	Histórias de Usuário e Engenharia de Requisitos Ágil	27
2.2.3	Estrutura Operacional do Scrum	29
2.2.4	Escalabilidade Organizacional e SAFe	31
2.2.5	Qualidade do Backlog em Ambientes Ágeis em Escala	32
2.3	Gestão à Vista, Dashboards e ETL	33
2.3.1	Gestão à Vista como Instrumento de Alinhamento e Tomada de Decisão	33
2.3.2	Dashboards como Evolução Digital da Gestão à Vista	34
2.4	Inteligência Artificial e Modelos de Linguagem (LLMs)	37
2.4.1	Conceitos básicos sobre IA generativa e LLMs	37
2.4.2	Limites e desafios no uso de IA	39
2.4.3	Aplicações da IA no desenvolvimento de software e no planejamento de backlog	40
2.5	Revisão da Literatura	40
2.5.1	Materiais e Métodos	41
2.5.2	Extração e Consolidação Inicial	42
2.5.3	Normalização dos Metadados	45
2.5.4	Tradução Automática e Síntese dos Resumos	45
2.5.5	Resultados da Análise Bibliométrica	46
2.5.6	Análise Bibliométrica do Eixo 1: Problemas Associados ao Product Backlog	47
2.5.7	Análise Bibliométrica do Eixo 2: Uso de LLMs na Análise de Requisitos	51
2.5.8	Síntese da Revisão, Lacuna de Pesquisa e Posicionamento deste Trabalho	54
3	METODOLOGIA	58
3.1	Caracterização da Pesquisa	58
3.2	Design Science Research como Estratégia Metodológica	60
3.3	Classe de Problema e Constructos da Pesquisa	63
3.3.1	Classe de problema	64
3.3.2	Constructos Fundamentais da Pesquisa	66
3.4	Derivação dos Requisitos do Artefato	67

3.5	Modelo Conceitual do Artefato	69
3.6	Estratégia de Demonstração e Avaliação do Artefato	72
4	DESENVOLVIMENTO DO ESTUDO	75
4.1	Contextualização e Motivação	76
4.2	Concepção do Artefato: Definição do Método de Avaliação da Qualidade do Backlog	76
4.3	Arquitetura do Sistema e Funcionamento do Artefato	83
4.4	Implementação do Artefato e Aplicação no Contexto Organizacio- nal	89
4.4.1	Heurísticas Construtivas do Método	90
4.4.2	Backlog Analyzer	92
4.4.3	Backlog Booster	95
4.4.4	Feedback Analyzer	96
4.4.5	Integração do artefato na rotina	99
4.4.6	Evolução e Processamento Incremental do Artefato	100
4.4.7	Análise Preliminar de Custos Operacionais	103
4.5	Síntese do Capítulo	104
5	RESULTADOS E DISCUSSÃO	106
5.1	Premissas e Delimitações da Avaliação da Qualidade do Backlog	107
5.2	Resultados da Aplicação do Artefato	108
5.3	Análise Longitudinal por <i>Program Increment</i>	115
5.4	Relação entre Qualidade do Backlog e Eficiência do Planejamento das Sprints	116
5.5	Percepção dos Desenvolvedores: Avaliação Subjetiva e Repre- sentatividade Amostral	119
5.6	Análise de Sentimento dos Comentários dos Desenvolvedores	120
5.7	Análise dos Resultados à luz dos GAPs Identificados	122
5.8	Discussão dos Resultados e Implicações Práticas	125
5.9	Síntese do Capítulo	131
6	CONCLUSÕES E TRABALHOS FUTUROS	133
6.1	Conclusões da Pesquisa	133
6.2	Limitações do Estudo	135
6.3	Trabalhos Futuros	139
	REFERÊNCIAS	143
	Glossário	152

1 INTRODUÇÃO

Nas últimas décadas, a intensificação da transformação digital tem provocado mudanças profundas na forma como as organizações planejam, executam e controlam seus projetos, especialmente no desenvolvimento de produtos e serviços intensivos em tecnologia. Conforme discutido por [Highsmith \(2004\)](#), a adoção de abordagens ágeis surge como resposta à necessidade de operar em ambientes caracterizados por elevada complexidade e incerteza organizacional, nos quais a inovação depende da capacidade de adaptação contínua dos processos de desenvolvimento. Diante disso, a transformação digital configura-se como um fenômeno organizacional abrangente que exige a integração estratégica de tecnologias digitais, impactando diretamente a criação de valor e a eficiência operacional ([Küpper et al., 2019](#)). Sob essa perspectiva, os métodos ágeis consolidaram-se como abordagens amplamente utilizadas ao priorizarem ciclos curtos de entrega, adaptação contínua e colaboração entre equipes multidisciplinares. De acordo com [Schwaber e Sutherland \(2020\)](#), estruturas como o Scrum são particularmente adequadas para contextos de trabalho complexos, nos quais os requisitos evoluem ao longo do processo de desenvolvimento. Além disso, a adoção desses métodos tem demonstrado vantagens em relação a abordagens tradicionais, ao reduzir desperdícios e melhorar a capacidade de resposta a requisitos voláteis ([Petersen; Wohlin, 2010](#)).

A ampla difusão dos métodos ágeis não elimina a necessidade de rigor técnico no planejamento do trabalho, uma vez que a efetividade dessas abordagens depende diretamente da qualidade dos artefatos utilizados para organizar e comunicar os requisitos do produto. Nessa direção, [Sedano, Ralph e Péraire \(2019\)](#) argumentam que o *Product Backlog* assume papel central ao atuar como o principal elo entre a estratégia organizacional e a execução operacional das equipes de desenvolvimento. Esse artefato, usualmente composto por histórias de usuário, critérios de aceitação e definições de pronto (DoR) e feito (DoD), orienta a priorização das demandas e a validação das entregas ao longo dos ciclos iterativos. Dessa forma, a qualidade e a organização do backlog tornam-se elementos críticos para a previsibilidade do trabalho, especialmente em contextos de desenvolvimento ágil em escala, nos quais o alinhamento entre múltiplas equipes é fundamental ([Scaled Agile, Inc., 2025b](#)). Conforme observado por [Schwaber e Sutherland \(2020\)](#), o backlog deve constituir uma lista ordenada e evolutiva

de tudo o que é necessário para o produto, funcionando como a principal fonte de requisitos para as equipes. Assim, a manutenção da qualidade desses itens de trabalho torna-se condição essencial para preservar a integridade do escopo e o sucesso das implementações em ambientes complexos (Amjad et al., 2017).

Entretanto, mesmo em organizações maduras na adoção de práticas ágeis, persiste um desafio recorrente relacionado à variabilidade na qualidade das descrições que compõem o *Product Backlog*. Conforme observado por Müter et al. (2019), a formulação de itens de trabalho frequentemente apresenta baixa qualidade devido à complexidade inerente à decomposição de requisitos e ao uso inadequado de templates padronizados. Como consequência, descrições ambíguas ou incompletas tendem a gerar interpretações divergentes entre os membros da equipe, comprometendo a clareza do escopo e dificultando o planejamento das atividades de desenvolvimento. Além disso, a alta volatilidade dos requisitos em ambientes ágeis pode intensificar esses problemas quando os itens de backlog não apresentam nível adequado de detalhamento. Nesses casos, mudanças frequentes no escopo reduzem a previsibilidade das entregas e frequentemente resultam em retrabalho e atrasos no cronograma (Alsalemi; Yeoh, 2016). Agravando esse cenário, Grapenthin et al. (2015) destacam que descrições superficiais dificultam a compreensão das implicações de negócio associadas aos itens do backlog, o que impede a identificação precoce de dependências e tarefas necessárias para a implementação das funcionalidades. Como resultado, parte do escopo é descoberta apenas durante a execução da sprint, comprometendo estimativas de progresso e aumentando os riscos do projeto. Adicionalmente, problemas estruturais na preparação dos itens, como critérios de aceitação incompletos e a ausência de definições formais de *Definition of Ready* (DoR) e *Definition of Done* (DoD), dificultam a avaliação objetiva da prontidão de um item para desenvolvimento, favorecendo a propagação de defeitos para etapas posteriores do processo (Jakobsen; Sutherland, 2009).

A partir disso, a avaliação da qualidade dos itens que compõem o *Product Backlog* permanece, em grande medida, dependente do julgamento subjetivo de especialistas, como Product Owners e Agile Masters. Embora a experiência desses profissionais seja fundamental para o planejamento ágil, a ausência de métodos formais e sistemáticos para avaliar a clareza, a completude e a consistência das descrições de backlog pode comprometer a eficiência do processo de planejamento (Rola; Kuchta,

2019). De acordo com Boerman et al. (2015), a carência de modelos de medição baseados em métricas objetivas limita a transparência do processo e reduz a percepção de controle por parte dos stakeholders envolvidos. Diante desse quadro, avaliações excessivamente baseadas em intuição tendem a introduzir vieses individuais que comprometem a confiabilidade dos diagnósticos organizacionais. Tal constatação reforça a importância do uso de indicadores estruturados e evidências quantitativas (fatos em vez de percepções) nas inspeções e retrospectivas conduzidas pelas equipes ágeis, de modo a aumentar a objetividade dos processos de melhoria contínua (Erdoğan; Pekaya; Gök, 2018).

Paralelamente a esses desafios, avanços recentes em Inteligência Artificial, especialmente no campo dos LLM, têm ampliado as possibilidades de análise automatizada de textos em diferentes contextos organizacionais. Conforme apontam Oswal, Kanakia e Suktel (2024), esses modelos apresentam elevada capacidade de compreensão semântica, síntese e classificação de informações, mesmo quando aplicados a domínios complexos e pouco estruturados (Alawaji; Hakami; Alshemaimri, 2024). No contexto da engenharia de requisitos, estudos recentes indicam que LLMs podem auxiliar na transformação de descrições informais em histórias de usuário mais estruturadas e na identificação de inconsistências em artefatos textuais (Oswal; Kanakia; Suktel, 2024). Sob a perspectiva de controle de qualidade, Ronanki, Cabrero-Daniel e Berger (2024) observam que avaliações realizadas com apoio dessas tecnologias podem apresentar níveis relevantes de convergência com julgamentos realizados por especialistas humanos. Entretanto, apesar desse potencial tecnológico, a aplicação sistemática dessas ferramentas em processos organizacionais ainda apresenta desafios relacionados à interpretação de dependências lógicas, à consistência das avaliações produzidas e à necessidade de supervisão humana para validação dos resultados (Brockenbrough; Salinas, 2024). Dessa forma, permanece em aberto como tais tecnologias podem ser integradas de maneira estruturada a processos de avaliação de artefatos utilizados no planejamento ágil.

Sob essa ótica, estudos recentes têm explorado o uso de arquiteturas baseadas em modelos de linguagem e sistemas multiagentes para apoiar atividades relacionadas à análise e ao refinamento de requisitos em processos de desenvolvimento de software. Conforme discutido por Zhang et al. (2024), tais abordagens indicam o potencial de utilização de agentes baseados em LLMs para apoiar tarefas como auditoria, classifica-

ção e revisão de itens de backlog. Entretanto, apesar dessas iniciativas iniciais, ainda são limitadas as abordagens que propõem métodos estruturados e sistemáticos para avaliar a qualidade de artefatos textuais utilizados no planejamento ágil, especialmente em contextos de desenvolvimento em escala. Considerando que o *Product Backlog* atua como artefato central de coordenação entre múltiplas equipes, a ausência de mecanismos analíticos consistentes para avaliar sua qualidade representa uma lacuna relevante tanto para a literatura quanto para a prática organizacional.

Diante desse cenário, a presente pesquisa propõe o desenvolvimento de um método estruturado para avaliação da qualidade de itens do *Product Backlog*, materializado por meio de um artefato computacional denominado *Backlog Analyzer*. O artefato foi concebido para empregar modelos de LLMs na análise de descrições textuais presentes em artefatos de planejamento ágil, por meio da orquestração de agentes computacionais responsáveis por examinar aspectos estruturais e semânticos desses itens. Conforme discutido por [Zhang et al. \(2024\)](#), sistemas compostos por agentes baseados em LLMs apresentam potencial para apoiar atividades de revisão e análise de requisitos que, tradicionalmente, dependem de avaliações intensivas realizadas por especialistas. Assim, o método proposto busca estruturar o processo de avaliação de artefatos de backlog, incluindo histórias de usuário, critérios de aceitação e definições de *Definition of Ready* (DoR) e *Definition of Done* (DoD), transformando um processo predominantemente qualitativo em um procedimento mais sistemático e comparável entre diferentes equipes e ciclos de planejamento ([Ronanki; Cabrero-Daniel; Berger, 2024](#)). Dessa forma, o método proposto nesta pesquisa busca contribuir para o desenvolvimento de abordagens sistemáticas de análise de artefatos textuais utilizados em processos organizacionais baseados em descrições de requisitos. Conforme discutido por [Sami et al. \(2025\)](#), modelos de LLM podem ser empregados em arquiteturas computacionais capazes de apoiar não apenas a geração de texto, mas também atividades analíticas relacionadas à priorização de demandas e à avaliação do valor de negócio de requisitos complexos. A presente pesquisa adota a abordagem de *Design Science Research* (DSR) como estratégia metodológica para o desenvolvimento e avaliação do artefato proposto. A DSR constitui uma abordagem de pesquisa orientada ao desenvolvimento e à avaliação de artefatos inovadores destinados à solução de problemas organizacionais relevantes, combinando rigor científico com utilidade prática ([Hevner et al., 2004](#); [Peppers et al., 2007](#)). Como referência metodológica principal, utiliza-se

o modelo de pesquisa em DSR apresentado por [Dresch, Lacerda e Antunes Junior \(2015\)](#), que estrutura o processo de desenvolvimento e avaliação do artefato ao longo da pesquisa. A descrição detalhada das etapas do método e de sua aplicação neste estudo é apresentada no Capítulo 3. Embora avanços recentes em LLMs indiquem novas possibilidades para análise de textos organizacionais, permanece em aberto como essas tecnologias podem ser empregadas de maneira estruturada para transformar processos de avaliação predominantemente subjetivos em mecanismos analíticos reproduzíveis, comparáveis e escaláveis.

Esta pesquisa investiga a ausência de métodos estruturados capazes de avaliar, de forma sistemática, consistente e escalável, a qualidade dos itens textuais que compõem o *Product Backlog* em ambientes ágeis em escala. Na perspectiva da *Design Science Research*, esse problema empírico pode ser compreendido como uma instância de uma classe de problema mais ampla, relacionada à avaliação da qualidade de artefatos textuais utilizados na especificação e comunicação de requisitos em processos de desenvolvimento de software.

1.1 Problema da Pesquisa

Apesar do reconhecimento do *Product Backlog* como um artefato central no planejamento em métodos ágeis, sua avaliação qualitativa ainda ocorre predominantemente de forma informal nas organizações. Conforme discutido por [Sedano, Ralph e Péraire \(2019\)](#), a análise da clareza, da consistência e da maturidade das descrições presentes nesse backlog frequentemente depende da experiência individual de especialistas, como Product Owners e Agile Masters, sem o apoio de critérios sistematizados de avaliação.

Essa dependência de julgamento empírico tende a introduzir variabilidade na interpretação dos requisitos e nas decisões de planejamento. Como apontam [Dilorenzo et al. \(2020\)](#) e [Jakobsen e Sutherland \(2009\)](#), diferentes equipes podem adotar critérios distintos para avaliar histórias de usuário ou definições de pronto, o que dificulta a comparação entre projetos e limita a consolidação do aprendizado organizacional.

Adicionalmente, a natureza dinâmica dos ambientes ágeis impõe desafios adicionais à análise da completude e da maturidade desses artefatos de forma sistemática. Segundo [Amjad et al. \(2017\)](#), a evolução contínua do escopo torna complexa a aplicação

de mecanismos tradicionais de avaliação de requisitos, especialmente quando se busca garantir consistência no planejamento ao longo do tempo.

Em decorrência, observa-se uma lacuna tanto na literatura quanto na prática organizacional quanto à existência de métodos estruturados e escaláveis capazes de avaliar, de forma sistemática, a qualidade das descrições presentes em artefatos de planejamento ágil, particularmente nos itens que compõem o *Product Backlog*.

Diante desse cenário, formula-se o seguinte problema de pesquisa: como avaliar, de maneira estruturada, consistente e escalável, a qualidade dos itens textuais que compõem o *Product Backlog* em ambientes ágeis em escala?

1.2 Questão de Pesquisa

Diante desse cenário, emerge a seguinte questão de pesquisa:

Como estruturar um método automatizado capaz de avaliar, de forma sistemática, a qualidade dos itens textuais que compõem o *Product Backlog* em ambientes ágeis em escala?

1.3 Objetivo Geral

O objetivo geral desta pesquisa consiste em desenvolver e avaliar um método estruturado para a avaliação da qualidade dos itens textuais do *Product Backlog* em ambientes ágeis em escala, materializado por meio de um artefato computacional denominado *Backlog Analyzer*, baseado na utilização de modelos de LLMs para análise de descrições textuais.

O método proposto busca permitir a mensuração estruturada da qualidade de itens do *Product Backlog* (incluindo histórias de usuário, critérios de aceitação, Definition of Ready (DoR) e Definition of Done (DoD)) com o propósito de reduzir a subjetividade das avaliações e apoiar processos de gestão e melhoria contínua no planejamento ágil.

1.3.1 Objetivos Específicos

Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

1. **Identificar e estruturar dimensões e critérios de qualidade aplicáveis a itens do *Product Backlog***, incluindo histórias de usuário, critérios de aceitação, Definition of Ready (DoR) e Definition of Done (DoD), com base na literatura de engenharia de requisitos e métodos ágeis.
2. **Projetar e desenvolver o artefato computacional denominado *Backlog Analyzer***, capaz de empregar modelos de LLMs de forma programática para avaliar descrições de backlog com base em critérios estruturados.
3. **Estruturar mecanismos de visualização e interpretação dos resultados produzidos pelo artefato**, permitindo a consolidação de indicadores de qualidade do backlog em dashboards de gestão à vista no contexto da gestão da rotina de um Agile Release Train (ART¹).
4. **Avaliar empiricamente a utilidade do método e do artefato propostos**, analisando seus efeitos na melhoria da qualidade estrutural do backlog e na percepção das equipes quanto à clareza, maturidade e consistência dos artefatos analisados.

1.4 Contribuições da Pesquisa

A principal contribuição desta pesquisa reside na integração entre práticas de gestão ágil, engenharia de requisitos e inteligência artificial baseada em modelos de LLMs, resultando na proposição de um método estruturado para avaliação da qualidade de itens do *Product Backlog*. Esse método é materializado por meio de um artefato computacional capaz de apoiar processos de planejamento e governança em ambientes ágeis em escala.

As contribuições deste estudo podem ser compreendidas em duas dimensões principais:

¹ No framework Scaled Agile Framework (SAFe), um Agile Release Train (ART) representa uma organização virtual composta por múltiplas equipes ágeis (tipicamente entre 50 e 125 profissionais) que trabalham de forma sincronizada para planejar, desenvolver e entregar incrementos de valor de um produto ou solução.

1.4.1 Contribuições Científicas

No âmbito científico, esta pesquisa contribui para o avanço do conhecimento na interface entre engenharia de requisitos, gestão ágil e inteligência artificial aplicada à análise de artefatos textuais utilizados no planejamento de software.

Primeiramente, o estudo amplia a discussão sobre a avaliação da qualidade de artefatos que compõem o *Product Backlog* ao propor um conjunto estruturado de dimensões analíticas, critérios de avaliação e escalas de mensuração capazes de transformar avaliações qualitativas em indicadores comparáveis.

Em segundo lugar, a pesquisa propõe um método estruturado para avaliação da qualidade e da maturidade de itens do *Product Backlog*, contribuindo para a formalização de práticas que, na maioria das organizações, ainda são conduzidas de forma predominantemente informal e dependente da experiência individual de especialistas.

Por fim, no contexto da Design Science Research, o estudo contribui ao demonstrar o potencial de utilização de modelos de LLMs como instrumentos analíticos para avaliação sistemática de artefatos textuais utilizados no planejamento ágil.

1.4.2 Contribuições para a Prática Organizacional

Do ponto de vista aplicado, a pesquisa oferece um artefato computacional denominado *Backlog Analyzer*, desenvolvido para apoiar organizações na avaliação sistemática da qualidade de itens do *Product Backlog*.

O artefato permite analisar descrições de backlog e produzir diagnósticos estruturados a partir da análise integral das histórias de usuário, considerando aspectos como clareza, completude e consistência dos artefatos utilizados no planejamento ágil. Diferentemente de abordagens tradicionais baseadas em amostragem, o método possibilita a avaliação de 100% dos itens analisados, o que se torna viável por meio do uso de modelos de LLMs e da automação do processo de análise.

Além disso, o método proposto possibilita a construção de diagnósticos comparáveis entre diferentes equipes e ciclos de planejamento, contribuindo para o fortalecimento de práticas de melhoria contínua em ambientes de desenvolvimento ágil em escala.

A aplicação do artefato em um contexto organizacional real permitiu observar seu potencial de apoio à gestão do backlog e ao alinhamento entre equipes responsáveis pelo desenvolvimento de produtos digitais.

1.5 Justificativa e Relevância da Pesquisa

A relevância desta pesquisa está associada à crescente complexidade organizacional observada em ambientes de desenvolvimento de software que adotam práticas de agilidade em escala. Nesses contextos, múltiplas equipes trabalham simultaneamente na construção e evolução de um mesmo produto ou plataforma digital, o que amplia a necessidade de mecanismos capazes de favorecer a clareza, a consistência e a rastreabilidade das atividades planejadas.

No âmbito da Engenharia de Produção, o *Product Backlog* pode ser compreendido como um elemento estruturante do planejamento operacional em sistemas de desenvolvimento de software. Esse artefato atua como um mecanismo de tradução entre os objetivos estratégicos do produto e a execução das atividades pelas equipes de desenvolvimento, influenciando diretamente a coordenação do trabalho, a priorização de demandas e a organização do fluxo de desenvolvimento.

Diante desse papel estruturante, aprimorar os mecanismos de avaliação da qualidade desses artefatos torna-se relevante para organizações que buscam aumentar a transparência do planejamento, reduzir ambiguidades na interpretação dos requisitos e melhorar a previsibilidade das entregas em ambientes de desenvolvimento complexos.

Paralelamente, avanços recentes em modelos de LLMs ampliam as possibilidades de análise automatizada de artefatos textuais. Esses modelos apresentam elevada capacidade de processamento semântico da linguagem natural, permitindo explorar abordagens computacionais capazes de apoiar a análise e a classificação de descrições de requisitos e itens de backlog.

Diante do exposto, esta pesquisa busca integrar conceitos da Engenharia de Produção, da gestão ágil e da inteligência artificial com o objetivo de desenvolver um método estruturado para avaliação da qualidade de artefatos textuais utilizados no planejamento ágil.

A revisão sistemática da literatura, conduzida no âmbito desta pesquisa e detalhada no Capítulo 2, identificou uma lacuna específica que fundamenta a relevância

do estudo: embora existam abordagens voltadas à *escrita* e à geração de histórias de usuário com apoio de LLMs, são escassos os métodos estruturados e escaláveis para a *avaliação sistemática* de artefatos textuais já existentes no backlog, com critérios explícitos e escalas mensuráveis. Essa lacuna, enunciada em detalhes ao final do Capítulo 2, constitui o ponto de partida e a principal motivação desta pesquisa.

Espera-se que o artefato resultante desta pesquisa contribua tanto para o avanço do conhecimento científico sobre mecanismos de avaliação de requisitos quanto para a prática organizacional, ao oferecer um instrumento capaz de apoiar organizações na análise sistemática da qualidade de seus artefatos de planejamento em ambientes de desenvolvimento ágil.

1.6 Delimitação da pesquisa

Esta pesquisa delimita-se à avaliação sistematizada da qualidade e da maturidade de itens do *Product Backlog* utilizados no planejamento de iniciativas de desenvolvimento de software em organizações que adotam práticas de agilidade em escala. O estudo concentra-se na análise de descrições textuais presentes em itens do *Product Backlog*, incluindo histórias de usuário, critérios de aceitação, Definition of Ready (DoR) e Definition of Done (DoD).

A aplicação empírica do método proposto foi conduzida em um contexto organizacional específico, envolvendo projetos digitais vinculados a uma organização do setor industrial brasileiro. A avaliação ocorreu no âmbito de 1 (um) *Agile Release Train (ART)*, utilizado como instância de análise para observar o comportamento do artefato desenvolvido em um ambiente real de planejamento ágil em escala. A pesquisa restringe-se à análise de dados internos relacionados à estrutura e ao conteúdo dos itens de backlog. Não foram considerados dados financeiros confidenciais, métricas de desempenho individual das equipes ou informações relacionadas à governança corporativa da organização. Da mesma forma, não fazem parte do escopo desta investigação análises sobre impactos organizacionais de longo prazo decorrentes da adoção do método proposto, tampouco avaliações comparativas com outros métodos ou ferramentas de análise de requisitos existentes no mercado. Por fim, destaca-se que os critérios de avaliação utilizados no estudo foram definidos a partir da literatura especializada sobre engenharia de requisitos, métodos ágeis e qualidade de backlog, sendo

posteriormente adaptados ao contexto operacional da organização analisada. Dessa forma, embora o método tenha sido aplicado em um contexto organizacional específico, a pesquisa busca contribuir com uma proposta metodológica passível de adaptação a contextos similares, não tendo como objetivo realizar generalizações estatísticas, mas sim generalizações analíticas relacionadas ao método e ao artefato proposto.

1.7 Estrutura do trabalho

Este trabalho está organizado em seis capítulos, estruturados de forma a conduzir o leitor desde a contextualização do problema de pesquisa até a apresentação das conclusões e contribuições do estudo. O Capítulo 1 apresenta a introdução da pesquisa, incluindo a contextualização do tema, a definição do problema de pesquisa, os objetivos do estudo, as contribuições da pesquisa e a delimitação da investigação. O Capítulo 2 apresenta a fundamentação teórica que sustenta o trabalho, abordando conceitos relacionados à gestão ágil, à engenharia de requisitos, à qualidade de backlog e às abordagens existentes para avaliação de artefatos utilizados no planejamento de software. O Capítulo 3 descreve o delineamento metodológico da pesquisa, apresentando a abordagem de *Design Science Research* adotada, os procedimentos de desenvolvimento do artefato e os critérios utilizados para sua avaliação. O Capítulo 4 apresenta o desenvolvimento do artefato proposto e sua aplicação no contexto organizacional analisado, descrevendo a estrutura do método e o processo de avaliação dos itens do *Product Backlog*. O Capítulo 5 apresenta e discute os resultados obtidos a partir da aplicação do método, analisando os diagnósticos gerados e suas implicações para o planejamento ágil. Por fim, o Capítulo 6 apresenta as considerações finais da pesquisa, sintetizando suas principais contribuições, limitações e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que sustenta a presente dissertação, organizando os principais conceitos, abordagens e referenciais necessários à compreensão do problema de pesquisa e da solução proposta. Diferentemente da revisão sistemática da literatura, que será apresentada na seção 2.5, a fundamentação teórica tem caráter conceitual e contextual, com o objetivo de estabelecer o arcabouço teórico que orienta a construção do método e do artefato desenvolvidos ao longo do trabalho. A fundamentação teórica está estruturada em quatro subcapítulos, que abrangem conceitos da Engenharia de Produção, métodos ágeis e agilidade em escala, bem como aspectos relacionados à Inteligência Artificial, gestão à vista, escalabilidade organizacional e engajamento humano. Essa estrutura busca refletir a natureza socio-técnica do problema investigado, integrando perspectivas técnicas, organizacionais e comportamentais relevantes para a compreensão do fenômeno analisado.

2.1 Engenharia de Produção, Gestão de Operações e Fluxo de Trabalho

A Engenharia de Produção estabelece como seus pilares fundamentais a direção, o projeto, a entrega e o desenvolvimento de sistemas produtivos (Slack; Brandon-Jones; Johnston, 2018). Sob essa perspectiva, essa disciplina orienta-se pela busca sistemática de objetivos clássicos de desempenho (qualidade, velocidade, confiabilidade, flexibilidade e custo) com vistas à maximização da eficiência organizacional. Embora historicamente vinculada à manufatura industrial, sua base conceitual transcende o ambiente fabril, sendo aplicável também a contextos intensivos em informação, como o desenvolvimento de software e projetos digitais, nos quais problemas de fluxo, variabilidade, capacidade e coordenação assumem papel igualmente central (Slack; Brandon-Jones; Johnston, 2018).

Nesse contexto ampliado, o diagnóstico do desempenho operacional depende, fundamentalmente, da capacidade de controlar a variabilidade e assegurar estabilidade processual. Para (Campos, 2003), a gestão da rotina constitui o mecanismo estruturante para contenção de dispersões, assegurando a manutenção da qualidade intrínseca do trabalho executado. Tal perspectiva converge de modo direto com os postulados da Física de Fábrica, segundo os quais a variabilidade não gerenciada atua inevitavelmente

na degradação da eficiência sistêmica, ampliando filas, aumentando tempos de espera e reduzindo a previsibilidade (Hopp; Spearman, 2001). Assim, como adverte Campos (2003), a padronização não deve ser compreendida como rigidez burocrática, mas como condição *sine qua non* para a estabilidade, na qual a melhoria contínua se configura como um processo estruturado de identificação de anomalias e de implementação de ações corretivas baseadas em dados.

Deste modo, a compreensão da dinâmica do fluxo de trabalho emerge como elemento central para o controle operacional. Diante disso, as chamadas “leis da física de fábrica” oferecem relações analíticas úteis para compreender o comportamento dos sistemas produtivos, dentre as quais se destaca a Lei de Little (Hopp; Spearman, 2001).

Essa premissa estabelece que o tempo médio de atravessamento (lead time) é diretamente proporcional ao estoque de trabalho em processo (WIP) e inversamente proporcional à taxa de saída (throughput), fornecendo base analítica para o gerenciamento racional de gargalos e para a compreensão dos efeitos da variabilidade sobre o desempenho do sistema (Hopp; Spearman, 2001).

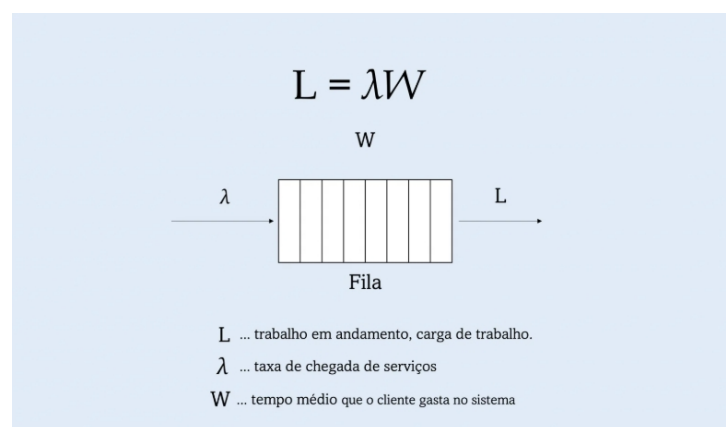


Figura 1 – Representação - Lei de Little

Fonte: (lunes, 2023)

Ao transpor esses fundamentos para o universo dos projetos digitais, observa-se que a fila de tarefas, usualmente estruturada na forma de backlog, desempenha função análoga ao estoque em processo na indústria. Conforme Slack, Brandon-Jones e Johnston (2018), todo sistema produtivo depende da adequada gestão de suas entradas e do balanceamento entre capacidade e demanda.

Dessa forma, a gestão do backlog, entendida como prática de controle do trabalho em processo em ambientes digitais, revela-se não apenas um elemento operacional dos métodos ágeis, mas uma aplicação dos fundamentos da gestão de operações

(Slack; Brandon-Jones; Johnston, 2018). O alinhamento entre fluxo, capacidade e variabilidade constitui, portanto, o elo conceitual entre esses dois domínios, reforçando a relevância da qualidade das informações que estruturam o backlog para o desempenho do sistema como um todo.

Essa aproximação entre sistemas industriais e sistemas digitais não deve ser atribuída a uma única corrente teórica da Engenharia de Produção. A interpretação do backlog como elemento regulador do fluxo dialoga de forma complementar com a gestão de operações, com a Física de Fábrica, ao enfatizar as relações entre variabilidade, filas, capacidade e desempenho sistêmico, com a Teoria das Restrições, especialmente na leitura de gargalos, restrições e buffers como elementos estruturantes do desempenho global, e com a abordagem Lean aplicada ao desenvolvimento de produtos (Reinertsen, 2009), ao destacar a gestão de filas, a limitação do trabalho em progresso, a redução de desperdícios e a geração de valor. Essa articulação evita uma filiação conceitual exclusiva e reforça a aderência da análise ao campo mais amplo da Engenharia de Produção.

Tabela 1 – Comparação entre sistemas industriais e sistemas digitais sob a ótica da Engenharia de Produção

Dimensão Analítica	Sistema Industrial	Sistema Digital (Software)
Natureza do Output	Produto físico tangível, materializado em bens manufaturados.	Produto informacional representado por funcionalidades ou incrementos de software.
Unidade de Fluxo	Peça, lote ou ordem de produção que percorre etapas físicas do processo produtivo.	Item de trabalho representado por histórias de usuário, tarefas ou funcionalidades no fluxo de desenvolvimento.
Estoque em Processo (WIP)	Inventário físico em processamento entre estações ou etapas do sistema produtivo.	Itens em desenvolvimento no fluxo de trabalho da equipe (ex.: quadro Kanban ou backlog ativo).
Variabilidade	Falhas de máquina, variabilidade de insumos, tempos de setup e interferências operacionais.	Mudanças de requisitos, dependências técnicas, retrabalho e incerteza na especificação das funcionalidades.
Restrição do Sistema (Gargalo)	Recurso produtivo cuja capacidade limita o desempenho global do sistema, conforme discutido na gestão de operações e na lógica de restrições.	Etapa do fluxo ou capacidade da equipe que limita o avanço do trabalho no desenvolvimento.
Buffer ou Pulmão	Estoque regulador posicionado antes da restrição para absorver flutuações do sistema e proteger o fluxo produtivo.	Fila priorizada de trabalho (backlog) que regula o fluxo e protege a capacidade das equipes.
Indicadores de Desempenho do Fluxo	Lead time, cycle time, throughput, nível de estoque (WIP) e utilização de capacidade.	Lead time, cycle time, throughput, WIP e métricas de fluxo do trabalho.
Lei Estruturante do Sistema	Lei de Little, princípios da Física de Fábrica e conceitos associados à gestão de restrições aplicados ao fluxo produtivo.	Lei de Little aplicada ao fluxo de trabalho e à gestão de tarefas em sistemas digitais.
Gestão da Rotina	Padronização operacional, gestão visual e controle estatístico de processo (CEP).	Gestão visual do trabalho (Kanban), Definition of Ready (DoR) e Definition of Done (DoD).
Objetivo Sistêmico	Eficiência operacional com estabilidade e previsibilidade do sistema produtivo.	Entrega contínua de valor com qualidade, previsibilidade e coordenação entre equipes.

Fonte: Elaborado pelo autor.

A partir dessa perspectiva, o backlog pode ser interpretado como um elemento estruturante do fluxo de trabalho em sistemas produtivos digitais. Assim como em ambientes industriais, a qualidade das informações que organizam o trabalho influencia diretamente a estabilidade do sistema, a previsibilidade das entregas e a eficiência operacional. Deste modo, a avaliação sistemática da qualidade das descrições presentes no backlog deixa de ser apenas uma prática de apoio à gestão ágil e passa a constituir um problema relevante de gestão de operações em ambientes de desenvolvimento de software.

2.2 Métodos Ágeis e Agilidade em Escala

2.2.1 Fundamentos da Agilidade

Segundo [Highsmith \(2004\)](#), os métodos ágeis emergiram como uma resposta às limitações dos modelos tradicionais de desenvolvimento de software, caracterizados por abordagens preditivas e fortemente baseadas em planejamento antecipado. Em contraste, os métodos ágeis fundamentam-se em princípios adaptativos e na entrega incremental de valor ao cliente. Diferentemente dos paradigmas tradicionais, centrados na aderência a escopo, prazo e orçamento previamente definidos, a agilidade desloca o foco da gestão para a geração contínua de valor ao usuário ([Cohn, 2005](#)). Nessa perspectiva, o sucesso de um projeto deixa de ser medido exclusivamente pela conformidade ao plano inicial e passa a ser avaliado pela capacidade do produto de resolver problemas reais e gerar benefícios sustentáveis ao negócio. De acordo com [Highsmith \(2004\)](#), o valor somente pode ser definido pelo cliente, entendido como o indivíduo ou grupo que efetivamente utiliza o produto para gerar resultados organizacionais. Sob uma perspectiva econômica, esse valor pode manifestar-se de diversas formas, como a geração de novas receitas, a ampliação do valor capturado de clientes existentes, a retenção de receitas que poderiam ser perdidas ou a obtenção de eficiências operacionais ([Cohn, 2005](#)). Dessa forma, a gestão ágil transcende a dimensão técnica do desenvolvimento de software e assume um papel estratégico no apoio à competitividade organizacional.

2.2.2 Histórias de Usuário e Engenharia de Requisitos Ágil

À primeira vista, pode parecer existir uma tensão entre a eliminação de burocracia (princípio central da filosofia Ágil) e a exigência de maior rigor na definição de requisitos em métodos ágeis. Entretanto, essa aparente contradição resulta de uma interpretação superficial da natureza desses conceitos. Enquanto a burocracia refere-se à produção de artefatos documentais que não agregam valor direto ao cliente, o rigor técnico atua como mecanismo de prevenção de desperdícios e retrabalho ([Highsmith, 2004](#)). Nos modelos tradicionais de desenvolvimento, a burocracia frequentemente manifesta-se por meio de especificações extensivas e detalhadas que buscam antecipar

exaustivamente todas as possíveis necessidades futuras. Em contraposição a essa abordagem, as histórias de usuário surgem como uma forma mais leve e colaborativa de especificação de requisitos. Conforme descrito por Cohn (2003), as histórias de usuário são estruturadas com base no modelo dos “Três Cs”, proposto por Ron Jeffries, composto por: Cartão (*Card*), que representa a descrição inicial da funcionalidade; Conversa (*Conversation*), que corresponde ao diálogo colaborativo entre stakeholders para detalhamento do requisito; e Confirmação (*Confirmation*), que estabelece critérios objetivos de aceitação que permitem verificar se a funcionalidade foi implementada corretamente.



Figura 2 – Modelo dos Três Cs das histórias de usuário
 Fonte: Elaborado pelo autor com base em (Menzinsky, 2022).

Diversas práticas complementares foram desenvolvidas para apoiar a qualidade das histórias de usuário. Entre as mais difundidas encontra-se o modelo INVEST, sistematizado por Cohn (2004), que estabelece características desejáveis para histórias bem estruturadas, tais como independência, negociabilidade, valor de negócio, estimabilidade, tamanho adequado e testabilidade. De forma complementar, abordagens derivadas do *Behavior-Driven Development* (BDD) têm sido utilizadas para estruturar critérios de aceitação verificáveis, frequentemente utilizando a linguagem Gherkin baseada no padrão *Given–When–Then* (North, 2006). Essas práticas contribuem para reduzir ambiguidades nas especificações e favorecer uma comunicação mais clara entre stakeholders técnicos e de negócio. Além dessas abordagens, a literatura de engenharia de software apresenta modelos formais para avaliação da qualidade de sistemas. Entre os mais difundidos encontra-se o modelo definido na norma ISO/IEC 25010 (International Organization for Standardization, 2011), que estabelece características

e subcaracterísticas de qualidade de software, incluindo confiabilidade, usabilidade, eficiência e manutenibilidade. Tais modelos servem frequentemente como referência para a definição de critérios técnicos de avaliação em processos de desenvolvimento e melhoria de software. Para além desses modelos formais, a compreensão adequada dos requisitos representa outro fator crítico de qualidade. Conforme demonstram [Lindland, Sindre e Sølvsberg \(1994\)](#), falhas de compreensão nos requisitos podem gerar perdas significativas de eficiência, uma vez que obrigam as equipes a interromper o fluxo de desenvolvimento para corrigir interpretações equivocadas ou retrabalhar funcionalidades.

2.2.3 Estrutura Operacional do Scrum

Essa lógica de redução de desperdícios e entrega incremental de valor se materializa de forma operacional na estrutura do Scrum. Conforme descrito por [Schwaber e Sutherland \(2020\)](#), o Scrum organiza o desenvolvimento em ciclos iterativos e incrementais denominados Sprints, nos quais ideias são transformadas em incrementos funcionais do produto. Cada Sprint atua como um contêiner temporal de duração fixa que organiza quatro eventos formais: Sprint Planning, Daily Scrum, Sprint Review e Sprint Retrospective. Esses eventos são projetados para operacionalizar os pilares empíricos da transparência, inspeção e adaptação que caracterizam o controle de processos complexos.



Figura 3 – Ciclo de trabalho no Scrum

Fonte: Elaborado pelo autor com base em (Scaled Agile, Inc., 2025a).

Segundo Schwaber e Sutherland (2020), o objetivo central de cada Sprint é a criação de um Incremento utilizável do produto, que representa um avanço validado em direção à meta final do sistema e deve ser integrado aos incrementos previamente desenvolvidos. Para garantir a qualidade e consistência técnica desse incremento, o trabalho realizado deve atender a uma Definition of Done (DoD) clara e compartilhada entre os membros da equipe. A DoD funciona como um mecanismo de transparência e qualidade embutida, assegurando que o incremento esteja devidamente testado e integrado antes de ser considerado concluído. Em ambientes caracterizados por elevada incerteza, o desenvolvimento ágil reconhece que planos devem ser tratados como hipóteses sujeitas a validação empírica contínua (Highsmith, 2004). Ao final de cada Sprint, mudanças de mercado, novos aprendizados técnicos ou oportunidades identificadas são incorporados ao Product Backlog, permitindo que o planejamento do produto seja continuamente ajustado. Nesse processo, o Product Owner exerce papel central ao repriorizar os itens do backlog com base nas informações mais recentes, buscando maximizar a geração de valor do produto ao longo do tempo (Schwaber; Sutherland, 2020). A Figura 4 apresenta o ciclo conceitual de interação entre o Product Backlog, uma etapa de avaliação da qualidade dos itens do backlog e o fluxo de desenvolvimento ágil. O modelo evidencia que essa etapa de avaliação da qualidade é de suma importância no planejamento ágil, pois influencia o planejamento das sprints e

retroalimenta continuamente o backlog a partir do feedback gerado pelos incrementos entregues.

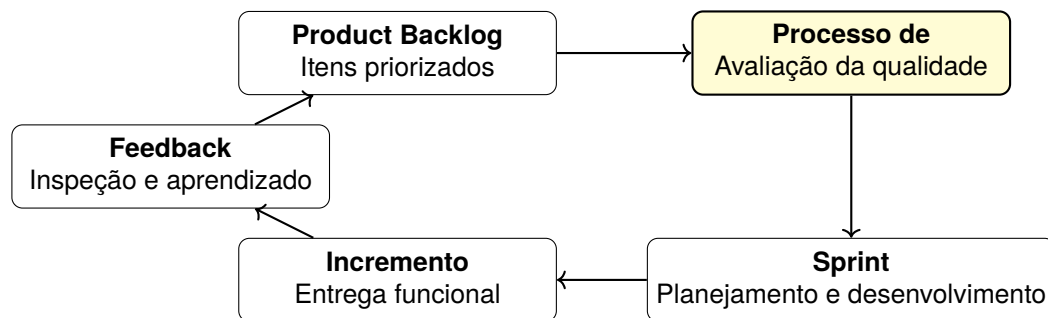


Figura 4 – Ciclo conceitual de avaliação do backlog e execução do desenvolvimento ágil

Fonte: Elaborado pelo autor com base em (Highsmith, 2004; Schwaber; Sutherland, 2020).

2.2.4 Escalabilidade Organizacional e SAgile

À medida que métodos ágeis passaram a ser adotados em organizações de maior porte, envolvendo múltiplas equipes e sistemas de software cada vez mais complexos, surgiu a necessidade de mecanismos capazes de coordenar o trabalho em escala. Segundo Paasivaara (2017), o Scaled Agile Framework (SAFe) foi desenvolvido com o objetivo de fornecer uma estrutura organizacional que permita alinhar múltiplos times de desenvolvimento a uma estratégia comum, mantendo ao mesmo tempo princípios de agilidade e entrega contínua de valor. No centro dessa estrutura encontra-se o conceito de Agile Release Train (ART), definido como uma organização virtual composta por múltiplas equipes ágeis que trabalham de forma sincronizada para planejar, desenvolver e entregar incrementos de valor de um produto ou solução (Scaled Agile, Inc., 2025a). Em geral, um ART reúne entre 50 e 125 profissionais organizados em equipes multifuncionais que compartilham uma mesma visão de produto e um roadmap comum. A Figura 5 apresenta uma visão simplificada da estrutura de coordenação em ambientes de agilidade em escala. Nesse modelo, múltiplas equipes ágeis são organizadas em torno de um ART, responsável por sincronizar o trabalho entre os times e garantir a entrega integrada de incrementos de valor ao longo dos ciclos de planejamento e execução.

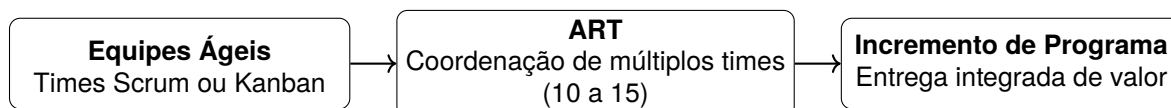


Figura 5 – Estrutura simplificada de coordenação em ambientes de agilidade em escala

Fonte: Elaborado pelo autor com base em (Scaled Agile, Inc., 2025a).

Entretanto, a coordenação de múltiplas equipes trabalhando simultaneamente em um mesmo produto amplia significativamente a complexidade da gestão técnica e organizacional. Conforme observado por [Staron, Meding e Baniasad \(2019\)](#), métricas e práticas originalmente concebidas para equipes isoladas podem tornar-se insuficientes quando aplicadas em contextos de desenvolvimento em larga escala. Diante desse quadro, o monitoramento eficaz do desenvolvimento passa a exigir indicadores adicionais capazes de identificar distorções críticas, como a expansão não controlada do escopo (*scope creep*) e o acúmulo de defeitos que escapam para fases posteriores de integração entre equipes.

2.2.5 Qualidade do Backlog em Ambientes Ágeis em Escala

À medida que métodos ágeis passam a ser adotados em ambientes organizacionais complexos, o Product Backlog assume um papel cada vez mais central na coordenação do trabalho entre equipes. Em contextos de agilidade em escala, inconsistências, ambiguidades ou descrições incompletas nos itens de backlog podem comprometer significativamente a previsibilidade do planejamento e aumentar o retrabalho durante as sprints. A heterogeneidade e a falta de rigor no refinamento das especificações de backlog podem assumir proporções sistêmicas, dificultando a compreensão compartilhada entre equipes e comprometendo a integridade dos incrementos produzidos ([Lindland; Sindre; Sølvsberg, 1994](#)). Esse problema torna-se particularmente relevante em estruturas organizacionais compostas por múltiplos times interdependentes, nas quais a qualidade da informação exerce papel crítico na coordenação do trabalho. Dessa forma, a existência de mecanismos estruturados capazes de apoiar a avaliação da qualidade desses artefatos torna-se um elemento relevante para a gestão eficiente do fluxo de desenvolvimento, especialmente em ambientes de agilidade em escala. Além de favorecer o alinhamento entre equipes e stakeholders, tais mecanismos podem fornecer evidências analíticas que apoiem a tomada de decisão e a melhoria contínua dos processos de planejamento. Nessa direção, instrumentos

de monitoramento organizacional, como dashboards e sistemas de gestão à vista, bem como técnicas computacionais baseadas em inteligência artificial para análise de artefatos textuais, surgem como abordagens promissoras para apoiar a avaliação sistemática da qualidade do backlog. As seções seguintes apresentam os fundamentos conceituais relacionados a esses instrumentos e tecnologias, que sustentam a proposta metodológica desenvolvida nesta pesquisa.

2.3 Gestão à Vista, Dashboards e ETL

À medida que organizações adotam práticas de desenvolvimento ágil em escala, torna-se cada vez mais relevante monitorar continuamente indicadores relacionados ao fluxo de trabalho, à qualidade dos artefatos e à eficiência dos processos. Em particular, a gestão do *Product Backlog* e de seus itens exige mecanismos capazes de tornar visíveis padrões, desvios e oportunidades de melhoria ao longo do tempo. A partir disso, abordagens baseadas em gestão visual e análise estruturada de dados organizacionais desempenham papel central ao permitir que informações operacionais sejam apresentadas de forma clara e acessível para diferentes níveis de decisão.

2.3.1 Gestão à Vista como Instrumento de Alinhamento e Tomada de Decisão

A gestão à vista constitui um dos fundamentos centrais da filosofia enxuta, atuando como mecanismo de transparência operacional e sincronização entre equipes. Em ambientes industriais, o controle visual permite tornar imediatamente perceptíveis desvios, fragilidades e anomalias do processo produtivo, favorecendo respostas corretivas rápidas e estimulando a disciplina operacional (Ohno, 1988). Segundo Ohno (1988), o gerenciamento visual é essencial para trazer à superfície as fraquezas da produção, permitindo que a organização identifique problemas e atue sobre eles de forma imediata.

Nessa perspectiva, a visibilidade dos desvios torna-se condição fundamental para a melhoria contínua, possibilitando inclusive a delegação de autoridade operacional para a interrupção de processos quando anomalias são identificadas (Ohno, 1988). Assim, a gestão visual não se limita à exposição de dados, mas constitui um

mecanismo organizacional que favorece a compreensão compartilhada do desempenho e a coordenação das ações corretivas.

Em ambientes organizacionais contemporâneos, esses princípios têm sido progressivamente operacionalizados por meio de sistemas digitais de monitoramento visual, como dashboards analíticos e scorecards gerenciais. Conforme destacam [Kaplan e Norton \(1996\)](#), tais instrumentos funcionam como uma linguagem capaz de traduzir a estratégia organizacional em indicadores operacionais compreensíveis para diferentes níveis da organização. Dessa forma, dashboards e sistemas de medição visual reduzem o distanciamento entre diretrizes estratégicas e ações cotidianas, contribuindo para o alinhamento organizacional e para o acompanhamento sistemático do desempenho.

De acordo com [\(Few, 2006\)](#), a eficácia desses instrumentos depende de um design informacional que privilegie a comunicação clara de informações relevantes, evitando a sobrecarga visual e a inclusão de elementos desnecessários. Portanto, a gestão visual consolida-se como um mecanismo de alinhamento organizacional que promove o comprometimento coletivo com o desempenho esperado.

2.3.2 Dashboards como Evolução Digital da Gestão à Vista

Com a digitalização dos sistemas produtivos e organizacionais, os painéis físicos de controle evoluíram para dashboards analíticos interativos, capazes de consolidar grandes volumes de dados provenientes de diferentes fontes e apresentá-los de forma sintética para análise gerencial [\(Khan, 2024\)](#). Segundo [Few \(2006\)](#), um dashboard consiste em uma apresentação visual das informações mais importantes para o alcance de objetivos específicos, organizada de modo a permitir o monitoramento contínuo do desempenho organizacional.

Sob essa ótica, a eficácia dessas ferramentas está diretamente relacionada à qualidade do design informacional utilizado na apresentação dos dados. Conforme argumenta [Knaflitz \(2019\)](#), dashboards eficazes devem priorizar clareza perceptual, hierarquização de informações e redução de ruídos visuais, permitindo que os usuários identifiquem rapidamente padrões, tendências e desvios relevantes.

A Figura 6 ilustra exemplos de boas práticas no uso de marcadores e legendas em visualizações analíticas, demonstrando como elementos gráficos podem ser utilizados estrategicamente para direcionar a atenção do usuário às informações mais

relevantes, evitando a poluição visual e favorecendo a interpretação adequada dos dados.



Figura 6 – Uso estratégico de marcadores e legendas - Manter foco nas informações relevantes (não poluir)

Fonte: (Knafllic, 2019), p. 103.

Adicionalmente, conforme destaca [Tufte \(2001\)](#), a integridade gráfica constitui um princípio fundamental para garantir que representações visuais não distorçam a interpretação dos dados. A representação gráfica deve preservar a proporcionalidade entre os valores numéricos e suas representações visuais, evitando manipulações ou simplificações que possam induzir interpretações equivocadas.

Sob a perspectiva operacional, dashboards analíticos integram-se diretamente à gestão da rotina organizacional ao fornecer suporte informacional para reuniões periódicas, ciclos de melhoria contínua e acompanhamento de metas ([Slack; Brandon-Jones; Johnston, 2018](#)). Ao reduzir a latência informacional e ampliar a rastreabilidade histórica dos indicadores, tais ferramentas contribuem para o aumento da maturidade analítica das organizações ([Khan, 2024](#)).

Entretanto, a eficácia desses sistemas de monitoramento visual depende da confiabilidade e da consistência dos dados que alimentam os indicadores apresentados. Diante desse cenário, processos de Extract, Transform and Load (ETL) desempenham papel fundamental ao possibilitar a integração e o tratamento de dados provenientes de fontes heterogêneas ([Kimball; Ross, 2013](#)). O processo de ETL compreende três etapas principais: extração de dados de sistemas operacionais, transformação dessas informações para garantir consistência e padronização semântica, e carregamento final em estruturas analíticas, como data warehouses ou sistemas de business intelligence.

A Figura 7 apresenta uma representação dos elementos centrais da arquitetura de dados proposta por Kimball e Ross (2013), na qual pipelines de ETL estruturam o fluxo de dados desde as fontes operacionais até os sistemas analíticos utilizados para geração de relatórios e dashboards.

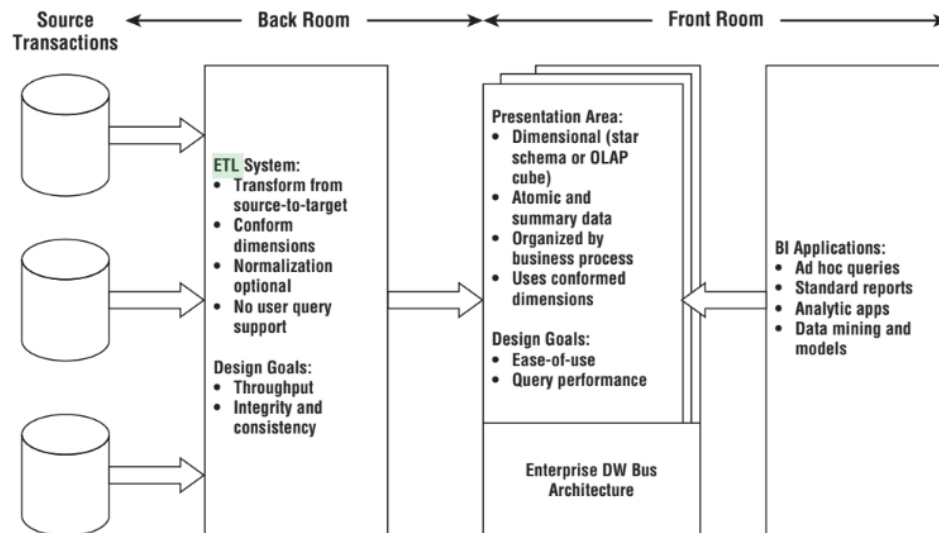


Figura 7 – Elementos centrais da arquitetura Data Warehouse (DW)/Business intelligence (BI) de Kimball

Fonte: (Kimball; Ross, 2013), p. 54.

De acordo com Kimball e Ross (2013), a etapa de transformação assume papel particularmente crítico, pois é responsável por assegurar a integridade semântica e a consistência estrutural dos dados utilizados nas análises. Complementarmente, Inmon (2005) destaca que arquiteturas analíticas bem estruturadas são essenciais para garantir confiabilidade informacional e permitir análises comparativas consistentes ao longo do tempo.

No contexto desta pesquisa, os princípios de ETL estruturam o fluxo de tratamento das informações utilizadas na avaliação da qualidade do backlog. Inicialmente, os dados referentes aos itens de backlog são extraídos das ferramentas de gestão utilizadas pelas equipes ágeis, como plataformas de gerenciamento de projetos e sistemas de gestão de requisitos. Em seguida, essas informações passam por uma etapa de transformação analítica, na qual o artefato computacional desenvolvido neste estudo (denominado *Backlog Analyzer*) emprega LLM para analisar as descrições textuais e gerar indicadores estruturados de qualidade. Por fim, os resultados produzidos são carregados em dashboards de gestão à vista, permitindo que equipes e gestores acompanhem de forma visual e sistemática a qualidade do planejamento ágil.

2.4 Inteligência Artificial e Modelos de Linguagem (LLMs)

Nos métodos ágeis de desenvolvimento de software, os artefatos que estruturam o planejamento, como histórias de usuário, critérios de aceitação e definições de pronto, são predominantemente expressos em linguagem natural. A qualidade dessas descrições textuais influencia diretamente a clareza das atividades executadas pelas equipes e a previsibilidade das entregas. Em decorrência, avanços recentes em Inteligência Artificial, particularmente no campo dos LLMs, têm ampliado significativamente as possibilidades de análise automatizada de textos complexos. Essas tecnologias apresentam elevada capacidade de compreensão semântica, classificação e geração de conteúdo textual, tornando-se ferramentas promissoras para apoiar processos de avaliação e melhoria de artefatos de planejamento em ambientes ágeis.

2.4.1 Conceitos básicos sobre IA generativa e LLMs

Os LLMs representam uma classe avançada de sistemas computacionais projetados para compreender e produzir conteúdo textual. De acordo com Alawaji et al., essa capacidade funcional deriva estritamente do aprendizado de padrões extraídos de volumes massivos de dados e de código. Como resultado desse treinamento em larga escala, torna-se possível a execução de diversas tarefas fundamentais de Processamento de Linguagem Natural (NLP), otimizando a análise e a geração de texto.

Conforme Alawaji, Hakami e Alshemaimri (2024) evidenciam, a base estrutural dessas ferramentas reside na arquitetura Transformer, que impulsionou a modelagem de linguagem moderna. Somado a isso, a eficácia dessa arquitetura provém da sua capacidade de capturar dependências distantes, permitindo que a camada de autoatenção acesse diferentes porções da sequência de entrada. Por meio da computação de matrizes de consulta, chave e valor, Oswal, Kanakia e Suktel (2024) explicam que o modelo pondera essas informações simultaneamente, gerando representações ricas e interligadas. Como resultado, o modelo extrai relações contextuais em sequências textuais de forma altamente eficiente, otimizando o processamento da linguagem. Em complemento, o treinamento em larga escala potencializa a habilidade do modelo em apreender a semântica e as intenções embutidas nos textos.

No funcionamento técnico desses sistemas, o processamento da linguagem natural inicia-se sistematicamente com a etapa de tokenização, na qual o texto de entrada é segmentado em unidades menores. Segundo [Oswal, Kanakia e Suktel \(2024\)](#)., essas unidades são posteriormente convertidas em matrizes vetoriais numéricas, conhecidas como embeddings, que permitem ao algoritmo processar semanticamente as informações estruturadas. Na arquitetura Transformer, o fluxo de dados depende de componentes de codificação e decodificação para interpretar dependências textuais complexas. O codificador (encoder) transforma a sequência de tokens em representações ocultas densas, ao passo que o decodificador (decoder) utiliza essa base contextual para prever e gerar a sequência de saída mais provável. O desenvolvimento dessas capacidades de abstração divide-se em fases, iniciando com o pré-treinamento autossupervisionado em grandes *corpora* textuais para a aquisição de proficiência geral em linguagem ([Unlu et al., 2024](#)).

Conforme [Unlu et al. \(2024\)](#), reforçam, a execução desse ajuste fino (fine-tuning) converte modelos genéricos em ferramentas altamente especializadas. Dessa forma, o sistema adquire a aderência técnica e o vocabulário necessários para atuar em contextos organizacionais delimitados.

Além disso, técnicas recentes como o In-Context Learning (ICL) permitem que os modelos executem tarefas complexas a partir de poucos exemplos inseridos diretamente no prompt de comando. Conforme detalham [Moradi Dakhel et al.](#), essa abordagem opera como uma estrutura que dispensa novos ciclos de treinamento, o que reduz de maneira significativa os custos computacionais associados à adaptação da rede neural para novas tarefas.

Como consequência direta desses avanços técnicos, observa-se uma mudança estrutural na dinâmica de trabalho da engenharia de software ([OPENAI, 2024](#)). Os profissionais estão transitando de uma atuação puramente focada na escrita de código para assumirem o papel de colaboradores integrados aos sistemas de inteligência artificial. Somado a isso, [OPENAI \(2024\)](#) ressaltam que a automação de rotinas transfere a atenção humana para aspectos mais criativos e analíticos do ciclo de desenvolvimento. Dessa forma, os desenvolvedores concentram seus esforços na definição de arquiteturas de alto nível, no estabelecimento de objetivos para a IA e na interpretação das soluções tecnológicas geradas. Em complemento, essa reconfiguração de papéis consolida uma parceria técnica que otimiza as decisões de projeto e eleva o rigor na engenharia dos

artefatos de software (OPENAI, 2024).

2.4.2 Limites e desafios no uso de IA

Apesar dos avanços tecnológicos contínuos, a incorporação de modelos generativos no ciclo de vida de desenvolvimento de software apresenta restrições estruturais severas. Segundo Moradi Dakhel et al. (2024), o desafio mais proeminente é o fenômeno das alucinações, em que o algoritmo gera saídas linguisticamente plausíveis, porém incorretas e de difícil verificação prática. Somado a isso, a limitação arquitetural da janela de contexto restringe criticamente a quantidade de informação que o modelo consegue analisar de maneira simultânea. Como resultado, a avaliação de ecossistemas complexos ou de grandes conjuntos de requisitos interdependentes torna-se tecnicamente inviável sem o particionamento dos dados (Moradi Dakhel et al., 2024).

No domínio da Engenharia de Requisitos, essas barreiras tornam-se particularmente evidentes. Embora os algoritmos consigam processar amplos volumes textuais, essas ferramentas falham na identificação de requisitos não funcionais e na interpretação de nuances que exigem conhecimento tácito do domínio (Arora; Grundy; Abdelrazek, 2024). Em complemento a essas deficiências de análise semântica, ocorrem severas quebras de rastreabilidade durante a evolução dos artefatos do projeto. Conforme aponta Kim (2024), avaliações práticas demonstram que modelos baseados em LLM frequentemente geram inconsistências graves entre as decisões de modelagem arquitetural e a implementação final do código-fonte. Essa falta de alinhamento estrutural compromete a integridade do ciclo de desenvolvimento e exige etapas adicionais de mitigação.

Além disso, como o desempenho desses sistemas depende estritamente das bases de treinamento, vieses nos dados originais podem propagar vulnerabilidades e falhas de forma sistêmica. Portanto, a validação ativa por especialistas humanos permanece indispensável para garantir a precisão técnica, a coerência e a segurança das soluções geradas (Kim, 2024).

2.4.3 Aplicações da IA no desenvolvimento de software e no planejamento de backlog

A incorporação de Inteligência Artificial Generativa no desenvolvimento de software tem provocado mudanças profundas na forma como as equipes projetam e mantêm sistemas digitais. Segundo [OPENAI \(2024\)](#) Nguyen-Duc, Abrahamsson e Khomh, essas tecnologias redefinem as práticas tradicionais ao automatizarem tarefas cognitivas complexas, o que permite aos profissionais atuarem como colaboradores focados em decisões arquiteturais. Do ponto de vista operacional, estimativas demonstram que o uso sistemático dessas ferramentas pode reduzir de 20% a 45% os custos anuais de engenharia. Essa otimização expressiva ocorre principalmente por meio da automação de atividades de base, como a geração inicial e a refatoração de código ([OPENAI, 2024](#)).

No contexto da engenharia de requisitos e do planejamento ágil, os LLMs apoiam diretamente a construção e o refinamento do backlog. Para assegurar o rigor técnico desses artefatos, [Zhang et al. \(2024\)](#) evidenciam que sistemas compostos por múltiplos agentes autônomos conseguem operar de forma coordenada na avaliação de descrições textuais. Esses agentes validam se as histórias de usuário atendem a critérios de qualidade rigorosos, como os estabelecidos pelo framework INVEST, o que abrange a garantia de que os requisitos sejam passíveis de estimativa. Somado a isso, essas ferramentas oferecem suporte central a Product Owners e equipes de desenvolvimento na geração de especificações detalhadas ([Zhang et al., 2024](#)). Dessa forma, consolidam uma análise sistemática capaz de identificar ambiguidades e elevar o alinhamento das demandas de negócio aos objetivos do projeto ([Kim, 2024](#)).

Apesar desse potencial analítico, a participação humana permanece insubstituível para resolver conflitos semânticos severos e garantir a rastreabilidade entre o design e a implementação do código ([Kim, 2024](#)). Portanto, essa evolução viabiliza uma parceria técnica na qual os especialistas atuam como arquitetos e supervisores absolutos das soluções geradas ([OPENAI, 2024](#)).

2.5 Revisão da Literatura

Esta seção apresenta a revisão da literatura relacionada ao tema desta pesquisa, com o objetivo de identificar evidências do problema associado à qualidade de artefatos

de backlog em ambientes ágeis e mapear abordagens tecnológicas emergentes voltadas ao uso de inteligência artificial no tratamento de requisitos textuais.

Para organizar essa investigação, a revisão foi estruturada em duas frentes principais: a primeira dedicada à identificação de estudos que abordam problemas relacionados à qualidade e à maturidade de artefatos de backlog, e a segunda voltada à análise de pesquisas que exploram o uso de inteligência artificial, particularmente LLMs, na análise de requisitos e descrições textuais.

2.5.1 Materiais e Métodos

A condução da revisão da literatura foi apoiada por um processo estruturado de coleta, organização e análise dos artigos selecionados. Para apoiar essas etapas, foi desenvolvido um pipeline modular e reproduzível em linguagem Python, responsável por automatizar parte das atividades relacionadas à coleta, organização e tratamento das referências bibliográficas.

O programa realiza a extração de metadados das bases bibliográficas consultadas e promove a normalização e consolidação dessas informações em uma base estruturada. Como parte desse processo, títulos e resumos dos artigos são submetidos a procedimentos de tradução automática, permitindo uniformizar a análise textual independentemente do idioma original das publicações. O pipeline também realiza a agregação e análise das palavras-chave associadas aos artigos, possibilitando identificar termos recorrentes e apoiar a expansão recursiva das estratégias de busca. Dessa forma, novos termos relevantes podem ser identificados a partir do próprio conjunto inicial de resultados, ampliando a capacidade exploratória da revisão.

Adicionalmente, ferramentas baseadas em LLMs são utilizadas para apoiar a análise preliminar dos resumos dos artigos, gerando sínteses estruturadas que destacam elementos como o problema de pesquisa, o tema central e os principais tópicos abordados em cada estudo. A partir do conteúdo textual dos resumos, o pipeline também aplica heurísticas para sugerir a identificação do setor econômico ou contexto de aplicação associado às pesquisas analisadas.

O sistema também permite o cadastro prévio de temas de interesse e conjuntos de palavras-chave associadas a cada tema. A partir dessas definições, o pipeline executa rotinas de análise textual que estimam, por meio de heurísticas e com apoio de LLMs,

a probabilidade de cada artigo estar relacionado aos temas definidos. Esse mecanismo contribui para as etapas iniciais de triagem da literatura, permitindo identificar de forma mais eficiente os estudos potencialmente relevantes para a pesquisa.

A lógica de classificação e filtragem progressiva dos artigos foi inspirada nos princípios metodológicos do ProKnow-C (Lacerda; Ensslin; Ensslin, 2012), que propõe a construção estruturada de um portfólio bibliográfico a partir de critérios de relevância definidos pelo pesquisador. No contexto desta pesquisa, o pipeline computacional foi utilizado como apoio operacional para executar parte dessas etapas de análise e organização da literatura.

A decisão final de inclusão ou exclusão dos estudos permaneceu sob responsabilidade do pesquisador. Dessa forma, os recursos computacionais foram utilizados como apoio ao processo de revisão, sem substituir o julgamento analítico necessário à composição do portfólio bibliográfico. Essa abordagem aproxima o processo de revisão de práticas de **literature mining**, nas quais técnicas computacionais são empregadas para apoiar a extração e organização sistemática de conhecimento a partir da literatura científica. Como produto final desse processamento, o sistema gera uma planilha estruturada contendo os principais metadados, indicadores bibliométricos e informações analíticas associadas aos artigos, permitindo ao pesquisador aplicar filtros e realizar análises comparativas de forma mais eficiente.

A Figura 8 sintetiza o fluxo geral de apoio computacional utilizado na revisão, destacando a articulação entre as etapas automatizadas do pipeline e a triagem conduzida pelo pesquisador. As etapas específicas de coleta e consolidação dos dados bibliográficos são detalhadas na subseção seguinte.

2.5.2 Extração e Consolidação Inicial

A etapa de extração dos artigos foi realizada a partir de bases bibliográficas amplamente utilizadas na literatura científica das áreas de engenharia de software e sistemas de informação. As buscas foram conduzidas em bases como Scopus, Web of Science e IEEE Xplore, selecionadas por sua relevância e abrangência na indexação de periódicos e conferências internacionais.

A estratégia de busca foi definida a partir da combinação de termos associados aos dois eixos investigados nesta pesquisa: (i) problemas relacionados à qualidade

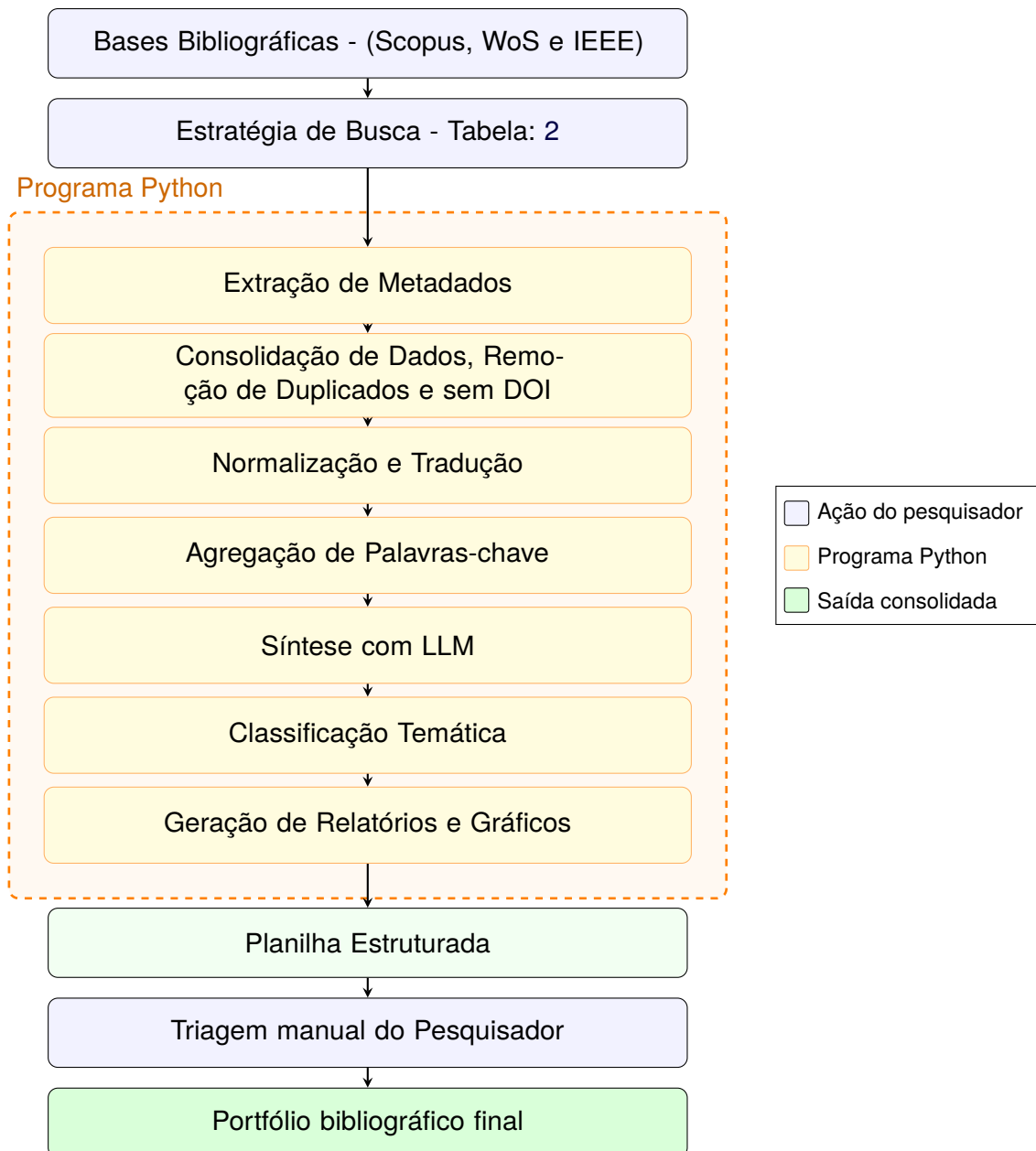


Figura 8 – Fluxo do processo de revisão da literatura com destaque para as etapas automatizadas pelo programa Python.

e estruturação de artefatos de backlog e (ii) uso de técnicas de inteligência artificial, particularmente LLMs, na análise de requisitos textuais. As expressões de busca foram aplicadas nos campos de título, resumo e palavras-chave das bases bibliográficas consultadas. A Tabela 2 apresenta a estrutura da estratégia de busca adotada.

Tabela 2 – Estratégia de busca utilizada na revisão da literatura

Eixo de investigação	Bases de dados	Campos pesquisados	Expressão de busca	Filtros
Problemas relacionados à qualidade de backlog	Scopus, Web of Science, IEEE Xplore	TITLE-ABS-KEY	("product backlog"OR "user stor**") AND ("requirement*"OR "quality"OR "clarity"OR "software requirement**")	Sem restrição
Uso de IA e LLMs na análise de backlog	Scopus, Web of Science, IEEE Xplore	TITLE-ABS-KEY	("user stor*"OR "product backlog") AND ("large language model*"OR LLM OR "generative AI"OR "generative language model"OR ChatGPT OR GPT-4 OR "GPT-3.5")	2023–2026

Fonte: Elaborado pelo autor (2026).

O recorte temporal aplicado ao segundo eixo da revisão (publicações entre 2023 e 2026) foi definido com o objetivo de capturar estudos recentes relacionados ao uso de LLMs, cuja difusão acadêmica e aplicada se intensificou com os avanços recentes da inteligência artificial generativa.

Os resultados obtidos nas diferentes bases foram inicialmente exportados em formato estruturado contendo metadados dos artigos, incluindo título, autores, resumo, palavras-chave, ano de publicação e fonte do documento. Esses registros foram posteriormente consolidados em uma base única de análise, etapa na qual foram eliminadas duplicidades provenientes da indexação de um mesmo artigo em múltiplas bases.

Após a consolidação inicial, o conjunto de artigos foi submetido às etapas de processamento descritas na subseção anterior, incluindo normalização de metadados, tradução automática de textos e organização das informações em planilhas estruturadas para análise posterior.

2.5.3 Normalização dos Metadados

Após a etapa de consolidação inicial, os registros bibliográficos foram submetidos a procedimentos de normalização com o objetivo de padronizar os campos de metadados e facilitar as etapas subsequentes de análise. Essa etapa buscou garantir consistência na estrutura das informações provenientes de diferentes bases bibliográficas, reduzindo inconsistências associadas a variações de formatação e nomenclatura entre as fontes.

Entre as atividades realizadas, destacam-se a padronização de campos como título, autores, ano de publicação, palavras-chave e identificadores digitais (DOI), bem como o tratamento de caracteres especiais e inconsistências de codificação. Esse processo permitiu estruturar os registros em um formato uniforme, adequado para processamento automatizado e análise posterior.

Além disso, foram aplicados procedimentos de limpeza dos dados com o objetivo de eliminar registros incompletos ou inconsistentes que pudessem comprometer a qualidade da base analisada. Como resultado, obteve-se um conjunto de dados bibliográficos padronizado, apto para as etapas subsequentes de análise textual e organização da literatura.

2.5.4 Tradução Automática e Síntese dos Resumos

Considerando que os artigos recuperados nas bases bibliográficas podem estar publicados em diferentes idiomas, foi aplicada uma etapa de tradução automática dos títulos e resumos com o objetivo de uniformizar a análise textual do conjunto de documentos. Esse procedimento permitiu reduzir possíveis barreiras linguísticas durante o processo de triagem e análise dos estudos identificados na revisão.

Após a tradução, os resumos dos artigos foram submetidos a um processo de síntese assistido por LLMs. Essa etapa teve como objetivo gerar representações estruturadas dos conteúdos dos artigos, destacando elementos relevantes como o problema de pesquisa, o tema central e os principais tópicos abordados em cada estudo.

As sínteses geradas foram utilizadas como apoio à análise preliminar dos trabalhos recuperados, facilitando a identificação de padrões temáticos e a classificação

inicial dos artigos. É importante destacar que o uso de LLMs teve caráter exclusivamente auxiliar nesse processo, permanecendo a interpretação final dos conteúdos e a decisão sobre a relevância dos estudos sob responsabilidade do pesquisador.

Essa etapa contribuiu para tornar o processo de análise da literatura mais sistemático e eficiente, especialmente diante do volume de artigos analisados, permitindo estruturar de forma mais clara as informações necessárias para as etapas subsequentes da revisão.

2.5.5 Resultados da Análise Bibliométrica

A partir do conjunto de artigos obtido após as etapas de extração, consolidação e normalização dos metadados, foi realizada uma análise bibliométrica com o objetivo de compreender as características da produção científica relacionada aos dois eixos investigados nesta pesquisa. Essa análise buscou identificar padrões temporais de publicação, distribuição geográfica dos estudos, principais palavras-chave utilizadas e temas recorrentes associados ao backlog do produto e ao uso de técnicas de inteligência artificial na análise de requisitos textuais.

Os resultados são apresentados a seguir de forma separada para cada eixo investigado, permitindo compreender tanto a maturidade da literatura sobre os desafios associados à construção do backlog do produto quanto o estado da arte das pesquisas que exploram o uso de modelos de linguagem nesse contexto.

No primeiro eixo, voltado à identificação de estudos relacionados aos desafios associados ao backlog do produto, o processo resultou em um conjunto consolidado de artigos únicos após as etapas de exportação, deduplicação e processamento automatizado. De modo análogo, no segundo eixo, dedicado ao uso de modelos de linguagem na análise de requisitos, obteve-se um conjunto específico de estudos recentes, posteriormente submetido à avaliação do pesquisador. Os quantitativos detalhados de cada etapa estão apresentados na Tabela 3.

A seguir são apresentados os resultados da análise bibliométrica referentes ao primeiro eixo da revisão, voltado à identificação de estudos que discutem desafios associados ao backlog do produto.

2.5.6 Análise Bibliométrica do Eixo 1: Problemas Associados ao Product Backlog

A aplicação da estratégia de busca associada ao primeiro eixo da revisão, voltado à identificação de estudos relacionados ao uso e aos desafios do *Product Backlog*, resultou inicialmente na recuperação de 254 registros nas bases bibliográficas consultadas.

Após a aplicação dos filtros de exportação e consolidação dos metadados (artigos sem DOI) provenientes das bases Scopus, Web of Science e IEEE Xplore, foram obtidos 223 registros aptos para processamento inicial. A etapa subsequente de consolidação realizada pelo pipeline computacional permitiu identificar registros duplicados decorrentes da indexação de um mesmo artigo em múltiplas bases bibliográficas. Como resultado desse procedimento foram removidas duplicidades, resultando em um conjunto final de 146 artigos únicos considerados para análise no primeiro eixo da revisão. Ainda no processo de filtragem dos artigos para o acervo bibliográfico da pesquisa neste eixos foram selecionados artigos com mais de 10 citações, pois trata-se de um tema amplamente avaliado e a intensão com este filtro é a de obter os artigos mais relevantes.



Figura 9 – Fluxograma de seleção de artigos Eixo 1

Por fim os artigos com possibilidade de avaliação foram os que disponibilizaram no formato PDF para poder avaliar o conteúdo. Foram ao todo 35 artigos disponíveis.

A distribuição temporal dessas publicações indica um crescimento gradual do interesse acadêmico pelo tema ao longo dos últimos anos.

Observa-se uma concentração maior de trabalhos a partir de 2018, período que coincide com a consolidação das metodologias ágeis em diferentes contextos organizacionais e com a difusão de práticas baseadas em Scrum e desenvolvimento iterativo. Esse movimento sugere que o backlog do produto passou a receber maior atenção acadêmica à medida que as organizações ampliaram a adoção de métodos ágeis em projetos de desenvolvimento de software.

revelou que um dos tópicos mais recorrentes na literatura está relacionado aos desafios associados à construção e à gestão do backlog do produto. Entre os problemas discutidos destacam-se dificuldades relacionadas à qualidade das descrições de *user stories*, inconsistências na definição de critérios de aceitação, limitações nos processos de refinamento e desafios na manutenção da clareza e rastreabilidade dos requisitos ao longo do ciclo de desenvolvimento.

Os principais estudos identificados que discutem esses desafios são apresentados na Tabela 3. Esses trabalhos abordam diferentes perspectivas relacionadas à conceituação do backlog, à estruturação das *user stories*, ao uso de métricas para monitoramento do desenvolvimento ágil e às dificuldades associadas à decomposição e evolução de requisitos em ambientes iterativos.

Tabela 3 – Principais estudos que discutem desafios associados ao *Product Backlog*

Autores	Tipo de estudo	Principais contribuições
Conceituação e papel do backlog em ambientes ágeis		
(Sedano; Ralph; Péraire, 2019)	Grounded Theory	Define o backlog como um modelo de trabalho informal e um objeto de fronteira essencial para a comunicação da equipe.
(Jakobsen; Johnson, 2008)	Estudo de caso	Introduz checklists de inspeção baseados na disciplina do CMMI para avaliar a qualidade e a completude das histórias.
(Jakobsen; Sutherland, 2009)	Estudo de caso	Destaca a importância de equilibrar a preparação do backlog com a execução das sprints usando um checklist de 'ready-ready'.
Estruturação e qualidade das <i>user stories</i>		
(Dilorenzo et al., 2020)	Métodos mistos	Cria uma taxonomia formal com as facetas 'módulo' e 'operação' para classificar os itens do backlog.
(Gilson; Irwin, 2018)	Proposta de framework	Propõe um modelo de ML para extrair automaticamente atributos de qualidade implícitos nas histórias.
(Müter et al., 2019)	Análise linguística	Analisa estruturalmente as tarefas e propõe diretrizes baseadas em famílias de 'verbos de ação'.
Gestão do backlog, métricas e monitoramento		
(Boerman et al., 2015)	Estudo de caso (GQM)	Define um modelo com 11 métricas derivadas diretamente dos dados do backlog.
(Amjad et al., 2017)	Revisão e experimento	Desenvolve o método A-SPSRI para medir objetivamente a completude do escopo.
(Erdoğan; Pekkaya; Gök, 2018)	Estudo de caso	Análises estatísticas aplicadas a métricas do backlog para melhorar retrospectivas e estimativas.
Planejamento, decomposição e evolução de requisitos		
(Grapenthin et al., 2015)	Pesquisa-ação	Apresenta o método 'Interaction Room' para análise colaborativa profunda.
(Petersen; Wohlin, 2010)	Estudo de caso	Evidencia que o uso de um backlog flexível diminui drasticamente o desperdício de requisitos.
(Rola; Kuchta, 2019)	Proposta de método	Baseado em números fuzzy para que os especialistas estimem o esforço.
(Alsalemi; Yeoh, 2016)	Survey industrial	Investiga volatilidade de requisitos e desafios na gestão de mudanças e rastreabilidade.
(Küpper et al., 2019)	Estudo multi-método	Identifica que o backlog é prático para orientar e suportar iniciativas contínuas de SPI.

Fonte: Elaborado pelo autor.

Em conjunto, os resultados do primeiro eixo evidenciam que a literatura reconhece o backlog do produto como artefato central da coordenação em ambientes ágeis, mas também aponta limitações persistentes relacionadas à qualidade das descrições, à consistência semântica e ao refinamento sistemático das informações nele registradas.

Essa constatação reforça a necessidade de mecanismos capazes de apoiar a avaliação estruturada da qualidade dos itens de backlog, especialmente em contextos organizacionais nos quais múltiplas equipes dependem de entendimento compartilhado e comparabilidade entre artefatos.

2.5.7 Análise Bibliométrica do Eixo 2: Uso de LLMs na Análise de Requisitos

O segundo eixo da revisão teve como objetivo identificar estudos que exploram o uso de inteligência artificial, particularmente LLMs, no tratamento de requisitos textuais associados ao desenvolvimento de software.

A aplicação da estratégia de busca resultou em um conjunto específico de registros recuperados nas bases bibliográficas consultadas, posteriormente submetido ao processo de consolidação e remoção de duplicidades realizado pelo pipeline computacional. Os quantitativos detalhados dessa etapa encontram-se sintetizados na Figura 13.

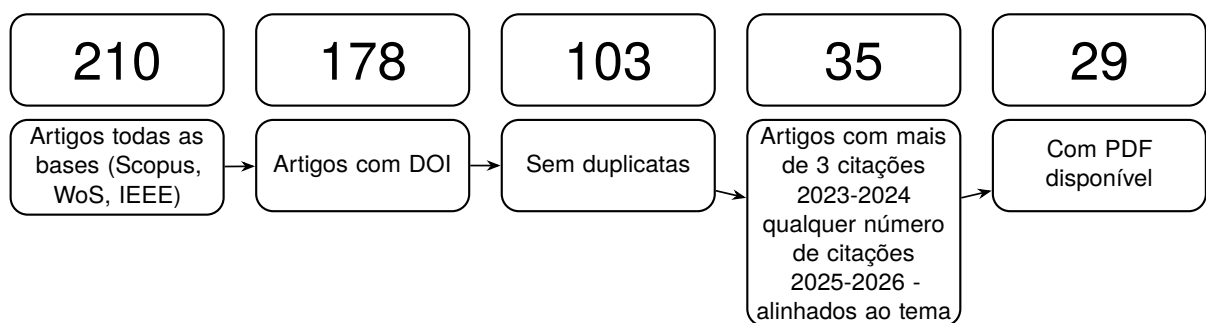


Figura 13 – Fluxograma de seleção de artigos Eixo 2

A distribuição temporal dessas publicações evidencia que se trata de um campo de investigação recente.

Observa-se um crescimento expressivo no número de trabalhos publicados a partir de 2024, seguido de uma ampliação significativa da produção científica nos anos subsequentes. Esse comportamento acompanha a rápida difusão dos modelos de linguagem generativa e o aumento do interesse da comunidade científica em explorar suas aplicações em diferentes atividades relacionadas à engenharia de software.

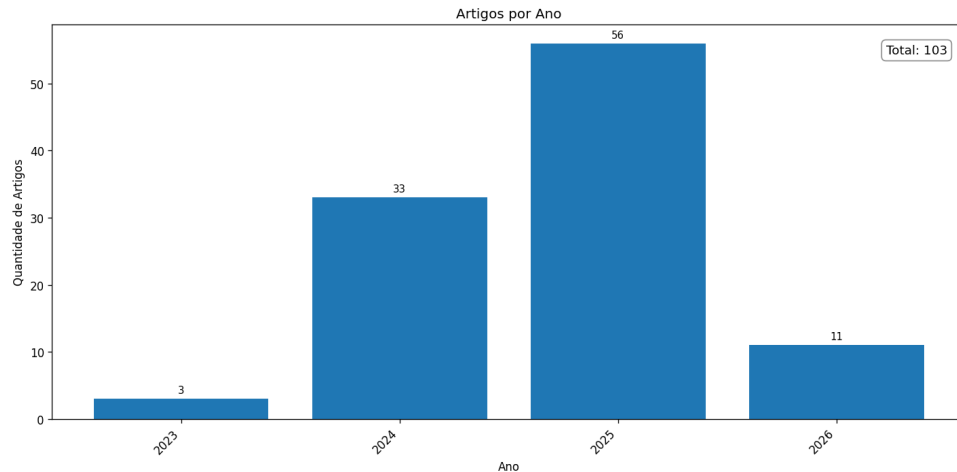


Figura 14 – Evolução anual das publicações de artigos Eixo 2

Fonte: Elaborado pelo autor

A análise das palavras-chave mais recorrentes reforça essa tendência. Entre os termos mais frequentes identificados na base analisada destacam-se *Large Language Models*, *User Stories*, *Requirements Engineering*, *Natural Language Processing* e *ChatGPT*. Esses resultados indicam que a literatura recente tem investigado de forma crescente o uso de técnicas de processamento de linguagem natural e inteligência artificial para apoiar o tratamento de requisitos descritos em linguagem natural.



Figura 15 – Nuvem de palavra Eixo 2

Fonte: Elaborado pelo autor

Localidades por País (Mapa)

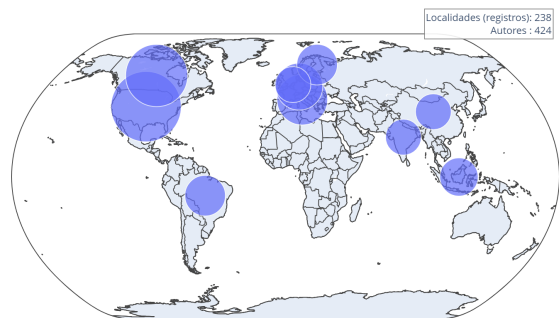


Figura 16 – Países publicações - Eixo 2

Fonte: Elaborado pelo autor

Do ponto de vista metodológico, observa-se predominância de estudos voltados à proposição de métodos e abordagens conceituais. Também aparecem com frequência pesquisas que propõem frameworks ou modelos conceituais para utilização de LLMs na engenharia de requisitos, além de estudos de caso que exploram aplicações dessas tecnologias em ambientes organizacionais ou educacionais.

Trabalhos focados especificamente no desenvolvimento de artefatos computacionais ou em avaliações experimentais mais robustas ainda são menos numerosos, o

que sugere que a área se encontra em fase inicial de maturação científica.

A análise da distribuição geográfica dos estudos revela que a pesquisa sobre o uso de LLMs na engenharia de requisitos apresenta caráter global. Entre os países com maior número de publicações destacam-se Estados Unidos, Canadá, Brasil, Finlândia e Países Baixos.

Apesar do crescimento recente da literatura relacionada ao uso de modelos de linguagem na engenharia de software, observa-se que a aplicação direta dessas tecnologias na análise da qualidade de artefatos de backlog ainda é relativamente limitada.

Embora existam estudos voltados à geração ou avaliação de *user stories*, bem como pesquisas que investigam o uso de LLMs na extração ou estruturação de requisitos, ainda são escassas as abordagens que utilizam essas tecnologias de forma sistemática no contexto da gestão de backlog em ambientes ágeis.

Os principais estudos identificados nesse eixo da revisão são sintetizados na Tabela 4, que apresenta uma visão comparativa dos conceitos explorados, dos resultados reportados e das limitações identificadas na literatura.

Tabela 4 – Estudos recentes sobre modelos de linguagem na análise de requisitos: Perspectiva Técnica

Artigo	Abordagem Técnica	Resultados positivos	Limitações identificadas
Sistemas Multiagentes e Colaboração entre LLMs			
(Zhang et al., 2024)	Sistema multiagente para aprimoramento automático de <i>user stories</i> fundamentado no framework INVEST.	Melhorias em clareza e valor de negócio em contexto industrial real.	Histórias tornaram-se excessivamente longas; necessidade de supervisão humana constante.
(Lotfi; Nafil, 2025)	Arquitetura de múltiplos agentes (Papéis de PO e QA) para avaliar e reescrever US via INVEST e QUS.	Ganhos em completude e testabilidade, preservando a intenção original.	Subjetividade nas métricas geradas pelo LLM; falha em entradas muito vagas ou técnicas.
Engenharia de Prompt: Chaining, Few-Shot e Meta-Prompting			
(Sharma; Chaturvedi; Tripathi, 2024)	Uso de <i>prompt chaining</i> (4 etapas sequenciais) para extrair US estruturadas de textos informais.	Precisão de 94% na extração de entidades (papéis e funções).	Inferências contextuais geraram falsos positivos; perda de conhecimento tácito.
(Santos et al., 2025)	Avaliação comparativa entre <i>Free-form</i> e <i>Meta-Few-Shot Prompting</i> para geração de US.	Técnica <i>Meta-Few-Shot</i> atingiu 88,57% de sucesso, superando métodos manuais.	Dificuldade em garantir atomicidade sintática; risco recorrente de alucinações.

Continua na próxima página...

(Brockenbrough; Salinas, 2024)	Uso de IA generativa baseada em <i>prompts</i> educacionais para criação de US e rastreabilidade.	Redução no tempo de elaboração e ganhos em gramática e coesão textual.	Inconsistências estruturais entre execuções e dificuldade em gerar <i>Definition of Done</i> .
Modelos Híbridos: NLP Tradicional, BERT(Gemini) e Machine Learning			
(Alshammari, 2025)	Comparação entre NLP tradicional e LLMs para conversão de requisitos em modelos estruturados.	Desempenho superior a ferramentas tradicionais em extração conceitual com baixo custo.	Latência elevada e dependência de bases de dados anotadas manualmente para validação.
(Chuur; Ittoo; Heng, 2024)	Classificação de taxonomia de US comparando ML Clássico, <i>embeddings</i> BERT e LLMs (Mistral/Llama).	<i>Features</i> do BERT otimizaram o ML clássico em classes minoritárias.	LLMs (<i>few-shot</i>) performaram muito abaixo do ML clássico no desbalanceamento de dados.
Benchmarks e Avaliação de Modelos (LLM vs. Humano)			
(Ronanki; Cabrero-Daniel; Berger, 2024)	Estudo comparativo de <i>benchmarking</i> entre ChatGPT, ferramenta AQUA e avaliadores humanos.	ChatGPT superou ferramentas especializadas com alta concordância humana.	Instabilidade nas respostas (<i>non-determinism</i>) e risco de uso por leigos.
(Svensson et al., 2019)	Método diagnóstico de dimensões (escopo, maturidade) aplicado como <i>framework</i> de avaliação.	Viabilidade e rapidez na avaliação diagnóstica para suporte à decisão.	Não automatiza a correção; limitações de validade externa para diferentes domínios.

2.5.8 Síntese da Revisão, Lacuna de Pesquisa e Posicionamento deste Trabalho

A análise conjunta dos dois eixos investigados nesta revisão permite compreender de forma mais abrangente como a literatura científica tem tratado o backlog do produto no contexto do desenvolvimento ágil e quais são as iniciativas recentes voltadas à utilização de técnicas de inteligência artificial para apoiar o tratamento de requisitos textuais.

Os estudos analisados no primeiro eixo evidenciam que o backlog do produto ocupa um papel central nos processos de desenvolvimento ágil, atuando como o principal artefato responsável por organizar, priorizar e comunicar os requisitos do sistema ao longo das iterações. Entretanto, a literatura também demonstra que a construção e manutenção desse artefato estão associadas a diferentes desafios práticos.

Entre os problemas mais recorrentes destacam-se dificuldades relacionadas à qualidade das descrições de *user stories*, inconsistências na definição de critérios de aceitação, limitações nos processos de refinamento e desafios associados à manu-

tenção de clareza e rastreabilidade das informações de requisitos ao longo do ciclo de desenvolvimento.

Esses resultados indicam que, embora o backlog seja amplamente reconhecido como elemento central na gestão ágil de produtos, sua utilização efetiva ainda depende de mecanismos capazes de apoiar a qualidade e a consistência das informações nele registradas.

De forma mais ampla, os estudos analisados também revelam a presença de uma tensão conceitual recorrente entre a flexibilidade inerente às metodologias ágeis e o nível de disciplina tradicionalmente associado à engenharia de requisitos. Enquanto as abordagens ágeis enfatizam adaptação contínua e comunicação informal entre equipes, a engenharia de requisitos enfatiza a necessidade de clareza, consistência e rastreabilidade na especificação das funcionalidades do sistema.

Por outro lado, os estudos analisados no segundo eixo demonstram que a literatura tem avançado de forma consistente na proposição de abordagens voltadas à avaliação e à melhoria da qualidade de artefatos ágeis, com potencial para apoiar tarefas como geração, análise e avaliação de requisitos descritos em linguagem natural. Destacam-se, nesse sentido, modelos fundamentados em critérios como o INVEST e práticas derivadas do *Behavior-Driven Development* (BDD), amplamente utilizados como referência para a estruturação de histórias de usuário e critérios de aceitação, além de *frameworks* conceituais, métricas de qualidade e instrumentos baseados em *checklists* voltados à redução da subjetividade inerente à escrita e à interpretação dos requisitos. Paralelamente, verifica-se uma crescente incorporação de técnicas de processamento de linguagem natural (NLP) e, mais recentemente, de LLMs, ampliando as possibilidades de análise automatizada de artefatos textuais e de identificação de inconsistências. Apesar desse avanço, a literatura também aponta limitações importantes, como instabilidade nas respostas dos modelos, necessidade de supervisão humana e ausência de métodos estruturados que integrem essas tecnologias de forma sistemática aos processos de gestão de backlog.

Essas limitações se somam a lacunas mais amplas de escalabilidade e de integração dessas soluções em contextos organizacionais reais. Em ambientes de agilidade em escala, como aqueles estruturados pelo SAFe, surgem desafios adicionais relacionados à coordenação entre múltiplas equipes e à necessidade de consistência nas práticas de planejamento, de modo que muitos estudos permanecem restritos

a experimentos controlados ou aplicações pontuais, com limitada evidência de uso contínuo ao longo de diferentes ciclos. Observa-se, ainda, a ausência de abordagens que integrem, de forma unificada, mecanismos de avaliação automatizada, geração de *feedback* acionável e instrumentos de governança visual, evidenciando a oportunidade de desenvolvimento de soluções mais abrangentes, capazes de transformar as descrições textuais do backlog em informações analíticas úteis ao processo de gestão de produtos e integradas ao fluxo de trabalho organizacional.

A análise consolidada da literatura permite enunciar com precisão a lacuna que esta pesquisa se propõe a endereçar: **existem abordagens para apoiar a escrita de histórias de usuário e estudos sobre o uso de LLMs em atividades de engenharia de requisitos, mas são escassos os métodos estruturados e escaláveis voltados à avaliação sistemática de artefatos textuais já existentes no backlog**, com critérios explícitos, escalas mensuráveis e capacidade de operação contínua em ambientes com múltiplas equipes. Essa lacuna é simultaneamente teórica, pela ausência de método formalizado, e prática, pela ausência de instrumento operacionalizável em escala, e fundamenta a proposição do *Backlog Analyzer* como resposta prescritiva no contexto da *Design Science Research*.

Diante desse cenário, o backlog do produto emerge como o artefato central em que essa tensão se manifesta, evidenciando a necessidade de mecanismos capazes de avaliar, de forma sistematizada, a qualidade das descrições dos requisitos nele registrados ao longo do planejamento ágil. Esta pesquisa propõe, assim, o desenvolvimento de um método estruturado para essa avaliação, com apoio de LLMs, materializado por meio do artefato *Backlog Analyzer*. Ao propor e avaliar esse artefato, a pesquisa busca aproximar o potencial das tecnologias baseadas em LLMs das necessidades práticas de gestão de backlog em ambientes de desenvolvimento ágil, contribuindo para o avanço do estado da arte na avaliação sistemática de artefatos textuais de requisitos.

Na perspectiva da *Design Science Research*, a lacuna identificada pode ser compreendida como pertencente a uma classe de problema mais ampla, cuja decomposição em elementos constitutivos permite explicitar a instância empírica investigada, os artefatos textuais envolvidos, os problemas observados e suas consequências organizacionais. A Figura 17 apresenta essa síntese conceitual, elaborada a partir da literatura e da modelagem analítica conduzida neste estudo.

Classe de Problema

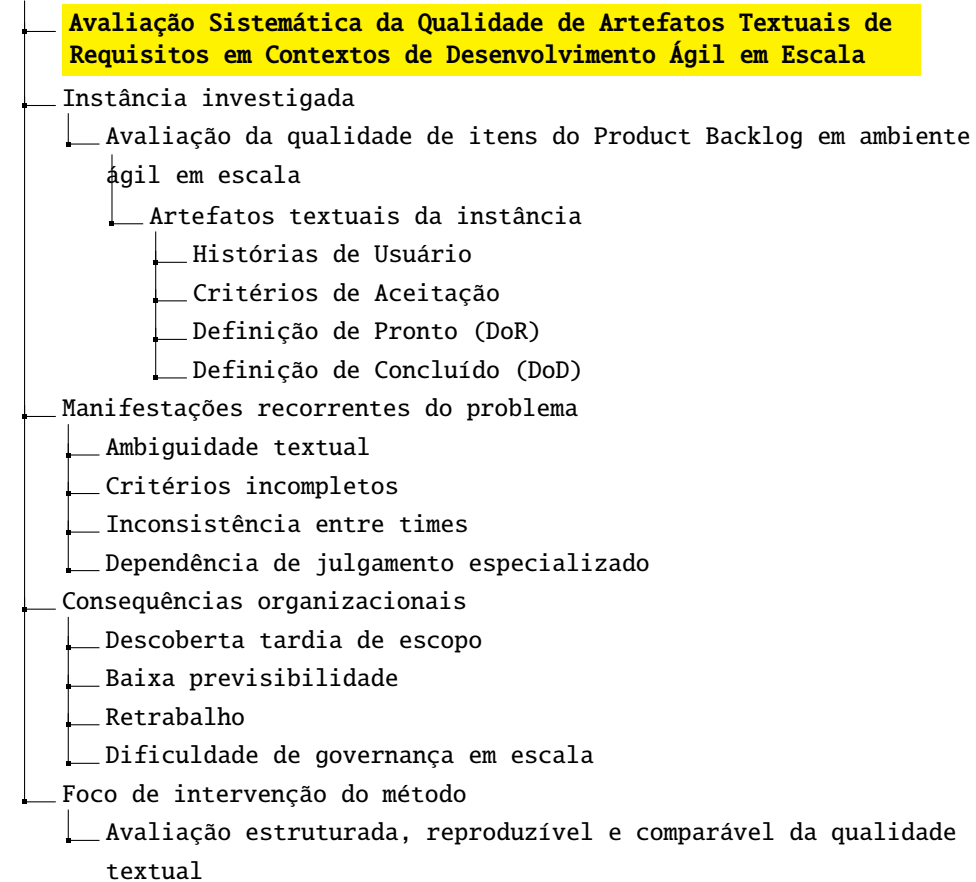


Figura 17 – Estrutura conceitual da classe de problema

Fonte: Elaborado pelo autor.

A classe de problema sintetizada na Figura 17 evidencia a necessidade de uma intervenção estruturada, reproduzível e comparável, direcionando a estratégia de pesquisa apresentada no Capítulo 3.

3 METODOLOGIA

3.1 Caracterização da Pesquisa

Com base na lacuna identificada no Capítulo 2 — a escassez de métodos estruturados e escaláveis para a avaliação sistemática da qualidade de artefatos textuais do *Product Backlog* —, este capítulo apresenta o percurso metodológico adotado para o desenvolvimento e a avaliação do artefato proposto, o *Backlog Analyzer*.

Diante desse contexto, esta pesquisa adota a abordagem de *Design Science Research* (DSR), uma vez que o objetivo central do estudo consiste no desenvolvimento e avaliação de um artefato destinado a apoiar a solução do problema identificado.

A lógica metodológica da pesquisa pode ser sintetizada conforme a Figura 18, que apresenta o encadeamento entre o problema identificado, a lacuna evidenciada na literatura, o desenvolvimento do artefato proposto e sua posterior avaliação em contexto organizacional.

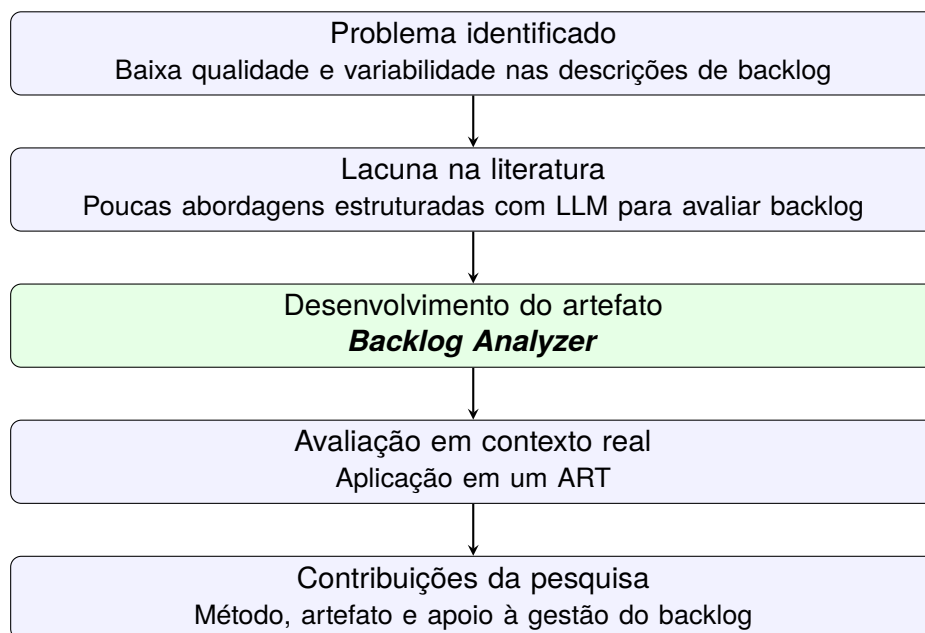


Figura 18 – Encadeamento metodológico da pesquisa à luz da Design Science Research.

Fonte: Elaborado pelo autor.

A presente pesquisa caracteriza-se como de natureza aplicada, pois tem como propósito a proposição, implementação e avaliação de uma solução estruturada para um problema em contexto organizacional real, associado à avaliação da qualidade do *Product Backlog* em ambientes ágeis. Diferentemente de estudos limitados exclusivamente à compreensão e descrição teórica de fenômenos, este trabalho busca intervir

sistematicamente sobre a prática. Sob essa ótica, [Dresch, Lacerda e Antunes Junior \(2015\)](#) destacam que a chamada Ciência do Artificial concentra-se na concepção de soluções capazes de modificar situações existentes, contrastando com abordagens científicas voltadas predominantemente à explicação de fenômenos.

O estudo também apresenta uma natureza simultaneamente exploratória e prescritiva, adotando uma abordagem qualitativa voltada à compreensão da interação entre o artefato proposto e o contexto organizacional em que será aplicado. Conforme argumenta [Wieringa \(2014\)](#), a pesquisa em *Design* não se limita à análise de tecnologias isoladas, mas busca compreender como essas soluções interagem com processos, pessoas e estruturas organizacionais.

Embora o método desenvolvido incorpore elementos quantitativos, como escalas de mensuração e indicadores estruturados, o objetivo central da pesquisa é prescrever formas de intervenção capazes de mitigar problemas observados no contexto organizacional analisado. Para materializar essa proposta, desenvolve-se um método estruturado de avaliação da qualidade do backlog que, de acordo com a taxonomia de artefatos proposta na literatura de Sistemas de Informação, é instanciado por meio de um sistema computacional. Essa instanciación baseada em Inteligência Artificial busca transformar percepções frequentemente subjetivas sobre a qualidade das descrições de backlog em diagnósticos estruturados, rastreáveis e reproduzíveis.

Além de sua contribuição para a área de engenharia de requisitos e para o uso de modelos de linguagem em análise textual, esta pesquisa também se insere no domínio da Engenharia de Produção ao abordar problemas relacionados à gestão e à qualidade de processos organizacionais. A avaliação estruturada de artefatos de backlog pode ser compreendida como um mecanismo de apoio à governança do processo de desenvolvimento de software, permitindo transformar descrições textuais frequentemente subjetivas em indicadores estruturados que apoiam a tomada de decisão, a melhoria contínua e o aprendizado organizacional. Dessa forma, o artefato proposto contribui para o desenvolvimento de instrumentos analíticos voltados à gestão de processos intensivos em conhecimento, tema amplamente discutido na literatura de Engenharia de Produção.

Em síntese, a pesquisa fundamenta-se na abordagem de *Design Science Research*, uma vez que seu escopo ultrapassa a mera análise do fenômeno e envolve a construção e avaliação empírica de um artefato voltado à solução do problema

identificado. O rigor desse arcabouço metodológico permite assegurar que a solução proposta apresente utilidade prática no contexto organizacional estudado e, ao mesmo tempo, contribua para o avanço do conhecimento científico na área (Vom Brocke; Hevner; Maedche, 2020).

3.2 Design Science Research como Estratégia Metodológica

A presente pesquisa adota a abordagem de *Design Science Research* (DSR) como estratégia metodológica central. Segundo Dresch, Lacerda e Antunes Junior (2015), trata-se de um método estruturado voltado à criação e à avaliação de artefatos destinados à resolução de problemas práticos. Diferentemente de abordagens predominantemente descritivas ou explicativas, características das ciências tradicionais, a DSR possui caráter construtivo e prescritivo, buscando não apenas compreender fenômenos organizacionais, mas também intervir sistematicamente sobre eles por meio de soluções viáveis.

Nessa perspectiva, Hevner et al. (2004) e Vom Brocke, Hevner e Maedche (2020) destacam que pesquisas orientadas ao *design* devem equilibrar rigor científico e relevância prática. Isso implica articular fundamentos conceituais consolidados com problemas organizacionais concretos, produzindo conhecimento aplicável à realidade investigada. Diante disso, a contribuição científica materializa-se no desenvolvimento de artefatos que, conforme a taxonomia clássica da área de Sistemas de Informação, podem assumir a forma de constructos, modelos, métodos ou instâncias (March; Smith, 1995; Dresch; Lacerda; Antunes Junior, 2015).

Para Wieringa (2014), a validade científica desses artefatos decorre tanto de sua fundamentação teórica quanto da avaliação de sua utilidade em contextos reais de aplicação. Assim, a pesquisa em *design* busca compreender como soluções tecnológicas interagem com processos organizacionais, atores humanos e estruturas institucionais.

Para operacionalizar a condução metodológica desta investigação, adotou-se como referência o processo de pesquisa em DSR apresentado por Dresch, Lacerda e Antunes Junior (2015), aqui interpretado à luz das necessidades desta pesquisa. Em termos operacionais, o processo investigativo compreende: (i) identificação do problema; (ii) conscientização do problema com apoio da literatura; (iii) definição da

classe de problema e do tipo de artefato; (iv) proposição da solução; (v) definição de sua estrutura conceitual; (vi) desenvolvimento da instanciação; (vii) demonstração e avaliação empírica; (viii) explicitação das aprendizagens decorrentes da aplicação; e (ix) consolidação das conclusões da pesquisa.

Embora apresentadas de forma sequencial, essas etapas constituem um processo iterativo de investigação. Conforme argumenta [Wieringa \(2014\)](#), a compreensão do problema e o refinamento do artefato evoluem de maneira conjunta ao longo da pesquisa, exigindo revisões sucessivas entre concepção, desenvolvimento e avaliação da solução.

A Figura 19 apresenta o ciclo da *Design Science Research* adotado nesta pesquisa.

O ciclo metodológico descrito também orienta a organização estrutural desta dissertação. Inicialmente, o Capítulo 1 apresenta a identificação e delimitação do problema de pesquisa, evidenciando as lacunas existentes na avaliação da qualidade do *Product Backlog* em ambientes ágeis em escala.

Na sequência, o Capítulo 2 desenvolve a fundamentação teórica que sustenta a conscientização do problema, abordando conceitos relacionados à engenharia de requisitos, gestão de backlog, práticas de desenvolvimento ágil e fundamentos da *Design Science Research*. Esses elementos fornecem a base conceitual necessária para a definição dos requisitos e princípios de construção do artefato proposto.

A proposição conceitual do artefato e a definição de seus elementos estruturantes são apresentadas no Capítulo 3, no qual se descreve a estratégia metodológica adotada, a classe de problema investigada e o modelo conceitual que orienta o desenvolvimento da solução.

O Capítulo 4 apresenta o detalhamento da implementação e da instanciação computacional do artefato, incluindo a arquitetura do sistema, os mecanismos de processamento utilizados e a operacionalização do método proposto.

A demonstração e avaliação do artefato em contexto organizacional real são apresentadas no Capítulo 5, no qual se analisa sua aplicação em um ambiente de desenvolvimento ágil em escala. Por fim, o Capítulo 6 consolida as conclusões da pesquisa, discutindo as contribuições científicas e práticas do estudo, bem como suas limitações e possibilidades de pesquisas futuras.

Além da correspondência entre as etapas da *Design Science Research* e a

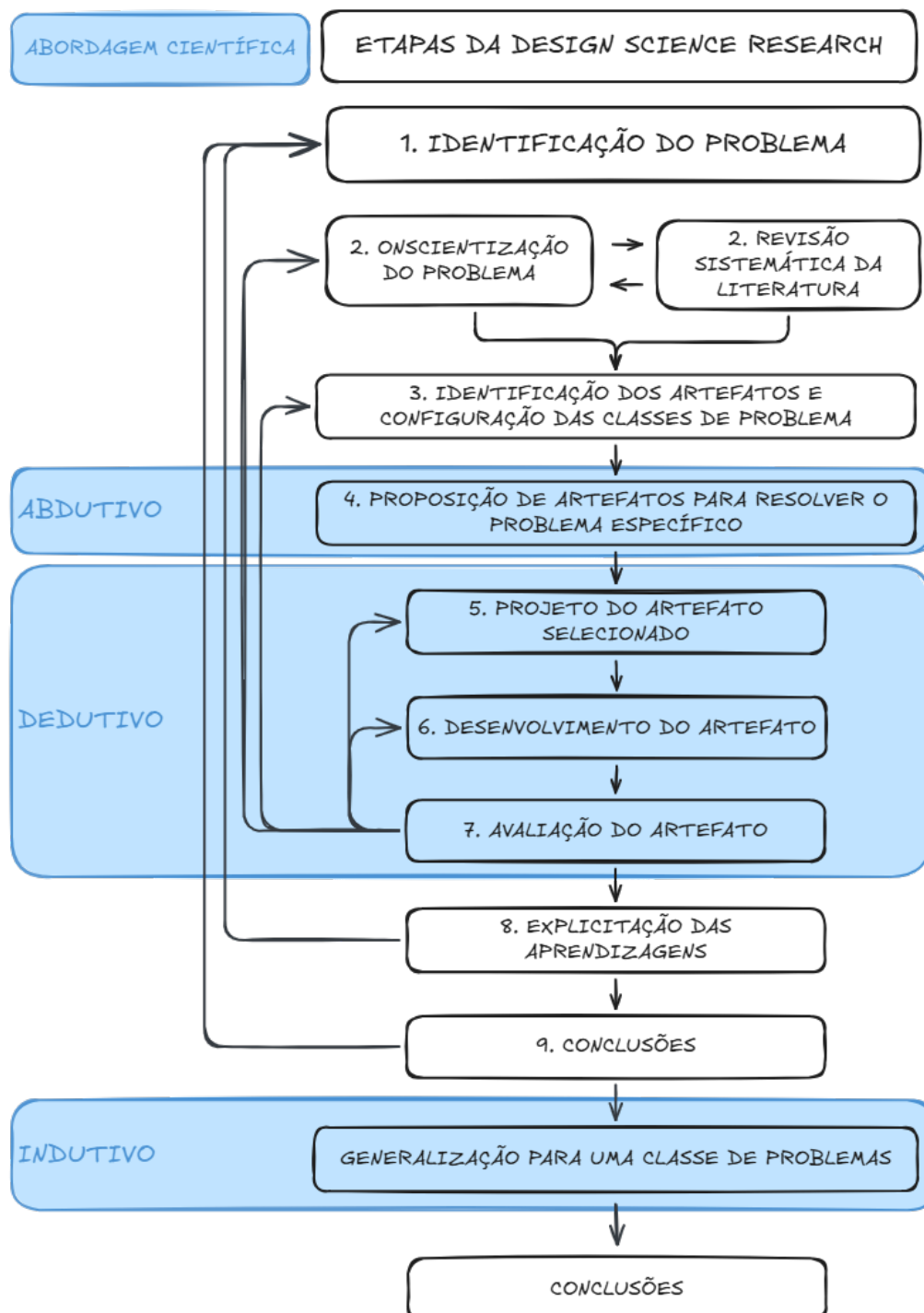


Figura 19 – Ciclo da Design Science Research adotado na pesquisa

Fonte: Adaptado de Dresch, Lacerda e Antunes Junior (2015).

estrutura desta dissertação, é importante explicitar o encadeamento lógico entre a classe de problema investigada, os constructos definidos, os requisitos do artefato, seu modelo conceitual, sua instanciação e sua posterior avaliação. A Figura 20 apresenta essa síntese do fluxo metodológico da pesquisa.

Tabela 5 – Correspondência entre as etapas da Design Science Research e sua aplicação na dissertação

Etapas DSR (Dresch et al.)	Capítulos / Seções	Evidências do estudo
1. Identificação do problema	Capítulo 1	Delimitação do problema relacionado à variabilidade e à baixa qualidade estrutural dos itens do <i>Product Backlog</i> em ambientes ágeis em escala, com impactos sobre clareza, previsibilidade e coordenação do trabalho.
2. Conscientização do problema e fundamentação teórica	Capítulo 2	Revisão da literatura sobre gestão de backlog, histórias de usuário, critérios de qualidade de requisitos, uso de inteligência artificial em engenharia de requisitos e fundamentos da <i>Design Science Research</i> , estabelecendo a base conceitual da pesquisa.
3. Definição da classe de problema e do artefato	Seção 3.3	Delimitação da classe de problema relacionada à avaliação sistemática da qualidade de itens do <i>Product Backlog</i> em ambientes ágeis em escala e definição do método proposto como artefato prescritivo para enfrentar esse problema.
4. Proposição do artefato	Seções 3.4 e 3.5	Derivação dos requisitos e definição do modelo conceitual do artefato <i>Backlog Analyzer</i> , concebido para analisar descrições de backlog e gerar indicadores estruturados de qualidade.
5. Projeto do artefato	Seção 3.5	Definição da arquitetura conceitual do artefato, incluindo seus componentes de entrada, processamento estruturado e geração de evidências para apoio à decisão.
6. Desenvolvimento da solução	Capítulo 4	Implementação do artefato por meio de um pipeline computacional baseado em modelos de linguagem, processamento textual e consolidação de resultados em bases estruturadas e dashboards analíticos.
7. Demonstração e avaliação empírica	Seção 3.6 e Capítulo 5	Definição da estratégia de demonstração e avaliação do artefato e análise de sua aplicação em um ART, incluindo métricas de qualidade do backlog e percepções dos participantes.
8. Explicitação das aprendizagens	Capítulo 5	Discussão dos resultados da aplicação do artefato, destacando implicações práticas, limitações observadas e fatores contextuais que influenciam a qualidade do backlog.
9. Conclusões	Capítulo 6	Síntese das contribuições científicas e práticas da pesquisa, discussão das limitações do estudo e reflexão sobre a possibilidade de generalização analítica do método para contextos organizacionais semelhantes.
10. Generalização	Seção 6.3	Proposição da extensão do método à instância “avaliação da qualidade de <i>Features</i> ” no contexto SAFe, pertencente à mesma classe de problema investigada.

3.3 Classe de Problema e Constructos da Pesquisa

Em pesquisas orientadas pela *Design Science Research* (DSR), a explicitação da classe de problema constitui um elemento central para delimitar o escopo da contribuição científica e evitar generalizações indevidas. Conforme argumentam Dresch, Lacerda e Antunes Junior (2015), diferentemente de um problema isolado e circuns-

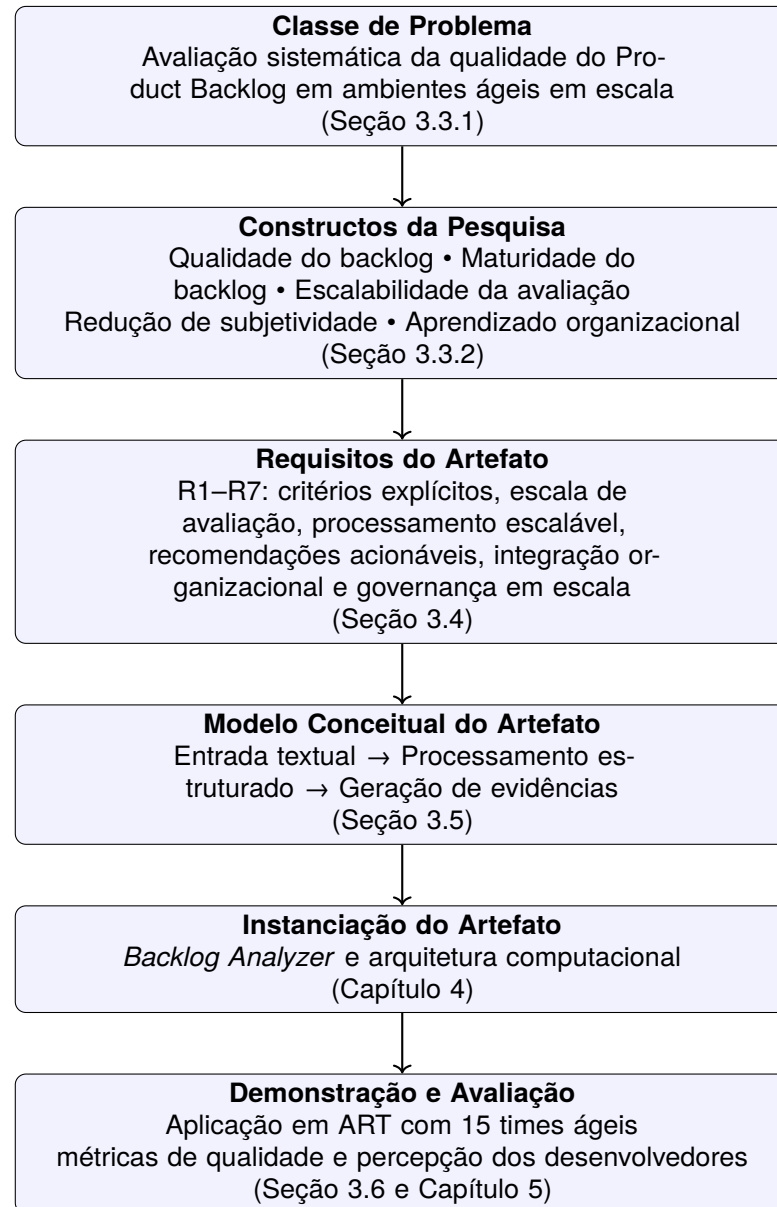


Figura 20 – Fluxo metodológico da pesquisa à luz da Design Science Research

tancial, a classe de problema representa um conjunto de situações organizacionais que compartilham características estruturais semelhantes e que demandam soluções fundamentadas em princípios comuns.

3.3.1 Classe de problema

Apoiado nessa premissa teórica, o presente estudo define sua classe de problema como a ausência de um método estruturado, padronizado e escalável para a avaliação da qualidade e da maturidade de itens de *Product Backlog* em ambientes ágeis em escala. Essa classe de problema manifesta-se tipicamente em organizações

que adotam práticas de agilidade em escala com múltiplas equipes, utilizam histórias de usuário como principal artefato de especificação e enfrentam variações significativas na qualidade dos itens de backlog. Em muitos desses ambientes, a avaliação da prontidão e da maturidade dos itens depende predominantemente de julgamentos subjetivos realizados por membros das equipes ou por papéis de liderança técnica. Como consequência, torna-se difícil estabelecer padrões consistentes de qualidade, monitorar a evolução do backlog ao longo do tempo ou comparar resultados entre diferentes equipes. A ausência de indicadores estruturados limita, portanto, a capacidade de aprendizado organizacional e de melhoria sistemática dos processos de planejamento. No contexto empírico desta pesquisa, essa situação foi observada em um ART composto por quinze times ágeis. Entretanto, a contribuição proposta não se restringe à resolução desse caso específico. O objetivo é propor um método aplicável a contextos organizacionais que compartilhem características estruturais semelhantes, como governança ágil em escala, utilização sistemática de histórias de usuário e necessidade de avaliação estruturada da qualidade do backlog. A contribuição científica desta pesquisa insere-se no domínio da generalização analítica, característica de estudos baseados em *Design Science Research*. Não se trata de generalização estatística para uma população ampla, mas da proposição de um método fundamentado teoricamente e demonstrado empiricamente em um contexto organizacional complexo. Assim, a estrutura metodológica proposta pode ser replicada em outros ARTs da mesma organização que compartilhem características estruturais semelhantes, como governança ágil em escala, uso sistemático de histórias de usuário e necessidade de avaliação padronizada da qualidade do *Product Backlog*. Ainda assim, sua aplicação em contextos organizacionais distintos requer análise das condições locais e eventual adaptação do método. Dessa forma, a validade da contribuição proposta está condicionada à preservação dos elementos estruturais que caracterizam a classe de problema investigada (Wieringa, 2014).

É importante delimitar explicitamente o escopo desta classe de problema para evitar generalizações indevidas. A presente pesquisa **não** se propõe a contribuir para o campo geral da avaliação de textos em linguagem natural, da classificação automática de documentos ou do processamento de linguagem natural como fins em si mesmos. Tampouco pretende oferecer uma solução genérica para qualquer tipo de artefato textual organizacional.

A classe de problema investigada é especificamente delimitada a três condições

simultâneas: **(i)** os artefatos avaliados são itens de *Product Backlog*, como histórias de usuário, critérios de aceitação, *Definition of Ready* e *Definition of Done*, não textos de natureza genérica; **(ii)** o contexto organizacional é de desenvolvimento ágil em escala, com múltiplas equipes coordenadas por uma estrutura como o SAFe, o que impõe requisitos de padronização, comparabilidade e governança ausentes em contextos de equipe única; e **(iii)** o problema central é a *avaliação sistemática* de artefatos já existentes no backlog, não a geração ou reescrita de novos artefatos.

Essa delimitação é relevante tanto para o posicionamento científico da pesquisa, evitando sobreposições indevidas com campos estabelecidos como Linguística Computacional ou Recuperação de Informação, quanto para orientar corretamente futuras aplicações e adaptações do método proposto.

3.3.2 Constructos Fundamentais da Pesquisa

Para operar sobre a classe de problemas delimitada, torna-se necessário estabelecer um vocabulário conceitual que permita descrever de forma estruturada os fenômenos investigados. Na *Design Science Research*, a construção de conhecimento prescritivo envolve o desenvolvimento de constructos que organizam a compreensão da realidade e orientam a concepção de artefatos destinados à solução de problemas organizacionais (March; Smith, 1995; Dresch; Lacerda; Antunes Junior, 2015).

A partir desses elementos, a presente investigação estabelece um conjunto de constructos centrais que compõem a base conceitual da pesquisa e orientam tanto a definição dos requisitos do artefato quanto os critérios utilizados em sua avaliação.

Qualidade do Product Backlog: entendida como o grau em que os itens de backlog apresentam clareza textual, critérios de aceitação testáveis, definição explícita de prontidão (*Definition of Ready*) e critérios claros de conclusão (*Definition of Done*).

Maturidade do Backlog: compreendida como o nível de consistência e padronização observado na elaboração dos itens, contribuindo para maior previsibilidade na execução e redução de ambiguidades.

Escalabilidade da Avaliação: capacidade de analisar múltiplos itens de backlog de forma consistente, mantendo critérios uniformes mesmo em ambientes com elevado volume de artefatos.

Redução de Subjetividade: diminuição da dependência exclusiva de julgamentos individuais por meio da utilização de critérios explícitos e escalas estruturadas de avaliação.

Aprendizado Organizacional Orientado por Evidências: utilização sistemática dos diagnósticos gerados para promover melhoria contínua na qualidade dos artefatos e no processo de planejamento.

Esses constructos orientam a definição dos requisitos de projeto do artefato e servem como referência para a interpretação dos resultados obtidos em sua aplicação. Conforme destaca [Wieringa \(2014\)](#), a explicitação de constructos conceituais é fundamental na Ciência do Design para permitir a descrição sistemática de fenômenos organizacionais e a formulação de soluções tecnológicas coerentes com o problema investigado.

Ao definir explicitamente os termos empíricos que orientam a análise da qualidade e da prontidão dos itens de backlog, o pesquisador reduz ambiguidades conceituais e estabelece uma linguagem comum entre o sistema desenvolvido e os diferentes atores organizacionais envolvidos. A explicitação formal desses constructos contribui, assim, para fortalecer o rigor metodológico da pesquisa e estabelecer uma base conceitual rastreável para a análise dos resultados apresentados nos capítulos subsequentes.

3.4 Derivação dos Requisitos do Artefato

Em pesquisas orientadas pela *Design Science Research* (DSR), a transição entre a identificação da classe de problema e o desenvolvimento do artefato ocorre por meio da explicitação sistemática de requisitos de projeto. Conforme argumenta [Hevner et al. \(2004\)](#), esses requisitos representam as propriedades essenciais que permitem alinhar a solução tecnológica às necessidades do ambiente organizacional, assegurando que o artefato responda de forma efetiva às lacunas previamente identificadas. Nessa direção, a formulação de requisitos de design constitui um passo intermediário fundamental entre a compreensão do problema e a concepção da solução. Para [Wieringa \(2014\)](#), a especificação desses requisitos deve estruturar um verdadeiro “argumento de contribuição”, no qual as propriedades projetadas no artefato são concebidas para produzir efeitos desejados quando aplicadas no contexto de uso. Aplicando esse princípio ao presente estudo, as limitações observadas na avaliação da qualidade

de artefatos de backlog foram traduzidas em propriedades funcionais e não funcionais que orientam o desenvolvimento da solução proposta. Dessa forma, os requisitos de design foram derivados da classe de problema e dos constructos conceituais definidos na seção anterior, estabelecendo os critérios que guiam a construção do artefato.

Com base nessa análise, foram definidos os seguintes requisitos para orientar o desenvolvimento do método proposto:

R1. Estruturação de Critérios Explícitos de Avaliação: O método deve definir critérios objetivos e claramente descritos para avaliação da qualidade das histórias de usuário, critérios de aceitação, *Definition of Ready* e *Definition of Done*, reduzindo ambiguidades interpretativas na análise dos itens de backlog.

R2. Escala Mensurável de Avaliação: O artefato deve adotar uma escala estruturada que permita traduzir percepções qualitativas em indicadores comparáveis, possibilitando análise evolutiva e comparação entre equipes.

R3. Capacidade de Processamento Escalável: A solução deve permitir a avaliação automatizada e escalável de múltiplos itens de backlog, mantendo consistência na aplicação dos critérios mesmo em contextos com elevado volume de artefatos.

R4. Geração de Recomendações Acionáveis: Além da atribuição de pontuações, o artefato deve produzir recomendações estruturadas de melhoria, apoiando o refinamento contínuo do backlog.

R5. Integração com a Rotina Ágil Existente: O método deve ser compatível com os rituais e fluxos já estabelecidos no ART estudado, evitando rupturas operacionais significativas.

R6. Transparência e Rastreabilidade das Avaliações: O artefato deve registrar justificativas associadas às pontuações atribuídas, permitindo auditoria, aprendizado organizacional e discussão fundamentada sobre os resultados obtidos.

R7. Apoio à Governança em Escala: A solução deve possibilitar a consolidação dos resultados em nível de ART, permitindo análises agregadas por time, período ou dimensão avaliada.

Esses requisitos funcionam como critérios estruturantes de projeto que orientam a concepção do artefato apresentado nos capítulos subsequentes. Sua explicitação reforça o rigor metodológico da pesquisa, evidenciando que o desenvolvimento da solução não ocorreu de forma arbitrária, mas foi derivado sistematicamente da classe de problema e dos constructos conceituais definidos neste estudo.

Além disso, esses requisitos constituem a base para o modelo conceitual do artefato, apresentado na seção seguinte, no qual são definidos os componentes e mecanismos responsáveis por operacionalizar essas propriedades no contexto da solução proposta.

Os requisitos definidos podem ser relacionados aos constructos centrais da pesquisa, evidenciando como a base conceitual apresentada na seção anterior orienta a concepção do artefato.

Tabela 6 – Mapeamento de Requisitos e Constructos

Requisito	Constructo relacionado
R1	Redução de Subjetividade
R2	Maturidade do Backlog
R3	Escalabilidade da Avaliação
R4	Aprendizado Organizacional Orientado por Evidências
R5	Maturidade do Backlog
R6	Redução de Subjetividade
R7	Aprendizado Organizacional Orientado por Evidências

3.5 Modelo Conceitual do Artefato

A transição dos requisitos definidos na seção anterior para a concepção estrutural do artefato exige a formulação de um modelo conceitual explícito que descreva como essas propriedades são operacionalizadas na solução proposta. Em pesquisas orientadas ao *design*, esse modelo atua como a arquitetura lógica da solução, decompondo o artefato em componentes inter-relacionados responsáveis por produzir os comportamentos necessários para mitigar o problema identificado (Wieringa, 2014).

Com base nessa perspectiva, o *Backlog Analyzer* foi estruturado em três camadas principais interdependentes: (i) entrada textual, (ii) processamento estruturado e (iii) geração de evidências para tomada de decisão.

Camada 1 – Entrada Textual

A primeira camada corresponde aos artefatos textuais do backlog que constituem a unidade básica de análise do método. Esses elementos representam as principais estruturas utilizadas na descrição de requisitos em ambientes ágeis e materializam os constructos relacionados à qualidade do *Product Backlog* e à maturidade do backlog.

Os artefatos considerados nessa camada incluem:

- Descrição da *User Story*;
- Critérios de Aceitação;
- *Definition of Ready* (DoR);
- *Definition of Done* (DoD).

Esses elementos constituem as entradas primárias do artefato e representam a base informacional utilizada para a avaliação da qualidade dos itens de backlog.

Camada 2 – Processamento Estruturado

A segunda camada corresponde ao núcleo metodológico do artefato. Nessa etapa, os textos são avaliados por meio da aplicação sistemática de critérios estruturados de qualidade derivados da literatura e operacionalizados por meio de uma escala ordinal de avaliação.

O processamento envolve três mecanismos principais:

- aplicação de critérios explícitos de qualidade;
- utilização de escala ordinal estruturada (0 a 5);
- uso programático de modelos de linguagem (LLMs) orientados por prompts padronizados.

Essa camada operacionaliza diretamente os requisitos R1 (critérios explícitos), R2 (escala mensurável), R3 (processamento escalável) e R6 (rastreabilidade), assegurando consistência na atribuição das pontuações e na geração de justificativas textuais para as avaliações realizadas.

Camada 3 – Geração de Evidências e Aprendizado

A terceira camada corresponde aos resultados produzidos pela instanciação computacional do artefato, responsáveis por transformar avaliações individuais em evidências organizacionais úteis para a tomada de decisão.

Os principais outputs gerados incluem:

- nota atribuída por dimensão avaliada;
- nota consolidada do item de backlog;
- recomendações estruturadas de melhoria;
- consolidação dos resultados em dashboards de gestão.

Essa camada está diretamente relacionada aos requisitos R4 (recomendações acionáveis), R5 (integração com a rotina ágil) e R7 (apoio à governança em escala), permitindo transformar avaliações individuais em diagnósticos organizacionais agregados.

Relação entre Constructos e Elementos do Modelo

O modelo conceitual articula os constructos definidos na seção anterior com os componentes operacionais do artefato. A qualidade do backlog é operacionalizada por meio da avaliação estruturada das dimensões textuais consideradas na análise. A maturidade do backlog é observada a partir da evolução das notas ao longo do tempo.

A redução de subjetividade decorre da aplicação sistemática de critérios explícitos de avaliação, enquanto a escalabilidade é viabilizada pelo uso programático de modelos de linguagem na análise de múltiplos itens de backlog. Por fim, o aprendizado organizacional emerge da consolidação histórica dos resultados e de sua utilização para apoiar processos de melhoria contínua.

A Figura 21 apresenta uma representação sintética do modelo conceitual adotado nesta pesquisa.

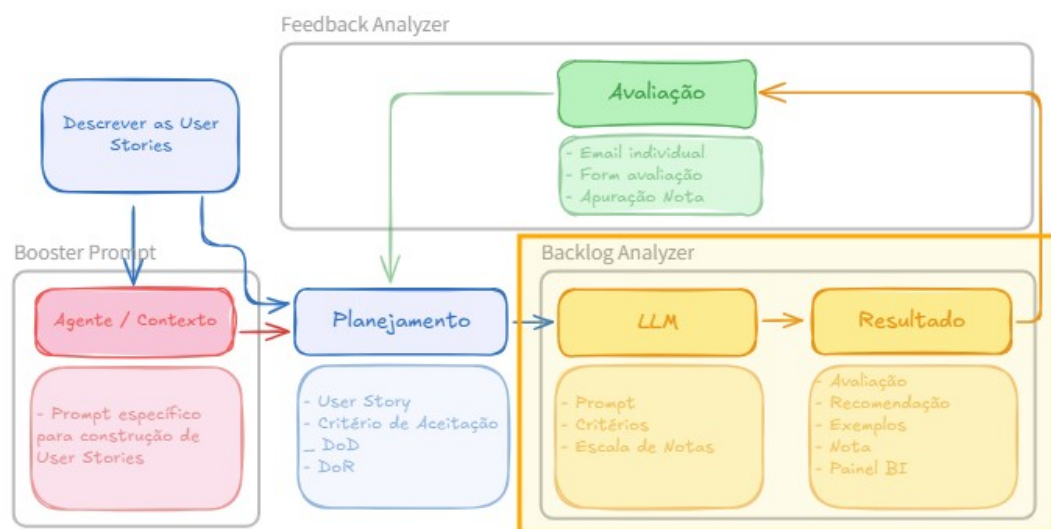


Figura 21 – Modelo conceitual do artefato Backlog Analyzer

Fonte: Elaborado pelo autor.

O modelo conceitual apresentado fornece a estrutura lógica que orienta a instânciação técnica detalhada no Capítulo 4, assegurando coerência entre fundamentação teórica, requisitos de projeto e implementação prática.

3.6 Estratégia de Demonstração e Avaliação do Artefato

Em pesquisas orientadas pela *Design Science Research*, a avaliação do artefato constitui uma etapa fundamental para verificar sua utilidade, consistência e viabilidade no contexto para o qual foi concebido (Hevner et al., 2004). Diferentemente de abordagens puramente técnicas de validação, a avaliação em DSR busca examinar em que medida o artefato desenvolvido é capaz de endereçar efetivamente a classe de problema definida.

Sob essa ótica, a avaliação não se restringe à análise do desempenho tecnológico da solução, mas considera também sua aplicabilidade prática e sua capacidade de produzir melhorias observáveis no ambiente organizacional em que é utilizado.

No presente estudo, a estratégia de avaliação foi estruturada combinando critérios clássicos da Design Science Research com evidências empíricas observáveis no contexto organizacional analisado.

Dimensões de Avaliação segundo a DSR

A avaliação do artefato foi orientada por três critérios centrais derivados da literatura de DSR:

Utilidade: verificação da capacidade do método de contribuir para a melhoria da qualidade estrutural dos itens de backlog.

Consistência: análise da estabilidade e padronização das avaliações produzidas pelo artefato, assegurando aplicação uniforme dos critérios definidos.

Viabilidade operacional: observação da possibilidade de integração do método à rotina do ART analisado, sem provocar rupturas significativas nos fluxos de trabalho existentes.

Dimensões Empíricas de Avaliação

Para operacionalizar esses critérios teóricos, foram definidas três dimensões empíricas de análise:

(1) Redução de Itens com Nota Zero: Foi monitorado o percentual de histórias de usuário que apresentavam ausência total de descrição estruturada (sem conteúdo

em Story, Critérios de Aceitação, DoR e DoD). Essa métrica representa um indicador mínimo de maturidade estrutural do backlog.

(2) Evolução da Nota Média do Backlog: A qualidade dos itens foi avaliada por meio de uma escala ordinal estruturada de 0 a 5, aplicada às quatro dimensões definidas no método. A análise considerou a evolução temporal das notas médias, permitindo observar padrões de melhoria ao longo do período de aplicação do artefato.

(3) Percepção dos Desenvolvedores: Foi aplicado um questionário estruturado aos membros das equipes ágeis, utilizando escala Likert de cinco pontos, com o objetivo de capturar a percepção dos participantes quanto à clareza, maturidade e consistência dos itens de backlog após a aplicação do método.

Procedimento de Demonstração

A demonstração do artefato ocorreu em um ART composto por 15 times ágeis, no qual o método foi aplicado de forma programática durante ciclos regulares de planejamento e refinamento.

O processo de aplicação envolveu:

- avaliação automatizada de itens de backlog por meio do Backlog Analyzer;
- consolidação das notas em repositório estruturado de análise;
- disponibilização dos resultados em dashboards de gestão;
- aplicação de questionário estruturado aos desenvolvedores participantes.

Esse procedimento permitiu observar o comportamento do artefato em um ambiente organizacional real, caracterizado por múltiplas equipes e elevado volume de itens de backlog.

Critério de Análise

A análise dos resultados considerou três perspectivas principais:

- comparação da proporção de itens com nota zero antes e após a aplicação do método;
- avaliação da evolução média das notas atribuídas aos itens de backlog ao longo do período de observação;

- cálculo das médias das respostas obtidas no questionário em escala Likert, utilizado como indicador da percepção dos usuários sobre a utilidade do método.

Importante destacar que a presente seção descreve exclusivamente os procedimentos metodológicos adotados para a demonstração e avaliação do artefato. Os resultados empíricos decorrentes da aplicação do método, bem como sua análise e discussão, são apresentados no Capítulo 5.

4 DESENVOLVIMENTO DO ESTUDO

O capítulo anterior apresentou a estratégia metodológica adotada nesta pesquisa, fundamentada na abordagem de *Design Science Research* (DSR), bem como a definição da classe de problema, dos constructos conceituais e dos requisitos que orientaram o desenvolvimento do artefato. Também foram descritos o modelo conceitual da solução proposta e os procedimentos previstos para sua demonstração e avaliação em contexto organizacional.

A partir dessas definições metodológicas, o presente capítulo descreve o processo de desenvolvimento e implementação do artefato proposto nesta pesquisa, materializado em um sistema computacional composto por módulos especializados para apoio à avaliação e melhoria da qualidade do Product Backlog. O objetivo é detalhar como os princípios conceituais definidos anteriormente foram materializados em uma solução computacional capaz de apoiar a avaliação estruturada da qualidade de itens de *Product Backlog* em ambientes ágeis.

O ponto de partida do estudo foi a análise crítica dos desafios associados à gestão do *Product Backlog* em ambientes de desenvolvimento ágil em escala. Conforme evidenciado na literatura e aprofundado na investigação conduzida nesta pesquisa, esses contextos frequentemente apresentam limitações relacionadas à subjetividade na avaliação da qualidade dos itens de backlog, à ausência de critérios padronizados de diagnóstico e à dificuldade de consolidar evidências que apoiem processos de aprendizado organizacional contínuo.

A análise desses desafios revelou lacunas relevantes tanto na literatura quanto na prática organizacional. Entre elas destacam-se a baixa padronização na avaliação da maturidade dos itens de backlog, a dificuldade de aplicar critérios de qualidade de forma consistente em ambientes com múltiplas equipes e a limitada escalabilidade dos métodos tradicionalmente utilizados para análise desses artefatos.

Diante desse cenário, investigou-se o potencial de utilização de Inteligência Artificial generativa, em especial *Large Language Models* (LLMs), como mecanismo tecnológico capaz de apoiar a avaliação automatizada de descrições textuais presentes no backlog. A adoção dessa abordagem permite explorar a capacidade desses modelos de interpretar linguagem natural e gerar avaliações estruturadas a partir de critérios previamente definidos.

Com base nessa premissa, este capítulo apresenta as decisões de projeto que conduziram à concepção do artefato proposto. São descritos os critérios utilizados para avaliação da qualidade do backlog, a definição da escala de mensuração adotada, a arquitetura da solução computacional e o fluxo de funcionamento do sistema no contexto organizacional estudado. Por fim, descreve-se a implementação prática do método e sua aplicação no ambiente real analisado, estabelecendo as bases para a apresentação e discussão dos resultados no capítulo seguinte.

4.1 Contextualização e Motivação

Os desafios que motivam esta pesquisa (subjetividade na avaliação de itens do *Product Backlog*, ausência de critérios padronizados para inspeção desses artefatos e baixa escalabilidade dos processos tradicionais de refinamento) foram caracterizados em detalhe na revisão da literatura (Capítulo 2), que constitui o embasamento empírico desta dissertação. Essa análise identificou, adicionalmente, que embora o uso de *Large Language Models* (LLMs) em Engenharia de Requisitos tenha avançado em atividades de *geração* e apoio à escrita de histórias de usuário, permanecem escassas as abordagens voltadas à *avaliação sistemática* de artefatos textuais já existentes no backlog, com critérios explícitos e escalas mensuráveis. Diante dessas evidências, as implicações para a concepção do artefato tornaram-se claras: era necessário projetar uma solução capaz de explicitar critérios estruturados de avaliação, utilizar escalas reproduzíveis e operar de forma escalável sobre grandes volumes de itens, princípios que orientam as decisões de projeto detalhadas na seção seguinte.

4.2 Concepção do Artefato: Definição do Método de Avaliação da Qualidade do Backlog

A concepção do artefato desenvolvido nesta pesquisa foi orientada pela necessidade de responder, de forma sistemática e fundamentada, às lacunas identificadas na seção anterior. Em consonância com a abordagem da *Design Science Research* (DSR), o desenvolvimento do artefato busca produzir conhecimento prescritivo voltado à solução de problemas práticos relevantes. Segundo [Dresch, Lacerda e Antunes Junior \(2015\)](#), a DSR concentra-se na construção e avaliação de artefatos capazes de modifi-

car situações existentes e alcançar resultados desejados. Nessa perspectiva, o artefato proposto materializa-se por meio de um método estruturado e replicável destinado à avaliação da qualidade de itens de *Product Backlog*.

O problema central que norteou o projeto reside na ausência de mecanismos capazes de avaliar, de forma objetiva e escalável, a qualidade e a maturidade dos itens que compõem o backlog. Conforme observado por [Sedano, Ralph e Péraire \(2019\)](#), o *Product Backlog* funciona como um objeto de fronteira complexo, emergindo de processos de *sensemaking*, entendidos como a construção ativa e coletiva de significado em contextos incertos. Em ambientes organizacionais caracterizados por múltiplas equipes e elevado volume de artefatos, essa natureza informal pode amplificar ambiguidades e dificultar o alinhamento entre estratégia e execução. Nesses contextos, falhas na preparação dos itens de trabalho podem resultar em eventos de planejamento caóticos e ineficientes ([Paasivaara, 2017](#)), além de comprometer a previsibilidade das entregas.

Adicionalmente, as práticas tradicionais de refinamento de backlog tendem a ser altamente dependentes da experiência tácita das equipes, configurando processos manuais, onerosos e suscetíveis a inconsistências ([Nasiri; Lahmer, 2024](#)). A literatura aponta que a ausência de critérios estruturados de avaliação dificulta a comparação entre equipes e limita o aprendizado organizacional sobre a qualidade dos artefatos produzidos. Diante disso, propostas de estruturação e taxonomia dos itens de backlog têm sido sugeridas como mecanismos capazes de reduzir ambiguidades e apoiar a mensuração sistemática desses artefatos ([Dilorenzo et al., 2020](#)).

A partir dessas evidências, a concepção do método proposto foi orientada pela definição explícita dos elementos do backlog a serem avaliados, bem como pela construção de critérios e escalas de mensuração capazes de traduzir avaliações qualitativas em indicadores estruturados. Essa abordagem busca preservar os princípios fundamentais dos métodos ágeis, como flexibilidade, comunicação contínua e adaptação incremental, ao mesmo tempo em que introduz um nível adicional de rigor analítico para apoiar processos de planejamento e governança do produto. Conforme argumenta [Cohn \(2004\)](#), as histórias de usuário constituem um mecanismo de comunicação conciso entre stakeholders; entretanto, sua eficácia depende da clareza textual e da explicitação do valor de negócio associado.

Nessa lógica, o método proposto estrutura a avaliação da qualidade do backlog

a partir de quatro dimensões centrais. A primeira refere-se à descrição da *User Story*, cuja formulação deve evidenciar o valor entregue ao usuário ou ao negócio (Cohn, 2004). A segunda dimensão corresponde aos *Crítérios de Aceitação*, responsáveis por explicitar o comportamento esperado do sistema e delimitar o escopo da entrega (North, 2006). A terceira dimensão considera a *Definition of Ready* (DoR), mecanismo que assegura que os itens estejam suficientemente compreendidos antes do início do desenvolvimento (Sedano; Ralph; Péraire, 2019). Por fim, inclui-se a *Definition of Done* (DoD), entendida como um compromisso organizacional que estabelece os critérios mínimos para considerar um incremento como concluído (Schwaber; Sutherland, 2020).

Para operacionalizar essas dimensões e refinar os requisitos em tarefas executáveis, foram adotados critérios inspirados no modelo INVEST, conforme as diretrizes clássicas apresentadas por Cohn (2004). Tais práticas são frequentemente complementadas por abordagens baseadas em análise linguística da decomposição de histórias (Müter et al., 2019) e por técnicas derivadas do *Behavior-Driven Development* (BDD). A aplicação combinada desses conceitos contribui para reduzir ambiguidades e promover maior verificabilidade dos requisitos, estabelecendo uma linguagem compartilhada entre os participantes do processo de desenvolvimento.

Para operacionalizar as dimensões do escopo e transformar requisitos em tarefas executáveis, a literatura recomenda a adoção de diretrizes clássicas para a estruturação de histórias de usuário, como o formato proposto por Cohn (2004) [101.57]. Complementarmente, a utilização de critérios de qualidade baseados no acrônimo INVEST tem se mostrado fundamental para assegurar que os itens do backlog sejam independentes, negociáveis, estimáveis e testáveis (Wake, 2003) [101.57].

Essas práticas estruturais podem ser enriquecidas por abordagens analíticas e comportamentais, que contribuem para aumentar a consistência técnica dos requisitos. Müter et al. (2019) evidenciam que a análise linguística da decomposição de histórias (com foco na organização das sentenças e no uso de verbos de ação elementares) contribui para reduzir a ocorrência de requisitos mal formulados [101.80]. Adicionalmente, o refinamento dos critérios de aceitação pode ser apoiado por técnicas oriundas do *Behavior-Driven Development* (BDD), cuja sintaxe padronizada orienta a equipe para a definição de comportamentos observáveis e passíveis de validação (Sedano et al., 2019) [101.15].

Como consequência da aplicação integrada dessas abordagens, observa-se

uma redução significativa da ambiguidade nos requisitos. Dessa forma, estabelece-se uma linguagem compartilhada entre os envolvidos, permitindo que o backlog atue efetivamente como um objeto de fronteira (*boundary object*) no processo de desenvolvimento [101.15].

Sob essa ótica, os critérios estruturados podem também ser interpretados à luz do modelo comportamental da linguagem Gherkin (baseado na estrutura *Contexto–Comportamento–Resultado*) amplamente utilizado para descrever cenários de teste e comportamento esperado do sistema. Conforme observado por [Sedano, Ralph e Péraire \(2019\)](#), a presença de critérios de aceitação bem definidos atua como um mecanismo de validação da completude dos requisitos. Contudo, a efetividade desses critérios depende da colaboração entre os principais participantes do processo de desenvolvimento. Além disso, [Ribeiro Dos Santos et al. \(2025\)](#) destacam a importância da reunião dos “3 Amigos” (Product Owner, desenvolvedores e testadores) como prática fundamental para assegurar a compreensão compartilhada dos requisitos.

Com base nos requisitos de projeto definidos na Seção 3.4, a concepção do artefato foi estruturada de modo a garantir a correspondência entre os requisitos identificados e os componentes efetivamente implementados no método proposto. A Tabela 7 apresenta o mapeamento entre os requisitos de design definidos para o artefato e os elementos de implementação que os materializam.

Tabela 7 – Rastreabilidade entre requisitos de projeto e componentes do artefato	
Requisito	Implementação no artefato
R1	Definição de critérios estruturados para avaliação da qualidade das histórias de usuário, critérios de aceitação, Definition of Ready e Definition of Done.
R2	Utilização de escala ordinal de avaliação variando de 0 a 5 para mensurar a maturidade dos itens de backlog.
R3	Pipeline automatizado de processamento textual utilizando modelos de linguagem para análise semântica dos itens do backlog.
R4	Geração automática de recomendações estruturadas de melhoria para os itens avaliados.
R5	Integração do método aos rituais existentes do ART, especialmente durante atividades de refinamento e planejamento.
R6	Registro de justificativas textuais associadas às avaliações geradas pelo modelo de linguagem.
R7	Consolidação dos resultados em indicadores agregados e dashboards para apoio à governança em escala.

Fonte: Elaborado pelo autor.

A operacionalização desses princípios resultou na definição de critérios estruturados e indicadores observáveis para avaliação dos itens do backlog, conforme

apresentado na Tabela 8.

A definição desses critérios foi orientada por heurísticas derivadas de boas práticas amplamente reconhecidas na engenharia de requisitos ágil. Tais heurísticas incluem princípios associados ao modelo INVEST para qualidade de histórias de usuário, abordagens derivadas do Behavior-Driven Development (BDD) para estruturação de critérios de aceitação e práticas organizacionais relacionadas à Definition of Ready e Definition of Done. A utilização dessas heurísticas permite transformar recomendações qualitativas da literatura em critérios estruturados de avaliação aplicáveis pelo método proposto.

Com base nessas heurísticas, foram definidos critérios específicos para cada uma das dimensões avaliadas no método proposto. Esses critérios foram organizados de modo a permitir a observação objetiva de aspectos relacionados à clareza, completude e verificabilidade das descrições presentes nos itens do backlog, conforme sintetizado na Tabela 8.

Embora as quatro dimensões avaliadas estejam relacionadas à qualidade dos itens de backlog, elas contemplam diferentes níveis de análise. As dimensões *User Story* e *Critérios de Aceitação* referem-se predominantemente à qualidade da especificação textual dos requisitos. Já as dimensões *Definition of Ready* (DoR) e *Definition of Done* (DoD) refletem aspectos relacionados à maturidade do processo de desenvolvimento e à governança técnica do backlog. A avaliação conjunta dessas dimensões permite capturar tanto a qualidade intrínseca das descrições quanto o nível de preparação organizacional necessário para sua execução.

A mensuração da qualidade dos artefatos foi operacionalizada por meio de uma escala ordinal de seis níveis, variando de zero a cinco. Essa escala foi concebida como uma adaptação de escalas do tipo Likert estendida, permitindo representar níveis crescentes de maturidade textual e técnica dos itens avaliados. Dessa forma, torna-se possível distinguir desde situações em que o critério é inexistente ou irrelevante até cenários nos quais ele está plenamente descrito, mensurável e tecnicamente verificável.

A utilização dessa escala permite realizar análises comparativas entre itens de backlog, acompanhar a evolução da qualidade ao longo do tempo e identificar padrões recorrentes nos artefatos produzidos pelas equipes. A escala de maturidade adotada nesta pesquisa é apresentada na Tabela 9.

É importante registrar que a definição da nota 3 como limiar de referência mínima

Tabela 8 – Critérios estruturados, indicadores observáveis e classificação INVEST para avaliação da qualidade do backlog

Dimensão Avaliada	Critérios de Avaliação	*	Indicadores Observáveis	Nota
User Story	• Clareza textual • Uso do formato padrão (Como / Quero / Para) • Explicitação do valor de negócio • Granularidade adequada para 1 Sprint	• N • N • V • S	• Texto objetivo e sem ambiguidade • Presença explícita de persona, ação e benefício • Benefício mensurável ou justificável • Escopo compatível com uma iteração	0–5
Critérios de Aceitação	• Objetividade • Testabilidade • Alinhamento com a Story • Cobertura de cenários positivos e negativos	• T • T • V • T	• Uso de formato estruturado (Dado / Quando / Então) • Critérios verificáveis tecnicamente • Relação direta com o objetivo da Story • Inclusão de casos alternativos ou exceções	0–5
Definition of Ready (DoR)	• Clareza do escopo • Entendimento compartilhado • Estimativa viável • Dependências identificadas • Riscos mapeados	• N • N • E • I • E	• História compreendida pela equipe • Story Points atribuídos • Dependências explícitas • Restrições ou impactos descritos	0–5
Definition of Done (DoD)	• Cobertura de testes • Revisão de código • Documentação • Critérios de qualidade • Integração ou implantação validada	• T • T • V • T • T	• Evidência de testes unitários ou funcionais • Registro de code review • Atualização de documentação • Critérios objetivos de aceite técnico	0–5

Fonte: Elaborado pelo autor (2025).

Nota:(*) Classificação INVEST: I (Independent), N (Negotiable), V (Valuable), E (Estimable), S (Small), T (Testable).

não decorreu de uma escolha teórica arbitrária, mas de uma **calibração empírica interna**. Na primeira medição sistemática realizada com o método proposto, a média geral dos itens avaliados situou-se em aproximadamente 1,0 em uma escala de 0 a 5, com mais de 57% dos itens recebendo nota 0, indicando ausência total de conteúdo estruturado em pelo menos um dos campos avaliados. Nesse mesmo levantamento, identificou-se que um dos times já alcançava sistematicamente média próxima a 3,0, caracterizando-se como o grupo com melhor desempenho observado no conjunto

Tabela 9 – Escala de maturidade utilizada para avaliação dos itens do backlog

Nota	Descrição
0	Inexistente ou irrelevante. O critério não é abordado ou não apresenta relação clara com o objetivo do item.
1	Muito superficial. O critério é citado de forma vaga, sem clareza ou detalhamento suficiente.
2	Poucos aspectos mencionados, com baixa objetividade e ausência de verificabilidade técnica.
3	Alguns aspectos abordados com linguagem clara, porém sem evidências suficientes de verificação técnica ou mensuração.
4	Abrangente e claro, com evidências parciais de verificabilidade técnica e alinhamento com boas práticas.
5	Abrangente, claro, mensurável e tecnicamente verificável, plenamente aderente às boas práticas ágeis.

Fonte: Elaborado pelo autor (2025).

analisado.

Essa constatação orientou a decisão de utilizar o desempenho desse time como **benchmark empírico interno**. Se uma equipe, com os recursos e o contexto organizacional disponíveis, já conseguia operar com backlog nesse nível, era razoável estabelecer a nota 3 como o parâmetro mínimo de qualidade operacional esperado para os demais times. Em outras palavras, a nota 3 não representa excelência técnica. Ela representa a *qualidade mínima operacional do backlog*, isto é, o limiar abaixo do qual as informações registradas tornam-se insuficientes para apoiar a execução consistente das histórias.

Essa abordagem de calibração por benchmark interno é coerente com a lógica da *Design Science Research*, segundo a qual o artefato deve ser progressivamente ajustado a partir das evidências empíricas coletadas no próprio contexto de aplicação (Dresch; Lacerda; Antunes Junior, 2015). Além disso, o uso da nota como indicador diagnóstico, e não como objetivo em si, constitui uma distinção importante. A discussão gerencial não deve concentrar-se na nota atribuída, mas sim na questão prática de verificar se o backlog contém informações suficientes para apoiar o trabalho das equipes.

Embora a escala proposta permita a avaliação sistemática dos itens de backlog, sua aplicação manual em ambientes organizacionais de grande escala pode tornar-se onerosa e suscetível a inconsistências. Dessa forma, buscou-se utilizar LLMs como mecanismo de apoio à análise semântica dos artefatos avaliados.

A aplicação do método foi projetada para ocorrer de forma automatizada, utilizando LLMs como mecanismo de apoio à análise semântica dos itens do backlog. A escolha dessa abordagem decorre da capacidade desses modelos em interpretar

linguagem natural e produzir avaliações contextualizadas. Estudos recentes indicam que tais modelos possuem elevada capacidade de generalização semântica e identificação de padrões textuais, características particularmente adequadas à natureza predominantemente textual dos artefatos ágeis (Can; Dalpiaz, 2025; Ribeiro Dos Santos et al., 2025).

Do ponto de vista operacional, o método prevê a ingestão de dados estruturados contendo histórias de usuário e seus respectivos artefatos associados. A avaliação é realizada por meio da aplicação de critérios estruturados via *prompts* especializados. Conforme discutido por Rahman et al. (2024), técnicas de *prompt engineering*, como refinamento iterativo e instruções estruturadas, são essenciais para aumentar a consistência das respostas geradas por modelos de linguagem. Essa abordagem permite que o método seja aplicado de forma escalável e reproduzível, superando limitações de tempo e inconsistências frequentemente observadas em processos manuais de avaliação de requisitos (Alawaji; Hakami; Alshemaimri, 2024). Além disso, estudos recentes indicam que LLMs podem apoiar a verificação de cobertura de implementação de histórias de usuário, fornecendo análises probabilísticas e justificativas textuais para as avaliações produzidas (Kolthoff et al., 2024).

Assim, a concepção do artefato traduz os *gaps* identificados na literatura em requisitos de projeto claros e operacionalizáveis, articulando fundamentos teóricos da engenharia de requisitos ágil com avanços recentes em processamento de linguagem natural. Conforme argumenta Dresch, Lacerda e Antunes Junior (2015), a contribuição central da Design Science Research reside justamente na construção de artefatos capazes de gerar conhecimento prescritivo e resolver problemas relevantes do mundo real. O método aqui proposto estabelece, portanto, as bases conceituais e operacionais para a implementação do sistema computacional apresentado na seção seguinte, na qual são descritas a arquitetura e o funcionamento do artefato desenvolvido.

4.3 Arquitetura do Sistema e Funcionamento do Artefato

A partir da definição conceitual do método de avaliação da qualidade do backlog, apresentada na seção anterior, o artefato foi materializado sob a forma de um sistema computacional estruturado e modular. Em consonância com Dresch, Lacerda e Antunes Junior (2015), a arquitetura do sistema foi concebida para garantir a coerência entre o

problema identificado, os requisitos de projeto e a solução implementada, assegurando não apenas a viabilidade técnica, mas também a escalabilidade e a reprodutibilidade da pesquisa científica orientada ao design (Hevner et al., 2004).

Durante o processo de desenvolvimento do artefato, verificou-se que a avaliação isolada da qualidade dos itens de backlog, embora útil como mecanismo de diagnóstico, não seria suficiente para promover melhorias sustentáveis no contexto organizacional analisado. A partir dessa constatação, o projeto do artefato evoluiu iterativamente, incorporando novos componentes destinados a apoiar também a construção inicial dos itens de backlog e a retroalimentação do processo após sua execução.

Essa evolução é consistente com o caráter iterativo da Design Science Research, na qual o artefato é progressivamente refinado à medida que novas compreensões sobre o problema emergem durante o processo de design (Dresch; Lacerda; Antunes Junior, 2015). Como resultado, o artefato final proposto nesta pesquisa configura-se como um sistema modular composto por três módulos inter-relacionados: Backlog Booster, Backlog Analyzer e Feedback Analyzer. Em conjunto, esses módulos implementam o método de avaliação e melhoria da qualidade do backlog proposto neste trabalho, permitindo apoiar tanto a construção inicial dos itens de backlog quanto a avaliação sistemática de sua qualidade e o aprendizado contínuo a partir da execução das histórias de usuário.

Assim, a arquitetura proposta adota uma organização modular, na qual cada componente desempenha um papel específico dentro do processo de avaliação e melhoria da qualidade do backlog. Essa estrutura permite que os módulos sejam utilizados de forma independente ou integrada, possibilitando tanto a adoção incremental da solução quanto sua adaptação a diferentes contextos organizacionais.

Essa organização modular reflete diferentes momentos do ciclo de vida das histórias de usuário no contexto do desenvolvimento ágil. Enquanto o Backlog Booster atua como suporte à construção e ao refinamento inicial dos itens de backlog, o Backlog Analyzer realiza a avaliação estruturada da qualidade desses artefatos, e o Feedback Analyzer incorpora aprendizados provenientes da execução das histórias durante as sprints. Dessa forma, os três módulos estabelecem um ciclo contínuo de apoio à melhoria da qualidade do backlog.

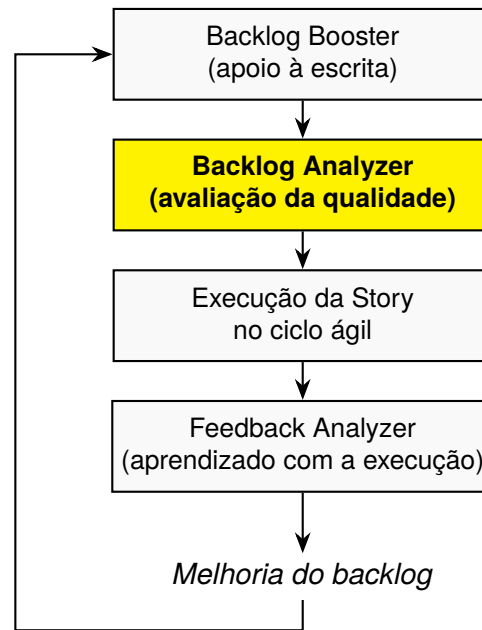


Figura 22 – Fluxo Iterativo do Ciclo de Melhoria do Backlog.

Tal característica, ilustrada no fluxo apresentado na Figura 22, é consistente com a literatura sobre preparação e gestão do backlog, que distingue atividades relacionadas à construção e refinamento dos itens daquelas associadas à execução das sprints (Jakobsen; Sutherland, 2009).

No núcleo da arquitetura encontra-se o módulo denominado Backlog Analyzer, responsável pela avaliação sistemática da qualidade dos itens do backlog. Segundo Zhang et al. (2024), o uso de agentes autônomos baseados em Modelos de Linguagem de Larga Escala (LLMs) demonstra eficácia na melhoria da clareza e completude desses artefatos. O módulo realiza a ingestão de dados estruturados contendo a descrição das histórias de usuário, seus Critérios de Aceitação, a Definition of Ready e a Definition of Done. A partir dessas informações, o sistema aplica os critérios de qualidade definidos no método e aciona o modelo de linguagem para realizar a análise semântica e sintática, uma prática que vem ganhando robustez na literatura recente de engenharia de requisitos (Zhang et al., 2024; Ronanki; Cabrero-Daniel; Berger, 2024).

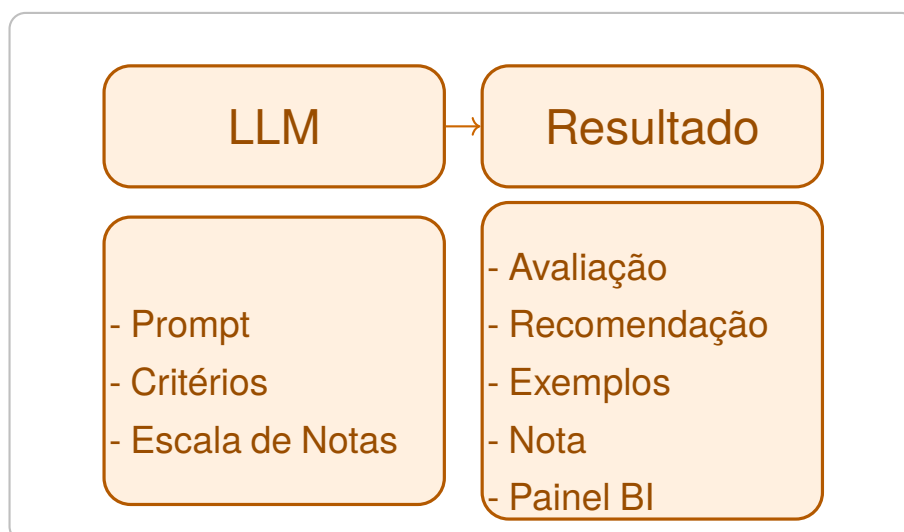


Figura 23 – Diagrama esquemático (simplificado) Backlog Analyzer

O funcionamento do Backlog Analyzer baseia-se na utilização de prompts especializados que instruem o modelo de linguagem a emular papéis críticos das metodologias ágeis, como o Product Owner e responsáveis pela garantia da qualidade do backlog. Essa abordagem fundamenta-se na simulação de papéis frequentemente observados em equipes ágeis, permitindo que o modelo de linguagem avalie os itens do backlog sob diferentes perspectivas de qualidade. Segundo [Tona et al. \(2021\)](#), a inclusão de um papel focado em qualidade contribui para assegurar que cada história de usuário contemple descrições técnicas mais rigorosas, a exemplo de diagramas de sequência ou contratos de API.

Dessa forma, o sistema avalia simultaneamente a aderência aos critérios de aceitação e aos "pontos de qualidade" baseados na norma ISO/IEC 25010, visando mitigar riscos de integração e dívida técnica ([Staron; Meding; Baniasad, 2019](#)). Como resultado, o analisador gera uma saída estruturada em formato JSON, que consolida uma pontuação técnica multidimensional acompanhada de justificativas detalhadas e recomendações práticas de melhoria.

Complementarmente ao módulo de avaliação, a arquitetura contempla o Backlog Booster, componente estratégico voltado ao suporte na construção e no refinamento das histórias de usuário. De acordo com [Ramasamy et al. \(2024\)](#), a utilização de técnicas baseadas em LLMs pode auxiliar as equipes a identificar funcionalidades e perspectivas frequentemente negligenciadas na análise manual. Nessa etapa, o módulo atua em uma etapa prévia à avaliação formal, oferecendo suporte à escrita técnica do backlog fundamentado em boas práticas ágeis. Tal abordagem permite que o sistema auxilie

na formulação de histórias claras e na definição de critérios de aceitação mensuráveis, mitigando defeitos inerentes à linguagem natural (Zhang et al., 2024). Adicionalmente, a ferramenta facilita a explicitação da Definition of Ready (DoR) e da Definition of Done (DoD), elementos cruciais para a verificação do cumprimento dos objetivos de cada história (Brockenbrough; Salinas, 2024). Segundo Lucassen et al. (2016), o objetivo central desse tipo de intervenção tecnológica é elevar a qualidade do backlog na origem, reduzindo ambiguidades e o consequente retrabalho em etapas subsequentes do ciclo de desenvolvimento. Dessa forma, a arquitetura busca assegurar uma estrutura textual robusta e uniforme, conforme preconizado pelos frameworks de qualidade para requisitos ágeis (Lucassen et al., 2016).

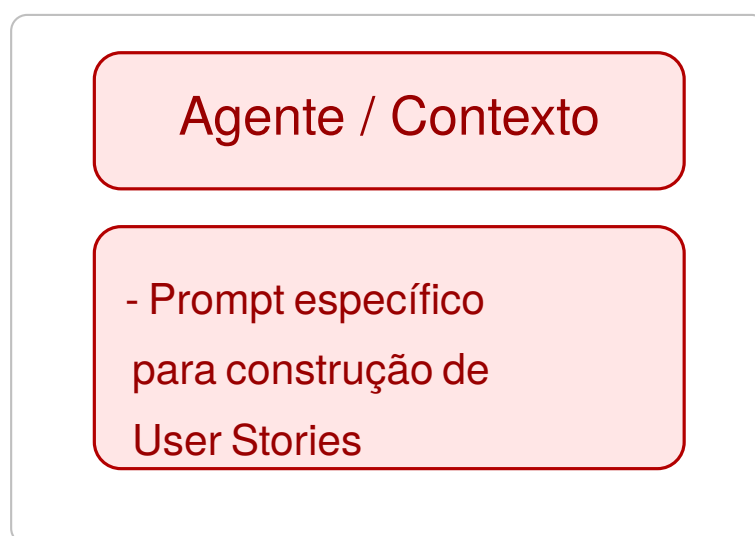


Figura 24 – Diagrama esquemático (simplificado) Agente

Fonte:Elaborado pelo autor.

A arquitetura inclui ainda o módulo denominado Feedback Analyzer, cuja função é incorporar a perspectiva dos desenvolvedores após a execução das histórias planejadas. Esse componente coleta feedback estruturado sobre aspectos como a clareza dos requisitos e a adequação do detalhamento, visto que histórias mal escritas frequentemente apresentam defeitos de qualidade que prejudicam o fluxo de trabalho. Para Lucassen et al. (2016), o sucesso de uma história depende não apenas do texto, mas da qualidade das conversas entre stakeholders e da precisão dos critérios de aceitação.

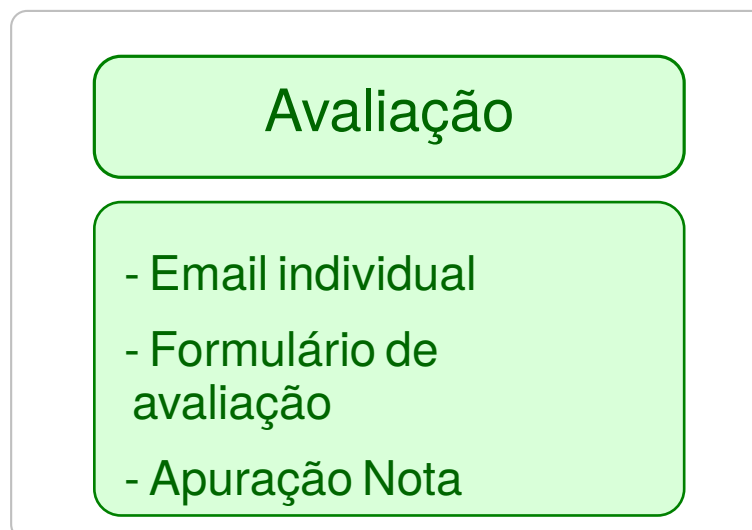


Figura 25 – Diagrama esquemático (simplificado) Feedback Analyzer
Fonte:Elaborado pelo autor.

Dessa forma, as informações obtidas complementam a avaliação técnica realizada pelo Backlog Analyzer, fortalecendo o caráter formativo do sistema. Conforme observado por [Erdoğan, Pekkaya e Gök \(2018\)](#), quando os membros da equipe participam ativamente da análise de métricas e fornecem sugestões, criam-se oportunidades reais para a melhoria contínua da produtividade e da qualidade do produto. Sob a ótica de [Sedano, Ralph e Péraire \(2019\)](#), essa retroalimentação é vital para mitigar a ambiguidade, um obstáculo comum que, se negligenciado, afeta negativamente o moral e o desempenho dos engenheiros e gestores.

Do ponto de vista do fluxo operacional, os três módulos podem ser utilizados de forma independente ou integrados em um ciclo contínuo de melhoria da qualidade do backlog. Segundo [Lucassen et al. \(2016\)](#), esse tipo de suporte automatizado é essencial para detectar defeitos sintáticos e pragmáticos, garantindo que os requisitos sigam critérios de qualidade como atomicidade e minimalidade. Outrossim, o Feedback Analyzer retroalimenta o processo com dados oriundos da execução, estabelecendo ciclos contínuos de diagnóstico e aprendizado que, conforme observado por [Erdoğan, Pekkaya e Gök \(2018\)](#), transformam métricas de produtividade em insumos valiosos para a melhoria do processo durante as retrospectivas. Dessa forma, essa integração promove a coevolução entre problema e solução, mitigando a ambiguidade que frequentemente compromete a moral das equipes ágeis ([Sedano; Ralph; Péraire, 2019](#)).

Os resultados gerados pelos módulos do sistema são armazenados em repositórios estruturados, o que permite a realização de análises históricas, comparações entre

equipes e a visualização detalhada por meio de painéis gerenciais. Segundo [Jakobsen e Sutherland \(2009\)](#), essa coleta periódica e sistemática de dados é fundamental para o controle estatístico do processo e para a eficácia das revisões mensais de projeto. Nessa perspectiva, a centralização das informações supera a fragmentação observada em abordagens baseadas em planilhas isoladas, viabilizando o acompanhamento evolutivo da qualidade e da maturidade do backlog ao longo do tempo ([Lucassen et al., 2016](#)).

Por sua vez, a arquitetura do sistema materializa o método proposto em uma solução computacional coerente, escalável e orientada à melhoria contínua. Para [Dresch, Lacerda e Antunes Junior \(2015\)](#), o desenvolvimento de um artefato sob a égide da Design Science Research deve visar não apenas a resolução de problemas reais, mas também a geração de valor e utilidade comprovada. Ao integrar a avaliação automatizada, o apoio à escrita e o feedback da execução, o artefato transcende a função de diagnóstico pontual, configurando-se como um robusto instrumento de governança do backlog em ambientes ágeis em escala ([Amjad et al., 2017](#); [Scaled Agile, Inc., 2025a](#)).

Essa arquitetura modular permite que o método proposto seja aplicado de forma incremental em diferentes contextos organizacionais, preservando os princípios de flexibilidade característicos dos métodos ágeis ao mesmo tempo em que introduz mecanismos estruturados de avaliação e aprendizado organizacional.

Na próxima seção, são detalhados os aspectos relacionados à implementação técnica do sistema e à sua aplicação no contexto organizacional estudado.

4.4 Implementação do Artefato e Aplicação no Contexto Organizacional

A implementação do artefato proposto teve como objetivo viabilizar, de forma prática e operacional, o método de avaliação e melhoria da qualidade do *Product Backlog* descrito nas seções anteriores. Em conformidade com a abordagem da *Design Science Research*, a implementação priorizou soluções tecnicamente viáveis, escaláveis e compatíveis com o contexto organizacional estudado, permitindo a instanciação do artefato em um sistema computacional aplicável à rotina das equipes ágeis.

O diagrama completo dos módulos interconectados pode ser observado na [figura 26](#)

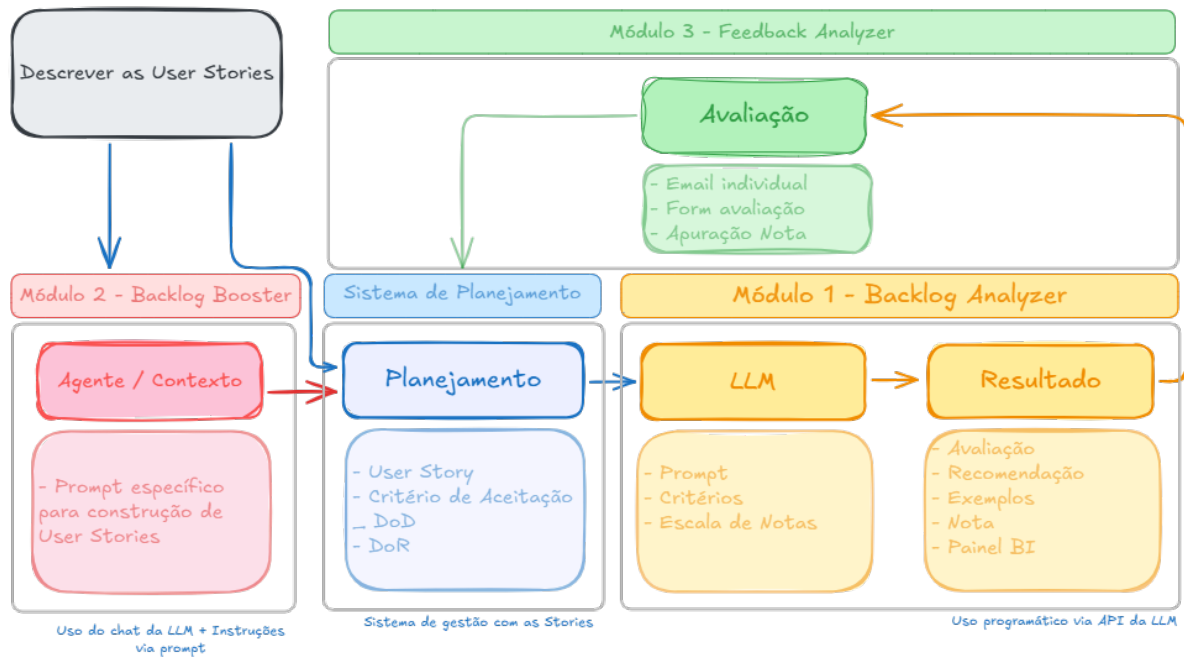


Figura 26 – Diagrama esquemático do Método de

Fonte:Elaborado pelo autor.

4.4.1 Heurísticas Construtivas do Método

A descrição dos componentes do artefato apresentada nas seções seguintes evidencia o *como* do método proposto. Igualmente relevante do ponto de vista da *Design Science Research* é explicitar o **porquê** de cada decisão de design, ou seja, as heurísticas construtivas que orientaram a concepção de cada elemento (Dresch; Lacerda; Antunes Junior, 2015). Essa explicação permite distinguir escolhas fundamentadas de escolhas arbitrárias e facilita a avaliação da transferibilidade do método a novos contextos.

Por que iniciar com o Backlog Analyzer: A primeira decisão de design foi criar um instrumento de diagnóstico antes de qualquer ferramenta de apoio à escrita. Essa sequência reflete o princípio de que, antes de prescrever melhorias, é necessário medir e compreender a situação atual. A avaliação sistemática da qualidade dos artefatos existentes permitiu quantificar a dimensão do problema e estabelecer um baseline objetivo, essencial para orientar as intervenções subsequentes e avaliar o impacto do método ao longo do tempo.

Por que criar o Backlog Booster: A primeira medição revelou um cenário crítico: a média geral dos itens avaliados situou-se em torno de 1,0 em uma escala de 0 a 5, com mais de 57% dos itens recebendo nota 0 em pelo menos um campo.

Esse diagnóstico indicou que o problema não era apenas de qualidade deficiente das descrições, mas de ausência quase total de conteúdo estruturado. O Backlog Booster foi criado como resposta direta a essa evidência: em vez de apenas diagnosticar o problema, o método passou a oferecer também apoio ativo à construção das histórias desde a origem, reduzindo o esforço cognitivo exigido das equipes e incentivando a adoção de boas práticas de escrita de requisitos. O Booster atua, portanto, antes da avaliação, como um *acelerador de qualidade* na origem.

Por que incorporar o Feedback Analyzer: A avaliação automatizada pelo LLM fornece um diagnóstico técnico-textual dos itens. A qualidade do backlog só pode ser plenamente avaliada em confronto com a perspectiva de quem executa o trabalho: os desenvolvedores. O Feedback Analyzer foi incorporado para fechar esse ciclo, capturando a percepção prática dos times após a execução das histórias e validando se as melhorias observadas nas notas correspondiam a ganhos reais de clareza e executabilidade. Esse componente também atua como salvaguarda contra o risco de *gaming*, a possibilidade de que os times aprendam a escrever para satisfazer o algoritmo sem que haja melhoria real na qualidade do planejamento.

Por que avaliar quatro dimensões: A escolha das quatro dimensões, Descrição da *User Story*, Critérios de Aceitação, *Definition of Ready* e *Definition of Done*, reflete as principais fontes de ambiguidade e incompletude identificadas na literatura e confirmadas empiricamente no contexto da pesquisa. Cada dimensão representa um aspecto distinto da prontidão do item: o valor de negócio (US), o comportamento esperado do sistema (CA), a verificação de pré-requisitos para início (DoR) e os critérios de conclusão técnica (DoD). A avaliação conjunta dessas dimensões permite capturar tanto a qualidade intrínseca das descrições quanto o grau de maturidade processual do backlog.

Por que adotar escala de 0 a 5: A escala ordinal de seis níveis foi concebida para equilibrar granularidade e interpretabilidade. Uma escala binária (presente/ausente) seria insuficiente para capturar nuances de maturidade; escalas muito amplas dificultariam a calibração dos critérios pelo modelo de linguagem e a interpretação pelos gestores. A escala de 0 a 5 mostrou-se suficientemente granular para distinguir níveis progressivos de qualidade, desde ausência total (nota 0) até plena verificabilidade técnica (nota 5), e suficientemente simples para ser aplicada de forma consistente e automática pelo LLM.

4.4.2 Backlog Analyzer

Do ponto de vista tecnológico, o sistema foi desenvolvido com base em uma arquitetura modular utilizando a linguagem Python, amplamente adotada em aplicações de análise de dados e integração com serviços de Inteligência Artificial. Essa escolha permitiu maior flexibilidade na manipulação de dados estruturados, na orquestração dos módulos do sistema e na integração com APIs externas. Para a análise semântica dos itens do backlog, foi utilizada a API de um modelo de linguagem de larga escala, responsável por processar os textos das histórias de usuário e retornar avaliações estruturadas conforme os critérios definidos.

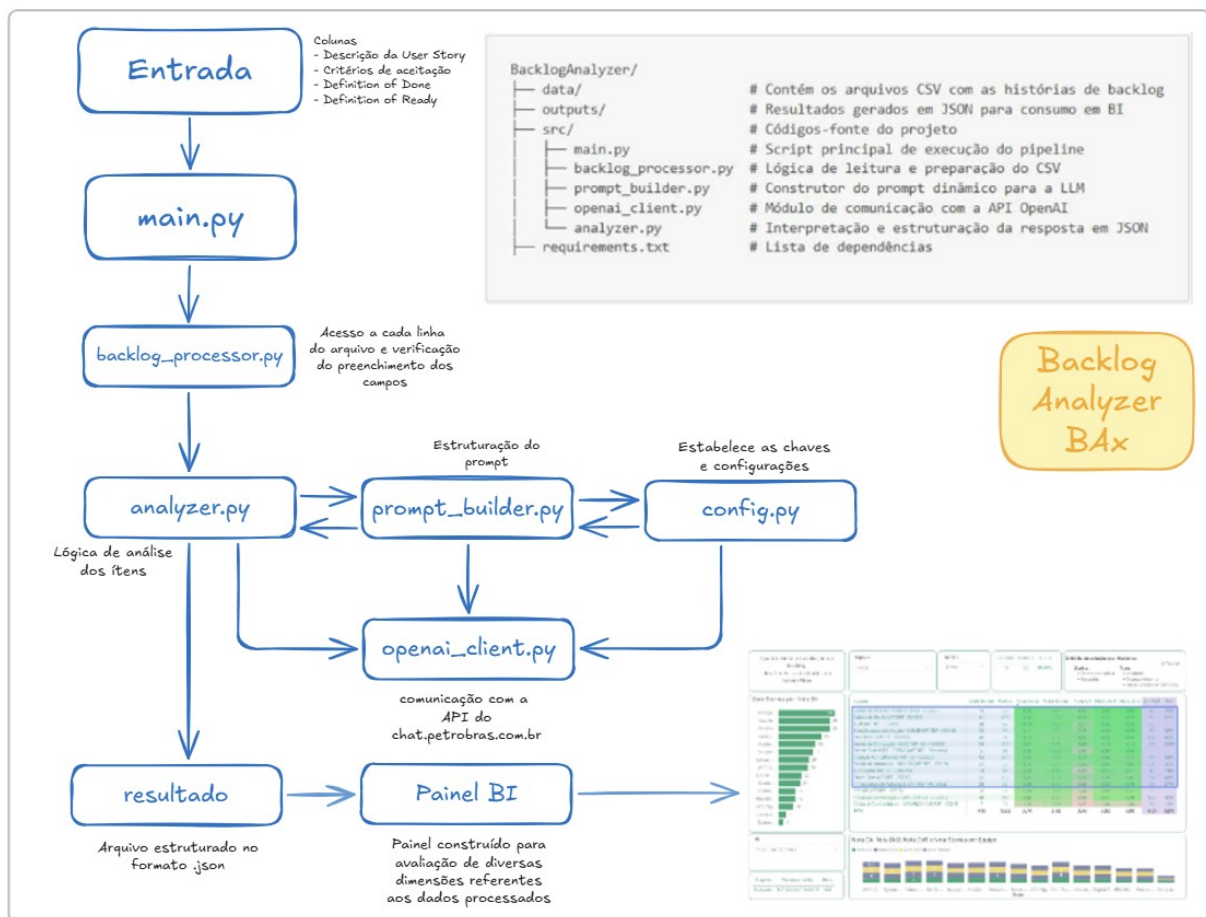


Figura 27 – Diagrama esquemático do Backlog Analyzer

Fonte:Elaborado pelo autor.

A implementação contempla uma etapa inicial de ingestão e pré-processamento dos dados, na qual os itens do backlog são importados a partir de arquivos estruturados, nos formatos .csv ou .xlsx. Esses arquivos contêm, para cada história de usuário, informações como descrição textual, Critérios de Aceitação, *Definition of Ready* e *Definition of Done*. Durante o pré-processamento, são realizadas operações de padronização

dos campos, tratamento de valores ausentes e normalização textual, assegurando consistência na entrada dos dados analisados pelo sistema.

Em seguida, os dados processados são submetidos ao módulo de avaliação, no qual ocorre a construção dinâmica dos prompts enviados ao modelo de linguagem. Esses prompts incorporam instruções explícitas sobre o papel assumido pelo modelo, os critérios de avaliação a serem considerados, a escala de pontuação adotada e o formato estruturado da resposta esperada. Essa etapa mostrou-se crítica para garantir a consistência e a qualidade das avaliações geradas, reforçando a importância da engenharia de prompts como elemento central da implementação.

A escolha dos hiperparâmetros seguiu diretrizes técnicas associadas à engenharia de prompts para tarefas de extração e classificação de informações textuais. De acordo com a documentação técnica da OpenAI (2024), o parâmetro *temperature* controla o nível de aleatoriedade das respostas geradas pelo modelo. Valores próximos de zero tornam a geração mais determinística, reduzindo variações entre execuções sucessivas.

No presente artefato, o valor de *temperature* foi configurado em 0.0 com o objetivo de maximizar a reprodutibilidade das avaliações realizadas pelo sistema. Nessa configuração, o modelo tende a selecionar consistentemente as respostas mais prováveis, reduzindo a variabilidade nas análises realizadas sobre um mesmo item de backlog.

De forma complementar, o parâmetro *top_p*, associado à técnica de amostragem por núcleo (*nucleus sampling*), foi ajustado para 0.1. Essa configuração restringe o conjunto de tokens candidatos às alternativas mais prováveis durante o processo de geração de texto. A combinação desses parâmetros contribui para reduzir variações semânticas indesejadas e aumentar a consistência das respostas produzidas pelo modelo.

Essa estratégia de configuração é particularmente importante no contexto do artefato proposto, pois permite que o modelo de linguagem seja utilizado como um mecanismo de avaliação estruturada, aproximando seu comportamento de um instrumento analítico padronizado, em vez de um sistema voltado à geração criativa de texto.

Embora a configuração de *temperature* = 0,0 e *top_p* = 0,1 maximize o determinismo das respostas, é importante reconhecer que LLMs possuem caráter inerente-

Tabela 10 – Configuração dos hiperparâmetros utilizados no módulo Backlog Analyzer

Parâmetro	Configuração Adotada	Justificativa Técnica
Modelo (engine)	gpt-4o	Modelo de linguagem de larga escala utilizado para análise semântica e avaliação textual dos itens do backlog.
Temperatura	0.0	Configuração que reduz a aleatoriedade das respostas, favorecendo comportamento determinístico e maior reprodutibilidade das avaliações.
Top_P	0.1	Aplicação de <i>nucleus sampling</i> restrita para limitar variações semânticas e aumentar a consistência das respostas geradas.
Política de tentativas	10	Mecanismo de resiliência para falhas de rede ou limites de taxa (<i>rate limiting</i>) da API. Implementado no código da aplicação.
Intervalo entre tentativas	3 segundos	Tempo de espera entre requisições consecutivas para evitar bloqueios por concorrência de chamadas à API.

Fonte: Elaborado pelo autor.

mente probabilístico e nunca são completamente determinísticos. Mesmo com esses parâmetros, podem ocorrer variações residuais nas avaliações produzidas para um mesmo item, especialmente em casos limítrofes entre dois níveis da escala ou quando o texto avaliado apresenta ambiguidades semânticas.

Atualizações do modelo pelo provedor, como mudanças de versão do *gpt-4o*, podem alterar o comportamento de geração de texto, introduzindo variações sistemáticas nos resultados ao longo do tempo, mesmo que o *prompt* e os hiperparâmetros permaneçam idênticos. Essa característica é uma limitação estrutural do uso de LLMs como mecanismo de avaliação e deve ser considerada na interpretação dos resultados.

A estratégia adotada para mitigar esse risco compreende três elementos: **(i)** fixação dos hiperparâmetros de reprodutibilidade em valores mínimos ao longo de toda a pesquisa; **(ii)** manutenção de um *prompt* padrão versionado, evitando alterações não documentadas nos critérios de avaliação; e **(iii)** priorização da análise de tendências e médias agregadas em vez de valores pontuais de itens individuais, uma vez que variações residuais tendem a ser diluídas em análises de maior granularidade. A não-determinabilidade residual é reconhecida como limitação metodológica e retomada na seção de limitações da pesquisa.

As respostas retornadas pelo modelo de linguagem são estruturadas em formato

JSON, contendo, para cada dimensão avaliada, a pontuação atribuída, uma justificativa textual e recomendações de melhoria. Essas respostas passam por um processo de validação e interpretação, no qual o sistema verifica a completude das informações e trata eventuais inconsistências ou omissões, garantindo a integridade dos resultados armazenados.

Os dados consolidados são então persistidos em repositórios estruturados, possibilitando análises históricas e comparativas. A partir desses repositórios, foram desenvolvidos painéis de visualização utilizando ferramentas de *Business Intelligence*, que permitem acompanhar a qualidade do backlog por equipe, sprint ou período de tempo. Esses painéis fornecem suporte à tomada de decisão gerencial, facilitando a identificação de padrões recorrentes, pontos críticos e oportunidades de melhoria.

A aplicação do artefato ocorreu em um contexto organizacional real, inserido em uma indústria de grande porte que adota metodologias ágeis em escala. O sistema foi utilizado tanto para apoiar a elaboração e o refinamento de histórias de usuário quanto para avaliar a qualidade dos artefatos produzidos e coletar feedback das equipes após a execução das histórias. Essa aplicação permitiu observar variações na qualidade dos artefatos e testar a escalabilidade da solução.

4.4.3 Backlog Booster

O módulo denominado Backlog Booster foi concebido para atuar em uma etapa anterior à avaliação da qualidade do backlog, oferecendo suporte à elaboração e ao refinamento das histórias de usuário. Enquanto o Backlog Analyzer desempenha um papel de diagnóstico da qualidade dos artefatos existentes, o Backlog Booster busca apoiar a construção inicial desses itens, auxiliando Product Owners e equipes ágeis na elaboração de descrições mais claras e completas.

Do ponto de vista operacional, o módulo utiliza modelos de linguagem de larga escala para analisar descrições preliminares de histórias de usuário e gerar sugestões de aprimoramento textual. A partir de uma entrada contendo uma descrição inicial do requisito, o sistema constrói prompts estruturados que orientam o modelo de linguagem a revisar a formulação da User Story, sugerir melhorias nos critérios de aceitação e indicar elementos relevantes associados à Definition of Ready e à Definition of Done. As respostas geradas pelo modelo são apresentadas em formato textual estruturado,

permitindo que os usuários comparem a versão original da história com as sugestões produzidas pelo sistema. Dessa forma, o Backlog Booster funciona como um mecanismo de apoio à escrita técnica do backlog, incentivando a adoção de práticas recomendadas da engenharia de requisitos ágil, como a explicitação do valor de negócio e a definição de critérios verificáveis.

No contexto organizacional estudado, o módulo foi utilizado principalmente durante atividades de refinamento do backlog. Nessas sessões, Product Owners e membros das equipes analisavam histórias preliminares e utilizavam as sugestões geradas pelo sistema como insumo para aprimorar a clareza e a completude das descrições antes de sua avaliação formal pelo Backlog Analyzer.

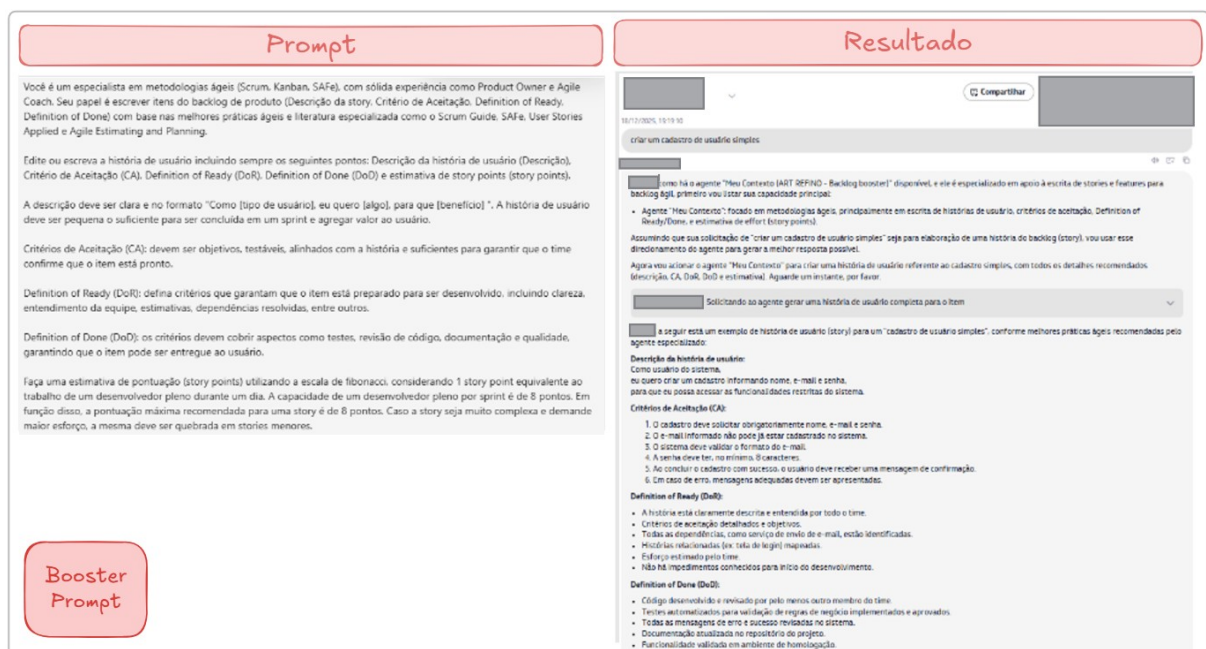


Figura 28 – Diagrama esquemático do Prompt Booster

Fonte: Elaborado pelo autor.

4.4.4 Feedback Analyzer

O módulo Feedback Analyzer foi desenvolvido com o objetivo de incorporar a perspectiva das equipes de desenvolvimento após a execução das histórias de usuário. Diferentemente dos módulos voltados à construção e à avaliação dos itens do backlog, esse componente atua em uma etapa posterior do ciclo de desenvolvimento, buscando transformar a experiência de implementação em insumos para melhoria contínua da qualidade dos requisitos.

Para esse propósito, o módulo coleta feedback estruturado dos desenvolvedores por meio de questionários organizados em escala Likert de cinco pontos. As questões abordam aspectos como clareza da descrição da história, adequação dos critérios de aceitação, suficiência das informações disponíveis para implementação e consistência entre o requisito descrito e a funcionalidade efetivamente desenvolvida. Os dados coletados são consolidados em bases estruturadas e posteriormente analisados pelo sistema, permitindo identificar padrões recorrentes relacionados à qualidade dos itens do backlog e compreender de que forma determinadas características das histórias influenciam o processo de desenvolvimento.

No contexto organizacional analisado, o Feedback Analyzer foi aplicado após a conclusão das sprints, permitindo que os desenvolvedores registrassem suas percepções sobre as histórias implementadas. Os resultados obtidos foram posteriormente comparados com as avaliações produzidas pelo Backlog Analyzer, contribuindo para validar a utilidade do método proposto e para identificar oportunidades adicionais de aprimoramento na elaboração dos itens do backlog.

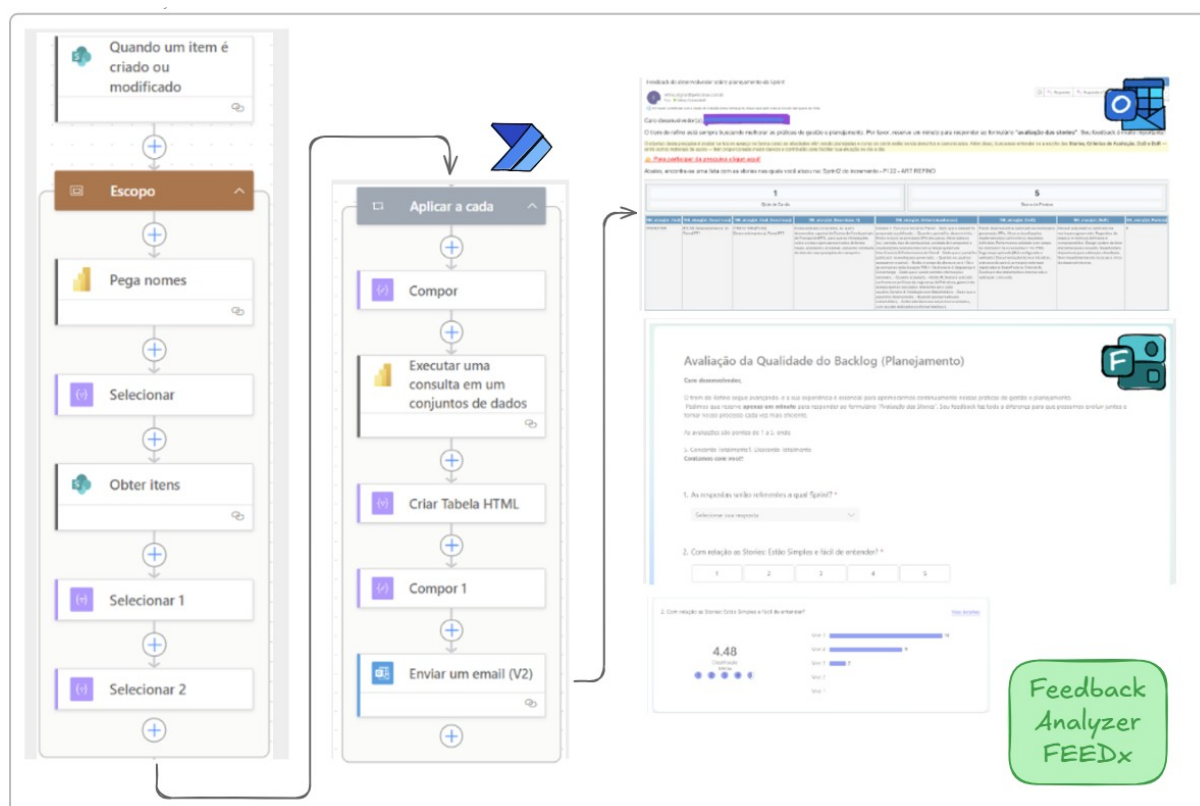


Figura 29 – Diagrama esquemático do Prompt Booster

Fonte:Elaborado pelo autor.

Além da coleta estruturada em escala Likert, o módulo incorpora um compo-

nente de análise qualitativa dos comentários livres registrados pelos desenvolvedores ao final de cada sprint. Para esse fim, foi desenvolvido um script em linguagem Python que implementa análise de sentimento baseada em léxico (*lexicon-based sentiment analysis*), sem uso de bibliotecas de aprendizado de máquina ou modelos pré-treinados. A abordagem léxica foi adotada em função do volume do corpus (112 comentários ao longo de dois *Program Increments*), da necessidade de controle explícito sobre o vocabulário específico do domínio em português e da exigência de interpretabilidade total do processo de classificação.

O algoritmo opera em três etapas. Na primeira, cada comentário é varrido em busca de ocorrências de palavras e expressões pertencentes a dois léxicos manuais definidos pelo pesquisador em português: um léxico de termos positivos (P) e um de termos negativos (N). Na segunda etapa, aplica-se tratamento de negações: palavras como "*não*", "*nem*" e "*nunca*", quando precedem imediatamente uma palavra positiva, invertem o sinal da ocorrência, que deixa de incrementar P e passa a incrementar N . Na terceira etapa, calcula-se um *score* de sentimento por comentário, conforme a Equação 1:

$$score = \frac{P - N}{P + N} \quad (1)$$

em que P e N representam, respectivamente, o total de ocorrências de palavras dos léxicos positivo e negativo identificadas no texto, após o tratamento de negações. O *score* resultante situa-se no intervalo $[-1, +1]$: valores positivos indicam predominância de tom afirmativo; valores negativos, predominância de tom crítico; e *score* = 0 denota equilíbrio ou ausência de palavras reconhecidas pelo léxico. Os comentários são classificados como **Positivos** quando $P > N$, **Negativos** quando $N > P$, e **Neutros** nos demais casos. Para análise por sprint ou por mês, calcula-se a média aritmética dos *scores* individuais de todos os comentários do período.

Adicionalmente, cada comentário é classificado em um único tema organizacional a partir de um dicionário de palavras-chave estruturado pelo pesquisador, composto por categorias como *Qualidade dos Cards*, *Refinamento/Planning*, *Uso de LLM/IA*, *Participação do PO* e *Escopo/Mudanças*, entre outras. A classificação temática é realizada por busca de substring no texto normalizado, com prioridade definida pela ordem do dicionário. Comentários enquadrados em expressões neutras padrão, como "*ok*" ou

"nada a comentar", recebem a categoria *Sem Observações* e são excluídos da análise de causa-raiz. Essa estratégia combinada de sentimento e tema permite identificar não apenas o tom dos comentários, mas também os aspectos específicos do processo de planejamento aos quais as percepções positivas e negativas estão associadas.

4.4.5 Integração do artefato na rotina

Durante a aplicação, observou-se que o artefato foi capaz de operar de forma integrada aos rituais ágeis existentes, sem impor alterações significativas no fluxo de trabalho das equipes. A avaliação automatizada forneceu subsídios objetivos para discussões tradicionalmente subjetivas, contribuindo para maior alinhamento conceitual entre os envolvidos e redução do esforço manual requerido para análise da qualidade das histórias.

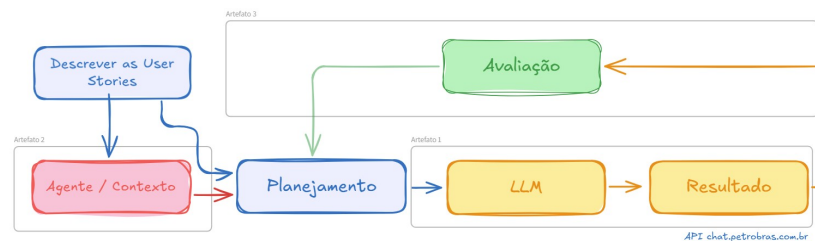


Figura 30 – Fluxograma resumido do artefato

Fonte: Elaborado pelo autor.

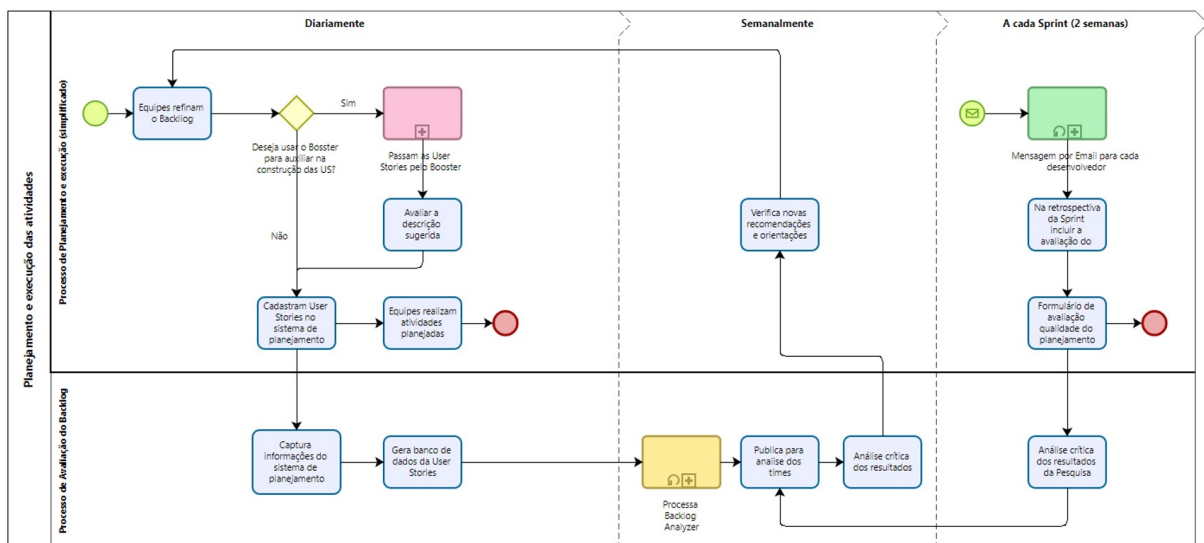


Figura 31 – Fluxograma das atividades de planejamento e utilização do artefato

Fonte: Elaborado pelo autor.

Assim, a implementação do artefato demonstrou a viabilidade técnica e operacional do método proposto, evidenciando sua adequação ao contexto organizacional

estudado. A integração do sistema aos rituais de planejamento e refinamento do backlog permitiu não apenas a geração de avaliações estruturadas sobre a qualidade dos itens analisados, mas também a produção de dados que subsidiam análises evolutivas sobre a maturidade do backlog ao longo do tempo.

Essas informações constituem a base empírica utilizada na etapa de avaliação do artefato, apresentada no capítulo seguinte, no qual são discutidos os resultados obtidos a partir da aplicação do método no contexto organizacional investigado.

4.4.6 Evolução e Processamento Incremental do Artefato

O desenvolvimento do artefato não ocorreu de forma linear, mas por meio de sucessivos ciclos de aprendizado e refinamento, consistentes com os princípios iterativos da *Design Science Research* (Dresch; Lacerda; Antunes Junior, 2015). A trajetória de evolução do método pode ser sintetizada em cinco fases distintas:

Fase 1: Validação manual do método (agosto de 2025). Nos estágios iniciais, as avaliações eram realizadas de forma pontual e predominantemente manual, a partir de exportações diretas dos dados de backlog. O objetivo central dessa etapa era validar a viabilidade conceitual do método, verificar se os critérios de avaliação eram aplicáveis ao contexto organizacional e compreender a estrutura dos dados disponíveis nos sistemas corporativos. Essa fase teve caráter exploratório e permitiu identificar os campos relevantes para a avaliação da qualidade dos cards, bem como as limitações decorrentes da dependência de extrações manuais.

Fase 2: Desenvolvimento da coleta automatizada. A partir da validação inicial, passou-se ao estudo da arquitetura informacional dos sistemas corporativos de gestão do backlog, incluindo APIs de acesso, formatos de dados, regras de atualização e mecanismos de integração disponíveis. Essa fase resultou no desenvolvimento dos primeiros pipelines de coleta automatizada, reduzindo a dependência de intervenções manuais recorrentes e criando uma base técnica mais estável para a execução contínua do método. Com isso, o artefato deixou de operar apenas como uma prova conceitual e passou a assumir características de uma solução operacional.

Fase 3: Processamento em lote (*full-batch*). Com a automação da coleta, o sistema passou a processar integralmente todas as histórias disponíveis a cada execução diária. Essa abordagem foi adequada para os primeiros ciclos de uso, pois permitiu

validar o comportamento do pipeline de ponta a ponta, desde a extração dos dados até a geração dos resultados de avaliação. Entretanto, à medida que o volume analisado ultrapassou aproximadamente 500 histórias, observou-se que o processamento integral diário não seria sustentável em termos de tempo de execução, consumo de recursos computacionais e quantidade de chamadas ao modelo de linguagem. Essa limitação tornou-se especialmente relevante diante da expectativa de expansão do método para novos times e trens ágeis.

Fase 4: Heurística de processamento incremental. Em resposta à limitação de escalabilidade identificada na etapa anterior, foi concebida e implementada uma heurística de processamento incremental. O sistema passou a identificar automaticamente alterações nos campos avaliados, incluindo descrição, Critérios de Aceitação, Definition of Ready e Definition of Done. Dessa forma, somente os itens com modificações detectadas em relação à execução anterior passaram a ser submetidos a novo processamento pela IA, enquanto os demais preservaram seus resultados anteriores. Essa estratégia reduziu significativamente o volume de chamadas ao modelo de linguagem e o tempo total de execução, tornando o processamento diário viável mesmo em um cenário de ampliação do número de cards monitorados.

Fase 5: Operacionalização diária e monitoramento do pipeline Após a estabilização da heurística incremental, a evolução seguinte consistiu na operacionalização da execução diária automatizada do pipeline. Nessa fase, o foco deixou de se restringir à viabilidade técnica do processamento e passou a incorporar o acompanhamento sistemático de seu desempenho em ambiente de produção. Para isso, foram criados indicadores específicos para monitorar o volume total de cards acompanhados, a quantidade de itens efetivamente submetidos ao reprocessamento pela IA, a taxa diária de reprocessamento e o comportamento operacional do sistema ao longo do tempo. Esses mecanismos de monitoramento somente puderam ser ativados após a consolidação do pipeline de tratamento de dados (*ETL*) e a validação do comportamento da solução em produção. Por esse motivo, as séries históricas apresentadas nesta pesquisa não abrangem toda a trajetória do projeto, iniciada em agosto de 2025, mas apenas o período a partir do qual a infraestrutura de monitoramento se encontrava plenamente operacional.

A efetividade da automação incremental pode ser observada diretamente na Figura 32. A linha laranja, que representa os cards efetivamente submetidos ao pro-

cessamento pela IA, mantém-se consistentemente abaixo da linha azul, que registra o volume total de cards monitorados no Kanban. Esse padrão evidencia que o sistema não reprocessa integralmente todos os itens a cada execução diária, mas encaminha para a IA apenas os cards com alterações relevantes nos campos avaliados.

No período analisado, de fevereiro de 2026 a junho de 2026, o sistema monitorou diariamente os cards presentes no quadro Kanban, totalizando 49.661 registros avaliados ao longo de 74 dias. Desse universo, apenas 2.870 cards foram efetivamente submetidos ao processamento pela IA, o que corresponde a uma taxa de processamento de 5,8%. Esse resultado demonstra o efeito prático da melhoria implementada no algoritmo incremental, ao reduzir de forma significativa a quantidade de chamadas ao modelo de linguagem e, conseqüentemente, o custo computacional do processo.

Dessa forma, a heurística incremental, concebida na fase anterior, passou a operar de maneira estável em uma rotina automatizada de produção. Além de comprovar a eficiência do processamento diário, os indicadores de monitoramento permitiram avaliar a viabilidade de expansão do método para outros ARTs. Considerando o universo atual de mais de 50 ARTs ativos na organização, com possibilidade de alcançar mais de 200 times em um cenário de adoção ampla, a manutenção do padrão incremental e automatizado constitui o principal mecanismo para tornar essa escalabilidade computacionalmente viável.

A trajetória descrita evidencia que o artefato amadureceu progressivamente como resultado da interação contínua entre a solução proposta e o ambiente organizacional real, em coerência com o caráter prescritivo e iterativo da *Design Science Research*. Cada fase contribuiu para transformar o método, inicialmente validado de forma manual e exploratória, em uma solução automatizada, monitorada e potencialmente escalável para uso organizacional ampliado.

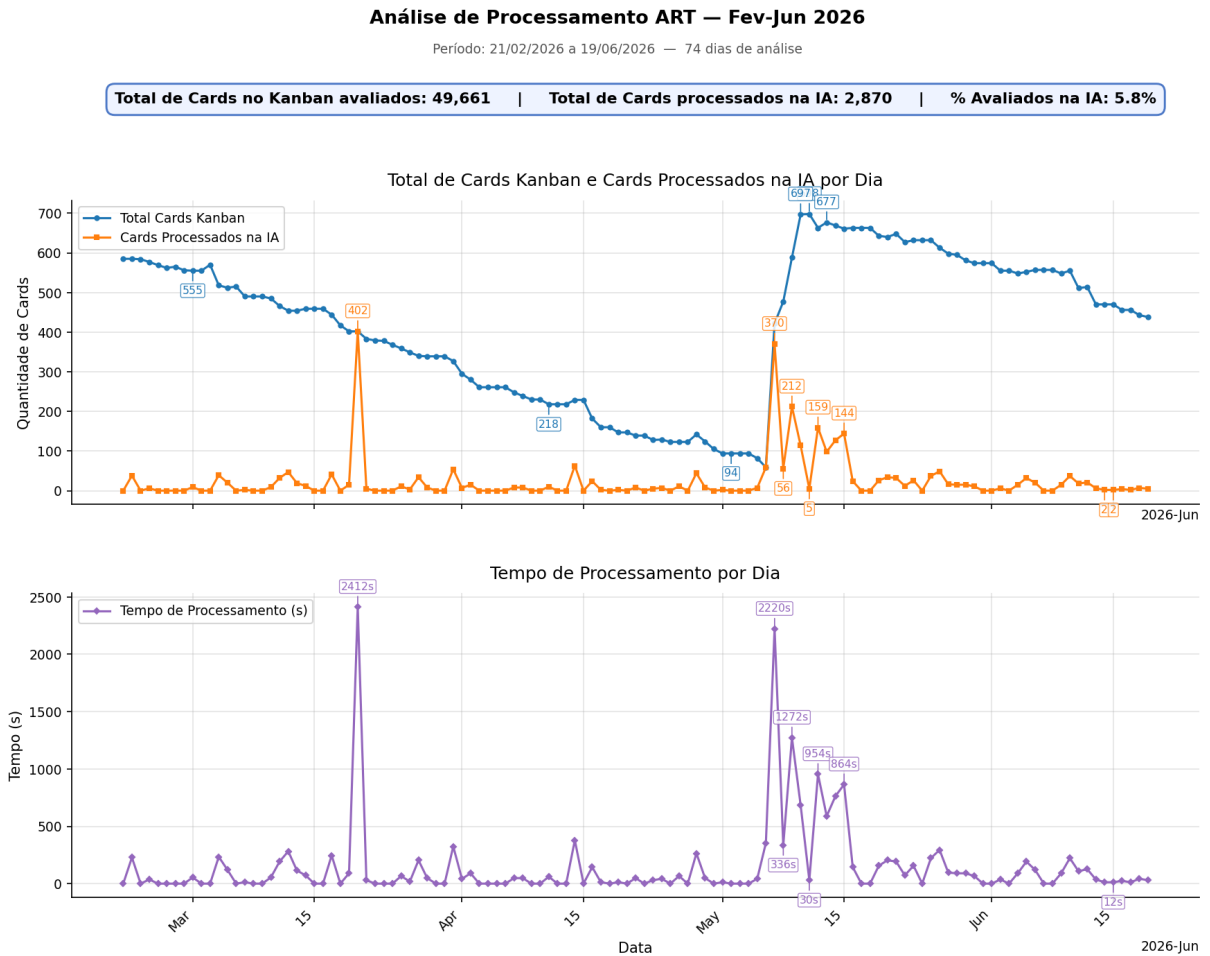


Figura 32 – Análise longitudinal de tempos de processamento
Fonte: Elaborado pelo autor.

4.4.7 Análise Preliminar de Custos Operacionais

A viabilidade operacional do método proposto não depende apenas de sua eficácia técnica, mas também de sua sustentabilidade econômica. A utilização de modelos de linguagem via API comercial implica custos proporcionais ao volume de texto processado (*tokens*) e à frequência de execução. Diante disso, foi realizada uma avaliação preliminar dos custos incorridos durante a fase de experimentação do método, compreendida entre junho e agosto de 2025.

Nesse período, os experimentos foram conduzidos utilizando diretamente a API comercial do modelo *gpt-4o*. O volume analisado durante essa fase correspondeu a **8.105 histórias de usuário**, processadas em 547 requisições à API com consumo total de 634.200 *tokens*, contemplando avaliação de quatro dimensões por item: Story, Critérios de Aceitação, DoR e DoD. O custo total acumulado durante o período de testes foi de **US\$ 3,93**, demonstrando potencial de viabilidade econômica mesmo para

volumes expressivos de análise.

Posteriormente, com a adoção da solução em ambiente corporativo, o processamento passou a utilizar infraestrutura institucional baseada em contratos corporativos de IA, na qual os custos individuais não são diretamente observáveis no contexto desta pesquisa. Os valores descritos devem, portanto, ser interpretados como referência empírica da fase de experimentação, não como estimativa dos custos correntes de operação.

O custo computacional observado representa uma parcela reduzida quando confrontado ao esforço manual equivalente para realizar avaliações semelhantes em um volume comparável de histórias. A implementação da heurística de processamento incremental descrita na subseção anterior reduz ainda mais o consumo de recursos nas execuções subsequentes, uma vez que apenas os itens efetivamente alterados são submetidos a reprocessamento. A possibilidade futura de migração para modelos de linguagem menores (*Small Language Models* – SLMs)¹ ou implantados localmente em infraestrutura corporativa representa uma alternativa de escalabilidade que pode reduzir substancialmente os custos operacionais de longo prazo, ampliando a viabilidade do método em diferentes perfis organizacionais.

4.5 Síntese do Capítulo

Este capítulo apresentou o desenvolvimento do estudo conduzido nesta pesquisa, descrevendo de forma sistemática o percurso adotado desde a contextualização do problema até a implementação e aplicação do artefato proposto. Em consonância com a abordagem da *Design Science Research*, o foco esteve na tradução dos *GAPs* identificados na literatura e na prática organizacional em uma solução estruturada, tecnicamente viável e orientada à resolução de um problema real.

Inicialmente, foram contextualizadas as dificuldades recorrentes associadas à gestão do *Product Backlog* em ambientes ágeis em escala, destacando-se limitações relacionadas à subjetividade na avaliação da qualidade dos itens, à ausência de critérios

¹ Small Language Models (SLMs), ou modelos pequenos de linguagem, são modelos de inteligência artificial voltados ao processamento e à geração de linguagem natural, porém com menor número de parâmetros, menor custo computacional e menor demanda de infraestrutura quando comparados aos *Large Language Models* (LLMs). Em geral, são empregados em tarefas mais específicas, em ambientes com restrição de recursos ou em aplicações que demandam menor latência, maior eficiência e possibilidade de execução local ou embarcada (Lu et al., 2024).

padronizados e mensuráveis, à baixa escalabilidade dos processos tradicionais e à carência de mecanismos sistemáticos de aprendizado organizacional. Esses *GAPs* forneceram a base conceitual e prática para a concepção do artefato desenvolvido ao longo do estudo.

Na sequência, foi apresentada a concepção do método de avaliação da qualidade do backlog, estruturado a partir de dimensões centrais da engenharia de requisitos ágil e operacionalizado por meio de critérios explícitos e de uma escala de mensuração capaz de capturar níveis crescentes de maturidade técnica e textual. Essa concepção buscou equilibrar o rigor analítico necessário à mensuração da qualidade com a flexibilidade inerente aos métodos ágeis.

O capítulo também detalhou a arquitetura do sistema computacional que materializa o método proposto, evidenciando sua organização modular e o papel desempenhado por cada componente no processo de avaliação, apoio à escrita e retroalimentação. A implementação técnica e a aplicação do artefato em um contexto organizacional real demonstraram a viabilidade operacional da solução, bem como sua capacidade de integrar-se aos rituais ágeis existentes sem impor rupturas significativas no fluxo de trabalho das equipes.

Dessa forma, o Capítulo 4 cumpriu o objetivo de articular os fundamentos teóricos e metodológicos apresentados nos capítulos anteriores com a materialização da solução proposta, estabelecendo um encadeamento lógico entre a identificação do problema, a concepção do método, a implementação do sistema e sua aplicação no contexto organizacional.

5 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos a partir da aplicação do artefato desenvolvido nesta pesquisa, bem como sua análise e discussão à luz dos objetivos do estudo e dos *GAPs* identificados nos capítulos anteriores. Em consonância com a abordagem da *Design Science Research*, os resultados aqui discutidos não se restringem a métricas puramente quantitativas, mas abrangem evidências relacionadas à utilidade, viabilidade, consistência e capacidade diagnóstica da solução proposta.

Os resultados derivam da aplicação prática do método e do sistema computacional em um contexto organizacional real, envolvendo a avaliação da qualidade de itens de *Product Backlog* provenientes de diferentes equipes ágeis. A análise considera tanto os dados gerados pelo sistema, tais como pontuações, classificações e recomendações. Quanto as implicações observadas no processo de planejamento, refinamento e tomada de decisão, a partir do uso das informações produzidas pelo artefato.

Em uma perspectiva alinhada à DSR, a avaliação do artefato concentra-se na verificação de sua capacidade de endereçar o problema prático que motivou a pesquisa, isto é, a ausência de um método estruturado, padronizado e escalável para avaliação da qualidade e maturidade do backlog. Assim, os resultados são analisados considerando critérios como redução da subjetividade, possibilidade de comparação entre equipes, apoio à identificação de fragilidades recorrentes e potencial para promoção de aprendizado organizacional.

O capítulo está estruturado de modo a apresentar, inicialmente, os resultados diretos da aplicação do artefato e a análise longitudinal da evolução da qualidade dos artefatos ao longo dos três *Program Increments*. Em seguida, são examinadas as associações entre a qualidade do backlog e os indicadores de eficiência do planejamento das sprints, bem como os resultados da pesquisa de percepção com desenvolvedores e a análise de sentimento dos comentários coletados. Esses resultados são então analisados à luz dos *GAPs* identificados e discutidos em termos de implicações práticas e teóricas. Por fim, são sintetizadas as principais evidências obtidas, preparando o terreno para as conclusões e trabalhos futuros apresentados no capítulo subsequente.

5.1 Premissas e Delimitações da Avaliação da Qualidade do Backlog

Os resultados apresentados neste capítulo derivam da aplicação do método aos artefatos textuais registrados no sistema corporativo de gestão ágil. Antes da apresentação dos dados, é necessário explicitar as premissas que delimitam o escopo desta avaliação, de modo a evitar interpretações equivocadas sobre o fenômeno analisado.

O método avalia a **qualidade das informações registradas** nos itens do *Product Backlog*, incluindo histórias de usuário, critérios de aceitação, *Definition of Ready* e *Definition of Done*. Não se trata, portanto, de uma avaliação da competência de planejamento das equipes. Em ambientes ágeis em escala, é comum que parte significativa do planejamento e do refinamento ocorra por meio de interações verbais e conhecimentos compartilhados que não são integralmente registrados nos cards do sistema. Dessa forma, o artefato mede o grau de **formalização e explicitação das informações no registro oficial**, que constitui o principal mecanismo de acesso ao contexto dos itens para outros atores organizacionais, como equipes dependentes, *Release Train Engineers*, lideranças técnicas e sistemas de automação.

Essa distinção é relevante por duas razões. Em primeiro lugar, uma nota baixa não implica necessariamente que a equipe desconheça o trabalho a ser realizado. Em muitos casos, ela indica apenas que o conhecimento permaneceu tácito e não foi adequadamente formalizado. Em segundo lugar, a evolução das notas ao longo do tempo reflete principalmente um processo de **explicitação progressiva do conhecimento compartilhado**. Assim, a melhoria observada não deve ser interpretada, necessariamente, como um aumento da competência técnica das equipes na atividade de planejamento.

O principal benefício proporcionado pelo método possui natureza informacional e organizacional. Entre os ganhos observados destacam-se o aumento da visibilidade sobre a qualidade dos artefatos, a padronização dos critérios de avaliação, a ampliação da transparência para a governança do ART e o fortalecimento da rastreabilidade para auditorias e análises históricas. Tais benefícios manifestam-se independentemente de as equipes já possuírem o conhecimento necessário para executar o trabalho. O que o método promove é a conversão desse conhecimento tácito em um ativo organizacional explícito, compartilhável e passível de mensuração.

5.2 Resultados da Aplicação do Artefato

Os resultados apresentados nesta seção referem-se à aplicação prática do artefato desenvolvido neste estudo em um contexto organizacional real que adota metodologias ágeis em escala. A aplicação teve como objetivo avaliar a viabilidade operacional do método proposto, bem como observar os dados e evidências produzidos a partir da avaliação sistemática da qualidade dos itens de *Product Backlog*.

A figura 33 mostra o dashboard de avaliação das notas de todos os times do ART antes da aplicação do método, enquanto a figura 34 mostra a situação do mesmo quadro após a aplicação do método.



Figura 33 – Avaliação inicial da qualidade dos itens de backlog antes da aplicação do método. Data: 25/08/2025, times analisados: 16, times com nota acima da nota 3: 1, média das notas: 0,98

Fonte: Elaborado pelo autor



Figura 34 – Avaliação inicial da qualidade dos itens de backlog após a aplicação do método. Data: 18/03/2026, times analisados: 15, times com nota acima da nota 3: 11, média das notas: 3,85

Fonte: Elaborado pelo autor

Com base nesse panorama comparativo entre o estado inicial do backlog e o patamar atingido após a adoção sistemática do método, as subseções seguintes detalham a análise das quatro dimensões avaliadas, a evolução longitudinal das notas ao longo dos três ciclos de planejamento e os padrões estruturais identificados no processo de refinamento.

O artefato foi aplicado sobre um conjunto de histórias de usuário provenientes de diferentes equipes ágeis, contemplando itens em distintos níveis de maturidade e estágios de refinamento. Para cada história avaliada, foram analisadas as quatro dimensões definidas no método: descrição da *User Story*, Critérios de Aceitação, *Definition of Ready* e *Definition of Done*. A avaliação resultou na atribuição de pontuações individuais por dimensão, acompanhadas de justificativas textuais e recomendações de melhoria.

Os resultados evidenciaram variações significativas na qualidade dos itens analisados, tanto entre equipes distintas quanto dentro de um mesmo backlog. Observou-se que histórias de usuário frequentemente apresentavam descrições compreensíveis em nível conceitual, mas careciam de detalhamento suficiente para suportar uma execução previsível, especialmente no que se refere à definição clara de critérios de aceitação e à explicitação das condições de prontidão e conclusão.

Referente aos Critérios de Aceitação, os dados indicaram que uma parcela relevante das histórias avaliadas possuía critérios genéricos ou pouco testáveis, dificultando a verificação objetiva do atendimento aos requisitos. Em diversos casos, os critérios estavam implícitos na descrição da história ou formulados de maneira ambígua, o que resultou em pontuações intermediárias na escala adotada.

Em relação à *Definition of Ready*, observou-se elevada heterogeneidade na sua aplicação. Algumas histórias apresentavam informações suficientes para serem consideradas prontas para execução, enquanto outras careciam de elementos básicos, como entendimento compartilhado do escopo, dependências técnicas resolvidas ou critérios claros de aceitação.

A avaliação da *Definition of Done* indicou, em muitos casos, sua presença apenas de forma implícita ou genérica, frequentemente associada a práticas organizacionais não explicitadas no item do backlog. Essa característica evidenciou fragilidades na explicitação dos critérios de conclusão e qualidade esperada das entregas.

Além das pontuações numéricas, o sistema gerou recomendações estruturadas de melhoria para cada item avaliado. Essas recomendações incluíram sugestões de reescrita das histórias de usuário, reformulação de critérios de aceitação para torná-los mais objetivos e testáveis, bem como orientações para explicitar condições mínimas de prontidão e critérios claros de conclusão.

Do ponto de vista operacional, a aplicação do artefato demonstrou capacidade de processar múltiplos itens de backlog de forma automatizada, mantendo consistência na aplicação dos critérios e na atribuição das pontuações. Os resultados foram consolidados em repositórios estruturados, permitindo sua visualização por meio de painéis gerenciais e viabilizando análises comparativas por equipe, período ou dimensão avaliada.

A análise dos dados também permitiu observar a evolução da qualidade estrutural dos itens avaliados ao longo do período de aplicação do método. Para facilitar a interpretação dos resultados, as pontuações atribuídas foram agrupadas em três faixas de qualidade: vermelho (nota 0), amarelo (notas 1 e 2) e verde (notas 3, 4 e 5).

A comparação entre diferentes momentos da coleta de dados revelou uma mudança significativa na distribuição das avaliações. Em agosto de 2025, aproximadamente 57,40% dos itens avaliados encontravam-se na faixa vermelha, indicando elevada incidência de histórias sem nenhuma descrição. Já na avaliação realizada em

fevereiro de 2026, essa proporção foi reduzida para 1,58%, evidenciando uma queda expressiva na ocorrência de itens considerados inadequados.

Tabela 11 – Comparação da distribuição das avaliações de qualidade do backlog

Data da Avaliação	● (Nota 0)	● (Notas 1 e 2)	● (Notas 3, 4 e 5)
25/08/2025	57,40%	16,96%	25,64%
25/02/2026	1,58%	6,85%	91,56%

Fonte: Elaborado pelo autor a partir dos dados gerados pelo Backlog Analyzer.

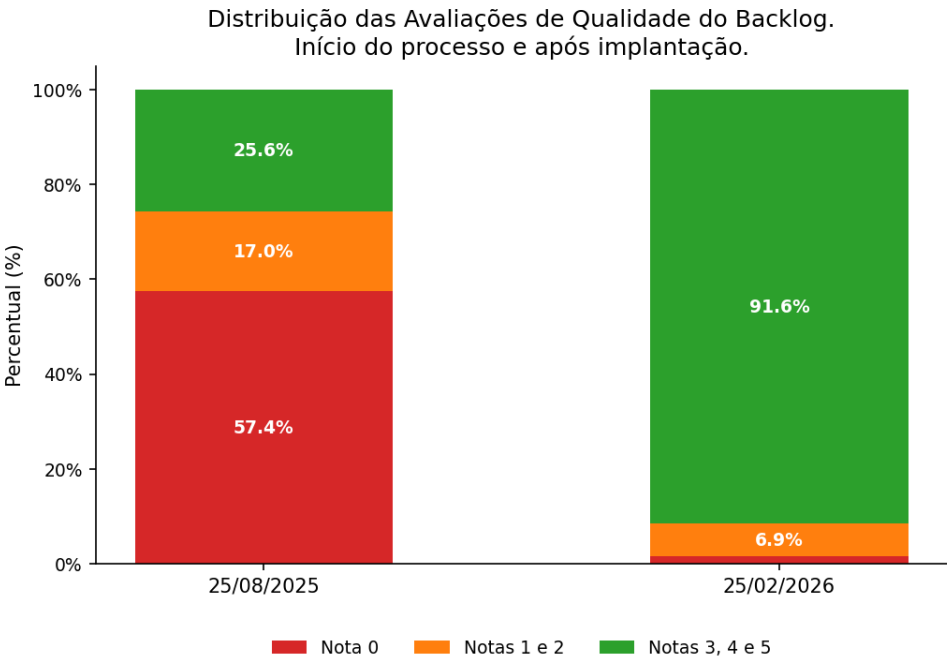


Figura 35 – Distribuição das avaliações de qualidade do backlog. Início do processo e após implantação.

Fonte: Elaborado pelo autor a partir dos dados gerados pelo Backlog Analyzer.

Paralelamente, observou-se um crescimento substancial da faixa verde, que passou de 25,64% para 91,56% dos itens avaliados. Esse resultado sugere que a maioria das histórias passou a apresentar características estruturais compatíveis com boas práticas de engenharia de requisitos em ambientes ágeis, incluindo maior clareza na descrição do problema, presença consistente de critérios de aceitação e maior alinhamento entre solicitação e entrega esperada.

Além da comparação entre os momentos inicial e final da aplicação do método, os dados também permitiram observar a evolução longitudinal da qualidade dos artefatos ao longo do tempo. No início da coleta sistemática dos dados, em julho de 2025, a nota geral média das avaliações era de 1,36, indicando um nível inicial de maturidade estrutural do backlog.

A partir de setembro de 2025 observa-se uma trajetória consistente de melhoria, com a nota média ultrapassando o valor 3, considerado neste estudo como um limiar indicativo de backlog estruturalmente adequado para orientar o trabalho das equipes de desenvolvimento.

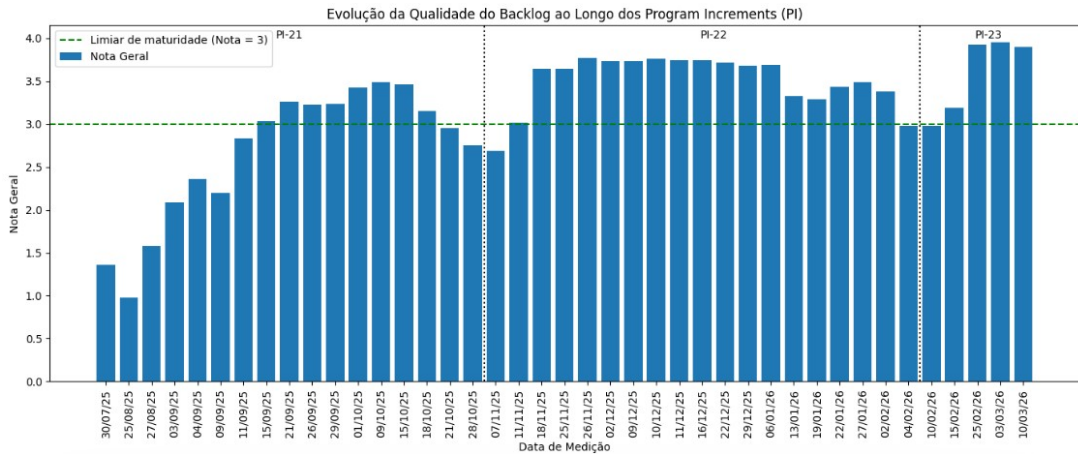


Figura 36 – Evolução das pontuações médias de qualidade dos itens do *Product Backlog* ao longo do período de aplicação do método

Outro indicador relevante refere-se à quantidade de equipes com nota média superior a 3. No início do período analisado, apenas um time atingia esse patamar. Ao longo dos meses seguintes, observa-se uma expansão gradual desse número, chegando a 13 equipes com notas superiores a esse nível em fevereiro de 2026.

A Figura 36 apresenta a evolução da nota média ao longo dos diferentes ciclos de planejamento (*Program Increments*). As linhas verticais indicam a transição entre os períodos PI-21, PI-22 e PI-23, permitindo observar a progressiva melhoria da qualidade estrutural dos artefatos de backlog ao longo do processo de adoção do método.

Durante o PI-21 observa-se uma fase inicial de aprendizado e adaptação das equipes aos critérios de avaliação propostos. No PI-22 a qualidade média estabiliza-se acima do limiar de maturidade estrutural, enquanto no PI-23 os valores aproximam-se de níveis elevados de qualidade, próximos de nota 4.

A evolução das notas médias observadas ao longo dos três ciclos é sintetizada na Tabela 12, que apresenta os principais indicadores estatísticos por *Program Increment*. Os dados evidenciam não apenas o crescimento do valor médio, mas também a redução expressiva da dispersão: o desvio padrão recuou de 0,73 no PI-21 para 0,22 no PI-23, indicando estabilização progressiva do processo de refinamento.

Tabela 12 – Evolução dos indicadores de qualidade do *Product Backlog* por *Program Increment*

PI	Período	Nota mín.	Nota máx.	Nota média	Desv. pad. / Tendência
PI-21	Ago–Out/2025	0,98	3,49	2,73	0,73 Crescente
PI-22	Nov/2025–Fev/2026	2,69	3,77	3,44	0,35 Estável (patamar elevado)
PI-23	Fev–Abr/2026	3,19	3,95	3,74	0,22 Estável (patamar mais elevado)

Fonte: Elaborado pelo autor a partir dos dados gerados pelo Backlog Analyzer.

Um padrão estrutural recorrente nos três ciclos merece destaque: a queda da nota nas semanas finais de cada PI, interpretada como reflexo da entrada de novos itens menos maduros e da redução do foco em refinamento ao final do ciclo. Para cada PI, foram calculadas a queda absoluta e a queda relativa em relação ao pico da nota, conforme as Equações 2 e 3:

$$\text{Queda absoluta} = N_{\text{pico}} - N_{\text{final}} \quad (2)$$

$$\text{Queda relativa} = \frac{N_{\text{pico}} - N_{\text{final}}}{N_{\text{pico}}} \times 100\% \quad (3)$$

onde N_{pico} é a nota máxima registrada no ciclo e N_{final} é a nota da última semana do PI. A Tabela 13 apresenta os valores calculados para cada ciclo, evidenciando a redução progressiva da magnitude desse fenômeno.

Tabela 13 – Padrão de queda da nota ao final de cada *Program Increment*

PI	Nota no pico	Nota final	Queda absoluta	Queda relativa
PI-21	3,49	2,75	–0,74	–21,2%
PI-22	3,77	2,98	–0,79	–20,9%
PI-23	3,95	3,50	–0,45	–11,4%

Fonte: Elaborado pelo autor a partir dos dados gerados pelo Backlog Analyzer.

A redução de –21% para –11% indica resiliência crescente do processo, embora o padrão não tenha sido completamente superado. A Figura 37 apresenta a distribuição completa das notas semanais por PI, evidenciando a progressão qualitativa acumulada ao longo dos três ciclos.

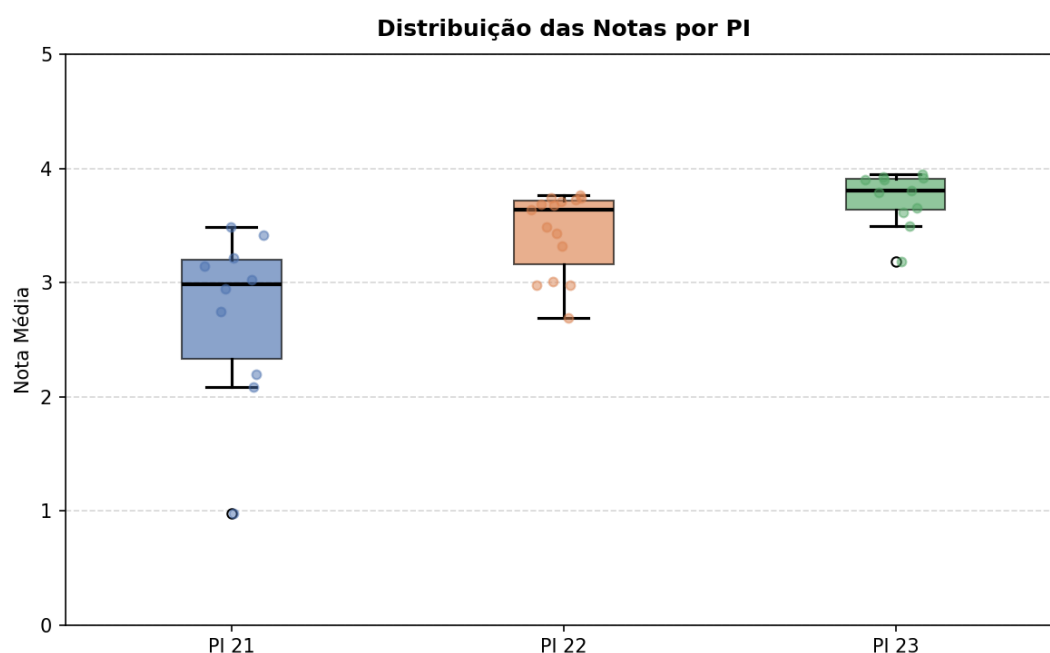


Figura 37 – Distribuição das notas semanais do backlog por *Program Increment* (boxplot)

Fonte: Elaborado pelo autor a partir dos dados gerados pelo Backlog Analyzer.

A progressão das caixas evidencia três fenômenos simultâneos: (i) a redução expressiva da dispersão, com o intervalo interquartil recuando de aproximadamente 1,1 no PI-21 para 0,3 no PI-23; (ii) a elevação da mediana de 2,99 para 3,81; e (iii) o deslocamento ascendente de toda a distribuição a cada PI. No PI-23, o valor mínimo absoluto (3,19) supera a média do PI-21 (2,73), ilustrando o salto qualitativo acumulado.

A redução progressiva da magnitude das quedas ao longo dos três ciclos não é um fenômeno casual. Por se tratar de um método de avaliação contínua, que monitora semanalmente a qualidade do backlog e disponibiliza os resultados em tempo real para a gestão do trem, a identificação do padrão de queda ao final de cada PI criou oportunidades concretas de intervenção. À medida que a nota começava a recuar nas semanas finais, a gestão passou a atuar junto aos times, reforçando o foco no refinamento de novos itens e adiantando ciclos de revisão que normalmente concentrar-se-iam apenas no início do PI seguinte. Essa capacidade de detecção antecipada e resposta adaptativa é, em si mesma, um dos mecanismos centrais do método: a avaliação automatizada e contínua não apenas mede a qualidade, mas alimenta um ciclo de melhoria que, ao longo dos ciclos, tende a tornar o processo progressivamente mais resiliente às perturbações estruturais do encerramento de cada PI.

5.3 Análise Longitudinal por *Program Increment*

A análise longitudinal dos dados coletados ao longo dos três ciclos de planejamento permite compreender a trajetória de evolução da qualidade do backlog de forma mais granular, identificando não apenas a tendência de crescimento da nota média, mas também as transformações qualitativas no comportamento do processo de refinamento. Conforme sintetizado na Tabela 12, a nota média avançou de 2,73 no PI-21 para 3,74 no PI-23, representando um ganho de +37% ao longo de 36 semanas.

Mais relevante que o crescimento absoluto é a transformação do padrão de comportamento da nota ao longo do tempo. No PI-21, observou-se correlação negativa entre a nota do backlog e o volume de cards em aberto ($r \approx -0,67$): a nota subia à medida que os itens eram consumidos durante a sprint, evidenciando que a qualidade dependia do esgotamento do backlog e não da melhoria estrutural dos artefatos. Nos PIs subsequentes, essa correlação desapareceu ($r \approx 0$), indicando que a nota passou a se sustentar independentemente do volume, sinal de maturação genuína do processo de refinamento.

A Figura 38 apresenta a evolução semanal da nota média e do volume de cards ao longo dos três PIs. A linha tracejada indica o objetivo de nota 3,0 (limiar considerado neste estudo como indicativo de backlog estruturalmente adequado para orientar o trabalho das equipes).

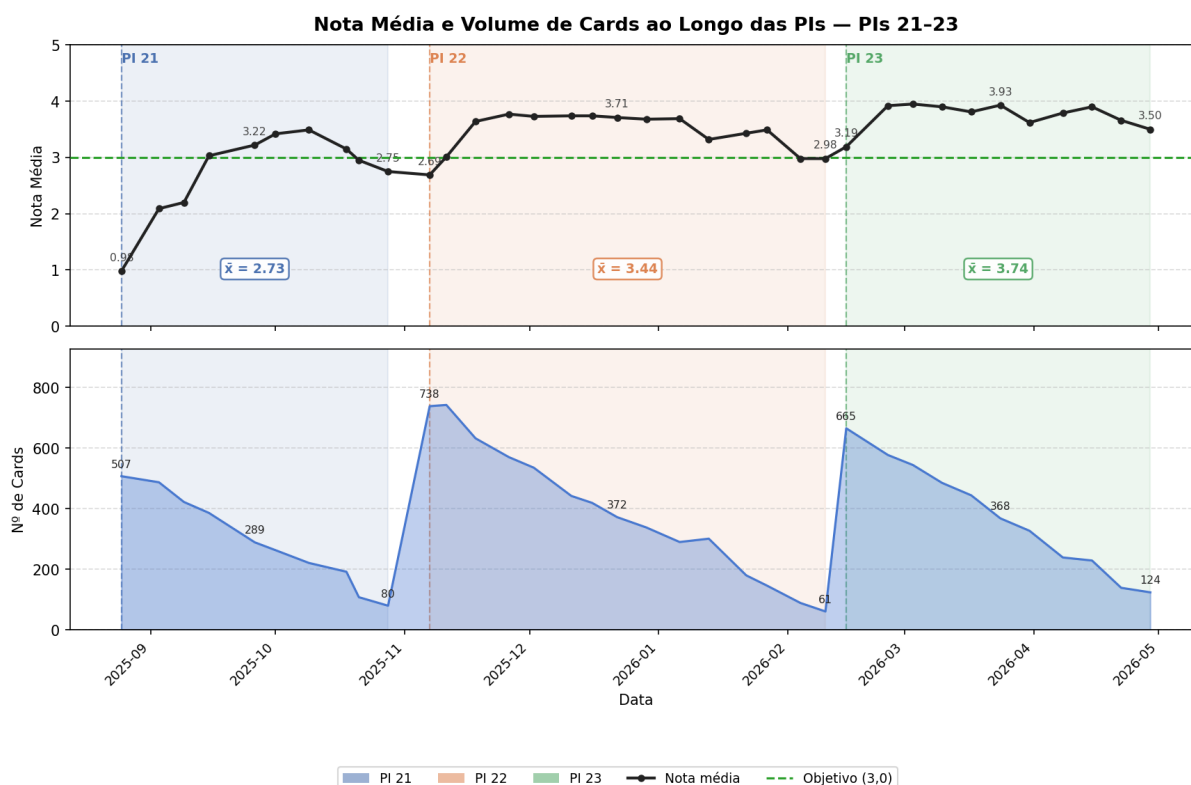


Figura 38 – Evolução semanal da nota média do backlog e volume de cards ao longo dos PI-21, PI-22 e PI-23

Fonte: Elaborado pelo autor a partir dos dados gerados pelo Backlog Analyzer.

A trajetória do PI-21 caracteriza uma fase de implantação e instabilidade inicial: o artefato é posto em operação com um backlog com nota 0,98 na primeira semana de coleta sistemática, e a nota sobe progressivamente à medida que os critérios de avaliação passam a orientar o processo de refinamento. O PI-22 representa a fase de consolidação: a nota inicia em patamar superior (2,69) e estabiliza-se em torno de 3,4–3,7, permanecendo acima do limiar de 3,0 por sete semanas consecutivas. A PI-23 evidencia a fase de maturidade: a nota mínima semanal (3,19) supera pela primeira vez a meta em todos os momentos do ciclo, e a nota máxima atinge 3,95, o valor mais elevado registrado no período.

5.4 Relação entre Qualidade do Backlog e Eficiência do Planejamento das Sprints

A análise cruzada entre os indicadores de qualidade do backlog e os indicadores de eficiência do planejamento das sprints permite identificar associações entre a melhoria da qualidade informacional dos artefatos e a previsibilidade das entregas. Para cada PI, foram calculados três indicadores: o percentual de alocação de capacidade

(% *Allocation*), o percentual de conclusão dos itens comprometidos (% *Concluded*) e o Indicador de Eficiência do Planejamento (IEP), definidos pelas Equações 4, 5 e 6, em que SP (*Story Points*) é a unidade de esforço relativo adotada pelo trem.

$$\% Allocation = \frac{SP_{comprometidos}}{SP_{capacidade}} \times 100 \quad (4)$$

$$\% Concluded = \frac{SP_{concluídos}}{SP_{comprometidos}} \times 100 \quad (5)$$

$$IEP = \frac{|\% Allocation| + |\% Concluded|}{2} \quad (6)$$

O IEP expressa, portanto, a fração da capacidade do time efetivamente convertida em entrega, com valor ideal igual a 1,0. A Tabela 14 apresenta a evolução integrada desses indicadores ao longo dos três ciclos.

Tabela 14 – Evolução integrada dos indicadores de qualidade do backlog e eficiência do planejamento por PI

Indicador	PI-21	PI-22	PI-23
Nota média do backlog	2,73	3,44	3,74
Desvio padrão da nota	0,73	0,35	0,22
IEP médio (Eq. 6)	88%	93%	97%
% <i>Allocation</i> média (Eq. 4)	112%	104%	98%
% <i>Concluded</i> média (Eq. 5)	88%	91%	96%

Fonte: Elaborado pelo autor a partir dos dados do sistema de planejamento ágil.

A trajetória conjunta dos indicadores sugere uma associação positiva entre a melhoria da qualidade do backlog e a melhoria da eficiência do planejamento das sprints. O IEP (indicador que combina % *Allocation* e % *Concluded* em um único índice, com valor ideal igual a 1,0) apresenta salto expressivo do PI-21 para o PI-23 (de 88% para 97%, +9%), período que coincide com a melhoria mais intensa da nota do backlog (de 2,73 para 3,44, +26%). Nos ciclos seguintes, tanto o IEP quanto a nota do backlog desaceleram o ritmo de crescimento, mantendo-se em patamares elevados: paralelismo consistente com a hipótese de que a qualidade informacional do backlog influencia a precisão do planejamento.

A Figura 39 apresenta a evolução dos indicadores de planejamento ao longo dos três Pis. Observa-se a convergência progressiva do % *Allocation* em direção à meta de 100% (de 112% para 98%), coerente com a hipótese de que histórias melhor definidas (com critérios de aceitação claros e nível de detalhamento adequado) permitem

estimativas de esforço mais precisas. O % *Concluded* atingiu 96% no PI-23, o mesmo PI em que a nota média do backlog alcançou seu maior valor (3,74) e em que a nota mínima semanal superou pela primeira vez a meta de 3,0 em todos os momentos do ciclo.

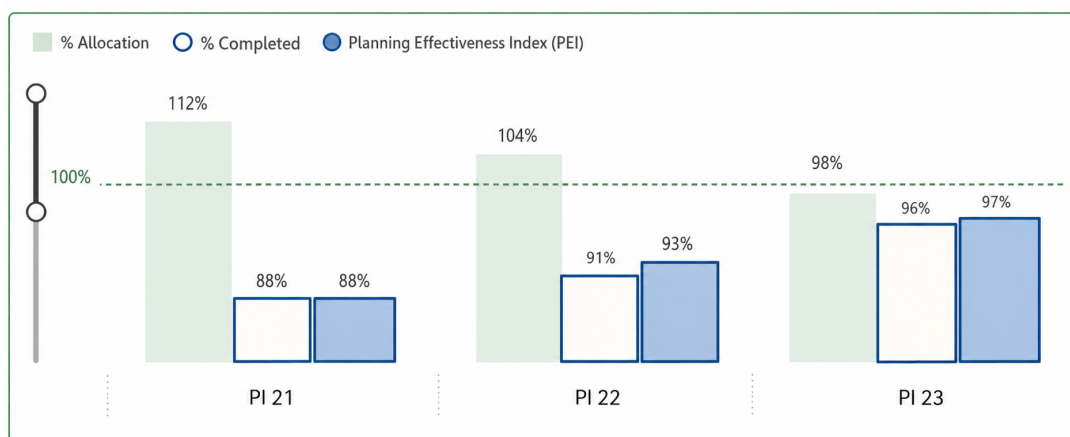


Figura 39 – Evolução de % Allocation e % Concluded por Program Increment

Fonte: Elaborado pelo autor a partir dos dados do sistema de planejamento ágil.

É importante ressaltar que a relação identificada é de **associação temporal**, e não de causalidade demonstrada. Ambos os indicadores (nota do backlog e IEP) podem estar sendo influenciados por um terceiro fator, como a maturidade geral do trem ao longo do tempo. Equipes que aprendem a refinar melhor o backlog tendem a aprimorar também outros aspectos do processo ágil, tornando difícil isolar o efeito específico da qualidade do backlog sobre a eficiência do planejamento. A demonstração formal dessa relação causal demandaria uma análise de regressão com controles adequados, o que está além do escopo desta pesquisa, mas constitui uma direção natural para investigações futuras.

Um método quase-experimental particularmente adequado para investigar essa relação causal em trabalhos futuros é o *Causal Impact*, proposto por (Brodersen et al.). O método utiliza uma abordagem bayesiana de séries temporais para estimar o efeito causal de uma intervenção, no caso, a adoção sistemática do método de avaliação do backlog, sobre uma variável de resultado de interesse, como o IEP ou o % *Concluded*. A construção de um grupo de controle sintético a partir de séries não afetadas pela intervenção permite estimar o contrafactual, ou seja, o que teria ocorrido com os indicadores na ausência da intervenção. Essa abordagem é particularmente viável neste contexto, dada a disponibilidade de múltiplas séries temporais de qualidade

de backlog e de desempenho de sprint provenientes de diferentes times e ARTs.

5.5 Percepção dos Desenvolvedores: Avaliação Subjetiva e Representatividade Amostral

Em complemento à avaliação objetiva realizada pelo artefato, foi conduzida uma pesquisa com os desenvolvedores diretamente envolvidos no processo de planejamento, com o objetivo de capturar a percepção subjetiva sobre a qualidade das histórias de usuário e dos critérios de aceitação. A coleta foi realizada por meio de formulário estruturado em escala Likert (1–5), abrangendo seis dimensões avaliativas, ao longo do PI-22 e do PI-23 (dezembro/2025 a abril/2026).

Os resultados indicam que as notas Likert mantiveram-se consistentemente acima de 4,0 em todos os meses avaliados, conforme apresentado na Tabela 15. Destaca-se que, mesmo nos períodos em que a nota objetiva do backlog registrou valores mais baixos, como no encerramento do PI-22 (nota 2,98), a percepção dos desenvolvedores permaneceu positiva. A queda observada nas dimensões de *Nível de detalhamento* e *Descrição suficiente* em abril/2026 (3,85) coincide com o período de encerramento do PI-23, em que a nota objetiva também recuou de 3,95 para 3,50, evidenciando convergência de sinal entre as duas fontes de avaliação.

Tabela 15 – Médias das avaliações Likert por mês e dimensão (histórias de usuário)

Mês	Simple	Tamanho	Detalh.	Descrição	Elem. mensur.	n
Dez/2025	4,32	4,37	4,11	4,18	≈4,2	38
Jan/2026	4,31	4,40	4,29	4,40	≈4,3	35
Fev/2026	4,56	4,33	4,44	4,33	≈4,7	9
Mar/2026	4,58	4,47	4,47	4,53	≈4,6	19
Abr/2026	4,54	4,62	3,85	3,85	≈4,1	13

Fonte: Elaborado pelo autor a partir dos dados da pesquisa com desenvolvedores.

A comparação entre as notas objetivas do backlog (avaliação por LLM, escala 0–5) e as notas subjetivas dos desenvolvedores (Likert 1–5) sugere que a avaliação automatizada é mais sensível às variações de qualidade. Duas hipóteses explicativas se complementam. A primeira refere-se ao viés de condescendência (*acquiescence bias*): respondentes em pesquisas voluntárias tendem a atribuir notas mais altas do que a situação efetiva justificaria, especialmente em contextos organizacionais em que a crítica pode gerar constrangimento. A segunda diz respeito ao escopo diferente de

avaliação: a nota objetiva do artefato considera todos os itens do backlog, enquanto a pesquisa com desenvolvedores avalia apenas as histórias da sprint em curso: os itens que, por definição, passaram pelo processo de refinamento e foram comprometidos, tendendo a ser os de maior qualidade do backlog total.

A Figura 40 apresenta a comparação das notas Likert entre os anos de 2025 e 2026, evidenciando a evolução da percepção dos desenvolvedores ao longo dos PIs.

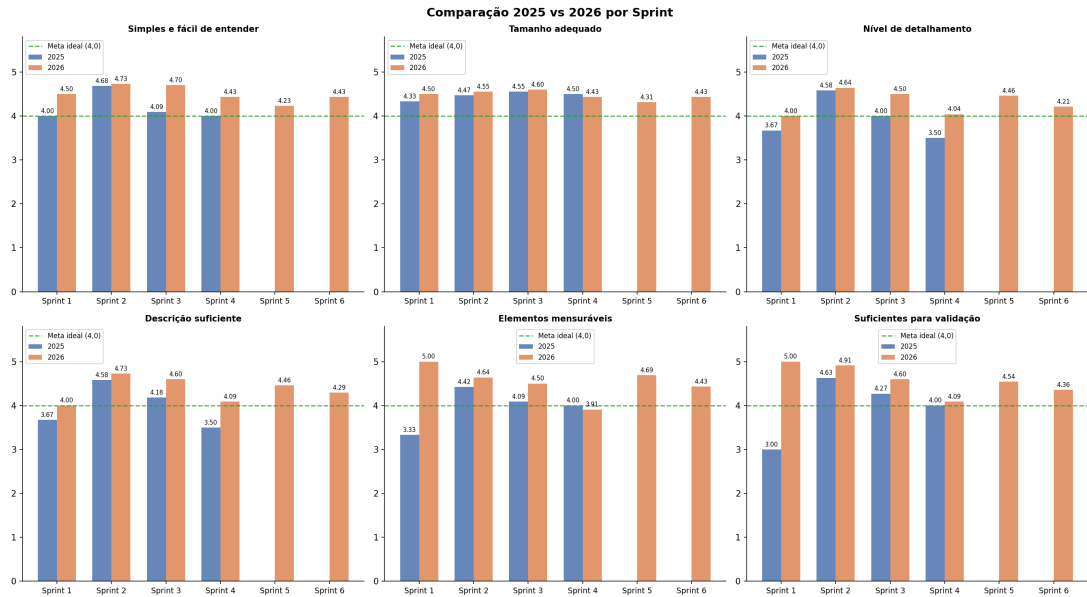


Figura 40 – Comparação das médias das avaliações Likert por sprint (2025 vs. 2026)
Fonte: Elaborado pelo autor a partir dos dados da pesquisa com desenvolvedores.

No que se refere à representatividade amostral, verificou-se que nenhuma sprint isolada atingiu o nível de representatividade considerado adequado (margem de erro $\leq \pm 15\%$). O resultado mais próximo foi de 23 respostas na quarta sprint de 2026, com taxa de adesão de 45,1% e margem de erro de $\pm 15,3\%$. O agrupamento por mês melhorou o quadro: dezembro/2025 e janeiro/2026 atingiram representatividade adequada. Essa limitação implica que as análises derivadas da pesquisa com desenvolvedores devem ser interpretadas como evidências qualitativas e de tendência, e não como inferências estatisticamente conclusivas sobre toda a população.

5.6 Análise de Sentimento dos Comentários dos Desenvolvedores

A pesquisa com desenvolvedores incluiu, além das questões em escala Likert, um campo de texto aberto para comentários livres. Os 112 comentários coletados ao longo dos PI-22 e PI-23 foram submetidos a análise de sentimento, com o objetivo de

identificar nuances qualitativas que a escala estruturada não captura. A distribuição geral indica predominância de comentários neutros (47%) e positivos (41%), com 11% de comentários negativos, conforme apresentado na Figura 41.

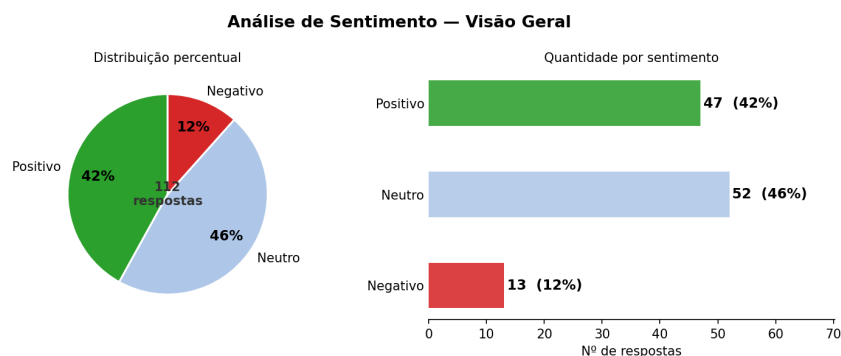


Figura 41 – Distribuição geral dos comentários por categoria de sentimento

Fonte: Elaborado pelo autor a partir dos dados da pesquisa com desenvolvedores.

A análise por sprint revelou um padrão oscilante, sem tendência linear positiva clara. Os mínimos de sentimento (registrados na quarta e na sexta sprints do período analisado) coincidem com os períodos de queda da nota objetiva do backlog: o encerramento do PI-22 e o final do PI-23, respectivamente. A recuperação do sentimento na quinta sprint coincide com o pico da nota objetiva no PI-23 (3,95). Essa convergência temporal sugere que o sentimento dos desenvolvedores sobre o processo de planejamento oscila de forma coerente com a qualidade objetiva do backlog, tornando a análise de sentimento um instrumento complementar relevante para o monitoramento do processo.

No que se refere aos temas negativos recorrentes, a Figura 42 apresenta os principais focos de insatisfação identificados nos comentários, em comparação com os temas de elogios. Destacam-se: qualidade dos cards, processo de refinamento e planejamento, participação do *Product Owner*, mudanças de escopo e uso de modelos de linguagem.

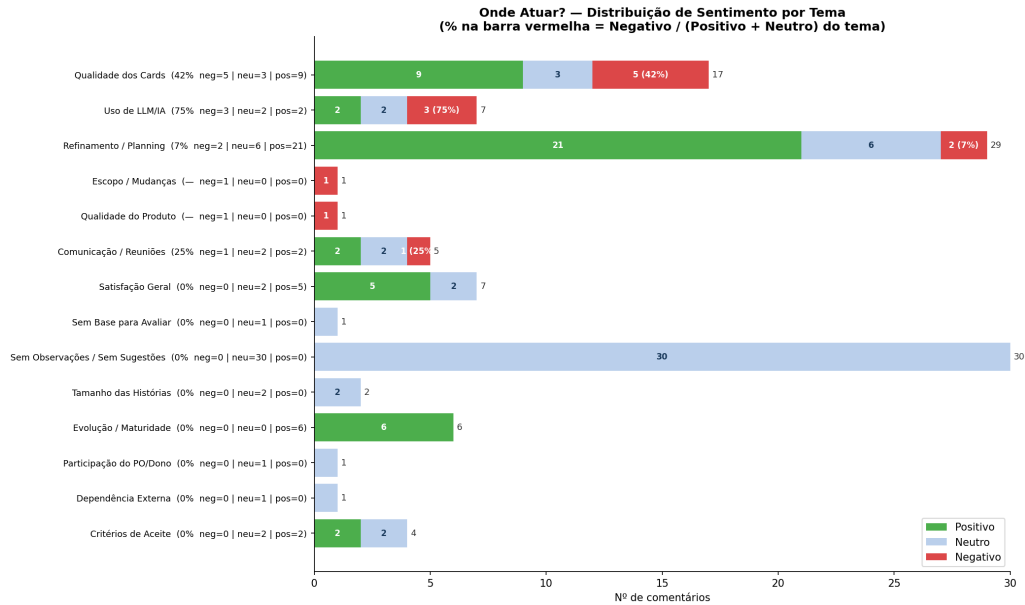


Figura 42 – Temas recorrentes nos comentários: comparação entre menções negativas e positivas
Fonte: Elaborado pelo autor a partir dos dados da pesquisa com desenvolvedores.

O tema relacionado ao uso de modelos de linguagem no processo de refinamento merece atenção especial no contexto desta pesquisa. Comentários negativos apontaram (3 respostas em 112) que critérios de aceitação gerados automaticamente pela LLM foram percebidos como texto de preenchimento desconectado do contexto real do card, quando não devidamente revisados pelo time. Esse achado reforça a importância de que o *Agile Master* e os desenvolvedores atuem como revisores críticos do conteúdo gerado, garantindo que a automação sirva de apoio ao refinamento e não substitua o julgamento coletivo da equipe. A avaliação automatizada deve, portanto, ser complementada por revisão humana e por ciclos de validação nas retrospectivas dos times.

5.7 Análise dos Resultados à luz dos GAPs Identificados

A análise dos resultados obtidos a partir da aplicação do artefato permite avaliar, de forma sistemática, em que medida o método proposto foi capaz de endereçar os *GAPs* identificados na gestão do *Product Backlog*. Em consonância com a abordagem da *Design Science Research*, essa análise não se limita à verificação de desempenho técnico, mas considera a capacidade do artefato de reduzir limitações práticas observadas no contexto organizacional estudado.

O primeiro *GAP* identificado, relacionado à subjetividade na avaliação da qualidade dos itens do backlog, foi diretamente endereçado pela explicitação de critérios de avaliação e pela utilização de uma escala estruturada de mensuração. Os resultados demonstraram que a aplicação sistemática desses critérios permitiu reduzir a dependência exclusiva do julgamento individual, uma vez que histórias semelhantes passaram a receber avaliações consistentes quando analisadas sob os mesmos parâmetros. Embora o processo de avaliação continue demandando interpretação contextual, a presença de critérios explícitos e justificativas associadas às pontuações contribuiu para maior transparência e reprodutibilidade do diagnóstico.

No que se refere ao *GAP* da ausência de escalas mensuráveis e reproduzíveis, os resultados indicaram que a adoção de uma escala ordinal de zero a cinco possibilitou traduzir percepções qualitativas em indicadores estruturados. A distribuição das pontuações observadas nas diferentes dimensões avaliadas evidenciou variações claras nos níveis de maturidade do backlog, tanto entre equipes distintas quanto entre itens de um mesmo backlog. Essa característica viabilizou análises comparativas mais consistentes e forneceu subsídios objetivos para discussões de planejamento e refinamento.

A análise também evidencia que o artefato contribuiu para mitigar o *GAP* relacionado à baixa escalabilidade dos processos tradicionais de avaliação. A capacidade do sistema de analisar múltiplos itens de backlog de forma automatizada permitiu ampliar significativamente o volume de histórias avaliadas, sem aumento proporcional do esforço humano envolvido. Esse aspecto mostrou-se particularmente relevante em um contexto ágil em escala, no qual a avaliação manual e aprofundada de grandes backlogs se revela operacionalmente inviável.

Outro *GAP* central identificado diz respeito à ausência de mecanismos sistemáticos de retroalimentação e aprendizado organizacional. Os resultados indicaram que as recomendações de melhoria geradas pelo sistema, associadas às justificativas das pontuações atribuídas, forneceram insumos claros e acionáveis para o refinamento das histórias de usuário. Diferentemente de abordagens pontuais e reativas, o artefato passou a oferecer um registro estruturado das fragilidades recorrentes, criando condições para a identificação de padrões e para a promoção de melhorias contínuas ao longo das sprints.

Adicionalmente, a análise dos resultados sugere que o artefato favoreceu o alinhamento conceitual entre os envolvidos no processo de planejamento. A utilização

de uma linguagem comum baseada em critérios explícitos e escalas compreensíveis reduziu ambiguidades interpretativas e facilitou a comunicação entre *Product Owners*, *Agile Masters* e equipes de desenvolvimento.

Entretanto, a análise também evidencia limitações inerentes à solução proposta. A avaliação automatizada permanece dependente da qualidade dos dados de entrada, de modo que histórias excessivamente incompletas ou mal estruturadas tendem a resultar em diagnósticos igualmente limitados. Além disso, o uso de modelos de linguagem requer cuidados relacionados à calibração dos *prompts* e à interpretação contextual dos resultados.

De forma geral, os resultados analisados nesta seção indicam que o artefato foi capaz de endereçar de maneira consistente os principais *GAPs* que motivaram o desenvolvimento do estudo. Ao combinar critérios explícitos, escalas mensuráveis e automação baseada em Inteligência Artificial, a solução proposta contribuiu para elevar o nível de objetividade, escalabilidade e capacidade diagnóstica da avaliação da qualidade do backlog. Esses achados constituem a base para uma discussão mais ampla acerca de suas implicações práticas e teóricas, apresentada na seção seguinte.

Tabela 16 – Síntese dos principais achados da pesquisa

Achado da pesquisa	Evidência observada	Implicação prática	Contribuição para a literatura
Redução da subjetividade na avaliação do backlog	Aplicação consistente de critérios estruturados e escala ordinal resultou em avaliações mais homogêneas entre itens analisados.	Permite diagnósticos mais transparentes e comparáveis sobre a qualidade das histórias de usuário.	Evidencia a viabilidade de mecanismos estruturados de avaliação da qualidade de artefatos ágeis.
Viabilidade de mensuração da maturidade do backlog	Distribuição das pontuações evidenciou diferenças claras de qualidade entre itens e entre equipes.	Possibilita análises comparativas e acompanhamento evolutivo da qualidade do backlog.	Contribui para a literatura ao propor uma forma operacional de mensurar qualidade em artefatos ágeis.
Escalabilidade da avaliação do backlog	Automação baseada em modelos de linguagem permitiu avaliar grande volume de histórias com baixo esforço adicional.	Viabiliza análise contínua de grandes backlogs em ambientes ágeis em escala.	Demonstra o potencial de uso de LLMs na análise de artefatos textuais de engenharia de software.
Geração de aprendizado organizacional	Registro estruturado de pontuações e recomendações permitiu identificar fragilidades recorrentes nas histórias.	Favorece melhoria contínua do backlog e refinamento das práticas de planejamento.	Amplia o entendimento sobre mecanismos de governança e melhoria contínua em ambientes ágeis.
Alinhamento conceitual entre papéis ágeis	Uso de critérios explícitos reduziu ambiguidades interpretativas entre Product Owners, Agile Masters e equipes.	Facilita comunicação e coordenação entre atores envolvidos no planejamento ágil.	Reforça a dimensão sociotécnica da qualidade dos artefatos de backlog.
Potencial de apoio da Inteligência Artificial na análise de backlog	Modelos de linguagem foram capazes de interpretar histórias e gerar avaliações contextualizadas.	Amplia a capacidade analítica das equipes sem substituir o julgamento humano.	Indica novas possibilidades de aplicação de IA na governança do desenvolvimento de software.

5.8 Discussão dos Resultados e Implicações Práticas

A discussão dos resultados obtidos a partir da aplicação do artefato desenvolvido neste estudo permite posicionar as evidências empíricas observadas em relação à literatura existente e aos objetivos da pesquisa. Em consonância com a abordagem da *Design Science Research*, esta seção busca compreender não apenas se o artefato funcionou, mas de que forma e em quais aspectos ele contribui para o avanço do conhecimento e da prática na gestão do *Product Backlog* em ambientes ágeis em escala.

Os resultados obtidos corroboram achados da literatura que apontam limitações estruturais nas práticas tradicionais de gestão do backlog, especialmente no que se refere à subjetividade e à informalidade inerentes às histórias de usuário. Conforme discutido na literatura de engenharia de requisitos ágil, a concisão e a flexibilidade das histórias de usuário, embora desejáveis, tendem a gerar ambiguidades quando aplicadas em contextos organizacionais complexos. A evidência empírica observada nesta pesquisa, caracterizada pela variação significativa na qualidade dos itens avaliados, reforça essa constatação e indica que a simples adoção de frameworks ágeis não é suficiente para assegurar a maturidade dos artefatos produzidos.

Um primeiro achado relevante refere-se ao impacto da melhoria da qualidade dos artefatos de backlog sobre o funcionamento sociotécnico do processo de desenvolvimento. A percepção dos participantes indica que a melhoria da qualidade das histórias contribuiu para reduzir fricções no processo de desenvolvimento, especialmente desacordos entre solicitantes e equipes, ambiguidades na interpretação das histórias e a necessidade de intervenções corretivas durante a sprint. Esse resultado sugere que a qualidade da descrição do trabalho exerce papel relevante na coordenação entre diferentes atores envolvidos no processo de desenvolvimento.

À luz desses achados, a principal contribuição do artefato proposto reside na introdução de um método estruturado capaz de transformar avaliações tradicionalmente subjetivas em diagnósticos mais objetivos e comunicáveis. A explicitação de critérios e o uso de uma escala mensurável permitiram tornar visíveis fragilidades que anteriormente permaneciam implícitas ou dispersas em percepções individuais. Essa característica dialoga diretamente com a literatura que defende a necessidade de maior rigor analítico na governança de produtos ágeis, sem comprometer os princípios de adaptação e aprendizado contínuo.

Do ponto de vista prático, os resultados sugerem que o artefato atua como um instrumento de mediação entre diferentes papéis envolvidos no processo ágil. Ao estabelecer uma linguagem comum baseada em critérios explícitos, o método contribui para reduzir desalinhamentos conceituais entre *Product Owners*, *Agile Masters* e equipes de desenvolvimento. Essa mediação mostra-se particularmente relevante em ambientes ágeis em escala, nos quais a heterogeneidade de práticas entre equipes tende a dificultar a consolidação de padrões mínimos de qualidade.

Além da mediação entre papéis, o método revelou uma contribuição de se-

gunda ordem que não estava explicitamente prevista no design original: a **elevação progressiva da maturidade dos times** no que se refere à especificação de artefatos de backlog. Ao expor de forma sistemática os critérios de qualidade esperados, o método criou um padrão comum de referência que passou a ser internalizado pelos times ao longo dos ciclos de aplicação. Times que, inicialmente, entregavam histórias de usuário sem critérios de aceitação verificáveis passaram, gradualmente, a incorporar esses elementos como parte natural do ritual de refinamento.

Esse efeito pedagógico emergiu de forma não intencional. Ao receberem de forma recorrente o *feedback* estruturado do *Backlog Booster*, os times passaram a antecipar os critérios de avaliação durante a escrita dos itens, reduzindo a necessidade de intervenções corretivas posteriores.

Dessa forma, o método não se limita à avaliação da qualidade dos registros. Sua utilização também contribui para orientar os times quanto aos elementos que caracterizam uma especificação de qualidade. Como consequência, boas práticas passam a ser progressivamente incorporadas à rotina de trabalho, uma vez que se tornam visíveis, mensuráveis e objeto de avaliação contínua.

Esse efeito é particularmente relevante em contextos ágeis em escala, nos quais a heterogeneidade das práticas adotadas pelas equipes constitui um desafio recorrente para a governança do ART.

Outro aspecto relevante refere-se às dificuldades associadas à mensuração direta dos ganhos operacionais decorrentes da melhoria do backlog. Embora os participantes reconheçam benefícios práticos decorrentes da melhoria das histórias de usuário, a mensuração direta desses ganhos foi limitada pela indisponibilidade de séries históricas estruturadas, especialmente registros sistemáticos de validação, homologação ou retrabalho associados às histórias.

Diante dessa limitação, a avaliação do artefato passou a apoiar-se em uma estratégia multimétodo, combinando diferentes tipos de evidência. Além dos indicadores quantitativos relacionados à evolução das notas de qualidade do backlog, foram considerados relatos e percepções de participantes diretamente envolvidos na gestão e execução do trabalho em escala. Essa combinação de evidências quantitativas e qualitativas permitiu construir uma avaliação mais abrangente do artefato, alinhada às recomendações metodológicas da *Design Science Research* para avaliação de soluções em contextos organizacionais reais.

Sob a perspectiva da Engenharia de Produção, o backlog pode ser interpretado como um mecanismo de organização do fluxo de trabalho em sistemas de desenvolvimento de software. Sob essa perspectiva, a qualidade da especificação das histórias de usuário exerce influência direta sobre a eficiência do processo produtivo, uma vez que descrições incompletas ou ambíguas tendem a gerar retrabalho, interrupções e ciclos adicionais de comunicação entre os membros da equipe.

Adicionalmente, a evolução recente das ferramentas de inteligência artificial voltadas ao desenvolvimento de software reforça a relevância do problema investigado nesta pesquisa. A partir de 2025 e, de forma mais intensa em 2026, observa-se o surgimento de agentes capazes de interpretar histórias de usuário e gerar automaticamente estruturas de implementação ou trechos de código. Diante desse quadro, a qualidade da especificação das histórias torna-se ainda mais crítica, pois descrições ambíguas podem comprometer não apenas o entendimento das equipes humanas, mas também a interpretação automatizada realizada por sistemas baseados em inteligência artificial.

Os resultados também sugerem evidências de utilidade organizacional do método proposto. A expansão progressiva de sua utilização para diferentes times e contextos organizacionais, bem como o interesse demonstrado por lideranças intermediárias, indicam que o artefato apresentou valor percebido além de seu contexto inicial de aplicação.

Após a validação do método no ART, o artefato passou a ser aplicado em outros ARTs com características operacionais distintas. Observou-se que sua implementação nesses novos contextos ocorreu sem necessidade de adaptações estruturais, mantendo-se os mesmos critérios de avaliação e a mesma lógica operacional originalmente proposta. Esse resultado sugere que o método apresenta baixo acoplamento às idiossincrasias do contexto inicial de desenvolvimento, indicando robustez conceitual e potencial de generalização para outros ambientes de desenvolvimento ágil em escala.

Esse achado reforça evidências de utilidade e transferibilidade do artefato em contextos organizacionais reais.

Entretanto, a discussão também evidencia cuidados necessários. A dependência da qualidade dos dados de entrada implica que o artefato não elimina a necessidade de práticas básicas de disciplina na construção do backlog. Além disso, a interpretação dos resultados requer maturidade organizacional para que as pontuações e recomendações sejam utilizadas como instrumentos de aprendizado e melhoria contínua, e não

como mecanismos de controle punitivo.

A análise integrada das diferentes fontes de evidência coletadas ao longo do estudo revela uma convergência relevante: os períodos de queda da nota objetiva do backlog, registrados nos encerramentos de cada PI, coincidem com os vales da avaliação subjetiva (dimensões de detalhamento e critérios de aceitação na escala Likert) e com os mínimos do *score* de sentimento dos comentários. Essa convergência multi-fonte fortalece a validade das evidências empíricas, uma vez que três instrumentos metodológicos independentes, avaliação automatizada por LLM, escala Likert e análise de sentimento, apontam para o mesmo padrão de comportamento. Cada fonte captura aspectos distintos: a avaliação objetiva é mais sensível e abrangente, cobrindo todo o backlog; a escala Likert é validadora da percepção prática dos desenvolvedores; e a análise de sentimento é diagnóstica, revelando os temas que efetivamente geram insatisfação no processo.

Um achado adicional que emerge da análise de sentimento merece destaque não apenas como alerta metodológico, mas sobretudo como evidência do valor do canal de escuta estruturada criado pelo método. Registros da pesquisa indicaram que, em determinados casos, critérios de aceitação foram gerados por modelos de linguagem e incorporados aos cards sem revisão humana adequada, com o objetivo de elevar artificialmente a pontuação do backlog. Essa vulnerabilidade não teria sido identificada sem a participação ativa dos desenvolvedores: são eles que, ao expressarem suas opiniões de forma estruturada, tornaram visível um comportamento que, de outra forma, permaneceria oculto nos dados objetivos.

É relevante situar esse achado no contexto da trajetória percorrida. No início da implantação do método, as dimensões avaliadas apresentavam notas próximas de zero, refletindo uma condição em que as atividades de refinamento já eram realizadas com elevado grau de subjetividade e sem critérios explícitos. A evolução registrada ao longo dos três PIs (de uma média de 2,73 para 3,74) demonstra que o processo de melhoria é real. O episódio do *gaming* não contradiz esse progresso; ao contrário, é um sintoma natural do amadurecimento do processo: times que ainda não internalizaram o valor intrínseco da qualidade tendem a buscar atalhos quando expostos a métricas visíveis. Identificar esse comportamento é, em si, uma conquista metodológica.

Esse ciclo de descoberta e ajuste é característico de processos de melhoria contínua estruturados no modelo PDCA: a própria existência de um mecanismo de

avaliação e retroalimentação cria as condições para que as disfunções emergjam e possam ser tratadas. Como encaminhamento concreto, a pesquisa de percepção com desenvolvedores, até aqui aplicada de forma piloto, será integrada de maneira estruturada ao rito de retrospectiva dos times, constituindo um canal permanente de diagnóstico qualitativo e de validação coletiva das notas geradas automaticamente. Dessa forma, o risco de otimização da métrica em detrimento da qualidade genuína pode ser mitigado pelo próprio processo ágil, transformando uma vulnerabilidade identificada em oportunidade de aprendizado organizacional.

Esse achado sugere uma diretriz relevante para a sustentabilidade do método: **a nota não deve ser tratada como um fim em si mesma, mas como um indicador da qualidade informacional do backlog e um instrumento de apoio à análise gerencial.** Quando interpretada dessa forma, a prática de *gaming* tende a perder relevância, pois avaliações complementares, como pesquisas de percepção, análises de sentimento e discussões em retrospectivas, permitem identificar divergências entre a pontuação obtida e a qualidade efetivamente percebida pelos participantes. Diante disso, a função da métrica não está associada ao alcance de determinado valor, mas à sua capacidade de estimular diagnósticos, discussões e ações de melhoria.

Para reduzir o risco de otimização da métrica em detrimento da qualidade efetiva dos artefatos, são propostas três salvaguardas complementares. A primeira consiste na realização de auditorias qualitativas periódicas por amostragem, com revisão humana dos itens que apresentam as maiores pontuações. A segunda envolve o cruzamento da nota obtida com os resultados da *sprint* correspondente, uma vez que histórias com elevada pontuação e baixo desempenho de entrega podem indicar inconsistências no processo de avaliação. A terceira refere-se à manutenção de mecanismos estruturados de escuta dos desenvolvedores, funcionando como fonte complementar de evidências em relação à avaliação automatizada. Em conjunto, essas medidas contribuem para incorporar o monitoramento desse risco às práticas regulares de governança do backlog.

De forma geral, os resultados discutidos nesta seção indicam que o artefato desenvolvido contribui tanto para o avanço da prática quanto para o debate acadêmico sobre a gestão da qualidade do backlog em ambientes ágeis. Ao integrar critérios estruturados, escalas mensuráveis e Inteligência Artificial, a solução proposta amplia o repertório de métodos disponíveis para lidar com um problema recorrente e ainda pouco sistematizado na literatura.

5.9 Síntese do Capítulo

Este capítulo apresentou os resultados obtidos a partir da aplicação do artefato desenvolvido neste estudo e discutiu suas implicações à luz dos *GAPs* identificados na gestão do *Product Backlog* em ambientes ágeis em escala. Em consonância com a abordagem da *Design Science Research*, os resultados foram analisados considerando não apenas aspectos técnicos, mas também sua utilidade prática, capacidade diagnóstica e potencial de contribuição para a melhoria dos processos organizacionais.

Inicialmente, foram descritos os resultados empíricos da aplicação do artefato, evidenciando variações significativas na qualidade e na maturidade dos itens de backlog avaliados, bem como a análise longitudinal da evolução das notas ao longo dos três *Program Increments*. Em seguida, foram examinadas as associações entre a qualidade do backlog e os indicadores de eficiência do planejamento das sprints, os resultados da pesquisa de percepção com desenvolvedores em escala Likert e a análise de sentimento dos comentários coletados.

De forma integrada, as evidências reunidas ao longo deste capítulo indicam que o artefato endereçou consistentemente as lacunas que fundamentaram a pesquisa, tornando a avaliação da qualidade do backlog mais objetiva, escalável e orientada por evidências. Mais do que a confirmação pontual de cada lacuna, o conjunto dos resultados revela um padrão de amadurecimento processual: a adoção do método não apenas gerou diagnósticos estruturados, mas criou as condições para que disfunções emergentes (como o risco de *gaming* da nota) fossem identificadas e tratadas pelo próprio mecanismo de retroalimentação. Esse ciclo de descoberta, ajuste e aprendizado é, em si, uma evidência da utilidade do artefato como instrumento de governança em ambientes ágeis complexos.

Dessa forma, o Capítulo 5 cumpriu o objetivo de analisar e discutir os resultados obtidos a partir da aplicação do artefato proposto, consolidando evidências empíricas (quantitativas e qualitativas) sobre sua utilidade, viabilidade e capacidade de apoiar a avaliação sistemática da qualidade do *Product Backlog*. A abordagem multimétodo adotada, combinando avaliação objetiva por LLM, percepção subjetiva em escala Likert e análise de sentimento, permitiu construir uma visão mais abrangente e robusta sobre o comportamento do processo ao longo dos três ciclos analisados. Sob a perspectiva metodológica da *Design Science Research*, esses resultados contribuem para a cons-

trução de conhecimento prescritivo acerca do uso de Inteligência Artificial no apoio à gestão de requisitos em ambientes ágeis, fornecendo o embasamento necessário para as conclusões finais apresentadas no capítulo seguinte.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo apresenta as conclusões finais da pesquisa, sintetizando os principais achados obtidos ao longo do desenvolvimento do estudo e discutindo suas contribuições científicas e práticas. Em alinhamento com a abordagem da *Design Science Research*, as conclusões aqui apresentadas decorrem da análise integrada do problema investigado, do artefato desenvolvido e dos resultados observados a partir de sua aplicação em um contexto organizacional real.

A pesquisa foi motivada pela identificação de lacunas recorrentes na gestão da qualidade do *Product Backlog* em ambientes ágeis em escala, caracterizadas em detalhe na revisão da literatura (Capítulo 2). A partir dessas lacunas, foi concebido e implementado um método estruturado, apoiado por Inteligência Artificial, com o objetivo de avaliar de forma sistemática a qualidade e a maturidade dos itens de backlog.

Neste capítulo, são inicialmente apresentadas as conclusões da pesquisa, destacando-se em que medida os objetivos propostos foram alcançados e quais evidências sustentam a relevância e a viabilidade do artefato desenvolvido. Em seguida, são discutidas as limitações do estudo, reconhecendo restrições metodológicas, técnicas e contextuais que impactam a generalização dos resultados. Por fim, são apontadas possibilidades de trabalhos futuros, indicando caminhos para a evolução do método e do sistema proposto, bem como oportunidades de aprofundamento da investigação científica sobre o tema.

Dessa forma, o Capítulo 6 encerra a dissertação ao consolidar as contribuições do estudo e ao posicionar seus resultados no contexto mais amplo da pesquisa em engenharia de produção, métodos ágeis e aplicação de Inteligência Artificial na gestão de produtos digitais.

6.1 Conclusões da Pesquisa

Esta pesquisa teve como objetivo principal propor, implementar e avaliar um método estruturado para a avaliação da qualidade e da maturidade do *Product Backlog* em ambientes ágeis em escala, apoiado por Inteligência Artificial. O estudo foi conduzido a partir da identificação de *GAPs* recorrentes observados tanto na literatura quanto na prática organizacional, relacionados à subjetividade das avaliações, à ausência de

critérios padronizados e mensuráveis, à baixa escalabilidade dos processos tradicionais e à limitação dos mecanismos de aprendizado organizacional.

À luz dos resultados obtidos, conclui-se que o objetivo geral da pesquisa foi alcançado. O artefato desenvolvido demonstrou ser capaz de transformar avaliações tradicionalmente informais e subjetivas em diagnósticos mais objetivos, estruturados e comunicáveis. A explicitação de critérios de avaliação, aliada à utilização de uma escala ordinal de mensuração, permitiu capturar diferentes níveis de maturidade dos itens de backlog, viabilizando análises comparativas e acompanhamento evolutivo da qualidade ao longo do tempo.

Essa capacidade foi verificada empiricamente ao longo de três *Program Increments* consecutivos. A nota média objetiva do backlog avançou de 2,73 no PI-21 para 3,74 no PI-23, representando um ganho de +37%, acompanhado pela redução do desvio padrão de 0,73 para 0,22, evidência de que o processo passou de instável para consistentemente estável. Paralelamente, o Indicador de Eficiência do Planejamento (IEP) elevou-se de 0,688 para 0,815 e o percentual médio de conclusão dos itens comprometidos (% *Concluded*) convergiu para 100% no PI-23. A associação temporal entre esses indicadores é consistente com a hipótese central da pesquisa: a melhoria da qualidade informacional do backlog tende a se refletir na previsibilidade das entregas.

No que se refere à contribuição prática, os resultados indicaram que o método proposto apresenta viabilidade operacional em contextos ágeis complexos, integrando-se aos rituais existentes sem impor alterações significativas no fluxo de trabalho das equipes. A aplicação do artefato forneceu subsídios objetivos para atividades de refinamento e planejamento, contribuindo para maior alinhamento conceitual entre Product Owners, Agile Masters e equipes de desenvolvimento, bem como para a redução de ambiguidades na interpretação da qualidade do backlog.

Do ponto de vista científico, a pesquisa contribui para o avanço do conhecimento ao propor uma abordagem que articula princípios da engenharia de requisitos ágil, escalas estruturadas de avaliação e o uso de modelos de linguagem de larga escala. Ao adotar a *Design Science Research* como método, o estudo evidencia como lacunas identificadas na literatura podem ser traduzidas em requisitos de projeto e materializadas em um artefato funcional, avaliável e replicável.

Outro aspecto relevante das conclusões diz respeito ao papel da Inteligência Artificial como elemento de apoio (e não de substituição) ao julgamento humano. Os

resultados indicam que os modelos de linguagem utilizados foram capazes de analisar artefatos textuais complexos e gerar avaliações contextualizadas, desde que orientados por critérios claros e prompts adequadamente definidos. Esse achado reforça a perspectiva de uso da Inteligência Artificial como instrumento complementar à tomada de decisão em ambientes organizacionais.

Um achado metodológico relevante refere-se à complementaridade entre as diferentes fontes de evidência utilizadas na avaliação do processo. A análise integrada demonstrou que a avaliação objetiva por LLM, a percepção dos desenvolvedores em escala Likert e a análise de sentimento dos comentários constituem fontes complementares: a avaliação automatizada é mais sensível e abrangente, cobrindo a totalidade do backlog de forma contínua; a escala Likert é validadora, confirmando que o nível de qualidade percebida no trabalho real é coerente com o diagnóstico automatizado; e a análise de sentimento é diagnóstica, revelando os temas que geram insatisfação real e apontando para ações corretivas específicas. A convergência entre essas fontes (especialmente nos períodos de queda da qualidade ao final de cada PI) fortalece a validade das evidências empíricas apresentadas.

Por fim, conclui-se que o artefato desenvolvido contribui para elevar o nível de governança do *Product Backlog*, ao oferecer um método estruturado, escalável e orientado à melhoria contínua da qualidade dos requisitos ágeis. As evidências apresentadas ao longo do estudo sustentam a relevância da solução proposta e indicam seu potencial de aplicação em diferentes contextos organizacionais que enfrentam desafios semelhantes na gestão de produtos digitais.

As limitações associadas ao estudo e as possibilidades de evolução do método e do sistema proposto são discutidas na próxima seção, de modo a contextualizar o alcance dos resultados e apontar caminhos para investigações futuras.

6.2 Limitações do Estudo

Embora os resultados apresentados nesta pesquisa indiquem a relevância e a viabilidade do artefato desenvolvido, é importante reconhecer as limitações inerentes ao estudo, de modo a contextualizar adequadamente o alcance das conclusões e evitar generalizações indevidas. A explicitação dessas limitações contribui para a transparência científica e para a correta interpretação dos achados.

Uma primeira limitação refere-se ao contexto organizacional no qual o artefato foi aplicado. A pesquisa foi conduzida em uma organização de grande porte que adota metodologias ágeis em escala, com características específicas de maturidade organizacional, estrutura de governança e complexidade operacional. Embora esse contexto seja representativo de ambientes ágeis complexos, os resultados obtidos não podem ser automaticamente generalizados para organizações de menor porte ou para contextos com diferentes níveis de adoção de métodos ágeis.

Outra limitação diz respeito à dependência da qualidade dos dados de entrada utilizados na avaliação. O método proposto analisa artefatos predominantemente textuais, como histórias de usuário e critérios de aceitação. Assim, itens de backlog excessivamente incompletos, ambíguos ou mal estruturados tendem a resultar em diagnósticos igualmente limitados, o que reforça a necessidade de práticas mínimas de disciplina na elaboração dos requisitos. O artefato, portanto, não substitui a responsabilidade dos atores envolvidos na elaboração inicial dos requisitos.

Do ponto de vista tecnológico, a utilização de modelos de linguagem de larga escala introduz limitações relacionadas à sensibilidade à formulação dos prompts e à variabilidade das respostas geradas. Embora tenham sido adotadas estratégias para padronizar as instruções e o formato das saídas, o comportamento dos modelos pode variar conforme atualizações da tecnologia ou mudanças nos parâmetros de configuração. Essa característica demanda monitoramento contínuo e ajustes periódicos para garantir a consistência das avaliações ao longo do tempo.

Cabe aprofundar essa limitação sob uma perspectiva mais específica. Mesmo com a utilização de hiperparâmetros destinados a reduzir a aleatoriedade das respostas, como *temperature* = 0 e *top_p* = 0,1, modelos de linguagem de larga escala apresentam um grau residual de não determinismo. Esse comportamento manifesta-se por meio de pequenas variações de avaliação em itens com características textuais muito semelhantes.

O fenômeno está associado a aspectos computacionais do processamento dos modelos, incluindo operações em ponto flutuante e execução paralela em unidades de processamento gráfico. Como consequência, duas execuções do mesmo *prompt*, realizadas sob condições aparentemente idênticas, podem produzir respostas ligeiramente diferentes. Para este estudo, a principal implicação é que comparações individuais entre avaliações realizadas em períodos distintos devem ser interpretadas

com cautela. O indicador mais confiável do comportamento do sistema encontra-se nas tendências agregadas observadas ao longo do tempo, e não nas variações pontuais de itens específicos.

Outra fonte potencial de variabilidade está associada às atualizações realizadas pelos fornecedores dos modelos. Modificações internas podem alterar sistematicamente o comportamento das avaliações, mesmo quando não há mudança de versão explicitamente comunicada aos usuários. Isso pode introduzir descontinuidades que se tornam difíceis de identificar em análises retrospectivas.

Para reduzir esse risco, o presente estudo adotou três estratégias complementares: (i) versionamento explícito do *prompt* e dos hiperparâmetros utilizados; (ii) monitoramento contínuo de estatísticas agregadas por período; e (iii) priorização da análise de tendências em detrimento de comparações absolutas entre avaliações individuais. Embora essas medidas contribuam para aumentar a estabilidade das análises, elas não eliminam completamente a possibilidade de variações sistemáticas não intencionais ao longo do tempo.

Um achado empiricamente relevante que emergiu da aplicação do método refere-se ao risco de *gaming* da nota: registros da pesquisa com desenvolvedores indicaram que, em determinados casos (1,8%), critérios de aceitação foram gerados por modelos de linguagem e incorporados aos cards sem revisão humana adequada, com o objetivo de elevar artificialmente a pontuação. Esse comportamento é, antes de tudo, uma evidência do funcionamento do canal de escuta estruturada criado pelo método: sem a participação ativa dos desenvolvedores, a vulnerabilidade permaneceria invisível nos dados objetivos. Situar esse achado na trajetória percorrida é igualmente importante: no início da implantação, as dimensões avaliadas apresentavam notas próximas de zero, refletindo um estado em que o refinamento era conduzido, potencialmente, com elevado grau de subjetividade e sem critérios explícitos. A tendência de buscar atalhos diante de métricas visíveis é sintoma natural do amadurecimento do processo, e identificá-la representa, em si, uma conquista metodológica. As salvaguardas propostas para mitigar esse risco de forma estrutural são detalhadas na seção de Discussão dos Resultados (Capítulo 5). Adicionalmente, a pesquisa não teve como objetivo realizar uma validação estatística rigorosa dos resultados, uma vez que a abordagem adotada se insere no escopo da *Design Science Research*, priorizando a utilidade e a viabilidade do artefato em um contexto real. Assim, as evidências apresentadas possuem caráter

predominantemente qualitativo e exploratório, o que limita inferências causais mais amplas sobre os impactos do método no desempenho das equipes ou nos resultados dos projetos.

Outra limitação refere-se ao envolvimento dos usuários na avaliação do artefato. Embora Product Owners e Agile Masters tenham utilizado os resultados gerados como apoio ao refinamento e ao planejamento, não foi conduzido um estudo longitudinal aprofundado para avaliar a adoção do método ao longo de múltiplos ciclos ou sua influência sustentada sobre a maturidade organizacional. Essa restrição temporal limita a observação de efeitos de longo prazo.

No que se refere à pesquisa de percepção com desenvolvedores, verificou-se uma limitação estrutural de adesão: nenhuma sprint isolada atingiu o nível de representatividade amostral considerado adequado (margem de erro $\leq \pm 15\%$), e apenas dois meses (dezembro/2025 e janeiro/2026) atingiram esse patamar no agrupamento por período. Essa restrição implica que as análises derivadas da escala Likert e da análise de sentimento possuem caráter predominantemente qualitativo e indicativo, não sendo adequadas para inferências estatísticas conclusivas sobre toda a população de desenvolvedores. O principal caminho de superação dessa limitação reside na incorporação do formulário de avaliação como item fixo das retrospectivas dos times, o que elevaria a taxa de adesão e ampliaria a representatividade das amostras em ciclos futuros.

Por fim, destaca-se que o artefato foi concebido como um instrumento de apoio à decisão, e não como um mecanismo prescritivo ou de controle. A interpretação e o uso dos resultados dependem da maturidade organizacional e da cultura das equipes. O estudo não explorou em profundidade os aspectos culturais e comportamentais associados à adoção do método, os quais podem influenciar significativamente sua efetividade em diferentes contextos.

O reconhecimento dessas limitações não compromete as contribuições do estudo, mas delimita seu escopo e reforça a necessidade de investigações complementares. Na próxima seção, são apresentados possíveis caminhos para trabalhos futuros, indicando oportunidades de evolução do método e aprofundamento da pesquisa.

6.3 Trabalhos Futuros

A partir dos resultados obtidos e das limitações identificadas nesta pesquisa, abrem-se diversas possibilidades para a continuidade e o aprofundamento do estudo, tanto no âmbito científico quanto no prático. Os trabalhos futuros podem contribuir para a evolução do método proposto, para a ampliação de sua aplicabilidade e para o fortalecimento das evidências empíricas relacionadas ao uso de Inteligência Artificial na gestão da qualidade do *Product Backlog*.

Uma primeira linha de investigação consiste na realização de estudos longitudinais que acompanhem a aplicação do artefato ao longo de múltiplos ciclos de planejamento e execução. Esse tipo de estudo permitiria avaliar a evolução da maturidade do backlog ao longo do tempo, bem como investigar a relação entre a melhoria da qualidade dos artefatos e indicadores de desempenho das equipes, como previsibilidade das entregas, retrabalho e estabilidade do fluxo de trabalho.

Uma contribuição metodológica de particular relevância para investigações futuras seria a aplicação de métodos quase-experimentais para investigar a relação causal entre a melhoria da qualidade do backlog e os indicadores de desempenho do ART. O *Causal Impact*, proposto por (Brodersen et al.), utiliza uma abordagem bayesiana de séries temporais para estimar o efeito de uma intervenção sobre uma variável de resultado, construindo um grupo de controle sintético a partir de séries não afetadas pela intervenção. No contexto desta pesquisa, a adoção sistemática do método de avaliação do backlog poderia ser tratada como a intervenção, e o IEP ou o % *Concluded* como as variáveis de resultado. A disponibilidade de séries temporais de qualidade de backlog e de desempenho de sprint provenientes de múltiplos times e ARTs, alguns dos quais ainda não adotaram o método, cria as condições necessárias para a construção do contrafactual e para a estimação do efeito causal. Essa linha de investigação transformaria a associação temporal documentada nesta dissertação em evidência de causalidade, fortalecendo substancialmente as bases científicas do método.

Outra possibilidade relevante refere-se à ampliação do contexto empírico da pesquisa. A aplicação do método em organizações de diferentes portes, setores e níveis de maturidade ágil permitiria avaliar sua adaptabilidade e generalização. Estudos comparativos entre contextos distintos poderiam fornecer evidências adicionais

sobre os fatores organizacionais que influenciam a efetividade do método e do sistema computacional proposto.

No âmbito da *Design Science Research*, um dos critérios de qualidade do artefato é sua capacidade de generalização para uma classe de problemas mais ampla, e não apenas para a instância investigada. A presente pesquisa abordou a instância “avaliação da qualidade de *User Stories* em ambiente ágil em escala”, pertencente à classe de problemas “Avaliação Sistemática da Qualidade de Artefatos Textuais de Requisitos”. A mesma classe de problema comporta, contudo, outras instâncias igualmente relevantes no contexto SAFe, sendo a avaliação da qualidade de *Features* uma extensão natural e prioritária.

As *Features* constituem o nível intermediário da hierarquia de requisitos do SAFe, situando-se entre as *Epics* e as *User Stories*, e agregam atributos textuais de maior complexidade, entre eles: descrição e contexto de negócio, objetivo e valor para o negócio, critérios de aceite, métricas de sucesso, dependências, riscos, estimativa de sprints e pontuação pelo método WSJF (*Weighted Shortest Job First*). A qualidade informacional desses artefatos tem impacto direto na previsibilidade do planejamento do PI (*Program Increment*) e na capacidade do trem de realizar entregas de valor consistentes. Aplicar o método proposto a essa instância demandaria a especificação de um novo conjunto de dimensões de avaliação e a construção de *prompts* calibrados para o perfil textual das *Features*, mas o núcleo do artefato (avaliação automatizada por LLM, escala estruturada, retroalimentação visual e canal de percepção dos usuários) pode ser reutilizado sem redesenho fundamental. Essa extensão representaria uma contribuição adicional à classe de problemas identificada, fortalecendo a validade do método como solução genérica para a gestão da qualidade de requisitos em ambientes ágeis em escala.

Do ponto de vista metodológico, trabalhos futuros podem explorar o refinamento dos critérios de avaliação e da escala de mensuração adotada. A incorporação de técnicas estatísticas ou psicométricas para validação da escala, bem como a análise da confiabilidade entre avaliações humanas e automatizadas, representa um caminho promissor para fortalecer o rigor científico do método.

Em relação à tecnologia empregada, há oportunidades para investigar a utilização de diferentes modelos de linguagem, bem como estratégias de ajuste fino (*fine-tuning*) ou adaptação ao domínio organizacional específico. Estudos futuros tam-

bém podem explorar a combinação de modelos de linguagem com técnicas adicionais de Inteligência Artificial, como aprendizado de máquina supervisionado ou sistemas baseados em regras, visando aprimorar a consistência e a explicabilidade das avaliações.

Uma linha de evolução tecnológica particularmente relevante para a sustentabilidade de longo prazo do método refere-se à possibilidade de migração dos modelos de linguagem de larga escala hospedados em infraestrutura de nuvem comercial para modelos menores, especializados e executáveis localmente (*Small Language Models*, SLMs). Essa transição responderia a duas limitações identificadas empiricamente: a dependência de infraestrutura externa de custo variável, cujo crescimento pode tornar o método operacionalmente inviável em cenários de alto volume de avaliações; e as restrições de privacidade e conformidade informacional associadas ao envio de dados corporativos a provedores externos. A execução local de modelos especializados no domínio de engenharia de requisitos ágeis (obtida por meio de *fine-tuning* em conjuntos de dados curados de histórias de usuário e critérios de aceitação) potencialmente permitiria manter a qualidade das avaliações com custo operacional previsível, latência reduzida e controle completo sobre os dados processados. A viabilidade dessa migração depende, contudo, da evolução das capacidades dos SLMs em tarefas de raciocínio contextual, bem como da disponibilidade de conjuntos de treinamento suficientemente representativos do domínio ágil em escala.

Outra linha de pesquisa refere-se à integração do artefato com ferramentas corporativas de gestão ágil, como sistemas de planejamento e acompanhamento de projetos. A automação da coleta de dados e a incorporação dos resultados diretamente nos fluxos de trabalho das equipes podem ampliar a adoção do método e reduzir ainda mais o esforço operacional associado à avaliação da qualidade do backlog.

Adicionalmente, investigações futuras podem aprofundar os aspectos humanos e organizacionais associados ao uso do artefato. Estudos qualitativos focados na percepção dos usuários, na aceitação do método e em seus impactos sobre a cultura de planejamento e aprendizado contínuo podem fornecer insights relevantes para a adoção sustentável da solução.

Por fim, os resultados desta pesquisa abrem espaço para a consolidação de um programa de pesquisa mais amplo, no qual o método proposto seja explorado como parte de uma abordagem integrada de governança ágil orientada por evidências. Essa perspectiva inclui a possibilidade de desdobramentos em nível de doutorado, bem como

a produção de artigos científicos e produtos tecnológicos derivados, contribuindo para o avanço do conhecimento na interface entre métodos ágeis, engenharia de requisitos e Inteligência Artificial.

REFERÊNCIAS

ALAWAJI, Batool; HAKAMI, Mona; ALSHEMAIMRI, Bader. Evaluating Generative Language Models with Prompt Engineering for Categorizing User Stories to its Sector Domains. In: 2024 IEEE 9TH INTERNATIONAL CONFERENCE FOR CONVERGENCE IN TECHNOLOGY, I2CT 2024. **IEEE Int. Conf. Converg. Technol., I2CT**. [S. l.]: Institute of Electrical e Electronics Engineers Inc., 2024. Journal Abbreviation: IEEE Int. Conf. Converg. Technol., I2CT. ISBN 979-8-3503-9447-4. DOI: [10.1109/I2CT61223.2024.10544242](https://doi.org/10.1109/I2CT61223.2024.10544242).

ALSALEMI, Ahmed Mubark; YEOH, Eng-Thiam. A survey on product backlog change management and requirement traceability in agile (Scrum). In: 2015 9TH MALAYSIAN SOFTWARE ENGINEERING CONFERENCE, MYSEC 2015. **Malays. Softw. Eng. Conf., MySEC**. [S. l.]: Institute of Electrical e Electronics Engineers Inc., 2016. p. 189–194. Journal Abbreviation: Malays. Softw. Eng. Conf., MySEC. ISBN 978-1-4673-8226-7. DOI: [10.1109/MySEC.2015.7475219](https://doi.org/10.1109/MySEC.2015.7475219).

ALSHAMMARI, Abdullah. **On the use of user interface specific domain Gherkin in behaviour driven development**. 2025. Tese (Doutorado) – University of Glasgow. DOI: [10.5525/gla.thesis.85190](https://theses.gla.ac.uk/85190/). Disponível em: <https://theses.gla.ac.uk/85190/>. Acesso em: 14 fev. 2026.

AMJAD, Sunnia et al. Calculating Completeness of Agile Scope in Scaled Agile Development. **IEEE Access**, Institute of Electrical e Electronics Engineers Inc., v. 6, p. 5822–5847, 2017. ISSN 2169-3536. DOI: [10.1109/ACCESS.2017.2765351](https://doi.org/10.1109/ACCESS.2017.2765351).

ARORA, Chetan; GRUNDY, John; ABDELRAZEK, Mohamed. Advancing Requirements Engineering Through Generative AI: Assessing the Role of LLMs. In: GENERATIVE AI for Effective Software Development. [S. l.]: Springer, 2024. p. 129–148. DOI: [10.1007/978-3-031-55642-5_6](https://doi.org/10.1007/978-3-031-55642-5_6).

BOERMAN, Martin P. et al. Measuring and monitoring agile development status. In: INTERNATIONAL WORKSHOP ON EMERGING TRENDS IN SOFTWARE METRICS, WETSOM. **Int. Workshop Emerg. Trends Softw. Metrics, WETSoM**. [S. l.]: IEEE Computer Society, 2015. 2015-August, p. 54–62. Journal Abbreviation: Int. Workshop Emerg. Trends Softw. Metrics, WETSoM. ISBN 978-1-4799-1934-5. DOI: [10.1109/WETSoM.2015.15](https://doi.org/10.1109/WETSoM.2015.15).

BROCKENBROUGH, Allan; SALINAS, Dominic. Using Generative AI to Create User Stories in the Software Engineering Classroom. In: 2024 36TH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING EDUCATION AND TRAINING (CSEE&T). **2024 36th International Conference on Software Engineering Education and Training (CSEE&T)**. [S. l.: s. n.], jul. 2024. p. 1–5. ISSN: 2377-570X. DOI: [10.1109/CSEET62301.2024.10662994](https://doi.org/10.1109/CSEET62301.2024.10662994). Disponível em:

<https://ieeexplore.ieee.org/abstract/document/10662994>. Acesso em: 19 fev. 2026.

BRODERSEN, Kay H. et al. Inferring causal impact using Bayesian structural time-series models. **The Annals of Applied Statistics**, v. 9, n. 1, p. 247–274, 2015. DOI: [10.1214/14-AOAS788](https://doi.org/10.1214/14-AOAS788). Disponível em: <https://doi.org/10.1214/14-AOAS788>.

CAMPOS, Vicente Falconi. **Daily Work Routine Management**. Belo Horizonte: Editora de Desenvolvimento Gerencial, 2003. ISBN 978-85-86948-45-9.

CAN, Ashley T. van; DALPIAZ, Fabiano. Locating requirements in backlog items: Content analysis and experiments with large language models. **Information and Software Technology**, Elsevier B.V., v. 179, 2025. ISSN 0950-5849. DOI: [10.1016/j.infsof.2024.107644](https://doi.org/10.1016/j.infsof.2024.107644).

CHUOR, Porchourng; ITTOO, Ashwin; HENG, Samedi. User Story Classification with Machine Learning and LLMs. In: CAO, Cungeng et al. (ed.). **Knowledge Science, Engineering and Management**. Singapore: Springer Nature, 2024. p. 161–175. ISBN 978-981-97-5492-2. DOI: [10.1007/978-981-97-5492-2_13](https://doi.org/10.1007/978-981-97-5492-2_13).

COHN, Mike. **Agile Estimating and Planning**. [S. l.]: Prentice Hall Professional Technical Reference, 2005. ISBN 978-0-13-147941-8.

COHN, Mike. **User Stories Applied: For Agile Software Development**. [S. l.]: Mountain Goat Software, ago. 2003. Apresentação no XP/Agile Universe 2003.

COHN, Mike. **User Stories Applied: For Agile Software Development**. [S. l.]: Addison Wesley, 2004. ISBN 0-321-20568-5. Disponível em: https://drive.google.com/drive/u/0/folders/1pu07tpgNnmS0cdtrigtBE6_TkfnIaHxP.

DILORENZO, Ednaldo et al. Enabling the Reuse of Software Development Assets through a Taxonomy for User Stories. **IEEE Access**, Institute of Electrical e Electronics Engineers Inc., v. 8, p. 107285–107300, 2020. ISSN 2169-3536. DOI: [10.1109/ACCESS.2020.2996951](https://doi.org/10.1109/ACCESS.2020.2996951).

DRESCH, Aline; LACERDA, Daniel Pacheco; ANTUNES JUNIOR, José Antonio Valle. **Design Science Research: método de pesquisa para avanço da ciência e tecnologia**. Porto Alegre: Bookman, 2015. ISBN 978-85-8260-299-7.

ERDOĞAN, Onur; PEKKAYA, Muhammed Emre; GÖK, Halime. More effective sprint retrospective with statistical analysis. In: JOURNAL OF SOFTWARE: EVOLUTION AND PROCESS, 5. **J. Softw. Evol. Process**. [S. l.]: John Wiley e Sons Ltd, 2018. v. 30. Journal Abbreviation: J. Softw. Evol. Process. DOI: [10.1002/smr.1933](https://doi.org/10.1002/smr.1933).

FEW, Stephen. **Information Dashboard Design: The Effective Visual Communication of Data**. Sebastopol, CA: O'Reilly Media, Inc., 2006. ISBN 0-596-10016-7.

GILSON, Fabian; IRWIN, Calum. From user stories to use case scenarios towards a generative approach. In: PROCEEDINGS - 25TH AUSTRALASIAN SOFTWARE ENGINEERING CONFERENCE, ASWEC 2018. **Proc. - Australas. Softw. Eng. Conf., ASWEC**. [S. l.]: Institute of Electrical e Electronics Engineers Inc., 2018. p. 61–65. Journal Abbreviation: Proc. - Australas. Softw. Eng. Conf., ASWEC. ISBN 978-1-7281-1241-1. DOI: [10.1109/ASWEC.2018.00016](https://doi.org/10.1109/ASWEC.2018.00016).

GRAPENTHIN, Simon et al. Improving task breakdown comprehensiveness in agile projects with an Interaction Room. In: INFORMATION AND SOFTWARE TECHNOLOGY. **Inf Software Technol**. [S. l.]: Elsevier B.V., 2015. v. 67, p. 254–264. Journal Abbreviation: Inf Software Technol. DOI: [10.1016/j.infsof.2015.07.008](https://doi.org/10.1016/j.infsof.2015.07.008).

HEVNER, Alan R. et al. Design science in information systems research. **MIS Quarterly: Management Information Systems**, Management Information Systems Research Center, v. 28, n. 1, p. 75–105, 2004. ISSN 0276-7783. DOI: [10.2307/25148625](https://doi.org/10.2307/25148625).

HIGHSMITH, Jim. **Agile Project Management: Creating Innovative Products**. Boston, MA: Addison-Wesley, 2004. ISBN 0-321-21977-5.

HOPP, Wallace J.; SPEARMAN, Mark L. **Factory Physics**. 2. ed. Boston: Irwin/McGraw-Hill, 2001. ISBN 978-0-02-562479-5.

INMON, W. H. **Building the Data Warehouse**. 4. ed. Indianapolis, Indiana: Wiley Publishing, Inc., 2005. ISBN 0-7645-9944-5.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/IEC 25010: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models**. Geneva: [s. n.], 2011.

IUNES, João Paulo. **Lei de Little**. Colaborae. Ago. 2023. Disponível em: <https://colaborae.com.br/blog/2023/08/22/lei-de-little/>. Acesso em: 26 fev. 2026.

JAKOBSEN, Carsten Ruseng; JOHNSON, Kent Aaron. Mature agile with a twist of CMMI. In: PROCEEDINGS - AGILE 2008 CONFERENCE. **Proc. - Agile Conf**. [S. l.: s. n.], 2008. p. 212–217. Journal Abbreviation: Proc. - Agile Conf. ISBN 978-0-7695-3321-6. DOI: [10.1109/Agile.2008.10](https://doi.org/10.1109/Agile.2008.10).

JAKOBSEN, Carsten Ruseng; SUTHERLAND, Jeff. Scrum and CMMI - Going from good to great: Are you ready-ready to be done-done? In: PROCEEDINGS - 2009 AGILE CONFERENCE, AGILE 2009. **Proc. - Agile Conf., AGILE**. [S. l.: s. n.], 2009. p. 333–337. Journal Abbreviation: Proc. - Agile Conf., AGILE. ISBN 978-0-7695-3768-9. DOI: [10.1109/AGILE.2009.31](https://doi.org/10.1109/AGILE.2009.31).

KAPLAN, Robert S.; NORTON, David P. **The Balanced Scorecard: Translating Strategy into Action**. Boston, Massachusetts: Harvard Business School Press, 1996. ISBN 0-87584-651-3.

KHAN, Arshad. **Visual Analytics for Dashboards: A Step-by-Step Guide to Principles and Practical Techniques**. Tracy, CA, USA: Apress, 2024. ISBN 979-8-8688-0118-1.

KIM, Dae-Kyoo. Comparing Proficiency of ChatGPT and Bard in Software Development. In: GENERATIVE AI for Effective Software Development. [S. l.]: Springer, 2024. p. 25–52. DOI: [10.1007/978-3-031-55642-5_2](https://doi.org/10.1007/978-3-031-55642-5_2).

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. 3. ed. Indianapolis, Indiana: John Wiley & Sons, Inc., 2013. ISBN 978-1-118-53080-1.

KNAFLIC, Cole Nussbaumer. **Storytelling com Dados: Um Guia Sobre Visualização de Dados para Profissionais de Negócios**. Tradução: João Tortello. Rio de Janeiro, RJ: Alta Books Editora, 2019. ISBN 978-85-508-1638-8.

KOLTHOFF, Kristian et al. Interlinking User Stories and GUI Prototyping: A Semi-Automatic LLM-Based Approach. In: 2024 IEEE 32ND INTERNATIONAL REQUIREMENTS ENGINEERING CONFERENCE (RE). **2024 IEEE 32nd International Requirements Engineering Conference (RE)**. [S. l.: s. n.], 2024. p. 380–388. ISSN: 2332-6441. DOI: [10.1109/RE59067.2024.00045](https://doi.org/10.1109/RE59067.2024.00045). Disponível em: <https://ieeexplore.ieee.org/document/10628517>. Acesso em: 18 fev. 2026.

KÜPPER, Steffen et al. How has SPI changed in times of agile development? Results from a multi-method study. **Journal of Software: Evolution and Process**, John Wiley e Sons Ltd, v. 31, n. 11, 2019. ISSN 2047-7481. DOI: [10.1002/smr.2182](https://doi.org/10.1002/smr.2182).

LACERDA, Rogério Tadeu De Oliveira; ENSSLIN, Leonardo; ENSSLIN, Sandra Rolim. Uma análise bibliométrica da literatura sobre estratégia e avaliação de desempenho. **Gestão & Produção**, v. 19, n. 1, p. 59–78, 2012. ISSN 0104-530X. DOI: [10.1590/S0104-530X2012000100005](https://doi.org/10.1590/S0104-530X2012000100005). Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-530X2012000100005&lng=pt&tlng=pt. Acesso em: 7 mar. 2026.

LINDLAND, Odd Ivar; SINDRE, Guttorm; SØLVBERG, Arne. Understanding Quality in Conceptual Modeling. **IEEE Software**, IEEE, v. 11, n. 2, p. 42–49, 1994. DOI: [10.1109/52.268955](https://doi.org/10.1109/52.268955).

LOTFI, Samia; NAFIL, Khalid. Automated Assessment and Enhancement of User Story Quality Using Large Language Models in Agile Development. In: 2025 INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS: THEORIES AND APPLICATIONS (SITA). **2025 International Conference on Intelligent Systems: Theories and Applications (SITA)**. [S. l.: s. n.], 2025. p. 1–8. DOI: [10.1109/SITA67914.2025.11273616](https://doi.org/10.1109/SITA67914.2025.11273616). Disponível em: <https://ieeexplore.ieee.org/document/11273616/>. Acesso em: 14 mar. 2026.

LU, Zhenyan et al. Small Language Models: Survey, Measurements, and Insights. **arXiv preprint arXiv:2409.15790**, 2024.

LUCASSEN, Garm et al. Improving agile requirements: the Quality User Story framework and tool. **Requirements Engineering**, Springer London, v. 21, n. 3, p. 383–403, 2016. ISSN 0947-3602. DOI: [10.1007/s00766-016-0250-x](https://doi.org/10.1007/s00766-016-0250-x).

MARCH, Salvatore T.; SMITH, Gerald F. Design and natural science research on information technology. **Decision Support Systems**, v. 15, n. 4, p. 251–266, 1995. ISSN 01679236. DOI: [10.1016/0167-9236\(94\)00041-2](https://doi.org/10.1016/0167-9236(94)00041-2). Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0167923694000412>. Acesso em: 2 mar. 2026.

MENZINSKY, Alexander. **Historias de Usuario**. 2022. Disponível em: https://mobilizaacademy.com/wp-content/uploads/2022/02/scrum_manager_historias_usuario.pdf. Acesso em: 27 fev. 2026.

MORADI DAKHEL, Arghavan et al. An Overview on Large Language Models. In: GENERATIVE AI for Effective Software Development. [S. l.]: Springer, 2024. p. 3–23. DOI: [10.1007/978-3-031-55642-5_1](https://doi.org/10.1007/978-3-031-55642-5_1).

MÜTER, Laurens et al. Refinement of User Stories into Backlog Items: Linguistic Structure and Action Verbs: Research Preview. In: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS). **Lect. Notes Comput. Sci.** [S. l.]: Springer Verlag, 2019. 11412 LNCS, p. 109–116. Journal Abbreviation: Lect. Notes Comput. Sci. ISBN 978-3-030-15537-7. DOI: [10.1007/978-3-030-15538-4_7](https://doi.org/10.1007/978-3-030-15538-4_7).

NASIRI, Samia; LAHMER, Mohammed. AI-driven methodology for refining and clustering Agile requirements. In: 2024 4TH INTERNATIONAL CONFERENCE ON INNOVATIVE RESEARCH IN APPLIED SCIENCE, ENGINEERING AND TECHNOLOGY, IRASET 2024. **Int. Conf. Innov. Res. Appl. Sci., Eng. Technol., IRASET**. [S. l.]: Institute of Electrical e Electronics Engineers Inc., 2024. Journal

Abbreviation: Int. Conf. Innov. Res. Appl. Sci., Eng. Technol., IRASET. ISBN 979-8-3503-0950-8. DOI: [10.1109/IRASET60544.2024.10548285](https://doi.org/10.1109/IRASET60544.2024.10548285).

NORTH, Dan. **Introducing BDD**. [S. l.: s. n.], 2006. Disponível em: <https://dannorth.net/blog/introducing-bdd/>.

OHNO, Taiichi. **Toyota Production System: Beyond Large-Scale Production**. Boca Raton, FL: Productivity Press / CRC Press, 1988. ISBN 978-0-915299-14-0.

OPENAI. **GPT-4o Capabilities and Configurations**. San Francisco: OpenAI, 2024. Documentação técnica oficial. Disponível em: <https://developers.openai.com/api/reference/resources/chat/subresources/completions/methods/create>. Acesso em: 10 ago. 2025.

OSWAL, Jay U.; KANAKIA, Harshil T.; SUKTEL, Devvrat. Transforming Software Requirements into User Stories with GPT-3.5 -: An AI-Powered Approach. In: 2ND INTERNATIONAL CONFERENCE ON INTELLIGENT DATA COMMUNICATION TECHNOLOGIES AND INTERNET OF THINGS, IDCIoT 2024. **Int. Conf. Intell. Data Commun. Technol. Internet Things, IDCIoT**. [S. l.]: Institute of Electrical e Electronics Engineers Inc., 2024. p. 913–920. Journal Abbreviation: Int. Conf. Intell. Data Commun. Technol. Internet Things, IDCIoT. ISBN 979-8-3503-2753-3. DOI: [10.1109/IDCIoT59759.2024.10467750](https://doi.org/10.1109/IDCIoT59759.2024.10467750).

PAASIVAARA, Maria. Adopting SAFe to scale agile in a globally distributed organization. In: 2017 IEEE 12th International Conference on Global Software Engineering (ICGSE). [S. l.]: IEEE, 2017. p. 36–40. Disponível em: https://ieeexplore.ieee.org/abstract/document/7976685/?casa_token=tZ8rFL_hHVwAAAA:lkyDg1UWRJfKf_X2bvRHR3soLHNFHZZT74s5QFLJnw-vFHLT46e9LUo7w_Kp0Dk91MybwxV7HPHLp. Acesso em: 14 fev. 2026.

PEFFERS, Ken et al. A Design Science Research Methodology for Information Systems Research. **Journal of Management Information Systems**, Routledge, v. 24, n. 3, p. 45–77, 2007. _eprint: <https://doi.org/10.2753/MIS0742-1222240302>. ISSN 0742-1222. DOI: [10.2753/MIS0742-1222240302](https://doi.org/10.2753/MIS0742-1222240302). Disponível em: <https://doi.org/10.2753/MIS0742-1222240302>. Acesso em: 24 fev. 2026.

PETERSEN, Kai; WOHLIN, Claes. The effect of moving from a plan-driven to an incremental software development approach with agile practices: An industrial case study. **Empirical Software Engineering**, v. 15, n. 6, p. 654–693, 2010. ISSN 1573-7616. DOI: [10.1007/s10664-010-9136-6](https://doi.org/10.1007/s10664-010-9136-6).

RAHMAN, Tajmilur et al. Take Loads Off Your Developers: Automated User Story Generation using Large Language Model. In: PROCEEDINGS - 2024 IEEE INTERNATIONAL CONFERENCE ON SOFTWARE MAINTENANCE AND EVOLUTION, ICSME 2024. **Proc. - IEEE Int. Conf. Softw. Maint. Evol., ICSME**. [S. l.]:

Institute of Electrical e Electronics Engineers Inc., 2024. p. 791–801. Journal Abbreviation: Proc. - IEEE Int. Conf. Softw. Maint. Evol., ICSME. ISBN 979-8-3503-9568-6. DOI: [10.1109/ICSME58944.2024.00082](https://doi.org/10.1109/ICSME58944.2024.00082).

RAMASAMY, Vijayalakshmi et al. Enhancing User Story Generation in Agile Software Development Through Open AI and Prompt Engineering. In: 2024 IEEE FRONTIERS IN EDUCATION CONFERENCE (FIE). **2024 IEEE Frontiers in Education Conference (FIE)**. [S. l.: s. n.], 2024. p. 1–8. ISSN: 2377-634X. DOI: [10.1109/FIE61694.2024.10893343](https://doi.org/10.1109/FIE61694.2024.10893343). Disponível em: <https://ieeexplore.ieee.org/document/10893343>. Acesso em: 18 fev. 2026.

REINERTSEN, Donald G. **The Principles of Product Development Flow: Second Generation Lean Product Development**. [S. l.]: Celeritas Publishing, 2009.

RIBEIRO DOS SANTOS, Shexmo et al. Guidelines for Adopting Behavior-Driven Development (BDD): A Case Study: in: 21ST INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS AND TECHNOLOGIES. **Proceedings of the 21st International Conference on Web Information Systems and Technologies**. Marbella, Spain: SCITEPRESS - Science e Technology Publications, 2025. p. 141–148. ISBN 978-989-758-772-6. DOI: [10.5220/0013681700003985](https://doi.org/10.5220/0013681700003985). Disponível em: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0013681700003985>. Acesso em: 14 fev. 2026.

ROLA, Paweł; KUCHTA, Dorota. Application of fuzzy sets to the expert estimation of Scrum-based projects. **Symmetry**, MDPI AG, v. 11, n. 8, 2019. ISSN 2073-8994. DOI: [10.3390/sym11081032](https://doi.org/10.3390/sym11081032).

RONANKI, Krishna; CABRERO-DANIEL, Beatriz; BERGER, Christian. ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box? In: LECTURE NOTES IN BUSINESS INFORMATION PROCESSING. **Lect. Notes Bus. Inf. Process.** [S. l.]: Springer Science e Business Media Deutschland GmbH, 2024. 489 LNBIP, p. 173–181. Journal Abbreviation: Lect. Notes Bus. Inf. Process. ISBN 978-3-031-48549-7. DOI: [10.1007/978-3-031-48550-3_17](https://doi.org/10.1007/978-3-031-48550-3_17).

SAMI, Malik Abdul et al. Early Results of an AI Multiagent System for Requirements Elicitation and Analysis. In: LECTURE NOTES IN COMPUTER SCIENCE (INCLUDING SUBSERIES LECTURE NOTES IN ARTIFICIAL INTELLIGENCE AND LECTURE NOTES IN BIOINFORMATICS). **Lect. Notes Comput. Sci.** [S. l.]: Springer Science e Business Media Deutschland GmbH, 2025. 15452 LNCS, p. 307–316. Journal Abbreviation: Lect. Notes Comput. Sci. ISBN 978-3-031-78385-2. DOI: [10.1007/978-3-031-78386-9_20](https://doi.org/10.1007/978-3-031-78386-9_20).

SANTOS, Reine et al. User Stories: Does ChatGPT Do It Better?: in: 27TH INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS. **Proceedings of the 27th International Conference on Enterprise Information**

Systems. Porto, Portugal: SCITEPRESS - Science e Technology Publications, 2025. p. 47–58. ISBN 978-989-758-749-8. DOI: 10.5220/0013365500003929. Disponível em: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0013365500003929>. Acesso em: 14 mar. 2026.

SCALED AGILE, INC. **SAFe Explained**. [S. l.]: Scaled Agile, Inc., 2025a. Disponível em: <https://scaledagile.com/resources/safe-explained-ebook/>.

SCALED AGILE, INC. **Scaled Agile Framework (SAFe)**. Acesso em: 01 fev. 2026. 2025b. Disponível em: <https://framework.scaledagile.com/>.

SCHWABER, Ken; SUTHERLAND, Jeff. **The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game**. [S. l.]: Attribution Share-Alike license of Creative Commons, nov. 2020. Disponível em: <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf>. Acesso em: 13 fev. 2026.

SEDANO, Todd; RALPH, Paul; PÉRAIRE, Cécile. The Product Backlog. In: 2019 IEEE/ACM 41ST INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING (ICSE). **2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)**. [S. l.: s. n.], 2019. p. 200–211. ISSN: 1558-1225. DOI: 10.1109/ICSE.2019.00036. Disponível em: <https://ieeexplore.ieee.org/document/8812076>. Acesso em: 18 fev. 2026.

SHARMA, Amol; CHATURVEDI, Amrita; TRIPATHI, Anil Kumar. From Problem Descriptions to User Stories: Utilizing Large Language Models through Prompt Chaining. In: 2024 15TH INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT). **2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)**. [S. l.: s. n.], 2024. p. 1–6. ISSN: 2473-7674. DOI: 10.1109/ICCCNT61001.2024.10724709. Disponível em: <https://ieeexplore.ieee.org/document/10724709/>. Acesso em: 14 mar. 2026.

SLACK, Nigel; BRANDON-JONES, Alistair; JOHNSTON, Robert. **Administração da Produção**. 8. ed. São Paulo, SP: Editora Atlas Ltda. / GEN, 2018. ISBN 978-85-97-01537-9.

STARON, Miroslaw; MEDING, Wilhelm; BANIASAD, Poupak. Information Needs for SAFe Teams and Release Train Management: A Design Science Research Study. In: PROCEEDINGS of the Design Science Research in Information Systems and Technology. [S. l.: s. n.], 2019. CEUR Workshop Proceedings (CEUR-WS.org), Vol-2476, paper 4.

SVENSSON, Richard Berntsson et al. BAM - Backlog Assessment Method. In: KRUCHTEN, Philippe; FRASER, Steven; COALLIER, François (ed.). **Agile Processes**

in Software Engineering and Extreme Programming. Cham: Springer International Publishing, 2019. p. 53–68. ISBN 978-3-030-19034-7. DOI: [10.1007/978-3-030-19034-7_4](https://doi.org/10.1007/978-3-030-19034-7_4).

TONA, Claudia et al. Scrumlity: An Agile Framework Based on Quality Assurance. In: 2021 9TH INTERNATIONAL CONFERENCE IN SOFTWARE ENGINEERING RESEARCH AND INNOVATION (CONISOFT). **2021 9th International Conference in Software Engineering Research and Innovation (CONISOFT).** [S. l.: s. n.], 2021. p. 88–96. DOI: [10.1109/CONISOFT52520.2021.00023](https://doi.org/10.1109/CONISOFT52520.2021.00023). Disponível em: <https://ieeexplore.ieee.org/document/9653207>. Acesso em: 18 fev. 2026.

TUFTE, Edward R. **The Visual Display of Quantitative Information.** 2. ed. Cheshire, Connecticut: Graphics Press, 2001.

UNLU, Huseyin et al. Predicting Software Functional Size Using Natural Language Processing: An Exploratory Case Study. In: PROCEEDINGS OF THE EUROMICRO CONFERENCE ON SOFTWARE ENGINEERING AND ADVANCED APPLICATIONS, EUROMICRO-SEAA, 2024. **Proc. Euromicro Conf. Softw. Eng. Adv. Appl., EUROMICRO-SEAA.** [S. l.]: Institute of Electrical e Electronics Engineers Inc., 2024. p. 188–193. Journal Abbreviation: Proc. Euromicro Conf. Softw. Eng. Adv. Appl., EUROMICRO-SEAA. ISBN 979-8-3503-8026-2. DOI: [10.1109/SEAA64295.2024.00036](https://doi.org/10.1109/SEAA64295.2024.00036).

VOM BROCKE, Jan; HEVNER, Alan; MAEDCHE, Alexander (ed.). **Design Science Research. Cases.** Cham: Springer International Publishing, 2020. (Progress in IS). ISBN 978-3-030-46780-7. DOI: [10.1007/978-3-030-46781-4](https://doi.org/10.1007/978-3-030-46781-4). Disponível em: <http://link.springer.com/10.1007/978-3-030-46781-4>. Acesso em: 2 mar. 2026.

WIERINGA, Roel. **Design Science Methodology for Information Systems and Software Engineering.** Berlin: Springer, 2014.

ZHANG, Zheyang et al. LLM-Based Agents for Automating the Enhancement of User Story Quality: An Early Report. In: LECTURE NOTES IN BUSINESS INFORMATION PROCESSING. **Lect. Notes Bus. Inf. Process.** [S. l.]: Springer Science e Business Media Deutschland GmbH, 2024. 512 LNBIP, p. 117–126. Journal Abbreviation: Lect. Notes Bus. Inf. Process. ISBN 978-3-031-61153-7. DOI: [10.1007/978-3-031-61154-4_8](https://doi.org/10.1007/978-3-031-61154-4_8).

GLOSSÁRIO

Agile Master

Profissional responsável por facilitar a aplicação dos princípios e práticas ágeis nas equipes de desenvolvimento. Atua como coach, removendo impedimentos e promovendo a melhoria contínua dos processos. Em contextos SAFe, o papel pode ser denominado *Scrum Master* ou *Team Coach*.

Agilidade em Escala

Aplicação de métodos ágeis em contextos organizacionais que envolvem múltiplas equipes trabalhando de forma coordenada em um mesmo produto ou portfólio. Requer frameworks específicos — como SAFe, LeSS ou Nexus — para alinhar o trabalho entre times, trens e programas.

ART — *Agile Release Train*

Unidade organizacional do SAFe composta por múltiplas equipes ágeis que colaboram de forma sincronizada para entregar incrementos de valor contínuos. O ART opera em ciclos denominados *Program Increments* (PI), com eventos de planejamento, revisão e retrospectiva em escala.

Artefato (DSR)

Na abordagem de *Design Science Research*, produto gerado pelo processo de pesquisa com o propósito de resolver um problema prático. Pode assumir a forma de constructo, modelo, método ou instanciação. O artefato é avaliado tanto pela sua utilidade prática quanto pela sua fundamentação teórica.

Análise de Sentimento

Técnica de Processamento de Linguagem Natural (NLP) que identifica e classifica a polaridade afetiva — positiva, negativa ou neutra — de trechos textuais. Nesta pesquisa, aplicada aos comentários dos desenvolvedores para identificar insatisfações recorrentes e temas críticos no processo de refinamento.

AQUSA — *Automatic Quality User Story Artisan*

Ferramenta acadêmica de apoio à escrita de histórias de usuário, que aplica técnicas de processamento de linguagem natural para detectar defeitos de qualidade e sugerir correções com base no framework QUS. Citada nesta pesquisa como referência de comparação com avaliações de qualidade realizadas por LLMs.

A-SPSRI — *Agile Software Project Scope Rating Index*

Método proposto para mensurar objetivamente a completude do escopo de projetos ágeis em escala, a partir de métricas derivadas dos dados do backlog. Citado na literatura revisada desta pesquisa como abordagem de medição da completude do escopo.

Backlog Refinement

Cerimônia ágil contínua na qual o *Product Backlog* é revisado, detalhado, estimado e priorizado em colaboração entre o *Product Owner* e a equipe de desenvolvimento. Tem como objetivo assegurar que os itens topo do backlog estejam suficientemente preparados para as próximas *sprints*.

BDD — *Behavior-Driven Development*

Abordagem de desenvolvimento de software que estende o TDD (*Test-Driven Development*) ao descrever o comportamento esperado do sistema em linguagem próxima da natural, utilizando a sintaxe *Dado/Quando/Então* (Gherkin). Contribui para tornar os critérios de aceitação mais objetivos e verificáveis.

Business Intelligence (BI)

Conjunto de tecnologias, processos e práticas voltados à coleta, ao armazenamento e à análise de dados organizacionais, com o objetivo de subsidiar a tomada de decisão. Nesta pesquisa, mencionado no contexto da arquitetura de Data Warehouse/BI proposta por Kimball, empregada na integração dos dados que alimentam os painéis de gestão à vista.

Boundary Object

Conceito oriundo da sociologia do conhecimento que descreve artefatos, documentos ou representações capazes de servir de interface entre diferentes grupos sociais com interesses distintos, possibilitando a coordenação sem exigir consenso pleno. O *Product Backlog* é tratado nesta pesquisa como um *boundary object* entre *Product Owners*, equipes e *stakeholders*.

Ciência do Artificial

Campo científico, formalizado por Herbert Simon, centrado no estudo e na concepção de artefatos destinados a transformar situações existentes em situações preferidas. Constitui a base epistemológica da *Design Science Research* e distingue-se das ciências naturais por seu caráter prescritivo e construtivo.

CMMI — *Capability Maturity Model Integration*

Modelo de referência para avaliação e melhoria da maturidade de processos organizacionais de desenvolvimento de software. Citado na literatura revisada como base disciplinar para checklists de inspeção da qualidade e da completude de histórias de usuário.

***Corpus* (pl. *corpora*)**

Conjunto amplo e estruturado de textos utilizado no treinamento ou na análise de modelos de linguagem. Nesta pesquisa, o termo aparece tanto no plural, ao descrever o pré-treinamento de LLMs em grandes *corpora* textuais, quanto no singular, ao referir-se ao conjunto de comentários analisados no módulo de análise de sentimento.

CrITÉRIOS de Aceitação

Conjunto de condições explícitas e verificáveis que uma história de usuário deve satisfazer para ser considerada concluída. Delimitam o escopo da entrega, reduzem ambiguidades e viabilizam a validação objetiva do comportamento esperado do sistema. Frequentemente formulados segundo a estrutura *Dado/Quando/Então* do BDD.

Cycle Time

Tempo decorrido entre o início efetivo do trabalho em um item e sua conclusão. Difere do *lead time* por não incluir o tempo de espera anterior ao início do trabalho. É um indicador de fluxo utilizado para monitorar a eficiência interna das equipes.

Daily Scrum

Evento diário de curta duração do Scrum, no qual a equipe de desenvolvimento sincroniza o andamento do trabalho e identifica impedimentos, operacionalizando os pilares de transparência e inspeção do framework.

Definition of Done (DoD)

Compromisso organizacional formalizado que estabelece os critérios mínimos de qualidade que um incremento de produto deve atender para ser considerado concluído. Inclui aspectos como testes realizados, documentação produzida e critérios técnicos satisfeitos, assegurando transparência e consistência nas entregas.

Definition of Ready (DoR)

Conjunto de critérios que um item do *Product Backlog* deve satisfazer para ser considerado suficientemente preparado para entrar em uma *sprint*. Engloba elementos como descrição clara do valor de negócio, critérios de aceitação definidos, dependências identificadas e estimativa de esforço realizada.

Design Science Research (DSR)

Estratégia metodológica voltada à construção e à avaliação rigorosa de artefatos destinados à resolução de problemas práticos relevantes. Distingue-se de abordagens meramente descritivas ao possuir caráter construtivo e prescritivo, equilibrando rigor científico e relevância organizacional. Adotada como método central desta pesquisa.

Embedding

Representação vetorial numérica de unidades textuais (palavras, sentenças ou documentos), gerada por modelos de linguagem a partir do processo de tokenização, que permite o processamento semântico das informações pelo algoritmo.

Engenharia de Requisitos Ágil

Área da Engenharia de Software dedicada à identificação, especificação, análise e validação de requisitos em contextos ágeis. Adapta as práticas tradicionais de engenharia de requisitos às características iterativas e colaborativas dos métodos ágeis, privilegiando a comunicação contínua e o refinamento progressivo dos artefatos.

Escala Likert

Instrumento psicométrico amplamente utilizado em pesquisas de percepção, no qual respondentes expressam seu nível de concordância ou discordância em relação a afirmações, geralmente em escala de 5 ou 7 pontos. Nesta pesquisa, utilizada para medir a percepção dos desenvolvedores sobre a qualidade do backlog ao longo dos ciclos analisados.

ETL — Extract, Transform and Load

Processo de integração de dados composto por três etapas: extração de dados de sistemas operacionais, transformação para garantir consistência e padronização semântica, e carregamento em estruturas analíticas, como data warehouses ou sistemas de *Business Intelligence*. Discutido nesta pesquisa no contexto de sistemas de gestão à vista e dashboards.

Feature (SAFe)

Nível intermediário da hierarquia de requisitos do SAFe, situado entre os *Epics* e as Histórias de Usuário. Reúne atributos textuais de maior complexidade, como descrição e contexto de negócio, critérios de aceite, métricas de sucesso e pontuação pelo método WSJF. Citada nesta pesquisa como possível extensão futura do método de avaliação de qualidade proposto.

Ajuste Fino (*Fine-tuning*)

Processo de refinamento adicional de um modelo de linguagem pré-treinado, por meio de treinamento em conjuntos de dados específicos de um domínio, com o objetivo de especializar seu vocabulário e sua aderência técnica a um contexto organizacional delimitado.

Física de Fábrica (*Factory Physics*)

Conjunto de princípios analíticos que descrevem o comportamento de sistemas produtivos em termos de fluxo, variabilidade, capacidade e estoque em processo. Provê fundamentos matemáticos para a gestão operacional, destacando-se a Lei de Little e o impacto da variabilidade na eficiência sistêmica.

GAP (lacuna)

No contexto desta pesquisa, refere-se a uma lacuna identificada na literatura científica ou na prática organizacional, que evidencia a ausência ou insuficiência de soluções existentes para um determinado problema. A identificação de GAPs motiva e justifica o desenvolvimento de novos artefatos na DSR.

Gherkin

Linguagem de especificação de comportamentos de software estruturada em torno de palavras-chave como *Given* (Dado), *When* (Quando) e *Then* (Então). Amplamente adotada em conjunto com o BDD para descrever critérios de aceitação de forma compreensível por desenvolvedores, testadores e especialistas de negócio.

Governança do *Product Backlog*

Conjunto de práticas, critérios e mecanismos institucionalizados que regulam a criação, o refinamento, a avaliação e a priorização dos itens do *Product Backlog*, com vistas a assegurar sua qualidade, rastreabilidade e alinhamento estratégico ao longo do ciclo de desenvolvimento.

GQM — *Goal-Question-Metric*

Método de definição de métricas orientado por objetivos, no qual indicadores mensuráveis são derivados de perguntas associadas a metas previamente estabelecidas. Citado na literatura revisada como abordagem para a definição de métricas de gestão do backlog.

História de Usuário (*User Story*)

Artefato textual conciso que descreve um requisito do produto a partir da perspectiva do usuário, estruturado no formato: "*Como [papel], quero [funcionalidade], para que*

[*benefício*]". Constitui a unidade básica de trabalho em metodologias ágeis como o Scrum e serve como mecanismo de comunicação entre equipes e *stakeholders*.

IEP — Indicador de Eficiência do Planejamento

Métrica desenvolvida nesta pesquisa para quantificar a eficiência do processo de planejamento ágil, relacionando a proporção de itens comprometidos efetivamente concluídos ao longo de um *Program Increment*. Valores próximos a 1,0 indicam alta previsibilidade e consistência no cumprimento dos compromissos de planejamento.

INVEST

Acrônimo que representa os critérios de qualidade desejáveis para histórias de usuário: **I**ndependent (Independente), **N**egotiable (Negociável), **V**aluable (Valiosa), **E**stimable (Estimável), **S**mall (Pequena) e **T**estable (Testável). Proposto por Bill Wake (2003) como guia para a elaboração de histórias de usuário de alta qualidade.

ISO/IEC 25010

Norma internacional que estabelece um modelo de características e subcaracterísticas de qualidade de software, incluindo confiabilidade, usabilidade, eficiência e manutenibilidade. Utilizada nesta pesquisa como referência para critérios técnicos complementares na avaliação de itens de backlog.

Kanban

Sistema de gestão visual do trabalho que utiliza quadros e cartões para representar o fluxo dos itens de trabalho por diferentes estágios do processo. Originou-se no Sistema Toyota de Produção e foi adaptado ao desenvolvimento de software. Fundamenta-se na limitação do WIP e na visualização do fluxo para identificar gargalos e promover a melhoria contínua.

Lead Time

Tempo total decorrido desde a entrada de um item no sistema de trabalho até a sua entrega final ao cliente ou usuário. Inclui tanto o tempo de espera quanto o tempo efetivo de processamento. É um indicador de desempenho de fluxo amplamente utilizado em contextos ágeis e na Física de Fábrica.

Lei de Little

Relação matemática fundamental da Física de Fábrica que estabelece que o tempo médio de atravessamento de um sistema (*lead time*) é igual ao estoque médio em processo (WIP) dividido pela taxa média de saída (*throughput*): $L = \lambda W$. Aplica-se tanto a sistemas industriais quanto ao fluxo de trabalho em desenvolvimento de software.

LLM — *Large Language Model*

Modelo de linguagem de larga escala treinado em grandes corpora textuais por meio de arquiteturas baseadas em *Transformers*. Possui elevada capacidade de compreensão semântica, geração e classificação de texto em linguagem natural. Nesta pesquisa, utilizado como motor de análise automatizada da qualidade dos itens de backlog.

Maturidade do Backlog

Grau de completude, clareza, consistência e verificabilidade dos itens que compõem o *Product Backlog*, avaliado em uma escala estruturada. Itens de maior maturidade apresentam descrição clara do valor de negócio, critérios de aceitação objetivos e condições de prontidão e conclusão explicitadas, contribuindo para a previsibilidade das entregas.

Métodos Ágeis

Conjunto de metodologias e frameworks para o desenvolvimento iterativo e incremental de software, caracterizados por ciclos curtos de entrega, colaboração contínua entre equipes e *stakeholders*, adaptação às mudanças e entrega de valor frequente. Entre os métodos mais difundidos estão o Scrum, o Kanban e o XP (*Extreme Programming*).

NLP — *Natural Language Processing* (Processamento de Linguagem Natural)

Campo da Inteligência Artificial dedicado ao desenvolvimento de técnicas e modelos capazes de interpretar, analisar e gerar linguagem humana de forma automatizada. Engloba tarefas como classificação de texto, extração de informações, análise de sentimento e geração de linguagem natural.

PI — *Program Increment*

Ciclo de planejamento e desenvolvimento do SAFe, geralmente com duração de 8 a 12 semanas, composto por um número fixo de *sprints*. Cada PI inicia-se com o *PI Planning* e encerra-se com uma cerimônia de *Inspect & Adapt*, promovendo o alinhamento e a sincronização entre todas as equipes do ART.

PI Planning

Cerimônia de planejamento em nível de programa no SAFe, realizada no início de cada *Program Increment*. Reúne todas as equipes do ART para definir objetivos, identificar dependências e comprometer-se com as entregas previstas para o ciclo. É considerado o principal evento de coordenação e alinhamento do SAFe.

Product Backlog

Lista ordenada e evolutiva que contém todos os requisitos, funcionalidades, melhorias

e correções necessárias ao produto. Gerenciada pelo *Product Owner*, constitui a única fonte de trabalho para as equipes Scrum e deve ser continuamente refinada para refletir mudanças nas necessidades do negócio e na visão do produto.

Product Owner (PO)

Papel ágil responsável por maximizar o valor do produto gerenciando e priorizando o *Product Backlog*. Representa os interesses dos *stakeholders* perante a equipe de desenvolvimento e é o principal responsável pela clareza e qualidade das descrições dos itens de backlog.

Prompt Engineering

Conjunto de técnicas e práticas para a formulação estratégica das instruções (*prompts*) fornecidas a modelos de linguagem, visando direcionar e padronizar as respostas geradas. A qualidade e a consistência dos resultados produzidos por LLMs são fortemente influenciadas pela precisão e estrutura dos prompts utilizados.

QUS — Quality User Story

Framework que define treze critérios de qualidade para histórias de usuário, organizados nas dimensões sintática, semântica e pragmática. Serve de base conceitual para ferramentas de avaliação automatizada, como o AQUISA, e é citado nesta pesquisa em conjunto com o INVEST como critério de reescrita de histórias de usuário por LLMs.

Requisito Ágil

Descrição de uma necessidade, funcionalidade ou capacidade que o produto deve possuir, expressa de forma concisa e orientada ao valor de negócio, em conformidade com os princípios dos métodos ágeis. Difere de especificações formais ao priorizar a comunicação e o entendimento compartilhado em detrimento do detalhamento exaustivo.

Retrospectiva

Cerimônia ágil realizada ao final de cada *sprint* ou *Program Increment*, na qual a equipe inspeciona seus processos, práticas e relacionamentos, identificando oportunidades de melhoria para os próximos ciclos. Constitui o principal mecanismo de aprendizado organizacional contínuo em contextos ágeis.

SAFe — Scaled Agile Framework

Framework de agilidade em escala que provê orientações para aplicar práticas Lean e ágeis em organizações de grande porte, estruturando o trabalho em múltiplos níveis hierárquicos: equipe, programa, portfólio e solução. Amplamente adotado em empresas que necessitam coordenar dezenas ou centenas de equipes ágeis.

Scrum

Framework leve para o desenvolvimento de produtos complexos, baseado em ciclos iterativos denominados *sprints*, papéis definidos (*Product Owner*, *Scrum Master*, Equipe de Desenvolvimento) e eventos estruturados (*Sprint Planning*, *Daily Scrum*, *Sprint Review*, Retrospectiva). Definido pelo *Scrum Guide* de Schwaber e Sutherland.

Sensemaking

Processo cognitivo e social por meio do qual indivíduos e equipes constroem coletivamente significado em situações de ambiguidade ou incerteza. No contexto desta pesquisa, refere-se à forma como o *Product Backlog* emerge da interpretação coletiva dos requisitos pelos membros das equipes, sendo por isso tratado como um *boundary object*.

SLM — *Small Language Model*

Modelo de linguagem de menor porte, com número reduzido de parâmetros, menor custo computacional e menor demanda de infraestrutura quando comparado aos LLMs. Empregado geralmente em tarefas específicas ou em ambientes com restrição de recursos, sendo apontado nesta pesquisa como direção futura para redução de custos e execução local do método.

Sprint

Ciclo de desenvolvimento de duração fixa — geralmente de uma a quatro semanas — no qual a equipe Scrum executa um conjunto selecionado de itens do *Product Backlog*, entregando ao final um incremento potencialmente utilizável do produto.

Sprint Planning

Cerimônia ágil realizada no início de cada *sprint* na qual a equipe seleciona os itens do *Product Backlog* a serem desenvolvidos no ciclo e planeja como o trabalho será executado. A qualidade dos itens selecionados — avaliada pelos critérios da DoR — é determinante para a eficiência desse evento.

Sprint Review

Evento realizado ao final de cada *sprint*, no qual a equipe apresenta o incremento desenvolvido a *stakeholders* para inspeção e coleta de feedback, subsidiando ajustes no *Product Backlog*.

Stakeholder

Qualquer indivíduo, grupo ou organização com interesse legítimo no produto, no processo de desenvolvimento ou nos resultados do projeto. Inclui usuários finais, clientes,

patrocinadores, gerentes e demais partes afetadas pelas decisões de desenvolvimento.

Story Points (SP)

Unidade relativa de estimativa ágil utilizada para expressar o esforço, a complexidade e a incerteza associados à implementação de uma história de usuário. Não corresponde diretamente a unidades de tempo, sendo calibrada pela equipe em relação a itens de referência conhecidos.

Throughput

Taxa de saída de um sistema produtivo, expressa pela quantidade de itens concluídos por unidade de tempo. Em contextos ágeis, corresponde à quantidade de histórias de usuário ou *story points* entregues por *sprint* ou por *Program Increment*. É um dos três componentes da Lei de Little.

Token / Tokenização

Unidade mínima de processamento textual (palavra, subpalavra ou caractere) resultante da segmentação do texto de entrada em modelos de linguagem. O consumo de tokens é também utilizado nesta pesquisa como parâmetro de mensuração dos custos operacionais associados ao uso de APIs comerciais de LLM.

Transformação Digital

Fenômeno organizacional que envolve a integração estratégica de tecnologias digitais nos processos, produtos e modelos de negócio de uma organização, impactando a criação de valor, a eficiência operacional e a cultura institucional. Constitui o contexto macro no qual a adoção de métodos ágeis e o desenvolvimento de produtos digitais se inserem.

Transformer

Arquitetura de rede neural profunda introduzida por Vaswani et al. (2017), baseada no mecanismo de atenção (*self-attention*), que revolucionou o campo do NLP e tornou possível o desenvolvimento de LLMs como GPT, BERT e SBERT. É a base tecnológica dos modelos de linguagem utilizados nesta pesquisa.

Velocidade (Velocity)

Métrica ágil que quantifica a quantidade de trabalho — geralmente expressa em *story points* — entregue por uma equipe ao longo de uma *sprint*. Utilizada para planejar capacidade e estimar o progresso futuro, a velocidade é influenciada diretamente pela qualidade dos itens do backlog e pela previsibilidade do processo.

WIP — *Work in Progress* (Trabalho em Processo)

Quantidade de itens de trabalho que se encontram simultaneamente em processamento em um sistema. Segundo a Lei de Little, o aumento não controlado do WIP eleva o *lead time* e reduz a previsibilidade das entregas. A limitação do WIP é um dos princípios centrais do Kanban e do pensamento Lean aplicado ao desenvolvimento de software.

WSJF — *Weighted Shortest Job First*

Método de priorização utilizado no SAFe para ordenar *Features* e *Epics* segundo a relação entre o valor de negócio esperado e o esforço/duração estimados. Um dos atributos textuais associados às *Features* na hierarquia de requisitos do SAFe.