



MUST UNIVERSITY  
MASTER OF SCIENCE IN BUSINESS DEVELOPMENT AND INNOVATION

LUCAS CAMPOS DE MAGALHÃES NUNES

**INTEGRAÇÃO RAG E MODEL CONTEXT PROTOCOL NA  
EXTENSÃO RURAL DIGITAL: UM AGENTE DE INTELIGÊNCIA  
ARTIFICIAL GENERATIVA PARA SUPORTE À RECUPERAÇÃO DE  
INFORMAÇÃO.**

FLORIDA — USA

Março/2026

**MUST UNIVERSITY**

70 SW 10th St, Deerfield Beach, FL 33441, USA | [info@mustedu.com](mailto:info@mustedu.com) | [www.mustedu.com](http://www.mustedu.com)

MUST University®: licensed by Florida Commission for Independent Education. License: 5593.

LUCAS CAMPOS DE MAGALHÃES NUNES

**INTEGRAÇÃO DE UMA ARQUITETURA RETRIEVAL-AUGMENTED  
GENERATION (RAG) AO MODEL CONTEXT PROTOCOL (MCP)  
PARA VIABILIZAR CONSULTAS CONVERSACIONAIS E  
APRIMORAR A RECUPERAÇÃO DE INFORMAÇÃO COM IA  
GENERATIVA NO ACERVO DE UMA PLATAFORMA PÚBLICA DE  
EXTENSÃO RURAL.**

Trabalho de Conclusão Final apresentado  
como requisito parcial para obtenção do título  
de Mestre no Curso de Master of Science in  
Business Development and Innovation da  
MUST University — Florida, USA.

Orientador: Prof. Dr. Nedisson Luis Gessi



FLORIDA — USA

Março/2026

**MUST UNIVERSITY**

70 SW 10th St, Deerfield Beach, FL 33441, USA | [info@mustedu.com](mailto:info@mustedu.com) | [www.mustedu.com](http://www.mustedu.com)

MUST University®: licensed by Florida Commission for Independent Education. License: 5593.

## **Dedicatória**

Dedico esta conquista primeiramente a Deus, pelo dom da vida.

Ao meu pai Renato e à minha mãe Suzana, por todo esforço que fizeram para que tivéssemos uma vida digna e de respeito, sempre com ética e amor ao próximo.

À minha amada esposa Laura, por todo o caminho que percorremos até aqui, com grande amor e respeito mútuo, além, claro, dos nossos dois filhos Davi e Mateus (*in memoriam*). Poder ser pai foi a maior transformação de toda a minha vida, e você me permitiu isso.

Ao meu filho Davi, com quem pude descobrir um outro tipo de amor: aquele sentimento sincero de pai para filho, mas também o sentido de uma grande responsabilidade.

Ao meu avô Pereira, por ter sido um grande sábio, pelos ensinamentos muitas vezes silenciosos em nossas caminhadas e por ser meu tutor espiritual enquanto esteve encarnado.

Aos meus irmãos Mariana, Pedro, Gustavo e Felipe, bem como à minha madrastra Lili, ao meu padrasto Ivo e a todos os meus familiares.

A todos os verdadeiros amigos presentes e àqueles que, embora nem sempre estejam ao nosso lado fisicamente, fazem-se sempre presentes em pensamentos.

## Lista de Figuras

Figura 01: Representação sistêmica da infraestrutura de governança Embrapa I/O .....	65
Figura 02: Escalas TRL e MRL aplicadas ao macroprocesso de ativos digitais .....	66
Figura 03: Modelos de Arquitetura (Boilerplates) para Integração de IA Generativa .....	67
Figura 04: Boilerplates disponíveis no Embrapa.oi para desenvolvimento com IA .....	68
Figura 05: Exemplo de Ecossistema de AaaS em PD&I .....	69
Figura 06: Exemplo de Ecossistema de AaaS institucionais .....	70
Figura 07: Página Ater+ Digital por Estados e Unidades da Embrapa .....	83
Figura 08: Exemplo de OCR em arquivo do corpus .....	84
Figura 09: Curadoria de documentos PDF .....	91
Figura 10: Trecho de código Javascript para “Validação de Integridade” de arquivos .....	92
Figura 11: Fluxo de inserção de arquivos PDF do corpus no Qdrant .....	94
Figura 12: Armazenamento das fontes por tema no Qdrant .....	95
Figura 13: Exemplo de trecho de conteúdo e os pontos correlacionados no Qdrant .....	96
Figura 14: Agente de IA orquestrando as chamados ao MCP por temas no n8n .....	98
Figura 15: Exemplo de uma chamada ao MCP server de Alho .....	99
Figura 16: Exemplo de uma consulta ao RAG de Alho (integrando IA e Qdrant) .....	99
Figura 17: Exemplo de uma pergunta e resposta sobre o tema Alho .....	100
Figura 18: Metodologia da análise e interpretação dos resultados .....	104
Figura 19: Google NotebookLM sobre Alho .....	105
Figura 20: Trecho da planilha com a resposta feita pelo agente de IA .....	106
Figura 21: Pipeline automatizado de auditoria do LLM-as-a-Judge .....	107
Figura 22: Prompt do Agente com critérios de Enriquecimento e formato de saída .....	109
Figura 23: Exemplo de análise e categorização do erro feito pelo “Agente Juiz” .....	112
Figura 24: Percurso metodológico para construção do “Agente Digital Ater+ Digital” .....	118
Figura 25: Captura de tela da página inicial do Ater+ Digital com o chatbot (protótipo) ...	123
Figura 26: Agente orquestrador no n8n com as respectivas chamadas para os MCPs .....	124
Figura 27: Consultor Digital recomendando ferramentas e apps disponíveis .....	125

Figura 28: Distribuição da Latência Operacional (Tempo de Resposta) do Agente .....	129
Figura 29: Ranqueamento de Qualidade por Domínio de Conhecimento .....	137
Figura 30: Variabilidade da Qualidade (Boxplot) por Domínio de Conhecimento .....	141
Figura 31: Variabilidade da Qualidade das Respostas por Perfil de Usuário .....	143
Figura 32: Dispersão entre Nota de Qualidade e Latência Operacional (Dados Brutos) .....	145
Figura 33: Matriz de Correlação de Spearman (Qualidade vs. Latência) .....	146
Figura 34: Mapa de Eficiência-Qualidade-Tempo por Domínio de Conhecimento .....	150

## Lista de Quadros

Quadro 01: Fundamentação da escolha metodológica .....	78
Quadro 02: Mapeamento e Estruturação do Corpus da Plataforma Ater+ Digital .....	81
Quadro 03: Distribuição e Volumetria do Corpus Documental (PDFs) Processado .....	86
Quadro 04: Critérios metodológicos de composição do corpus documental .....	87
Quadro 05: Inventário e Caracterização do Corpus Documental Processado .....	89
Quadro 06: Sistematização do Corpus e Categorização Temática .....	103
Quadro 07: Escala de Avaliação Likert do “Agente Juiz” .....	110
Quadro 08: Síntese dos Indicadores de Desempenho e Estabilidade .....	128
Quadro 09: Extremos de Desempenho Computacional e Qualidade por Domínio .....	130
Quadro 10: Métricas Descritivas de Qualidade por Domínio de Conhecimento .....	137
Quadro 11: Resumo dos Testes Estatísticos de Hipótese e Significância .....	147

## Lista de Abreviaturas e Siglas

AI — Artificial Intelligence (Inteligência Artificial)

API — Application Programming Interface (Interface de Programação de Aplicação)

ATER — Assistência Técnica e Extensão Rural

CAPES — Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CGI — Comitê Gestor da Internet

CNPq — Conselho Nacional de Desenvolvimento Científico e Tecnológico

CSV — Comma-Separated Values (Valores Separados por Vírgula)

EMBRAPA — Empresa Brasileira de Pesquisa Agropecuária

GCP — Google Cloud Platform

IA — Inteligência Artificial

IAGen — Inteligência Artificial Generativa

ICT — Instituição de Ciência e Tecnologia

JSON — JavaScript Object Notation

LLM — Large Language Model (Grande Modelo de Linguagem)

MCP — Model Context Protocol

MRL — Manufacturing Readiness Level

MVP — Minimum Viable Product (Produto Mínimo Viável)

NLTK — Natural Language Toolkit

PDI — Pesquisa, Desenvolvimento e Inovação

PDF — Portable Document Format

PoC — Proof of Concept (Prova de Conceito)

RAG — Retrieval-Augmented Generation (Geração Aumentada por Recuperação)

SaaS — Software as a Service (Software como Serviço)



SDK — Software Development Kit (Kit de Desenvolvimento de Software)

SLA — Service Level Agreement (Acordo de Nível de Serviço)

TCF — Trabalho de Conclusão Final

TRL — Technology Readiness Level (Nível de Maturidade Tecnológica)

UI — User Interface (Interface do Usuário)

## Resumo

Este trabalho apresenta o desenvolvimento e a avaliação do "Consultor Online Ater+ Digital", um artefato de Inteligência Artificial Generativa projetado para otimizar o acesso ao acervo técnico-científico da Embrapa no ecossistema da plataforma Ater+ Digital. A problemática da pesquisa fundamenta-se na complexidade de recuperação de informações em bases documentais densas por meio de métodos convencionais. Fundamentada no arcabouço da metodologia *Design Science Research* (DSR), a pesquisa propõe uma arquitetura que integra *Retrieval-Augmented Generation* (RAG) ao *Model Context Protocol* (MCP). A avaliação do artefato contemplou 4.600 interações distribuídas em 24 cadeias produtivas e temas transversais, voltadas para a extensão rural. As evidências de performance demonstraram uma taxa de sucesso de 100% e latência média de 6,45 segundos. A qualidade das respostas, auditada por um modelo de linguagem independente (*LLM-as-a-Judge*), atingiu mediana máxima (5,0) na escala Likert. Análises inferenciais indicaram diferença estatisticamente significativa, porém de magnitude desprezível, entre os perfis de agricultores e extensionistas ( $r = 0,0384$ ), bem como uma associação monotônica positiva, fraca, entre o tempo de processamento e a acurácia técnica ( $\rho = 0,1535$ ;  $p < 0,001$ ). Os achados indicam que a integração RAG-MCP contribuiu para mitigar alucinações e favorecer a soberania do conhecimento institucional, posicionando o artefato no nível de prontidão tecnológica e potencial uso ao público externo. A pesquisa fomenta a transformação digital da extensão rural, oferecendo uma solução robusta para democratizar o suporte à tomada de decisão no campo.

**Palavras-chave:** Inteligência Artificial Generativa, RAG, Model Context Protocol, Embrapa, Extensão Rural

## **Abstract**

*This study presents the development and evaluation of the “Ater+ Digital Online Advisor,” a Generative Artificial Intelligence artifact designed to optimize access to Embrapa’s technical-scientific collection within the Ater+ Digital platform ecosystem. The research problem is grounded in the complexity of retrieving information from dense documentary databases through conventional methods. Based on the Design Science Research (DSR) methodological framework, the study proposes an architecture that integrates Retrieval-Augmented Generation (RAG) with the Model Context Protocol (MCP). The artifact evaluation comprised 4,600 interactions distributed across 24 production chains and cross-cutting themes related to rural extension. Performance evidence demonstrated a 100% success rate and an average latency of 6.45 seconds. The quality of the responses, audited by an independent language model (LLM-as-a-Judge), reached the maximum median (5.0) on the Likert scale. Inferential analyses indicated a statistically significant but negligible difference between the profiles of farmers and extension agents ( $r = 0.0384$ ), as well as a weak positive monotonic association between processing time and technical accuracy ( $\rho = 0.1535$ ;  $p < 0.001$ ). The findings indicate that the RAG-MCP integration helped mitigate hallucinations and strengthen institutional knowledge sovereignty, positioning the artifact at a level of technological readiness and potential use for the external public. The study fosters the digital transformation of rural extension by offering a robust solution to democratize decision-support in the field.*

**Keywords:** *Generative Artificial Intelligence, RAG, Model Context Protocol, Embrapa. Rural Extension*

## Sumário

<b>1. Introdução</b>	14
<b>1.1 Contextualização e Relevância do Tema</b>	17
<b>1.2 Problema de Pesquisa e Objetivos</b>	20
<b>1.3 Justificativa e Contribuições do Estudo</b>	23
<b>1.4 Estrutura do Trabalho</b>	26
<b>2. Fundamentação Teórica</b>	28
<b>2.1 Contexto do Domínio e do Problema</b>	29
<b>2.1.1 Extensão rural e evolução para o digital</b>	29
<b>2.1.2 Embrapa, gestão do conhecimento e a plataforma Ater+ Digital</b>	33
<b>2.1.3 Necessidades de recuperação da informação na extensão rural digital</b>	38
<b>2.2 Fundamentos de Inteligência Artificial Aplicados ao Estudo</b>	41
<b>2.2.1 Inteligência artificial</b>	42
<b>2.2.2 Grandes Modelos de Linguagem: capacidades, limites e alucinações</b>	45
<b>2.2.3 Agentes conversacionais (chatbots)</b>	48
<b>2.3 Ferramentas e Tecnologias Habilitadoras</b>	52
<b>2.3.1 Recuperação Semântica e Embeddings</b>	53
<b>2.3.2 Bancos de dados vetoriais</b>	54
<b>2.3.3 Retrieval-Augmented Generation (RAG)</b>	57
<b>2.3.4 Orquestração e integração de serviços</b>	59
<b>2.3.5 Interoperabilidade e governança</b>	62
<b>2.4 Ambiente Institucional de Desenvolvimento e Implantação</b>	64
<b>2.4.1 Embrapa.io e a governança de ativos digitais</b>	65
<b>2.4.2 Arquitetura de referência para soluções de inteligência artificial</b>	69
<b>2.4.3 Aplicabilidade da arquitetura proposta ao artefato desenvolvido</b>	72
<b>3. Metodologia</b>	73
<b>3.1 Classificação e Natureza da Pesquisa</b>	77
<b>3.2 Delimitação do Estudo e Corpus Documental</b>	80
<b>3.3 Procedimentos Metodológicos e Desenvolvimento do Artefato</b>	91
<b>3.3.1 Extração, tratamento e segmentação dos dados</b>	92
<b>3.3.2 Vetorização (Embeddings) e indexação no Qdrant</b>	94
<b>3.3.3 Orquestração e protocolo MCP</b>	98
<b>3.4 Procedimentos de Análise e Interpretação dos Resultados</b>	103

<b>3.4.1 Critérios de julgamento e engenharia de prompt</b> .....	109
<b>3.4.2 Escala de avaliação e tratamento de alucinações</b> .....	111
<b>3.4.3 Procedimentos de tratamento estatístico dos dados</b> .....	115
<b>3.5 Síntese Operacional da Metodologia</b> .....	119
<b>4. Resultados e Discussão</b> .....	122
<b>4.1 Apresentação do Artefato</b> .....	124
<b>4.2. Análise de Desempenho Computacional (Performance)</b> .....	129
<b>4.3 Avaliação da Qualidade das Respostas (LLM-as-a-Judge)</b> .....	137
<b>4.4 Análise Inferencial e Significância Estatística</b> .....	144
<b>4.5 Discussão dos Achados: Qualidade-Tempo-Confiança</b> .....	150
<b>5. Considerações Finais e Conclusão</b> .....	155
<b>5.1 Contribuições Teóricas e Práticas</b> .....	156
<b>5.2 Limitações do Estudo</b> .....	157
<b>5.3 Integração dos Procedimentos e Reflexões Metodológicas</b> .....	160
<b>5.4 Agenda para Pesquisas Futuras</b> .....	161
<b>5.5 Considerações Institucionais e Síntese Final</b> .....	164
<b>6. Referências</b> .....	166

## 1. Introdução

A agricultura contemporânea atravessa um período de ampla reestruturação tecnológica, frequentemente denominada Agricultura 4.0. Este novo paradigma caracteriza-se pela integração massiva de recursos digitais, automação e análise de dados aos processos de gestão rural, com o intuito de otimizar a eficiência produtiva e a sustentabilidade ambiental (Embrapa Agricultura Digital, 2025).

Nesse cenário, a informação técnica validada torna-se um insumo tão imprescindível quanto a terra ou o capital, exigindo novos mecanismos de mediação entre os centros de pesquisa e os produtores rurais. A complexidade crescente dos sistemas de produção demanda que o conhecimento científico seja não apenas acessível, mas operacionalizável e contextualizado às realidades locais (Zuin et al., 2022).

Historicamente, o serviço de Assistência Técnica e Extensão Rural (ATER) desempenhou o papel primordial de transpor o conhecimento acadêmico para a realidade produtiva. No entanto, as limitações orçamentárias e logísticas do modelo convencional presencial impulsionaram a emergência da ATER Digital. Essa modalidade não objetiva substituir o extensionista, mas complementar sua atuação por meio de Tecnologias de Informação e Comunicação (TICs), permitindo expandir a capilaridade do atendimento e democratizar o acesso à informação em territórios vastos e heterogêneos como o brasileiro (Deus et al., 2024; Zuin et al., 2022).

No Brasil, a institucionalização dessa estratégia é consolidada pela Plataforma Ater+ Digital, uma iniciativa liderada pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) em cooperação com a Agência Nacional de Assistência Técnica e Extensão Rural (Anater) e parceiros. Estabelecida para subsidiar as redes de assistência técnica em âmbito nacional, a plataforma foi estruturada para prover um vasto acervo de vídeos, áudios, aplicativos e cursos, organizados para viabilizar tanto ações remotas quanto presenciais (Embrapa, 2024).

A estrutura inicial da plataforma, consolidada em abril de 2024, organizou o conhecimento em sete eixos temáticos principais: Apicultura, Caprinos e Ovinos, Feijão, Feijão-Caupi, Mudanças Climáticas, Nutrição e Saúde, e Sistemas Agroflorestais. Essa organização por cadeias produtivas objetivou fornecer informações em linguagem concisa e inteligível, validadas por corpos técnicos especializados, a fim de mitigar a escassez de conteúdo técnico fidedigno e de célere compartilhamento para os extensionistas (Embrapa, 2024).

A evolução da plataforma tem sido constante, refletindo a dinâmica do setor. Em maio de 2025, novos conteúdos foram incorporados para expandir o escopo de atendimento, incluindo quatro novos eixos: Mandioca, Manga, Peixes e Sistemas Agrícolas Tradicionais (SATs).

Além da expansão temática, instituiu-se uma página dedicada a "Ferramentas Digitais", agregando aplicativos e portais de gestão rural para acesso célere (Embrapa, s.d.-b). Esse crescimento contínuo do acervo, embora favorável, intensifica o desafio de navegar por um volume massivo de documentação técnica heterogênea.

A mera digitalização e oferta de documentos não asseguram, primordialmente, a apropriação do conhecimento. À medida que novos eixos e ferramentas são incorporados, os mecanismos convencionais de busca e a navegação por menus hierárquicos tornam-se insuficientes para sanar dúvidas complexas do cotidiano rural. O extensionista, frequentemente em campo e com tempo restrito, demanda respostas diretas e contextualizadas, em detrimento de listas extensas de documentos que exijam leitura e triagem manual.

O avanço da Inteligência Artificial (IA) descortina novas perspectivas para superar essa barreira de usabilidade. Ferramentas baseadas em Processamento de Linguagem Natural (PLN), como chatbots e assistentes virtuais, têm demonstrado potencial para automatizar o suporte e prover orientações em tempo real (Souza et al., 2025). Experiências singulares na agricultura,

a exemplo de sistemas para auxílio na incubação de aves ou no diagnóstico de fitossanidade, indicam que interfaces conversacionais podem reduzir significativamente a assimetria de informações (Oliveira Filho et al., 2021).

Entretanto, observa-se uma escassez de estudos que articulem, de forma sistemática, a aplicação de Modelos de Linguagem de Grande Escala (Large Language Models – LLMs) sobre bases de conhecimento institucionais robustas, como a da Embrapa.

Embora o interesse pela utilização de IA no agronegócio brasileiro tenha progredido aceleradamente, remanesce uma carência de projetos de aplicabilidade prática que transcendam a fase de prova de conceito, especialmente no contexto da extensão rural pública, em que a precisão da informação e a soberania dos dados são mandatórias (Nordheim et al., 2023).

Para preencher essa lacuna, a arquitetura *Retrieval-Augmented Generation* (RAG) apresenta-se como uma solução técnica promissora. Ao vincular a fluidez dos modelos generativos a bases de dados verificadas, torna-se exequível mitigar as "alucinações" inerentes às IAs genéricas (Google, 2026).

A integração dessa arquitetura por intermédio de protocolos de interoperabilidade, a exemplo do *Model Context Protocol* (MCP), viabiliza a estruturação de agentes capazes de processar os manuais, conteúdos e notas técnicas da Ater+ Digital e interagir com o usuário, transmutando o repositório estático em um consultor ativo.

Embora iniciativas de sistemas conversacionais voltados à extensão rural já tenham sido registradas na literatura — como o Poulbot (Oliveira Filho et al., 2021) e o *Virtual Agronomist* (Shepherd et al., 2025) —, tais soluções operam predominantemente sobre bases de conhecimento genéricas ou de escopo restrito, sem mecanismos formais de ancoragem ao acervo institucional de organizações públicas de pesquisa. Da mesma forma, estudos sobre arquiteturas RAG aplicadas ao setor público brasileiro ainda se concentram em domínios isolados, como saúde mental (Netto & Brito, 2025) e processos legislativos (Lima, 2025), sem

contemplar a complexidade multitemática e a escala documental característica de plataformas de extensão rural. Configura-se, portanto, uma lacuna científica quanto à integração do *Model Context Protocol* (MCP) como camada de orquestração agêntica orientada à soberania informacional institucional — configuração que o presente estudo se propõe a desenvolver e validar empiricamente.

Diante desse cenário, o presente estudo propõe o desenvolvimento de um agente conversacional, com o intuito de não apenas teorizar sobre a IA na agricultura, mas construir e validar um artefato funcional que otimize a recuperação de informações na Plataforma Ater+ Digital, por intermédio de IA generativa.

A pesquisa justifica-se pela premente necessidade de modernizar os instrumentos de ATER, assegurando que o vasto conhecimento gerado pela pesquisa agropecuária alcance, efetivamente, os tomadores de decisão no campo (Hevner et al., 2004; Embrapa, 2025a).

### **1.1 Contextualização e Relevância do Tema**

A modernização da agricultura brasileira tem sido historicamente marcada pela dualidade entre a elevada incorporação tecnológica em *commodities* de exportação e os desafios de inclusão produtiva na agricultura familiar.

Nesse cenário, a Agricultura 4.0 impõe uma nova camada de complexidade, em que a posse de dados e a capacidade de processar informações tornam-se determinantes para a competitividade e sustentabilidade (Vieira Filho et al., 2016). No entanto, essas inovações não se difundem espontaneamente entre os atores envolvidos; tal processo depende de sistemas robustos de Assistência Técnica e Extensão Rural (ATER) capazes de transpor o conhecimento científico em práticas agronômicas aplicáveis.

A relevância da ATER pública é evidenciada pela lacuna de cobertura no campo. Dados do Censo Agropecuário indicam que uma parcela minoritária dos estabelecimentos rurais

brasileiros recebe orientação técnica regular, o que perpetua desigualdades regionais e produtivas (IPEA, 2022). Para mitigar essa limitação, governos e instituições têm buscado na digitalização uma estratégia para expandir o alcance das políticas públicas.

A ATER Digital surge, nesse cenário, não apenas como um recurso tecnológico, mas como um novo instrumento de ação pública, com o intuito de complementar o atendimento presencial e transpor barreiras logísticas em um país de dimensões continentais (Pacífico, 2022; FAO, 2023).

No âmbito estadual e federal, iniciativas como o IDR-Paraná Digital e, centralmente, a Plataforma Ater+ Digital da Embrapa, materializam essa estratégia. Essas plataformas operam como amplos repositórios de conhecimento, organizando notas técnicas, cartilhas e cursos em eixos temáticos (IDR-Paraná, n.d.; Embrapa, 2024).

Contudo, a mera transposição de conteúdos analógicos para o ambiente virtual fomenta o fenômeno da sobrecarga cognitiva. Extensionistas relatam que, embora a informação esteja disponível, o tempo necessário para localizar, filtrar e interpretar documentos dispersos pode inviabilizar seu emprego tempestivo durante o atendimento ao produtor (Santos, 2023).

A emergência da Inteligência Artificial (IA) descortina caminhos para reestruturar essa dinâmica de acesso à informação. Estudos internacionais, a exemplo do *Virtual Agronomist* na África, demonstram que sistemas conversacionais podem atuar como intermediários eficazes, provendo suporte à decisão em tempo real para o manejo de culturas (Shepherd et al., 2025).

Da mesma forma, pesquisas recentes na Índia com o sistema *Agri Assist* indicam que a integração de IA na rotina agrícola pode otimizar o emprego de insumos e aperfeiçoar a gestão da propriedade, desde que a interface seja inteligível e funcional (Reddy et al., 2025).

No Brasil, experiências piloto, a exemplo do emprego de *chatbots* na cadeia de frangos de corte (Poulbot) e na produção orgânica no Rio Grande do Norte, sinalizam o potencial dessas ferramentas para democratizar o conhecimento técnico (Oliveira Filho et al., 2021; Souza et al.,

2025). Entretanto, a adoção massiva de *chatbots* na extensão rural pública confronta o desafio da fidedignidade. Revisões sistemáticas sobre a interação humano-computador alertam que a confiança nos sistemas de IA é exígua e depende diretamente da transparência e da exatidão das respostas fornecidas (Nordheim et al., 2023).

Tal aspecto é crítico no que tange aos Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs). Embora capazes de produzir textos fluentes, esses modelos, quando desprovidos de mecanismos de controle, podem gerar informações imprecisas ("alucinações") que, no contexto agrícola, acarretariam prejuízos econômicos ou ambientais (Embrapa Agricultura Digital, 2025). Por conseguinte, a relevância deste estudo reside na proposição de uma arquitetura tecnológica — fundamentada em *Retrieval-Augmented Generation* (RAG) — que restrinja a base de conhecimento do *chatbot* exclusivamente aos documentos técnicos validados pela pesquisa agropecuária, assegurando a soberania e a segurança da informação.

Além do aspecto técnico, a pesquisa possui relevância social ao investigar como “agentes digitais sociais” podem ser integrados às organizações de trabalho. A inserção de um assistente virtual na rotina da extensão rural transcende uma simples atualização de *software*, configurando um processo de maturação organizacional que reestrutura a mediação do conhecimento (Zunino, 2022).

Ao viabilizar o acesso célere a recomendações sobre fitossanidade e práticas culturais, a ferramenta proposta provê autonomia ao extensionista, permitindo que este dedique maior tempo ao diagnóstico e ao relacionamento interpessoal com o produtor, reduzindo a carga despendida na busca processual por informações (Lima, 2024).

O estudo também preenche uma lacuna na literatura nacional ao aplicar a metodologia de *Design Science Research* (DSR) na construção de artefatos de IA no setor público. Enquanto diversas investigações focam na análise teórica da digitalização ou em modelos preditivos

isolados, este trabalho propõe a construção e a validação de um artefato funcional, enfrentando os desafios reais de interoperabilidade e curadoria de dados em uma organização pública de grande porte (Hevner et al., 2004; Santos & Zonin, 2024).

Verifica-se, por conseguinte, que a temática é relevante e oportuna, pois articula a fronteira tecnológica (IA Generativa e RAG) com uma demanda latente do Estado brasileiro: a modernização da ATER para assegurar a segurança alimentar e a sustentabilidade. A proposição de um assistente virtual contextualizado, alinhado às diretrizes da FAO e às políticas nacionais, representa um avanço concreto para transmutar o vasto acervo da pesquisa agropecuária em valor tangível para a sociedade.

A revisão do estado da arte evidencia que, embora o campo da Inteligência Artificial aplicada à agricultura venha avançando consistentemente, persiste um vazio específico na confluência entre três dimensões: a orquestração agêntica de múltiplos domínios temáticos, a soberania institucional dos dados e a validação empírica em larga escala em contexto de extensão rural pública. Experiências como o *Agri Assist* (Reddy et al., 2025) e iniciativas nacionais com RAG (Borges, 2025; Lima, 2025) demonstram a viabilidade técnica da abordagem, mas não operam sobre a heterogeneidade documental e a envergadura institucional características da Embrapa. É precisamente nesse interstício — entre o potencial demonstrado pela literatura e a ausência de soluções validadas para contextos institucionais públicos multitemáticos — que o presente estudo se posiciona, propondo uma arquitetura que integra RAG ao MCP como resposta concreta a essa demanda não atendida.

## **1.2 Problema de Pesquisa e Objetivos**

A expansão dos serviços de extensão rural digital, materializada na Plataforma Ater+ Digital, resultou na consolidação de um vasto repositório de conhecimento técnico. A estruturação inicial em sete eixos e a subsequente adição de novos temas e ferramentas digitais

criaram um ambiente rico em informações validadas para o suporte à produção agropecuária (Embrapa, 2024; Embrapa, 2025a). No entanto, o crescimento contínuo desse acervo impõe desafios relacionados à arquitetura da informação e à recuperabilidade dos dados.

A literatura aponta que, em amplas bases de conhecimento governamentais, a navegação fundamentada exclusivamente em menus hierárquicos ou em busca por palavras-chave tende a desencadear sobrecarga cognitiva, exigindo do usuário um esforço excessivo de triagem e síntese (Souza et al., 2025).

Para o extensionista, cuja atuação frequentemente ocorre em condições de campo e sob restrições temporais, a celeridade no acesso à informação é crítica. Estudos sobre a percepção dos agentes de ATER indicam que a tecnologia deve atuar como um facilitador imediato da tomada de decisão, em detrimento de uma barreira burocrática adicional (Santos, 2023).

O modelo vigente de busca da plataforma, ao retornar listas de documentos ou *links* genéricos, frequentemente falha em prover respostas contextualizadas às demandas específicas do cotidiano rural. Essa fricção na recuperação da informação pode ensejar a subutilização do conhecimento científico disponível, corroborando a assimetria informacional que a própria ATER Digital objetiva mitigar (Deus et al., 2024).

Adicionalmente, a emergência de ferramentas de Inteligência Artificial (IA) Generativa no setor agrícola suscita questionamentos acerca da fidedignidade. Embora *chatbots* tenham demonstrado potencial para democratizar o acesso à informação, revisões sistemáticas alertam que a utilidade dessas ferramentas depende diretamente da exatidão técnica e da transparência das fontes utilizadas (Nordheim et al., 2023). Modelos genéricos, desconectados de bases de dados institucionais controladas, apresentam riscos elevados de alucinação e imprecisão conceitual, o que é mandatório mitigar no contexto da extensão rural pública (Embrapa Agricultura Digital, 2025).

Diante desse cenário, o problema de pesquisa delinea-se a partir da necessidade de converter o repositório passivo da Embrapa em um sistema ativo de suporte. A questão central que orienta este estudo é: Como desenvolver um agente conversacional utilizando Inteligência Artificial Generativa e arquitetura *Retrieval-Augmented Generation* (RAG) para otimizar a recuperação de informações e o suporte técnico na Plataforma Ater+ Digital?

Para responder a essa indagação, o estudo estabelece objetivos que alinham a construção do artefato tecnológico aos princípios da *Design Science Research* (Hevner et al., 2004). O objetivo geral consiste em desenvolver e validar um protótipo de agente conversacional (*chatbot*) fundamentado em *Retrieval-Augmented Generation* (RAG) e integrado ao MCP, visando otimizar a recuperação e a disponibilização de informações técnicas do acervo da Plataforma Ater+ Digital.

Para a consecução dessa meta, os objetivos específicos compreendem, primeiramente, a preparação e estruturação do *corpus* documental da plataforma, assegurando a integridade dos dados e sua conversão em base vetorial para viabilizar a recuperação semântica. Subsequentemente, projeta-se a arquitetura do agente mediante a plataforma n8n para a orquestração de fluxos e integração com Modelos de Linguagem de Grande Escala (LLMs), garantindo a interoperabilidade e a soberania dos dados (Embrapa for Developers, n.d.).

O percurso metodológico prevê ainda a implementação da arquitetura RAG integrada ao MCP para a geração de respostas ancoradas em fontes oficiais, seguida pela avaliação do desempenho em testes controlados que verifiquem a fidedignidade, a acurácia semântica, a latência e a estabilidade frente aos mecanismos convencionais de busca. O estudo propõe diretrizes e recomendações de aperfeiçoamento fundamentadas nos resultados da avaliação, com o intuito de elevar a efetividade da recuperação da informação e a fidedignidade às fontes institucionais.

Tais objetivos articulam-se com a metodologia empregada, caracterizada como pesquisa aplicada de natureza tecnológica, que orienta a produção de conhecimento voltada à solução de problemas concretos identificados no ambiente organizacional.

A investigação estrutura-se para transcender a análise teórica, concentrando-se no desenvolvimento e na validação de um artefato funcional — o agente conversacional — capaz de intervir na realidade da extensão rural. Assim, o percurso metodológico delineado assegura que cada etapa, da curadoria de dados à avaliação de desempenho, contribua de forma efetiva para a otimização dos processos de recuperação da informação na Plataforma Ater+ Digital, garantindo a utilidade e a aderência da ferramenta ao contexto dos usuários finais.

### **1.3 Justificativa e Contribuições do Estudo**

A justificativa para a realização deste estudo ancora-se, primordialmente, na necessidade premente de modernizar os instrumentos de mediação de conhecimento na extensão rural pública brasileira. A transição para a agricultura digital não deve restringir-se à mecanização ou à utilização de sensores no campo; ela deve abranger, com igual rigor, a gestão do conhecimento que orienta a tomada de decisão.

O modelo vigente de disponibilização de informações na Plataforma Ater+ Digital, embora robusto em conteúdo, reproduz a lógica de repositórios estáticos, transferindo para o extensionista o ônus da busca e da síntese documental (Deus et al., 2024). Diante da limitação temporal e da multiplicidade de demandas que caracterizam o trabalho de campo, a ausência de ferramentas céleres de consulta pode resultar na subutilização do acervo técnico da Embrapa, comprometendo a eficácia da política pública de ATER (Santos, 2023).

Sob a ótica tecnológica, o estudo justifica-se pela oportunidade de mitigar os riscos associados à adoção de Inteligência Artificial (IA) no setor público, especificamente na extensão rural. Enquanto Modelos de Linguagem de Grande Escala (*Large Language Models*

– LLMs) ganham popularidade, sua aplicação direta na agricultura apresenta riscos relevantes devido à propensão a "alucinações" e à ausência de rastreabilidade das fontes, o que infringe princípios éticos e regulatórios institucionais (Embrapa Agricultura Digital, 2025; Nordheim et al., 2023).

A proposição de uma arquitetura *Retrieval-Augmented Generation* (RAG) responde a esse desafio técnico, favorecendo a implementação da inovação com soberania de dados; isto é, orientando a geração de respostas a partir de conhecimento científico validado pela pesquisa agropecuária nacional. (Embrapa Agricultura Digital, 2025).

A pesquisa alinha-se, ademais, às diretrizes estratégicas para o fortalecimento da ATER Digital, conforme preconizado por organismos internacionais, a exemplo da FAO, e por políticas nacionais de inovação. A capacidade de personalizar o atendimento em larga escala, provendo respostas contextualizadas para distintas cadeias produtivas — da bovinocultura à viticultura —, representa um avanço qualitativo na prestação do serviço (FAO, 2023; Embrapa, 2024).

Ao desenvolver um artefato capaz de processar linguagem natural e fornecer orientações precisas, o projeto contribui para mitigar a lacuna de cobertura da assistência técnica, permitindo que o extensionista atenda a um maior contingente de produtores com superior assertividade; concomitantemente, viabiliza que o produtor rural acesse informações fidedignas mais próximas à sua linguagem e ao seu contexto.

No que tange às Contribuições Teóricas, este trabalho colabora para o avanço do conhecimento em Sistemas de Informação aplicados ao agronegócio, especificamente na interseção entre Inteligência Artificial e Gestão do Conhecimento. A literatura contemporânea carece de investigações que documentem a implementação empírica de arquiteturas *Retrieval-Augmented Generation* (RAG) em contextos de elevada especificidade técnica e vocabulário controlado, como se observa na agronomia tropical. Ao sistematizar o processo de curadoria de

dados e a integração por intermédio do MCP, este estudo oferece um modelo de referência para futuras iniciativas de digitalização em organizações de Pesquisa e Desenvolvimento (P&D) (Hevner et al., 2004; Embrapa for Developers, n.d.).

As Contribuições Práticas materializam-se na entrega do protótipo funcional do agente conversacional, outrora mencionado. Para a Embrapa e seus parceiros, o artefato representa uma ferramenta de "usabilidade do conhecimento", transmutando manuais densos em diálogos acionáveis.

Tal implementação impacta diretamente a rotina operacional de extensionistas e produtores rurais, reduzindo o tempo despendido em pesquisas bibliográficas e ampliando a fidedignidade técnica das recomendações (Lima, 2024). Ademais, a arquitetura proposta, fundamentada em padrões abertos, favorece a escalabilidade e a manutenção evolutiva da solução no âmbito do ecossistema de inovação da empresa.

Do ponto de vista das Contribuições Sociais, o estudo fomenta a democratização do acervo científico. Ao facilitar o acesso à informação qualificada, a ferramenta provê autonomia aos agentes de ATER que atuam na ponta, especialmente em regiões desprovidas de especialistas temáticos presenciais. Esse suporte tecnológico especializado reflete-se no aprimoramento das práticas produtivas adotadas pelos agricultores familiares, potencializando a segurança alimentar, a geração de renda e a sustentabilidade ambiental — objetivos basilares de qualquer intervenção de extensão rural (Zuin et al., 2022; IPEA, 2022).

Em vez de limitar-se a diagnosticar limitações e vulnerabilidades na Plataforma Ater+ Digital, o estudo adota uma postura prescritiva e propositiva, validando o artefato por intermédio de ciclos rigorosos de *design*, desenvolvimento e avaliação. Tal abordagem reitera o papel da academia não meramente como observadora, mas como agente de inovação tecnológica (Hevner et al., 2004).

O estudo apresenta, outrossim, relevância estratégica ao explorar a interoperabilidade de sistemas por intermédio do MCP. Em um cenário em que a fragmentação de ferramentas digitais é recorrente, a validação de um protocolo que viabiliza a conexão do assistente a distintas fontes de dados e funcionalidades — a exemplo de planilhas de custos ou sistemas agrometeorológicos — indica uma tendência de serviços integrados na infraestrutura Embrapa I/O. Tal abordagem evita a formação de silos de informação e fomenta uma experiência de usuário mais fluida e coesa (Embrapa for Developers, n.d.).

Os achados corroboram que a investigação contribui para a consolidação de uma cultura de confiança digital na agricultura. Ao priorizar a transparência das fontes e a exatidão técnica em detrimento da mera fluência conversacional, o projeto estabelece parâmetros de qualidade para o emprego de IA na extensão rural.

Essa postura é fundamental para mitigar a resistência de usuários e gestores quanto à adoção de tecnologias emergentes, pavimentando o caminho para que a Plataforma Ater+ Digital se consolide como referência na prestação de suporte ao campo (Nordheim et al., 2023; Embrapa, 2025a).

## **1.4 Estrutura do Trabalho**

A estrutura deste Trabalho de Conclusão Final (TCF) foi delineada para assegurar a consistência interna e uma progressão lógica entre a delimitação do problema, a fundamentação teórica, o desenvolvimento tecnológico e a análise dos resultados. O estudo perpassa o cenário macro da agricultura digital até a validação do artefato proposto, viabilizando uma compreensão integrada sobre o impacto da Inteligência Artificial (IA) na extensão rural.

A presente investigação estrutura-se em cinco capítulos interdependentes, articulados para garantir a unidade temática e o rigor metodológico. O Capítulo 1 – Introdução estabelece a base do estudo ao apresentar a contextualização do tema e delimitar o problema de pesquisa

voltado à recuperação da informação na Plataforma Ater+ Digital; nele, definem-se os objetivos e explicitam-se a justificativa e as contribuições sociobiotécnicas da proposta.

Na sequência, o Capítulo 2 – Fundamentação Teórica aprofunda as bases conceituais necessárias à compreensão do artefato, discutindo a evolução da ATER Digital no Brasil sob a égide da Embrapa, os óbices da sobrecarga cognitiva em acervos de larga escala e os fundamentos da Inteligência Artificial Generativa, com ênfase na arquitetura *Retrieval-Augmented Generation* (RAG) e nos protocolos de interoperabilidade, a exemplo do *Model Context Protocol* (MCP).

A dimensão procedimental é detalhada no Capítulo 3 – Metodologia, que descreve a natureza aplicada da pesquisa e o percurso tecnológico para a construção do agente conversacional. Este capítulo pormenoriza as etapas de curadoria do *corpus* documental, a estruturação da base de dados vetorial e o desenho da arquitetura de orquestração via n8n, explicitando os critérios de seleção de ferramentas e os protocolos de validação que garantem a replicabilidade do desenvolvimento.

Os resultados dessa implementação são analisados no Capítulo 4 – Resultados e Discussão, no qual o artefato é apresentado e confrontado com testes de desempenho, incluindo métricas de acurácia, latência e fidedignidade às fontes. O capítulo promove ainda uma discussão crítica sobre a usabilidade e o potencial de mitigação de alucinações da IA, articulando os achados empíricos à literatura sobre confiança digital. O Capítulo 5 – Considerações Finais sintetiza as conclusões do estudo, retomando os objetivos propostos para destacar as implicações teóricas e práticas da solução, além de apontar limitações e recomendações para a evolução do ecossistema digital da Embrapa.

À guisa de fechamento, o presente estudo mantém estrito alinhamento com a agenda acadêmica da MUST University, voltada à formação de líderes inovadores aptos a integrar tecnologia e estratégia, robustecendo a articulação entre a produção científica e a práxis

organizacional. A investigação contribui, outrossim, para o avanço das práticas de gestão do conhecimento no agronegócio, ao propor soluções fundamentadas em evidências que impulsionam a transformação digital e a eficiência dos serviços de suporte tanto ao extensionista quanto ao produtor rural.

## 2. Fundamentação Teórica

A fundamentação teórica deste estudo estrutura-se sobre a convergência de dois eixos temáticos centrais: a evolução dos serviços de extensão rural ante a transformação digital e a aplicação de arquiteturas avançadas de Inteligência Artificial (IA) para a gestão do conhecimento. Inicialmente, o capítulo contextualiza a trajetória da Assistência Técnica e Extensão Rural (ATER) no Brasil, analisando como a estratégia de "ATER Digital" tem sido implementada como instrumento de ação pública para ampliar a capilaridade do atendimento e democratizar o acesso à informação técnica (Deus et al., 2024; Pacífico, 2022).

Nesse âmbito, examina-se o papel da Embrapa e da Plataforma Ater+ Digital na sistematização do conhecimento agropecuário, identificando os desafios críticos de recuperabilidade da informação e a sobrecarga cognitiva que emergem com a expansão dos acervos digitais — fenômenos que demandam novas formas de mediação tecnológica para o suporte ao extensionista (Embrapa, 2024; Zuin et al., 2022).

No segundo eixo, estabelecem-se as bases conceituais para a solução tecnológica, explorando as potencialidades e as limitações dos Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) e da IA Generativa, com ênfase nos riscos de alucinação e na imperatividade da precisão técnica em contextos sensíveis (Campos, 2025; Embrapa Agricultura Digital, 2025). Discute-se a adoção da arquitetura *Retrieval-Augmented Generation* (RAG) como mecanismo para assegurar a fidedignidade das respostas, operacionalizada por

intermédio de padrões de interoperabilidade, como o MCP), e ancorada em bancos de dados vetoriais (Netto & Brito, 2025; Google, 2026).

Por conseguinte, o referencial alinha o desenvolvimento do artefato às diretrizes de governança da infraestrutura Embrapa I/O, garantindo que a solução proposta seja aderente aos macroprocessos institucionais de geração de ativos digitais (Carromeu et al., n.d.; Embrapa for Developers, n.d.).

## **2.1 Contexto do Domínio e do Problema**

A compreensão do problema de pesquisa exige, primordialmente, a contextualização do domínio em que este se insere, caracterizado pela interseção entre as políticas públicas de desenvolvimento rural e a transformação digital. Esta seção aborda a trajetória evolutiva dos serviços de Assistência Técnica e Extensão Rural (ATER) em direção ao paradigma digital, analisando como esses instrumentos têm sido reconfigurados para ampliar a capilaridade do atendimento e democratizar o acesso ao conhecimento no campo (Deus et al., 2024; Pacífico, 2022).

Nesse cenário, examina-se o papel estratégico da Embrapa na gestão do conhecimento agropecuário por intermédio da Plataforma Ater+ Digital, que atua como um *hub* centralizador de tecnologias e subsídios técnicos validados (Embrapa, 2024).

Dessarte, discutem-se as demandas críticas de recuperação da informação que emergem da expansão incremental desses acervos, evidenciando como a saturação de dados e a complexidade de navegação impõem óbices à usabilidade. Tais fatores demandam novas arquiteturas de mediação para assegurar a efetividade do suporte aos extensionistas (Zuin et al., 2022).

### **2.1.1 Extensão rural e evolução para o digital**

A trajetória da Assistência Técnica e Extensão Rural (ATER) no Brasil reflete as transformações sociopolíticas e tecnológicas que balizaram o desenvolvimento agrário nacional. Historicamente fundamentada no modelo difusionista — o qual priorizava a transferência vertical de tecnologias oriundas de centros de pesquisa para o campo —, a ATER vivenciou uma reconfiguração substantiva com a promulgação da Lei nº 12.188/2010, que instituiu a Política Nacional de ATER (PNATER). Essa legislação redirecionou o escopo para a agricultura familiar e para metodologias participativas, nas quais a construção do conhecimento ocorre de forma dialógica entre extensionistas e produtores (Pacífico, 2022).

Contudo, a persistência de gargalos estruturais — a exemplo do déficit de recursos humanos e da vasta extensão territorial — tem limitado a universalização do serviço. Tal cenário é corroborado pelos dados do Censo Agropecuário de 2017, os quais demonstram que a maioria dos estabelecimentos rurais ainda carece de orientação técnica regular (IPEA, 2022; Deus et al., 2024).

Ante essas limitações logísticas e orçamentárias, a digitalização emerge não como um substituto, mas como uma estratégia complementar para ampliar a capilaridade e a tempestividade dos serviços de extensão. O conceito de "ATER Digital" consolida-se como o emprego intencional de Tecnologias de Informação e Comunicação (TICs) para mediar processos de ensino-aprendizagem e suporte técnico.

Diretrizes da FAO (2023) destacam que a digitalização permite transpor barreiras físicas, provendo aos agricultores acesso a informações de mercado, previsões agrometeorológicas e diagnósticos fitossanitários em tempo real. Essa transição para o ambiente digital, contudo, exige uma mudança de paradigma: de um sistema centrado na entrega de insumos para um sistema focado na gestão do conhecimento e na curadoria de informações validadas (Zuin et al., 2022).

A implementação de serviços digitais de extensão tem se diversificado globalmente, provendo modelos que oscilam de plataformas de vídeo a assistentes fundamentados em Inteligência Artificial (IA). Um exemplo internacional proeminente é o modelo adotado pela organização *Digital Green*, que utiliza vídeos curtos protagonizados por agricultores locais para disseminar boas práticas na África e na Ásia, evidenciando que a tecnologia alcança maior efetividade quando contextualizada socioculturalmente (FAO, 2023).

Paralelamente, o emprego de agentes conversacionais tem angariado destaque. Na Índia, o sistema *Agri Assist* ilustra como a integração de IA pode auxiliar produtores no diagnóstico de fitopatologias e na recomendação de fertilizantes por intermédio de interfaces acessíveis em *smartphones*, democratizando o acesso a conhecimentos agronômicos complexos (Reddy et al., 2025).

Outro caso emblemático no cenário internacional é o *Virtual Agronomist*, desenvolvido para subsidiar pequenos produtores na África. Este sistema utiliza uma arquitetura de *chatbot* para orientar decisões de manejo de culturas, fundamentando-se em princípios de agronomia digital para prover recomendações personalizadas.

A experiência demonstra que ferramentas conversacionais possuem o potencial de reduzir a assimetria informacional, desde que transponham desafios relacionados à conectividade e ao letramento digital dos usuários (Shepherd et al., 2025). Tais exemplos corroboram a tendência global de transição de repositórios estáticos de informação para sistemas interativos e responsivos.

No contexto brasileiro, iniciativas estaduais têm buscado alinhar-se a essa tendência global. O Instituto de Desenvolvimento Rural do Paraná (IDR-Paraná), por exemplo, implementou o projeto "IDR-Paraná Digit@l", o qual modernizou a infraestrutura de Tecnologia da Informação (TI) e desenvolveu aplicações móveis para agilizar o atendimento e o registro de atividades de campo. A estratégia paranaense integra o provimento de dispositivos

móveis (*tablets*) para extensionistas ao desenvolvimento de sistemas *web* que viabilizam a comunicação remota e o acesso a bases de dados, visando otimizar o tempo operacional dos técnicos e qualificar a gestão das propriedades assistidas (IDR-Paraná, n.d.).

Para além das plataformas de gestão, o Brasil também registra experiências pontuais no desenvolvimento de artefatos de IA customizados para cadeias produtivas específicas. O *Poulbot*, um *chatbot* direcionado à área de incubação artificial de frangos de corte, exemplifica como a tecnologia pode ser verticalizada para atender a nichos produtivos. Desenvolvido em ambiente acadêmico, o *Poulbot* utiliza processamento de linguagem natural (PLN) para dirimir dúvidas técnicas de produtores, evidenciando o potencial de agentes virtuais em atuar em domínios de conhecimento especializado nos quais a presença física perene de um extensionista seria inviável (Oliveira Filho et al., 2021).

Embora as experiências descritas sinalizem o potencial das tecnologias conversacionais na extensão rural, uma análise comparada revela distinções estruturais relevantes. Enquanto o *Virtual Agronomist* (Shepherd et al., 2025) e o *Agri Assist* (Reddy et al., 2025) foram concebidos para contextos de baixa institucionalidade documental — nos quais o conhecimento é gerado e validado ao longo da própria interação —, o *Poulbot* (Oliveira Filho et al., 2021) e as iniciativas brasileiras operam sobre acervos técnicos preexistentes, de autoria institucional consolidada. Essa distinção não é trivial: ela implica que a arquitetura tecnológica adequada para cada contexto difere substancialmente. Sistemas desenhados para ambientes de baixa formalização documental não são diretamente transponíveis a organizações como a Embrapa, cuja soberania informacional exige mecanismos de ancoragem factual ao acervo institucional — demanda que os modelos descritos por Shepherd et al. (2025) e Reddy et al. (2025) não endereçam em sua concepção original. É precisamente essa lacuna que justifica a abordagem RAG-MCP proposta no presente estudo.

Entretanto, a massificação dessas tecnologias impõe desafios prementes relacionados à competência digital dos agentes de ATER. A mera disponibilização de ferramentas não assegura sua apropriação; é imperativo que os extensionistas desenvolvam novas competências para atuar como **infomediários**, filtrando e validando as informações que subsidiam o produtor. Investigações sobre a percepção desses profissionais indicam que, conquanto haja o reconhecimento da relevância da Agricultura 4.0, subsistem lacunas significativas na formação para o uso pedagógico e crítico desses recursos (Santos, 2023). A ATER Digital, neste sentido, deve ser indissociável de políticas de letramento digital e de uma infraestrutura de conectividade resiliente.

A evolução da extensão rural para o paradigma digital configura-se como um processo irreversível e multifacetado, o qual amalgama a modernização da infraestrutura pública — conforme observado no caso do IDR-Paraná — à inovação disruptiva de agentes inteligentes, a exemplo do *Agri Assist* e do *Poulbot*. Nesse cenário, a Plataforma Ater+ Digital da Embrapa posiciona-se como um eixo estruturante, assumindo o desafio de integrar tais tecnologias para assegurar que o conhecimento técnico especializado flua de maneira fidedigna e tempestiva aos tomadores de decisão no campo (Embrapa, 2024).

### **2.1.2 Embrapa, gestão do conhecimento e a plataforma Ater+ Digital**

A Empresa Brasileira de Pesquisa Agropecuária (Embrapa), vinculada ao Ministério da Agricultura e Pecuária (Mapa), consolidou-se ao longo de cinco décadas como um dos pilares basilares da revolução agrícola tropical. Fundada em 1973, a instituição desempenhou papel decisivo na transição do Brasil de importador de alimentos a um dos maiores exportadores mundiais de *commodities* agrícolas.

Esse salto produtivo alicerçou-se em um modelo de pesquisa científica aplicada, voltado à adaptação de tecnologias às condições edafoclimáticas locais e à estruturação de um Sistema

Nacional de Inovação Agropecuária (SNIA) robusto (Vieira Filho & Gasques, 2016). A relevância da Embrapa transcende as fronteiras nacionais, sendo reconhecida globalmente por sua contribuição para a segurança alimentar e para o desenvolvimento de práticas de agricultura sustentável e de baixa emissão de carbono (Embrapa, 2018).

Contudo, a manutenção desse protagonismo na era da sociedade da informação exige a atualização contínua dos processos de gestão do conhecimento. A Política de Inovação da Embrapa estabelece que a inovação não deve restringir-se à geração de novas cultivares ou maquinários, mas deve abranger, outrossim, a forma como esse conhecimento é transladado ao setor produtivo (Embrapa, 2022).

O documento diretriz enfatiza a imperatividade de "institucionalizar conceitos, modelos e métricas para a inovação e a avaliação de impactos", ampliando a capacidade institucional de interagir com distintos atores da cadeia produtiva (Embrapa, 2022, p. 8). Nesse contexto, a transferência de tecnologia (TT) evolui de um modelo unidirecional para ecossistemas digitais que favorecem a conectividade e o compartilhamento dinâmico de saberes.

A materialização dessa estratégia no âmbito da extensão rural pública consubstancia-se na Plataforma Ater+ Digital. Lançada com o propósito de apoiar de forma on-line a implementação de serviços de Assistência Técnica e Extensão Rural (Ater), a plataforma transcende a função de mero repositório, configurando-se como um ambiente virtual robusto para a convergência de conteúdos técnicos em múltiplos formatos — incluindo vídeos, áudios, aplicativos, cartilhas, infográficos e cursos (Embrapa, s.d.-a).

Sua concepção visa superar o desafio histórico de acesso à informação e capilaridade no campo, dado que somente 18,2% dos agricultores familiares brasileiros têm acesso a serviços de Ater, segundo o Censo Agropecuário de 2017 (Embrapa, 2024). Dessa forma, opera como uma ferramenta de suporte voltada a facilitar ações remotas e presenciais dos extensionistas,

integrando tecnologias da informação e da comunicação para levar digitalmente informações mais precisas a produtores e técnicos (Embrapa, s.d.-b; Embrapa, 2024).

A arquitetura da informação da Plataforma Ater+ Digital rompe com a lógica departamental tradicional ao organizar o conhecimento em hubs virtuais, estruturados a partir de "cadeias agropecuárias" e "temas transversais". Tal estrutura otimiza a navegabilidade para o usuário final, que encontra, em um ponto único de acesso, informações de qualidade sobre domínios específicos — a exemplo de Apicultura, Caprinos e Ovinos, Feijão-Caupi, Mandioca e Mudanças Climáticas (Embrapa, s.d.-a; Embrapa, s.d.-b; Embrapa, 2024).

Segundo as diretrizes da iniciativa, a plataforma atua com uma "curadoria especializada" para prover informações organizadas e validadas por especialistas, oferecendo conteúdos em diversos formatos adaptados à realidade do produtor para robustecer a capacidade de resposta dos serviços de Ater (Embrapa, 2024).

Um diferencial crítico da Plataforma Ater+ Digital reside em seu modelo de parceria e governança multi-institucional. A iniciativa é uma ação conjunta que envolve a Embrapa, a Agência Nacional de Assistência Técnica e Extensão Rural (Anater), o Instituto Interamericano de Cooperação para a Agricultura (IICA), o Banco Interamericano de Desenvolvimento (BID), a Associação Brasileira das Entidades de Assistência Técnica e Extensão Rural, Pesquisa Agropecuária e Regularização Fundiária (Asbraer), além dos Ministérios do Desenvolvimento Agrário e Agricultura Familiar (MDA) e da Agricultura e Pecuária (Mapa) (Embrapa, 2025a; IICA, 2025).

A gestão estratégica envolve um Comitê Gestor que, segundo suas diretrizes, atua para que o "lançamento desses novos Hubs [fortaleça] a proposta da Ater+ Digital de disponibilização de conteúdos confiáveis e em diferentes formatos". A expectativa do comitê com essa integração de parceiros é intensificar e ampliar o uso da plataforma para potencializar

as estratégias de assistência técnica e extensão rural em formato digital no país (Embrapa, 2025a).

O processo de alimentação e curadoria da plataforma segue fluxos rigorosos para assegurar a qualidade técnica, o que a distingue de fontes de informação não verificadas disponíveis na rede mundial de computadores.

As equipes de projeto, compostas por pesquisadores, analistas, extensionistas, incumbem-se de manter os conteúdos atualizados ante as novas descobertas científicas ou demandas de campo. Conforme estabelecido no fluxo de processos da plataforma: "A Unidade da Embrapa que coordena o tema na Plataforma Ater+ Digital deverá manter e aprimorar os conteúdos disponibilizados, tendo a responsabilidade de atualizar as informações de maneira contínua. Essa atividade é considerada essencial para manter o interesse do público-alvo" (Embrapa, 2025a).

Entretanto, o próprio êxito da iniciativa em agregar um volume massivo de informações impõe desafios contínuos de usabilidade e formatação. À medida que novos hubs virtuais são implementados — a exemplo dos lançamentos sobre Manga, Peixes e Sistemas Agrícolas Tradicionais (SATs) —, a densidade de recomendações e conhecimentos cresce (Embrapa, 2025a). Para o extensionista que demanda uma resposta célere durante o atendimento de campo, a leitura de extensos arquivos textuais em dispositivos móveis pode tornar-se inviável.

Esse fenômeno de sobrecarga informacional, se não gerido adequadamente por formatos acessíveis, pode engendrar um gargalo de usabilidade, visto que, em ambientes digitais, o envio de textos longos ou arquivos difíceis de baixar, "em vez de ajudar, atrapalha a sua interação com o produtor rural" (Zuin et al., 2022, p. 101).

Para garantir que a informação seja recuperada em tempo oportuno, a plataforma foca em estruturar seus conteúdos de forma responsiva (sem a necessidade de baixar aplicativos

pesados que consomem memória do celular) e em formatos curtos adaptados para o compartilhamento via WhatsApp (Embrapa, s.d.-a; Zuin et al., 2022).

Nesse cenário, reconhece-se a imperatividade de evoluir para interfaces mais inteligentes e interativas. O planejamento estratégico da Plataforma Ater+ Digital contempla a integração com tecnologias emergentes, como o desenvolvimento de ambientes em realidade virtual e metaverso para capacitação na apicultura, além de sistemas de busca via Interfaces de Programação de Aplicações (APIs), como o sistema "500 Perguntas e 500 Respostas", que entrega soluções imediatas a partir de palavras-chave (Embrapa, s.d.-a).

A interoperabilidade com sistemas governamentais correlatos também atua para transmutar o repositório em um ecossistema dinâmico, interligando bases de dados do Agrofit e Agroinsumos, bem como centralizando documentos e o acesso a políticas públicas no aplicativo Meu Imóvel Rural (Embrapa, s.d.-a; Embrapa, s.d.-b). A gestão do conhecimento passa a demandar ferramentas capazes de contextualizar a informação assegurando que o vasto acervo científico da instituição gere valor efetivo e interativo na extremidade da cadeia produtiva (Embrapa, s.d.-a; Zuin et al., 2022).

A Embrapa e sua Plataforma Ater+ Digital representam a base de conhecimento soberana sobre a qual este estudo se debruça. A densidade e a fidedignidade dos dados nela depositados constituem o ativo mais valioso para a estruturação de qualquer solução de Inteligência Artificial. O desafio contemporâneo não reside mais na geração ou na digitalização da informação — etapas já satisfatoriamente endereçadas pela instituição —, mas sim na sua recuperação precisa e contextualizada. Tal lacuna justifica o desenvolvimento de agentes conversacionais fundamentados em arquiteturas *Retrieval-Augmented Generation* (RAG).

### 2.1.3 Necessidades de recuperação da informação na extensão rural digital

A consolidação de plataformas digitais de Assistência Técnica e Extensão Rural (ATER), a exemplo da Ater+ Digital, resultou na centralização de um volume massivo de documentos técnicos, conteúdo multimídia e diretrizes agronômicas.

Conquanto esse fenômeno represente um avanço na preservação e disponibilização do conhecimento, ele introduz o desafio da "infobesidade" ou saturação informacional. Para o extensionista que atua em campo, a disponibilidade de centenas de manuais técnicos torna-se inócua se o tempo despendido para localizar, filtrar e extrair o subsídio específico exceder a janela de oportunidade para a tomada de decisão (Deus et al., 2024).

A literatura aponta que a eficácia da extensão rural digital não reside meramente no armazenamento de dados, mas na capacidade dos Sistemas de Recuperação da Informação (SRI) de proverem respostas precisas e contextualizadas às demandas complexas dos sistemas produtivos (Santos, 2023). Os mecanismos de busca tradicionais, predominantes em portais governamentais e institucionais, operam majoritariamente fundamentados na correspondência léxica de palavras-chave (*keyword matching*). Essa abordagem apresenta limitações severas no contexto agropecuário, no qual a terminologia pode variar regionalmente — fenômeno da sinonímia — e onde um mesmo termo pode assumir acepções distintas conforme a cultura ou o bioma, caracterizando a polissemia.

Como destaca Lima (2025), sistemas que não processam a intenção semântica do usuário tendem a retornar listas extensas de documentos irrelevantes ou, inversamente, falham em recuperar conteúdos pertinentes que não contenham a grafia exata da consulta. Para um extensionista que busca por "controle de lagarta no milho", uma busca booleana elementar pode omitir documentos cruciais que abordem o tema sob descritores como "manejo integrado de pragas em *Zea mays*", engendrando lacunas de conhecimento.

Adicionalmente, a navegação fundamentada em menus hierárquicos, prevalente na arquitetura de informação de portais como a Ater+ Digital, impõe uma carga cognitiva elevada ao usuário. O extensionista vê-se compelido a conhecer previamente a estrutura taxionômica da plataforma — discernindo, por exemplo, se a incidência de uma praga específica está catalogada sob o menu de "Sanidade Vegetal", "Entomologia" ou inserida no *hub* de uma cultura específica (Embrapa, 2025a).

Estudos sobre usabilidade em sistemas de informação agrícola indicam que essa fricção na navegabilidade desestimula a adoção da ferramenta, induzindo os profissionais a recorrerem a fontes exógenas e, amiúde, carentes de validação técnica, como grupos de mensagens instantâneas ou motores de busca genéricos, nos quais a fidedignidade da informação é incerta (Souza et al., 2025).

A necessidade de recuperação da informação no meio rural é, outrossim, intrinsecamente dependente de variáveis contextuais. Uma recomendação agrônômica válida para o Bioma Cerrado pode ser inadequada para o Semiárido; de forma análoga, as práticas de manejo oscilam conforme a escala de produção — familiar ou empresarial — e o patamar tecnológico da propriedade. Por conseguinte, um sistema de recuperação eficiente para a ATER deve ser capaz de filtrar e personalizar o conhecimento, transmutando-se de um repositório passivo em um assistente proativo. Experiências internacionais, a exemplo do *Virtual Agronomist* na África, corroboram que a incorporação de interfaces conversacionais permite ao usuário refinar sua consulta por intermédio do diálogo, provendo nuances do seu contexto para obter uma orientação customizada — patamar de especificidade inalcançável em sistemas de busca estáticos (Shepherd et al., 2025).

Nesse cenário, a emergência de Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) integrados a bancos de dados vetoriais oferece uma nova perspectiva para a recuperação da informação. Distintamente da busca por palavras-chave, a

busca semântica, viabilizada por *embeddings*, permite que o sistema decodifique a intenção contida na indagação e recupere fragmentos de documentos que respondam conceitualmente à dúvida, ainda que não haja coincidência léxica entre os termos (Borges, 2025; Lima, 2025).

Para a Embrapa, tal avanço representa a possibilidade de otimizar o acesso ao conhecimento perenizado em décadas de publicações científicas, permitindo que um extensionista formule consultas em linguagem natural e receba uma síntese fidedigna, fundamentada em excertos específicos de circulares técnicas, manuais da instituição e de parceiros.

Contudo, a aplicação dessas tecnologias na extensão rural exige rigorosos critérios de confiabilidade. A recuperação da informação não deve ser meramente fluida; ela deve ser factualmente precisa e rastreável. O risco de "alucinações" inerente aos modelos generativos — fenômeno no qual a IA produz informações plausíveis, porém apócrifas — assume caráter crítico em domínios sensíveis como a agricultura, em que uma recomendação equivocada de defensivos ou de manejo pode acarretar severos prejuízos econômicos e ambientais (Embrapa Agricultura Digital, 2025).

Por conseguinte, subsiste a necessidade premente de arquiteturas que amalgamentem a fluidez da IA Generativa à segurança da Recuperação Aumentada de Geração (*Retrieval-Augmented Generation* – RAG), assegurando que cada resposta sintetizada pelo sistema esteja estritamente ancorada em fontes oficiais da Plataforma Ater+ Digital.

As necessidades de recuperação da informação na extensão rural digital transcendem a mera digitalização de acervos. Elas demandam a transição de um modelo de "disponibilização" para um paradigma de "mediação inteligente", no qual a tecnologia atua para mitigar a complexidade e o tempo de acesso ao conhecimento técnico.

Superar as barreiras da busca léxica e da navegabilidade estática constitui condição *sine qua non* para que a Plataforma Ater+ Digital cumpra sua missão de capilarizar a inovação

tecnológica da Embrapa, de forma ágil e segura, a todos os recantos do Brasil rural (Deus et al., 2024; Embrapa, 2024).

## 2.2 Fundamentos de Inteligência Artificial Aplicados ao Estudo

A fundamentação tecnológica para o desenvolvimento de soluções inovadoras na extensão rural exige a compreensão dos avanços recentes no campo da Inteligência Artificial (IA), com ênfase na transição paradigmática impulsionada pela IA Generativa. Distintamente dos sistemas convencionais focados em classificação ou predição numérica — a exemplo da IA discriminativa —, a IA Generativa introduz a capacidade de sintetizar, sumarizar e processar dados não estruturados, como a linguagem natural, viabilizando novas modalidades de interação humano-computador (Saraiva et al., 2024).

Esta seção explora a transição conceitual da IA clássica para os modelos generativos, situando seu potencial para perpassar a complexidade do conhecimento agrônômico, ao mesmo tempo em que delinea as fronteiras técnicas e os riscos regulatórios de sua aplicação em ambientes institucionais (Embrapa Agricultura Digital, 2025).

Nesse contexto, os Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs) assumem papel central como motores de inferência semântica, aptos a interpretar e gerar texto a partir de extensos *corpora* documentais. Contudo, sua aplicação em domínios críticos — nos quais a precisão informacional é imperativa — requer uma análise rigorosa de suas capacidades e, sobretudo, de suas limitações intrínsecas, como a propensão a "alucinações" (Campos, 2025). Para operacionalizar tais modelos na realidade da extensão rural, preconiza-se a arquitetura de Agentes Conversacionais, que atuam como interfaces de mediação orquestradas, gerenciando o fluxo de diálogo e assegurando que a tecnologia seja acessível e aderente às demandas operacionais dos usuários finais (Netto & Brito, 2025; Bolivar Huayhua, 2025).

### 2.2.1 Inteligência artificial

A Inteligência Artificial (IA) pode ser compreendida como um campo multidisciplinar da ciência da computação dedicado ao desenvolvimento de sistemas aptos a realizar tarefas que, intrinsecamente, requerem a cognição humana. Tais competências incluem o raciocínio lógico, a aprendizagem contínua, a percepção sensorial, a resolução de problemas complexos e o processamento de linguagem natural (PLN).

Historicamente, a IA evoluiu de abordagens simbólicas — fundamentadas em regras lógicas explícitas e representações formais do conhecimento (IA Clássica) — para abordagens conexionistas, alicerçadas em grandes volumes de dados e modelos estatísticos, popularizadas pelo Aprendizado de Máquina (*Machine Learning* – ML). Segundo Russell e Norvig (2020), o escopo central da área reside na construção de agentes racionais capazes de perceber seu ambiente e agir de modo a maximizar suas probabilidades de êxito no atingimento de objetivos predefinidos.

No contexto da agricultura, a aplicação predominante da IA nas últimas décadas concentrou-se em modelos de Aprendizado de Máquina (*Machine Learning* – ML) supervisionados e não supervisionados, voltados a tarefas analíticas e preditivas. Esses sistemas, frequentemente denominados "IA Discriminativa", são treinados para classificar dados — a exemplo da identificação de fitossanidades em imagens — ou para estimar variáveis quantitativas, como a produtividade de uma safra fundamentada em séries históricas climáticas (Embrapa Agricultura Digital, 2025).

Conquanto sejam extremamente eficazes para otimizar processos produtivos e subsidiar a agricultura de precisão, esses modelos operam em escopos delimitados: analisam dados estruturados para extrair padrões estatísticos, carecendo da capacidade intrínseca de gerar

conteúdos originais ou de sustentar interações complexas em linguagem natural (Saraiva et al., 2024).

A Inteligência Artificial Generativa (IA Generativa ou *GenAI*) representa uma transição paradigmática em relação às abordagens convencionais. Distintamente dos modelos discriminativos, que se restringem a distinguir classes de dados preexistentes, os modelos generativos aprendem a distribuição de probabilidade subjacente aos dados de treinamento para sintetizar novas instâncias que emulam o conjunto original. Isso implica que a IA Generativa transcende a análise passiva; ela detém a capacidade de gerar conteúdos inéditos — sejam textos, imagens, áudios ou códigos de programação — com elevado grau de coerência semântica e verossimilhança (Campos, 2025; Google, 2026).

O avanço recente da IA Generativa foi impulsionado pela evolução das Redes Neurais Profundas (*Deep Learning*) e, especificamente, pela arquitetura *Transformer*, introduzida por Vaswani et al. (2017). Tal arquitetura viabilizou o processamento paralelo de grandes volumes de dados sequenciais e a captura de dependências de longo alcance em textos, superando as limitações intrínsecas às Redes Neurais Recorrentes (RNNs) e às redes de Longo Prazo e Curta Memória (LSTM) (Borges, 2025). A capacidade dos *Transformers* de operar mediante mecanismos de "atenção" — para ponderar a relevância de diferentes segmentos de uma entrada de dados — possibilitou a emergência dos Modelos de Linguagem de Grande Escala (*Large Language Models* – LLMs), que constituem o eixo central das modernas aplicações de texto generativo (Bolivar Huayhua, 2025).

Os LLMs são submetidos ao treinamento em *corpora* massivos de texto, que abrangem vasta parcela do conhecimento disponível em repositórios digitais, acervos bibliográficos e periódicos acadêmicos. Durante esse processo, tais modelos não se limitam à memorização factual; eles internalizam a estrutura sintática e semântica da linguagem, bem como padrões de inferência e conhecimento de mundo.

Essa característica confere à IA Generativa uma versatilidade sem precedentes, habilitando um único modelo a performar múltiplas tarefas para as quais não foi explicitamente otimizado — fenômenos de aprendizado *zero-shot* ou *few-shot* —, tais como tradução idiomática, sumarização, questionamento e resposta (*Question Answering*) e redação criativa (Campos, 2025; Integração de LLMs..., 2025).

Diferentemente dos sistemas de busca convencionais, que recuperam documentos pautados na correspondência de palavras-chave específicas, a IA Generativa é apta a analisar, interpretar e consolidar informações de múltiplas fontes para sintetizar uma resposta única e contextualizada. Tal capacidade transmuta a interação humano-computador: o usuário deixa de emitir comandos sintáticos rígidos para estabelecer um diálogo semântico fluido, no qual a máquina atua como um copiloto cognitivo.

Este agente torna-se capaz de auxiliar na elaboração de estratégias e na resolução de dilemas técnicos complexos, operando de forma sinérgica às necessidades do extensionista (Embrapa Agricultura Digital, 2025; Netto & Brito, 2025).

Contudo, a adoção da IA Generativa não é isenta de riscos substanciais. A natureza estocástica desses modelos implica que a geração de respostas fundamenta-se na probabilidade estatística da linguagem, e não necessariamente na fidedignidade factual. Tal característica pode conduzir à elaboração de conteúdos verossímeis, porém incorretos, fenômeno tecnicamente designado como "alucinações".

Em domínios críticos como a extensão rural — nos quais uma orientação técnica equívoca pode resultar em severos prejuízos econômicos e ambientais ao produtor —, a mitigação desses riscos constitui um requisito de projeto basilar. Isso exige a implementação de arquiteturas que ancorem a geração de texto em bases de dados fidedignas e auditáveis (Embrapa Agricultura Digital, 2025; Campos, 2025).

Importa confrontar as perspectivas da literatura quanto ao papel da IA Generativa em ambientes institucionais sensíveis. Saraiva et al. (2024) adotam uma postura otimista, destacando a capacidade dos modelos generativos de processar dados não estruturados e acelerar a transferência de conhecimento científico. Campos (2025), por sua vez, é mais cauteloso: ainda que reconheça as potencialidades dos *LLMs*, enfatiza que sua adoção sem controles arquiteturais rigorosos representa um vetor de risco para a credibilidade institucional, especialmente em domínios nos quais a precisão técnica não admite margem de erro. A Embrapa Agricultura Digital (2025) ocupa uma posição intermediária, reconhecendo o potencial transformador da tecnologia ao mesmo tempo em que estabelece requisitos mandatórios de governança para sua aplicação no setor público. A convergência dessas perspectivas indica que a IA Generativa não se configura como uma solução autossuficiente, mas como uma capacidade que demanda uma camada de controle — papel que a arquitetura RAG-MCP assume no presente estudo.

A transição da IA discriminativa para a IA Generativa descortina novas fronteiras para a automação de tarefas cognitivas na agricultura. Se a primeira vaga de inovação em Inteligência Artificial permitiu que os sistemas "percebessem" o campo por meio de sensores e visão computacional, esta nova etapa facultou-lhes a capacidade de "interpretar" e "articular" o conhecimento, processando o vasto acervo agrônômico acumulado. O desafio atual reside em canalizar esse potencial produtivo mediante mecanismos de controle e governança informacional que assegurem a acurácia técnica indispensável ao serviço público de ATER.

### **2.2.2 Grandes Modelos de Linguagem: capacidades, limites e alucinações**

Os Modelos de Linguagem de Grande Escala (*Large Language Models* – *LLMs*) representam um avanço disruptivo na Inteligência Artificial, fundamentados na arquitetura de redes neurais profundas do tipo *Transformer*. Treinados em *corpora* textuais de escala massiva

— que abrangem vasta parcela do conhecimento público disponível na rede mundial de computadores —, esses modelos desenvolvem uma representação estatística complexa da linguagem. Tal estrutura permite-lhes não meramente predizer o próximo *token* em uma sequência, mas realizar tarefas de inferência, síntese e tradução com desempenho comparável ao humano em diversos domínios. A capacidade de "aprendizado em contexto" (*in-context learning*) faculta que modelos como o GPT-4 ou o Gemini executem novas tarefas a partir de instruções exíguas (*few-shot*), prescindindo de retreinamento específico, o que os qualifica como tecnologias de propósito geral ou "modelos de fundação" (Campos, 2025; Bolivar Huayhua, 2025).

A principal potencialidade dos LLMs reside na sua competência semântica. Distintamente dos sistemas baseados em palavras-chave, tais modelos conseguem capturar nuances, intenções e contextos implícitos em uma consulta.

No cenário da extensão rural, essa faculdade traduz-se na habilidade de interpretar indagações formuladas em linguagem natural coloquial — a exemplo de "meu feijão tá com as folhas amareladas, o que pode ser?" — e correlacioná-las a constructos técnicos de deficiência nutricional ou fitopatologia, ainda que os termos léxicos não coincidam. Essa flexibilidade cognitiva é imprescindível para mitigar a barreira de acesso ao conhecimento especializado, permitindo que usuários interajam de forma fluida com repositórios de alta complexidade (Netto & Brito, 2025; Embrapa Agricultura Digital, 2025).

No entanto, a arquitetura probabilística que confere fluidez aos LLMs constitui, paradoxalmente, a gênese de sua limitação mais crítica: a propensão a "alucinações". O termo designa a geração de conteúdo sintaticamente coerente e semanticamente plausível, porém factualmente errôneo ou dissociado da realidade objetiva. Visto que tais modelos são otimizados para maximizar a verossimilhança da sequência textual — e não necessariamente a

fidedignidade da informação —, eles podem fabricar citações, dados estatísticos ou prescrições técnicas com elevado grau de convicção.

Em domínios sensíveis como a agricultura, nos quais uma orientação equívoca sobre a dosagem de fitossanitários ou o zoneamento agrícola de risco climático pode acarretar prejuízos severos, a tolerância a essa modalidade de erro é nula (Embrapa Agricultura Digital, 2025; Campos, 2025).

Ademais das alucinações, os LLMs enfrentam gargalos operacionais substanciais, como a "janela de contexto" finita, que restringe a volumetria de dados passíveis de processamento em uma única interação. Conquanto modelos contemporâneos tenham expandido essa capacidade, o tratamento de manuais técnicos extensos ou de históricos prolixos de conversação ainda impõe desafios de custo computacional e de degradação da precisão — fenômeno conhecido como *Lost in the Middle*. Outra limitação crítica é o hiato temporal do conhecimento (*knowledge cutoff*): como o saber do modelo reside cristalizado nos parâmetros da rede neural estabelecidos durante o treinamento, ele é incapaz de acessar informações dinâmicas — como oscilações de mercado ou variáveis climáticas em tempo real — a menos que seja integrado a ecossistemas externos de dados (Google, 2026; Bolivar Huayhua, 2025).

A opacidade dos modelos ("caixa-preta") constitui outro fator limitante para a adoção institucional. Em sistemas de suporte à decisão pública, é imperativo que a resposta fornecida seja rastreável até sua fonte originária. Os LLMs puros, por sua natureza, diluem os dados de procedência durante o processo de treinamento, dificultando a atribuição de autoria e a validação técnico-científica da informação.

Tal lacuna gera insegurança jurídica e operacional, visto que o usuário carece de meios para discernir se a orientação provém de uma circular técnica validada pela Embrapa ou de repositórios não moderados na rede, nos quais a fidedignidade é incerta (Embrapa Agricultura Digital, 2025; Netto & Brito, 2025).

Diante desse quadro, torna-se evidente que, conquanto os LLMs possuam capacidades linguísticas extraordinárias, eles não podem atuar como oráculos autônomos no âmbito da extensão rural. Sua aplicação segura demanda a implementação de arquiteturas de controle que circunscrevam a geração de texto a um domínio de fidedignidade predefinido. Nesse estágio, a integração com bases de conhecimento externas, por intermédio da arquitetura *Retrieval-Augmented Generation* (RAG), apresenta-se como a solução metodológica para mitigar as alucinações e assegurar que a fluidez conversacional permaneça subordinada ao rigor técnico institucional.

### 2.2.3 Agentes conversacionais (chatbots)

Os agentes conversacionais — comumente designados como *chatbots* — constituem sistemas computacionais projetados para emular a interação humana por intermédio da linguagem natural, seja via texto ou voz. Historicamente, tais sistemas evoluíram de algoritmos baseados em correspondência de padrões elementares, a exemplo do programa ELIZA, desenvolvido por Joseph Weizenbaum na década de 1960, para arquiteturas complexas fundamentadas em Inteligência Artificial.

Sob o prisma conceitual, um agente conversacional atua como uma interface de mediação entre o usuário e um ecossistema de informação, incumbindo-se de decodificar a intenção do interlocutor para prover uma resposta ou ação fidedigna. Na atualidade, esses agentes transcendem a mera automação, sendo classificados como "robôs sociáveis", dado que sua eficácia e aceitação condicionam-se à capacidade de estabelecer uma interação fluida e socialmente competente com os usuários (Zunino, 2022; Oliveira Filho et al., 2021).

Sob a ótica arquitetural, os agentes conversacionais são categorizados em duas tipologias principais: os baseados em regras (*rule-based*) e os fundamentados em Inteligência Artificial (IA). Os modelos baseados em regras operam mediante algoritmos determinísticos e

árvores de decisão predefinidas, nos quais o usuário navega por fluxos sintáticos rígidos, selecionando opções delimitadas.

Conquanto sejam eficazes em triagens protocolares, tais sistemas carecem de plasticidade para processar ambiguidades ou consultas exógenas ao roteiro programado. Em contrapartida, os agentes baseados em IA empregam o Processamento de Linguagem Natural (PLN) para decodificar a semântica da entrada do usuário, viabilizando diálogos dinâmicos e não lineares (Oliveira Filho et al., 2021; Souza et al., 2025).

A arquitetura canônica de um agente conversacional moderno compreende módulos de Compreensão de Linguagem Natural (*Natural Language Understanding* – NLU), Gerenciamento de Diálogo e Geração de Linguagem Natural (*Natural Language Generation* – NLG). Em sistemas de vanguarda, essa estrutura é potencializada pela integração com Modelos de Linguagem de Grande Escala (LLMs), que desempenham as funções de NLU e NLG com fidedignidade superior.

Nesse cenário, emergem paradigmas de interação centrados na capacidade de inferência e planejamento (*reasoning and planning*), nos quais o agente não se limita a prover respostas, mas é apto a decompor solicitações complexas em etapas lógicas. Para tanto, utiliza protocolos como *Chain-of-Thought* (CoT) ou *ReAct* (*Reasoning and Acting*) para interagir com o ambiente e recuperar subsídios exógenos antes de formular a síntese final (Bolívar Huayhua, 2025; Google, 2026).

Para a aplicação em domínios de conhecimento especializado, como a agricultura, a arquitetura de referência tem convergido para o emprego da Geração Aumentada por Recuperação (*Retrieval-Augmented Generation* – RAG).

Distintamente de um modelo que se baseia exclusivamente em seu treinamento prévio (memória paramétrica), um agente RAG consulta uma base de conhecimento externa (memória não paramétrica) em tempo real. O fluxo operacional inicia-se com a conversão da indagação

do usuário em vetores numéricos (*embeddings*), seguida pela busca semântica em um banco de dados vetorial que armazena documentos técnicos fragmentados (*chunks*). Os segmentos recuperados são, então, injetados no *prompt* do LLM, instando-o a sintetizar a resposta fundamentada estritamente naquele contexto. Tal abordagem amplia significativamente a acurácia factual e a rastreabilidade da informação (Netto & Brito, 2025; Nunes et al., 2025).

A implementação desses agentes requer, adicionalmente, a orquestração de microsserviços e a integração via interfaces de programação de aplicações (APIs). O advento de plataformas *low-code* e de protocolos de interoperabilidade tem catalisado a conexão entre as interfaces de usuário (*front-end*) — como aplicativos de mensageria instantânea ou portais governamentais — e os motores de IA, permitindo que o agente consuma recursos de ferramentas institucionais.

No ecossistema da Embrapa, tal dinâmica viabiliza que o agente consulte não apenas acervos textuais, mas também dados estruturados oriundos de sistemas legados. Essa integração expande o escopo funcional do sistema: de um assistente de consultas passivas para um agente de execução, apto a automatizar tarefas complexas e subsidiar a tomada de decisão (Nunes et al., 2025; Embrapa for Developers, n.d.).

A usabilidade e a adoção de tais sistemas no meio rural dependem, invariavelmente, de padrões de *design* que contemplem as especificidades sociotécnicas do público-alvo. Estudos indicam que a linguagem empregada pelo *chatbot* deve ser adaptada ao léxico do produtor e do extensionista, mitigando tecnicismos herméticos ou fluxos de interação excessivamente mecanizados.

Experiências como o *Poulbot* — aplicado à avicultura — e projetos-piloto na produção orgânica demonstram que a humanização da interface e a transparência acerca das limitações do sistema constituem fatores críticos para o estabelecimento de confiança (*rapport*). O agente deve ser percebido como um suporte técnico qualificado e onipresente (disponível 24/7), que

suplementa a atuação humana sem pretender substituí-la (Oliveira Filho et al., 2021; Souza et al., 2025).

A literatura, contudo, não é unívoca quanto às condições para o sucesso desses sistemas. Zunino (2022) argumenta que a aceitação de agentes conversacionais em organizações está diretamente condicionada ao processo de maturação organizacional — ou seja, à capacidade da instituição de ressignificar seus fluxos de trabalho para acomodar o agente como um ator legítimo, e não como um mero substituto de tarefas. Essa perspectiva dialoga criticamente com os achados de Oliveira Filho et al. (2021) e Souza et al. (2025), que tendem a avaliar o sucesso do *chatbot* predominantemente por métricas de desempenho técnico (acurácia, tempo de resposta).

A tensão entre essas abordagens é produtiva: ela indica que a validação de um artefato conversacional em contexto institucional público não pode limitar-se à dimensão computacional, mas deve incorporar indicadores de confiança e apropriação pelo usuário — dimensão que o presente estudo endereça ao optar pela metodologia DSR, cuja avaliação do artefato contempla tanto o desempenho técnico quanto a utilidade percebida no contexto analisado.

Os agentes conversacionais representam, hodiernamente, a fronteira da interação humano-computador na extensão rural digital. A evolução das arquiteturas para modelos alicerçados em *Large Language Models* (LLMs) e *Retrieval-Augmented Generation* (RAG) permite transpor as limitações dos *chatbots* de primeira geração, provendo respostas contextuais, dinâmicas e consubstanciadas em evidências técnicas.

Contudo, o êxito dessa tecnologia não reside meramente na sofisticação algorítmica, mas na sua integração harmoniosa aos fluxos de trabalho da Assistência Técnica e Extensão Rural (ATER). Tal sinergia favorece que a inovação digital cumpra seu propósito axiológico:

ampliar a capacidade do extensionista e do produtor rural com informações precisas, acionáveis e tempestivas (Zunino, 2022; Shepherd et al., 2025).

### 2.3 Ferramentas e Tecnologias Habilitadoras

A operacionalização de agentes conversacionais aptos a interagir com bases de conhecimento complexas exige a articulação de um ecossistema específico de ferramentas e tecnologias habilitadoras. Distintamente de *chatbots* fundamentados em regras determinísticas, os sistemas baseados em Inteligência Artificial Generativa demandam uma arquitetura robusta que integre o processamento semântico da linguagem à recuperação célere da informação.

Nesse contexto, técnicas de representação vetorial (*embeddings*) e bancos de dados vetoriais assumem papel preponderante, permitindo que o sistema interprete semanticamente as consultas e recupere, com baixa latência, os segmentos documentais de maior relevância em vastos acervos não estruturados (Ribeiro et al., 2025; Lima, 2025). Essa infraestrutura de dados constitui o alicerce para a implementação da arquitetura *Retrieval-Augmented Generation* (RAG), que amalgama a fluidez dos modelos generativos à acurácia factual imperativa em domínios críticos como a extensão rural (Netto & Brito, 2025).

Para além dos componentes de IA, a construção de um artefato funcional em ecossistemas institucionais requer mecanismos eficazes de orquestração e governança. A integração entre Modelos de Linguagem de Grande Escala (LLMs), repositórios de dados e interfaces de usuário demanda o emprego de plataformas de orquestração de serviços — frequentemente fundamentadas em paradigmas *low-code* —, que otimizam o fluxo de dados e a manutenibilidade do sistema (Nunes et al., 2025).

Simultaneamente, a sustentabilidade tecnológica da solução está condicionada à adesão a padrões de interoperabilidade e diretrizes de governança de ativos digitais, como as preconizadas pela plataforma Embrapa I/O. Tais diretrizes favorecem que a inovação respeite

protocolos de segurança, auditabilidade e o ciclo de vida do software corporativo, contribuindo para que o agente conversacional seja não apenas inteligente, mas também aderente aos requisitos institucionais de operação e evolução tecnológica (Carromeu et al., 2023; Embrapa, 2022).

### 2.3.1 Recuperação Semântica e Embeddings

A recuperação da informação evoluiu substancialmente com a transição dos métodos baseados em correspondência exata de palavras-chave (*keyword matching*) para abordagens fundamentadas na semântica profunda. Enquanto a busca léxica tradicional está adstrita à presença literal dos termos da consulta nos documentos-alvo — o que frequentemente resulta em silêncios de recuperação decorrentes da sinonímia ou polissemia —, a recuperação semântica visa decodificar a ontologia e a intenção subjacentes à consulta do usuário. Segundo Lima (2025), essa abordagem viabiliza que um sistema identifique documentos pertinentes mesmo na ausência de sobreposição vocabular direta, analisando o contexto sintagmático em que as palavras são empregadas e as intrincadas relações conceituais entre elas.

O núcleo tecnológico que viabiliza essa compreensão reside no conceito de *embeddings* (ou vetores de incorporação). Os *embeddings* constituem representações numéricas de dados textuais — sejam palavras, sentenças ou documentos íntegros — projetadas em um espaço vetorial multidimensional. Nesse domínio, a proximidade geométrica entre dois vetores denota sua **similaridade semântica**: conceitos correlatos são agrupados em coordenadas adjacentes.

Modelos de redes neurais profundas, a exemplo do BERT (*Bidirectional Encoder Representations from Transformers*) e suas derivações, são treinados em vastos *corpora* para mapear a linguagem humana nessas representações matemáticas densas. Tal abordagem permite capturar nuances sintáticas e semânticas que transcendem as capacidades dos algoritmos estatísticos clássicos e esparsos, como o TF-IDF (Ribeiro et al., 2025; Godoy, 2024).

A superioridade dos *embeddings* reside na sua capacidade de mitigar a ambiguidade e a variabilidade intrínsecas à linguagem natural. Em um ecossistema técnico como o da extensão rural, essa faculdade é crucial. Por exemplo, um vetor gerado para o termo coloquial "praga no milho" situar-se-á, matematicamente, em uma vizinhança próxima ao vetor de "lagarta-do-cartucho" (*Spodoptera frugiperda*), permitindo que o sistema recupere uma circular técnica sobre o inseto ainda que o usuário ignore sua nomenclatura científica ou comum.

Borges (2025) destaca que essa capacidade de generalização é o que diferencia um assistente inteligente de um mecanismo de busca rudimentar, pois habilita o sistema a atuar como um **mediador cognitivo** que traduz a incerteza do usuário para a precisão terminológica dos acervos documentais.

A implementação de uma camada de recuperação semântica, fundamentada em *embeddings*, constitui o estágio basilar para a edificação de arquiteturas RAG eficazes. Ao transpor tanto o acervo de conhecimento da Embrapa quanto as indagações dos extensionistas para o domínio vetorial, torna-se viável a execução de cálculos de similaridade — a exemplo da **similaridade de cosseno** — para ranquear e extrair os fragmentos de informação de maior pertinência. Tal procedimento assegura que o Modelo de Linguagem receba um contexto denso e acurado para a síntese de suas respostas, mitigando a probabilidade de alucinações e elevando a assertividade do suporte técnico provido (Lima, 2025; Ribeiro et al., 2025).

### 2.3.2 Bancos de dados vetoriais

A viabilização da busca semântica em larga escala exige uma infraestrutura de armazenamento e recuperação apta a gerenciar a alta dimensionalidade dos *embeddings*. Os bancos de dados convencionais — sejam relacionais (SQL) ou orientados a documentos (NoSQL) — não são otimizados para o cômputo de similaridade entre vetores, o que torna as

operações de busca por vizinhança computacionalmente proibitivas à medida que a volumetria de dados se expande.

Para mitigar esse gargalo, emergiram os Sistemas de Gerenciamento de Bancos de Dados Vetoriais (*Vector Database Management Systems – VDBMS*), projetados especificamente para o armazenamento, indexação e consulta de representações vetoriais densas. Segundo Pan et al. (2023), tais sistemas distinguem-se por oferecerem algoritmos de indexação especializados que viabilizam a execução eficiente de consultas do tipo *Approximate Nearest Neighbor* (ANN), logrando um equilíbrio otimizado entre precisão e latência na recuperação da informação.

O funcionamento desses sistemas fundamenta-se na edificação de estruturas de dados que particionam o espaço vetorial, circunscrevendo a busca a subconjuntos promissores de vetores, em preterição à exaustiva varredura linear de toda a base. A técnica de maior proeminência na atualidade é o *Hierarchical Navigable Small World (HNSW)*, um algoritmo alicerçado em grafos que estabelece camadas hierárquicas de navegação. Essa estrutura facilita ao sistema "transitar" célere entre regiões do espaço vetorial até identificar os vizinhos mais próximos da indagação do usuário. Tal abordagem demonstra-se interessante para ecossistemas de extensão rural, como a Ater+ Digital, nos quais a latência de resposta deve ser minimizada para assegurar uma interação fluida, ainda que sob a consulta de volumosos acervos de documentos técnicos (Qdrant, n.d.; Ribeiro et al., 2025).

No âmbito das arquiteturas RAG (*Retrieval-Augmented Generation*), o banco de dados vetorial opera como a "memória não paramétrica" do sistema. Enquanto o Modelo de Linguagem de Grande Escala (LLM) detém um conhecimento estático, cristalizado em seus parâmetros durante o treinamento, o repositório vetorial facilita a inserção, atualização e exclusão dinâmica de informações, prescindindo do oneroso retreinamento da rede neural.

Conforme Netto e Brito (2025), essa característica é basilar para salvaguardar a soberania e a atualidade dos ativos informacionais, permitindo que novos dispositivos técnicos — como circulares ou alertas fitossanitários da Embrapa — sejam vetorizados e disponibilizados para consulta por meio do *chatbot* em regime de quase tempo real.

Para a implementação do artefato desenvolvido nesta pesquisa, selecionou-se o **Qdrant**, um motor de busca vetorial de código aberto (*open-source*) desenvolvido em linguagem **Rust**, notabilizado por seu elevado desempenho e escalabilidade. O Qdrant implementa uma versão otimizada do algoritmo HNSW, provendo funcionalidades avançadas de filtragem híbrida que permitem consubstanciar a busca semântica à aplicação de filtros de metadados convencionais — a exemplo da restrição de resultados ao bioma "Cerrado" ou ao recorte temporal de "2024".

Essa capacidade de *Payload Filtering* demonstra-se decisiva para a acurácia das respostas no domínio agropecuário, visto que viabiliza a circunscrição do contexto da IA a parâmetros agronômicos específicos, mitigando o risco de que prescrições técnicas válidas para uma ecorregião sejam transpostas erroneamente a outra (Qdrant, n.d.; Nunes et al., 2025). A seleção de um banco de dados vetorial profícuo envolve, por conseguinte, a análise criteriosa de variáveis como latência, *recall* (sensibilidade na recuperação de itens pertinentes), custo computacional e interoperabilidade com o ecossistema de desenvolvimento.

Ribeiro et al. (2025) salientam que a performance da recuperação semântica no idioma português depende, intrinsecamente, da sinergia entre o modelo de *embedding* empregado e a eficiência da indexação vetorial na preservação das relações semânticas. Nesse sentido, a infraestrutura vetorial transcende a função de um repositório passivo, configurando-se como um componente ativo que determina a fidedignidade e a autoridade da informação provida ao extensionista.

### 2.3.3 Retrieval-Augmented Generation (RAG)

A arquitetura *Retrieval-Augmented Generation (RAG)*, introduzida seminalmente por Lewis et al. (2020), representa um marco na evolução do Processamento de Linguagem Natural (PLN) ao propor um modelo híbrido que amalgama a fluidez dos modelos generativos pré-treinados à acurácia de sistemas de recuperação de informação.

Distintamente dos LLMs convencionais, que dependem exclusivamente de sua "memória paramétrica" (conhecimento cristalizado nos pesos da rede neural durante o treinamento), o RAG acopla ao modelo uma "memória não paramétrica" externa, consubstanciada em um índice vetorial denso. Essa abordagem faculta ao sistema a consulta a dados dinâmicos e específicos de domínio em tempo real, transpondo limitações críticas como o hiato temporal do conhecimento (*knowledge cutoff*) e a impossibilidade de acessar repositórios proprietários ou privados (Netto & Brito, 2025; Lewis et al., 2020).

O funcionamento do RAG estrutura-se em dois processos sequenciais e interdependentes: a recuperação (*retrieval*) e a geração (*generation*). Na etapa de recuperação, a indagação do usuário é transposta para um vetor numérico (*query embedding*) e confrontada semanticamente com um repositório vetorial que armazena o *corpus* documental previamente fragmentado (*chunks*). Os segmentos de maior relevância — identificados mediante métricas de similaridade vetorial — são extraídos. Na etapa subsequente, esses fragmentos são injetados no *prompt* de entrada do Modelo de Linguagem, constituindo um "contexto de fidedignidade" sobre o qual o modelo deve alicerçar sua resposta. Esse mecanismo de ancoragem (*grounding*) reduz drasticamente a incidência de alucinações, visto que compele a IA a operar estritamente dentro dos limites das evidências documentais providas (Borges, 2025; Lima, 2025).

A eficácia dessa arquitetura depende intrinsecamente da qualidade dos modelos de representação vetorial utilizados para indexar e recuperar a informação. Para o desenvolvimento

do artefato proposto neste estudo, optou-se pela utilização do modelo Google text-embedding-001 (Google Cloud, 2024).

Este modelo, disponibilizado via Vertex AI, destaca-se por sua alta dimensionalidade e otimização específica para tarefas de recuperação semântica e resposta a perguntas (Question Answering). Segundo a documentação técnica oficial, a versão 001 oferece desempenho superior em benchmarks de recuperação multilíngue, sendo capaz de capturar nuances semânticas complexas em língua portuguesa, o que é essencial para a interpretação correta do vocabulário técnico agrônomo da Embrapa (Google Cloud, 2024).

Além da acurácia semântica, a arquitetura RAG oferece vantagens substanciais em termos de auditabilidade e governança informacional. Distintamente de um modelo de "caixa-preta" (*black box*), no qual a gênese da informação é difusa, o sistema RAG permite rastrear com precisão qual documento ou segmento textual foi mobilizado para subsidiar a gênese da resposta.

Tal faculdade possibilita que o agente conversacional proveja não meramente a orientação técnica, mas também a atribuição de fonte originária (e.g., "Conforme a Circular Técnica 45 da Embrapa Milho e Sorgo..."), o que amplia a transparência e a confiabilidade do usuário no ecossistema. Essa característica converge diretamente com os princípios de **IA Responsável** e explicabilidade preconizados pelas diretrizes de governança da Embrapa (Embrapa Agricultura Digital, 2025; Netto & Brito, 2025).

A adoção do RAG, potencializada pelo modelo Google text-embedding-001, constitui a espinha dorsal da solução tecnológica para a Plataforma Ater+ Digital. Essa configuração permite transformar o vasto repositório estático da instituição em um ecossistema dinâmico de conhecimento, onde o extensionista pode dialogar com a base técnica da Embrapa de forma natural e obter respostas que são, ao mesmo tempo, cognitivamente acessíveis e cientificamente

rigorosas. A tecnologia atua, assim, como uma ponte segura entre a fronteira do conhecimento agrônômico e a prática de campo.

Não obstante as vantagens amplamente documentadas, a literatura aponta limitações operacionais do RAG que merecem atenção crítica. Borges (2025) e Lima (2025) convergem ao identificar que a eficácia da recuperação semântica é diretamente proporcional à qualidade da segmentação (*chunking*) e à homogeneidade estrutural dos documentos indexados: acervos com alta incidência de tabelas, imagens técnicas ou vocabulário regional atípico tendem a produzir recuperações subótimas, comprometendo a ancoragem factual da resposta gerada.

Ribeiro et al. (2025) acrescentam que modelos de *embedding* treinados em *corpora* predominantemente em inglês podem apresentar degradação semântica na língua portuguesa, especialmente no que tange ao vocabulário técnico agrônômico regional. Essa convergência crítica reforça que a qualidade do artefato RAG não é determinada exclusivamente pela sofisticação do modelo generativo, mas, em igual medida, pela curadoria documental que antecede a indexação — aspecto que o presente estudo trata explicitamente nos procedimentos metodológicos da atual pesquisa.

### 2.3.4 Orquestração e integração de serviços

Para viabilizar essa orquestração de forma ágil e escalável, a engenharia de *software* contemporânea tem adotado, crescentemente, o paradigma **Low-Code/No-Code** (LCNC). Essa abordagem permite o desenvolvimento de aplicações mediante interfaces visuais e componentes pré-configurados, mitigando significativamente a necessidade de codificação manual extensiva. Segundo Nelimarkka (2023), o emprego de plataformas *low-code* democratiza o desenvolvimento de soluções tecnológicas ao permitir que especialistas de domínio — no presente contexto, profissionais com expertise em extensão rural — participem ativamente da construção do artefato. Ademais, tal paradigma acelera o ciclo de prototipagem

e validação de Produtos Mínimos Viáveis (MVPs), otimizando o tempo de resposta às demandas do campo.

Nesse contexto, a plataforma **n8n** foi selecionada como a ferramenta de orquestração para este projeto. Distintamente de outras soluções de automação de mercado — como Zapier ou Make —, o n8n constitui uma plataforma de *workflow automation* de "código justo" (*fair-code*) e arquitetura aberta, projetada especificamente para suportar fluxos de trabalho de elevada complexidade técnica.

A literatura recente destaca o n8n como uma infraestrutura ideal para a "Automação Neural-Aumentada", visto que permite a integração nativa de nós de Inteligência Artificial. Tal característica possibilita que agentes autônomos executem tarefas de raciocínio, planejamento e execução (*chain-of-thought*) dentro de um ambiente visual controlado e auditável (Davis, 2023; Global Knowledge Academy, 2024).

A justificativa técnica para a seleção do n8n fundamenta-se em sua arquitetura baseada em nós (*node-based architecture*), que provê uma flexibilidade granular para a manipulação de objetos JSON e dados binários em cada estágio do processamento. Para a implementação do agente conversacional, o n8n viabiliza a estruturação de "**fluxos agênticos**" (*agentic workflows*), nos quais múltiplos agentes de IA podem interagir de forma síncrona ou assíncrona.

Por exemplo, um fluxo pode integrar um agente incumbido de classificar a intenção da indagação do extensionista, um segundo componente para a conversão da consulta em *embeddings* e um terceiro para a síntese e estruturação da resposta final. Tal modularidade simplifica a manutenção e a escalabilidade do sistema, facultando a substituição de modelos de LLM ou provedores de bancos de dados vetoriais sem a necessidade de refatoração do código-fonte da aplicação (Global Knowledge Academy, 2024).

Outro fator determinante para a adoção do n8n no ecossistema da Embrapa reside na sua capacidade de auto-hospedagem (*self-hosting*). Diferentemente de diversas plataformas

*low-code* que operam exclusivamente sob o modelo de *Software as a Service* (SaaS) em nuvens públicas, o n8n permite a implantação em servidores locais ou em nuvens privadas virtuais (*Virtual Private Clouds* – VPC).

Tal prerrogativa assegura a estrita conformidade com as diretrizes de governança de dados e segurança da informação da instituição, garantindo que fluxos de dados sensíveis ou estratégicos processados pelo agente conversacional permaneçam sob controle institucional, sem trafegar por infraestruturas de terceiros não auditáveis. Ademais, a auto-hospedagem mitiga vulnerabilidades de latência e a volatilidade de custos atrelada ao volume de execuções, conferindo previsibilidade operacional e estabilidade à Plataforma Ater+ Digital (n8n.io, n.d.; Carromeu et al., 2023).

É pertinente, contudo, situar a adoção do n8n em perspectiva crítica. Davis (2023) e Global Knowledge Academy (2024), embora endossem as capacidades da plataforma para automação neural-aumentada, reconhecem que fluxos agênticos de alta complexidade podem apresentar gargalos de desempenho em cenários de alta concorrência de requisições — limitação relevante para um serviço de extensão rural com potencial de escala nacional. Nelimarkka (2023), ao analisar o paradigma *low-code* de forma mais abrangente, alerta que a democratização do desenvolvimento, embora desejável, pode gerar débitos técnicos quando a lógica de negócio atinge níveis de complexidade que extrapolam as capacidades visuais da plataforma.

No contexto deste estudo, essas limitações foram consideradas e mitigadas pela opção de auto-hospedagem em infraestrutura controlada e pela arquitetura modular adotada, que preserva a capacidade de substituição de componentes sem refatoração profunda. A escolha do n8n deve, portanto, ser compreendida como adequada ao escopo e à fase atual do artefato, sem necessariamente configurar a solução definitiva para um ambiente de produção em escala plena.

A orquestração via n8n, dessa forma, não representa apenas uma escolha por conveniência de desenvolvimento, mas uma decisão arquitetural estratégica. Ela viabiliza a integração "fim a fim" (end-to-end) entre o front-end de chat, o cérebro cognitivo (LLM/RAG) e os sistemas legados da Embrapa, provendo a cola tecnológica necessária para transformar componentes isolados em um serviço de extensão rural digital coeso, auditável e de alto desempenho (Global Knowledge Academy, 2024; Davis, 2023).

### 2.3.5 Interoperabilidade e governança

A fragmentação dos sistemas de informação na agricultura constitui um dos principais óbices à transformação digital do setor. Dados agronômicos, climáticos e mercadológicos residem, frequentemente, em silos isolados, sob formatos proprietários e APIs heterogêneas, o que obstaculiza a consolidação de uma visão holística para a tomada de decisão. A interoperabilidade, por conseguinte, transcende o requisito técnico para configurar-se como uma premissa estratégica, permitindo que agentes de Inteligência Artificial correlacionem informações outrora dispersas.

Convencionalmente, a integração de LLMs com repositórios externos demandava o desenvolvimento de conectores customizados, resultando em arquiteturas rígidas e de onerosa manutenção. Para mitigar esse cenário, o projeto adota o MCP, um padrão aberto que normaliza a interface entre modelos de IA (*clientes*) e ecossistemas de dados (*servidores*), promovendo o desacoplamento entre a inteligência do modelo e a recuperação da informação (Model Context Protocol, n.d.; Hou et al., 2025).

O MCP estabelece uma arquitetura universal para o intercâmbio de contexto, permitindo que o agente conversacional se conecte a repositórios heterogêneos — como o banco vetorial **Qdrant**, APIs institucionais ou sistemas de arquivos locais — mediante uma interface

padronizada. Segundo a especificação do protocolo, o MCP opera sob um modelo triádico de cliente-host-servidor, que assegura o isolamento e a integridade das conexões.

O elemento *Host* atua como um orquestrador central que gerencia o ciclo de vida das sessões e impõe políticas de governança, garantindo que o LLM acesse estritamente os dados sob permissão explícita. Essa padronização mitiga a necessidade de refatoração do código do agente ante novas integrações, favorecendo a escalabilidade e a modularidade do ecossistema digital (Model Context Protocol, n.d.; Intelligent Automation..., 2025).

A adoção de padrões de interoperabilidade deve, contudo, estar subordinada a uma estrutura robusta de governança institucional. Conforme Embrapa (2022), a Política de Inovação da estatal preconiza que as iniciativas de transformação digital devem observar estritamente os princípios de segurança da informação, a salvaguarda da propriedade intelectual e a conformidade legal.

No âmbito da Plataforma Ater+ Digital, a governança é exercida por um Comitê Gestor multi-institucional, incumbido de dirimir normas, procedimentos e indicadores de desempenho (KPIs). O desenvolvimento do agente conversacional converge com tais diretrizes ao integrar-se à infraestrutura da plataforma Embrapa I/O, que implementa fluxos automatizados de *DevSecOps* para auditoria de código, mitigação de vulnerabilidades (*Common Vulnerabilities and Exposures* – CVEs) e telemetria de *logs*. Tal integração assegura que o ativo digital seja concebido em estrita aderência aos padrões de segurança e resiliência da empresa (Embrapa, 2025b; Carromeu et al., 2023).

Um aspecto crítico da governança em sistemas de IA reside na salvaguarda de dados pessoais e na ética algorítmica. O emprego de arquiteturas RAG orquestradas internamente — mediante o n8n em servidores próprios (*on-premises*) — faculta à Embrapa a manutenção da soberania sobre os dados transacionados, mitigando o tráfego de informações sensíveis de extensionistas ou produtores para ambientes de terceiros não auditáveis.

Complementarmente, a transparência do sistema é corroborada pela capacidade do MCP de prover metadados precisos acerca da proveniência das informações recuperadas, o que viabiliza a auditabilidade técnica das respostas geradas pelo modelo (Embrapa Agricultura Digital, 2025; Embrapa, 2022).

A integração entre a interoperabilidade técnica e a governança institucional estabelece um ecossistema favorável à inovação responsável. Enquanto o MCP soluciona o óbice da conectividade de dados — facultando ao assistente virtual a consulta a manuais técnicos e bases de dados em tempo real —, as diretrizes da Embrapa asseveram que tal acesso ocorra sob preceitos éticos e protocolos de segurança.

A governança da Ater+ Digital, ao estipular atribuições precípuas para a curadoria de conteúdo e a validação técnica, suplementa a arquitetura de *software*, garantindo que o motor de inferência do sistema seja alimentado exclusivamente por conhecimentos convalidados. Essa sinergia demonstra-se basilar para mitigar riscos reputacionais e assegurar que a tecnologia cumpra sua função social de robustecer a extensão rural pública (Embrapa, 2025b; Hou et al., 2025).

A arquitetura proposta para o artefato deste estudo transcende a mera justaposição de tecnologias contemporâneas; ela estabelece um modelo de referência para a edificação de sistemas de IA soberanos no setor público. Ao amalgamar a flexibilidade do MCP ao rigor da governança da Embrapa, o projeto demonstra a viabilidade de se alcançar a interoperabilidade sem o comprometimento do controle institucional. Tal convergência pavimenta o caminho para um cenário em que a Inteligência Artificial atue como uma infraestrutura transversal, resiliente e fidedigna, em prol do fortalecimento e da modernização da agricultura brasileira.

## **2.4 Ambiente Institucional de Desenvolvimento e Implantação**

A operacionalização de inovações tecnológicas na Embrapa transcende a mera codificação isolada, inserindo-se em um ecossistema institucional robusto, concebido para salvaguardar a sustentabilidade, a segurança e a escalabilidade das soluções digitais. Esta seção pormenoriza a plataforma Embrapa I/O, ambiente oficial de *DevOps* que materializa o Macroprocesso de Desenvolvimento de Ativos Digitais.

Tal infraestrutura estabelece um fluxo de governança padronizado que abrange desde a concepção científica até a entrega de produtos de *software* maduros ao setor produtivo (Carromeu et al., 2023; Embrapa, 2025c). Sob essa ótica, examina-se a Arquitetura de Referência em IA Generativa, proposta pela Gerência de Tecnologia da Informação (GTI) da Embrapa, que preconiza os padrões para a edificação de serviços sob o paradigma de "Agente como Serviço" (*Agent as a Service – AaaS*). Essa fundamentação assegura a conformidade técnica e a aderência do artefato desenvolvido às diretrizes corporativas de inovação aberta e descentralizada (Embrapa, 2025b).

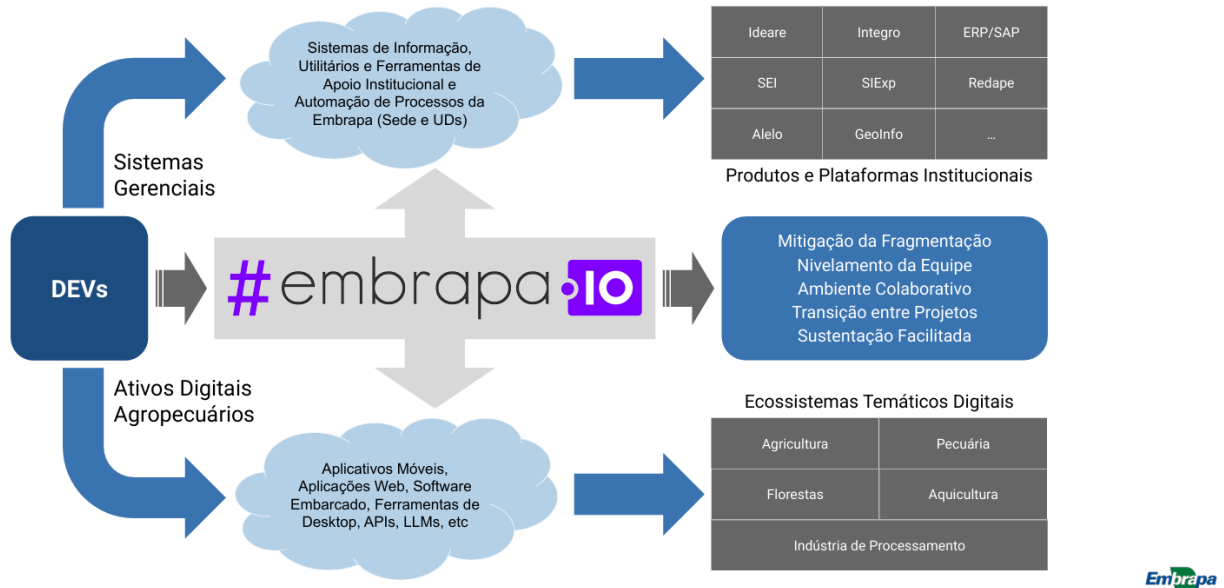
#### **2.4.1 Embrapa.io e a governança de ativos digitais**

A plataforma Embrapa I/O constitui o ecossistema oficial para o desenvolvimento, a operação e a governança de ativos digitais da Empresa Brasileira de Pesquisa Agropecuária. Concebida para mitigar a fragmentação de esforços e a descontinuidade de projetos de *software*, a plataforma opera sob o paradigma *DevOps*, integrando ferramentas de código aberto para automatizar o ciclo de vida das aplicações — da concepção à entrega contínua (*Continuous Delivery*).

Transcendendo a função de um mero repositório de código, o Embrapa I/O materializa uma estratégia institucional de transformação digital, provendo uma infraestrutura padronizada que habilita equipes multidisciplinares — internas e parceiras — a colaborar na edificação de

soluções tecnológicas aderentes às demandas complexas da agricultura tropical (Carroneu et al., 2023).

**Figura 01. Representação Sistêmica da Infraestrutura de Governança Embrapa I/O.**



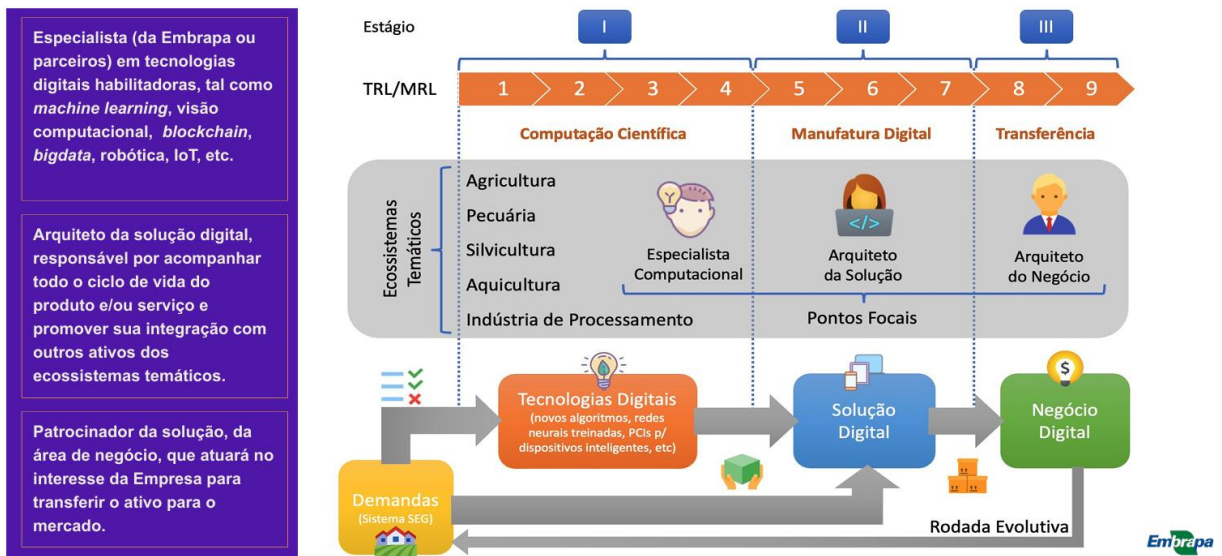
Fonte: Embrapa. (2025c).

A governança no âmbito da plataforma é regida pelo Macroprocesso de Desenvolvimento de Ativos Digitais, um arcabouço metodológico que categoriza a maturidade das soluções com esteio nas escalas *Technology Readiness Level (TRL)* e *Market Readiness Level (MRL)*.

Esse processo estratifica o desenvolvimento em três estágios distintos: a computação científica, na qual modelos e algoritmos são convalidados; a manufatura digital, em que tais modelos são encapsulados em produtos de *software* resilientes; e a transferência mercadológica, centrada na modelagem de negócios e na entrega de valor ao usuário final.

Tal estrutura assegura que apenas ativos tecnicamente exauridos e estrategicamente alinhados progridam para as etapas de produção, mitigando o risco de disponibilizar soluções incipientes ao setor agropecuário.

**Figura 02 – Escalas TRL e MRL Aplicadas ao Macroprocesso de Ativos Digitais.**



Fonte: Embrapa. (2025c).

Para salvaguardar a qualidade e a integridade dos ativos, o Embrapa I/O implementa *pipelines* automatizados de **CI/CD** (*Continuous Integration/Continuous Delivery*) que executam testes unitários, validação de estilo de código (*linting*) e varreduras de segurança. A plataforma incorpora práticas de **DevSecOps**, monitorando ativamente vulnerabilidades conhecidas (*Common Vulnerabilities and Exposures – CVEs*) e riscos de exposição de dados, além de prover painéis analíticos acerca da telemetria e do desempenho das aplicações em tempo real.

Esse rigor técnico demonstra-se imperativo para a sustentabilidade de projetos de longo curso, como a Plataforma Ater+ Digital, pois estabelece padrões de auditabilidade e manutenibilidade que prescindem da perenidade das equipes de desenvolvimento.

**Figura 03 – Modelos de Arquitetura (Boilerplates) para Integração de IA Generativa**



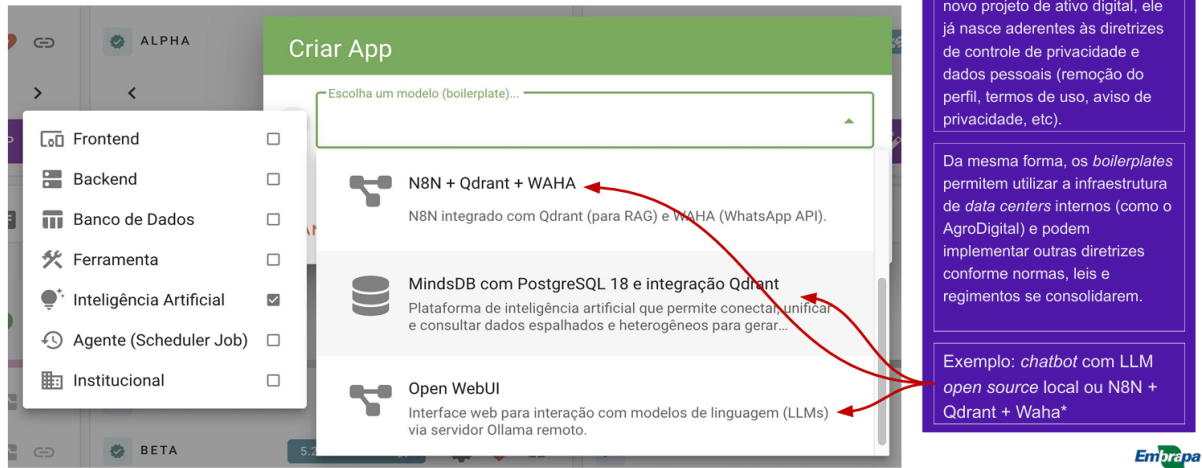
Fonte: Embrapa. (2025c).

A demais, a plataforma fomenta a inovação aberta mediante o conceito de *boilerplates* — modelos de arquitetura de *software* pré-configurados e otimizados para domínios específicos da agricultura digital e sistemas corporativos. Tais gabaritos arquiteturais provêm estruturas fundacionais validadas, que incorporam as melhores práticas de segurança e interoperabilidade da instituição.

No âmbito desta pesquisa, a utilização de um *boilerplate* específico para microsserviços e integração de IA permitiu que o desenvolvimento do agente conversacional se concentrasse na lógica de negócio e no refino do RAG, prescindindo da configuração manual de infraestruturas de base repetitivas e acelerando o ciclo de entrega de valor ao extensionista.

**Figura 04 – Boilerplates disponíveis no Embrapa.ai para desenvolvimento com IA**

**Boilerplates** para Uso de IA Generativa e desenvolvimento de **Ativos Digitais e Institucionais e Agropecuários**



Fonte: Embrapa. (2025c).

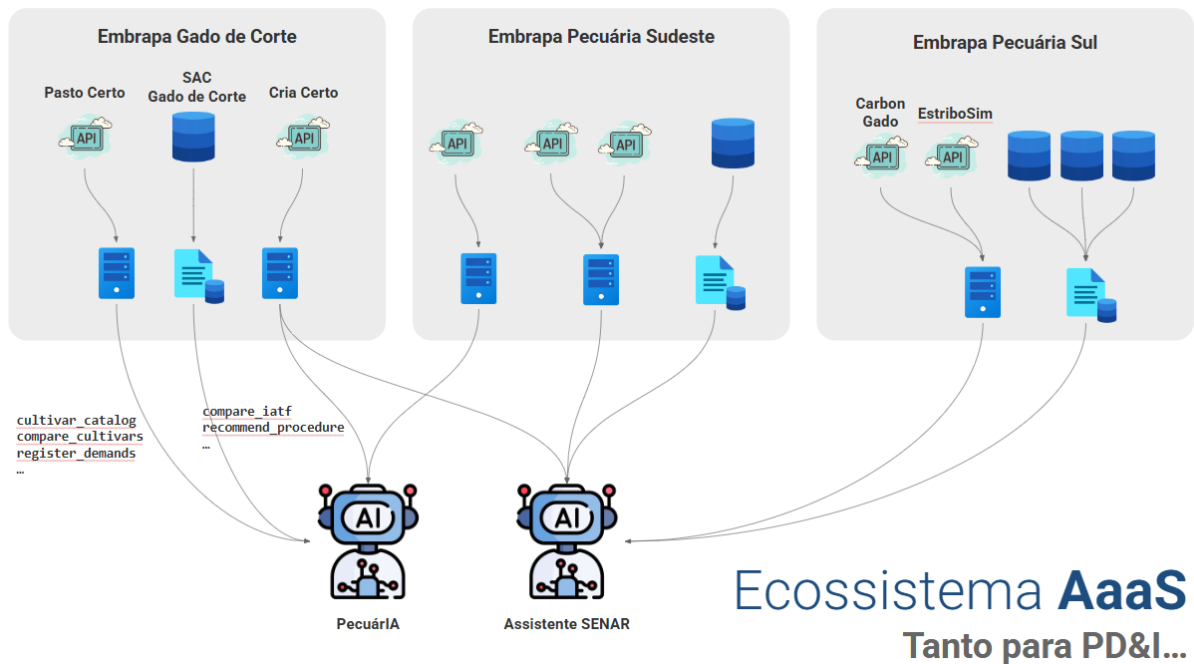
Tais modelos aceleram a fase embrionária de novos projetos e asseguram que, desde o primeiro *commit*, o ativo digital guarde conformidade com as diretrizes de arquitetura institucional. Ao centralizar a governança técnica e, simultaneamente, descentralizar a inovação, o **Embrapa I/O** atua como um catalisador para a edificação de um ecossistema digital agropecuário interoperável. Nesse ambiente, ativos como o agente conversacional podem ser desenvolvidos, convalidados e escalonados sob condições de segurança jurídica e resiliência operacional.

**2.4.2 Arquitetura de referência para soluções de inteligência artificial**

A Arquitetura de Referência em IA Generativa, estabelecida pela Gerência de Tecnologia da Informação (GTI) da Embrapa, preconiza um modelo descentralizado de "Agente como Serviço" (*Agent as a Service – AaaS*), projetado para operar sobre a infraestrutura de *clusters* do Embrapa I/O. Esse paradigma de arquitetura orientada a serviços permite que as competências de IA sejam instanciadas como microcomponentes independentes e escalonáveis, garantindo que o agente conversacional possa ser consumido por múltiplas

interfaces institucionais sem a necessidade de replicação de código. Tal abordagem favorece a otimização de recursos computacionais e a consistência das respostas em diferentes pontos de contato com o extensionista.

**Figura 05 – Exemplo de Ecossistema de AaaS em PD&I**

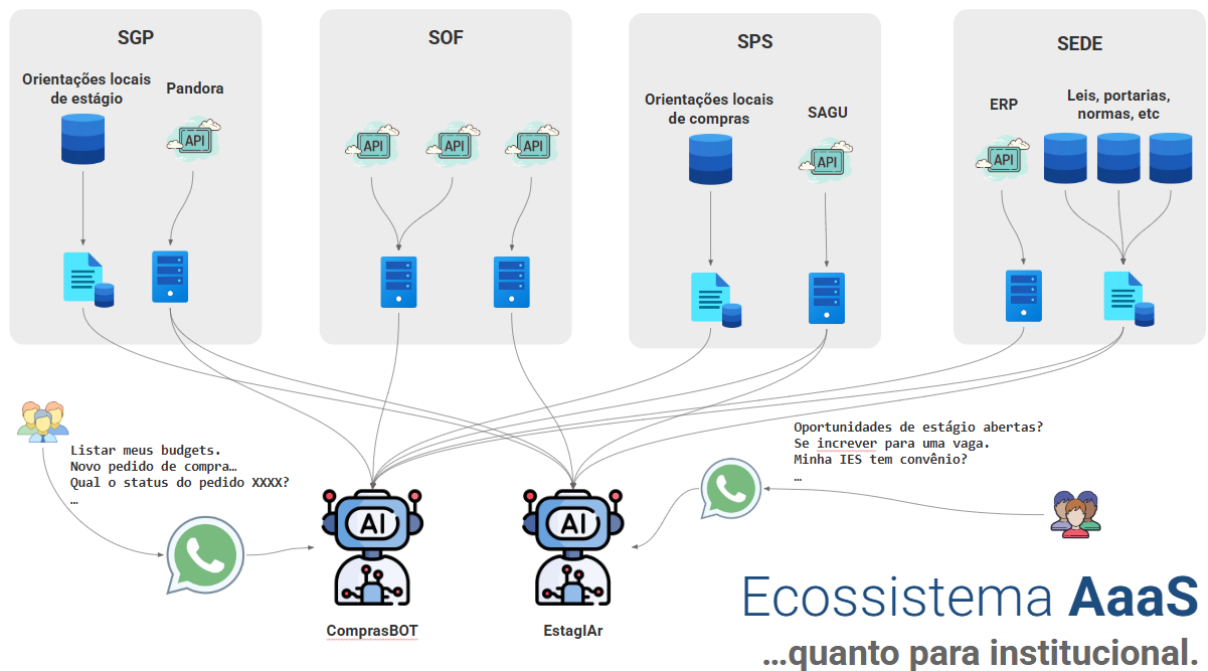


Fonte: Embrapa. (2025c).

Distintamente de abordagens monolíticas, essa arquitetura adota uma filosofia modular e agnóstica, fundamentada na integração de componentes de código aberto (*open source*) para a composição de ecossistemas de inteligência.

As premissas basilares incluem o emprego de **contêineres** para assegurar a portabilidade entre ambientes, a independência em relação a provedores de modelos de linguagem — mitigando o risco de aprisionamento tecnológico (*vendor lock-in*) — e a priorização de ferramentas de baixo custo para a validação célere de Provas de Conceito (PoCs) e Produtos Mínimos Viáveis (MVPs) (Embrapa, 2025b; Carromeu et al., n.d.).

**Figura 06 – Exemplo de Ecosystema de AaaS institucionais**



Fonte: Embrapa. (2025c).

Sob a perspectiva técnica, a referência arquitetural estrutura-se mediante a orquestração de fluxos de trabalho que integram o padrão RAG (*Retrieval-Augmented Generation*) e o protocolo MCP. Nesse delineamento, o agente conversacional não opera de forma isolada; ele atua como um orquestrador que consulta servidores de contexto (MCP Servers) — incumbidos da conexão com bancos de dados, APIs corporativas e repositórios documentais — previamente à submissão do *prompt* enriquecido ao Modelo de Linguagem de Grande Escala (LLM).

Essa segregação de responsabilidades permite que a lógica de negócio e os ativos informacionais permaneçam sob o controle estrito da **Embrapa**, enquanto o LLM desempenha exclusivamente a função de motor de inferência semântica e síntese textual (Embrapa, 2025b; Nunes et al., 2025).

A arquitetura de referência endereça, de forma pragmática, os riscos inerentes à IA Generativa, estabelecendo diretrizes para a soberania de dados e a estrita observância à Lei

Geral de Proteção de Dados (LGPD). O modelo preconiza, preferencialmente, a execução local de LLMs (*on-premises*) em servidores da infraestrutura própria da **Embrapa** para o processamento de informações sensíveis ou estratégicas, mitigando vulnerabilidades de vazamento de dados para APIs públicas de terceiros. Na eventualidade de o uso de modelos externos ser imperativo, a arquitetura impõe camadas de anonimização e mecanismos de controle de fluxo, assegurando que a inovação tecnológica ocorra sob um perímetro de segurança jurídica e institucional robusto (Embrapa Agricultura Digital, 2025; Embrapa, 2025b).

### 2.4.3 Aplicabilidade da arquitetura proposta ao artefato desenvolvido

A aplicabilidade da Arquitetura de Referência em IA Generativa da Embrapa ao artefato desenvolvido nesta pesquisa materializa-se na adoção integral do modelo de "Agente como Serviço" (*Agent as a Service – AaaS*), no qual o assistente conversacional é concebido como um sistema modular orquestrado sobre a infraestrutura do Embrapa I/O. O desenvolvimento do protótipo derivou de *boilerplates* institucionais padronizados para soluções RAG (*Retrieval-Augmented Generation*).

Tal abordagem favoreceu, desde a sua gênese, o alinhamento da ferramenta aos requisitos de interoperabilidade e segurança preconizados para ativos digitais públicos (Embrapa, 2025b). Essa estratégia de alinhamento permite que o artefato opere de forma descentralizada nos *clusters* da plataforma, beneficiando-se de *pipelines* automatizados de integração e entrega contínuas (CI/CD) para favorecer a manutenibilidade e a evolução ágil do *software* (Carroneu et al., n.d.).

No plano da implementação técnica, a arquitetura proposta evidencia capacidade de integrar o vasto acervo da Plataforma Ater+ Digital à inteligência inferencial dos LLMs, sem abdicar da governança de dados. A utilização do n8n como orquestrador central, operando em

ambiente *self-hosted* (auto-hospedado), concretiza a premissa de soberania digital, favorecendo que o fluxo de dados sensíveis e a lógica de negócio permaneçam sob controle institucional estrito.

A interlocução entre o agente e a base de conhecimento — estruturada no banco vetorial Qdrant — é mediada pelo MCP, o que confere ao sistema a granularidade necessária para conectar-se a fontes heterogêneas de informação ou substituir componentes tecnológicos sem a necessidade de refatoração profunda do código-fonte, conforme demonstrado em estudos preliminares de arquiteturas abertas para a Embrapa (Nunes et al., 2025; Embrapa, 2025b).

A aderência a essa arquitetura de referência projeta o artefato não apenas como uma solução pontual para o problema da recuperação de informações, mas como um componente com potencial de evolução, integração e expansão no ecossistema de inovação institucional.

A estrutura fundamentada em microsserviços e contêineres favorece a expansão prospectiva das capacidades do agente, podendo viabilizar, por exemplo, a integração com APIs de previsão climática ou sistemas de crédito rural mediante a implementação de novos servidores MCP e integrações pertinentes. Por conseguinte, a aplicação prática dessa arquitetura no desenvolvimento do assistente conversacional atua como um caso de uso demonstrativo (*showcase*) para a viabilidade de sistemas de IA soberanos e auditáveis na agricultura tropical, pavimentando o caminho para a transição da Ater+ Digital em uma plataforma de serviços inteligentes de elevado valor agregado.

### 3. Metodologia

O delineamento metodológico desta pesquisa fundamenta-se na natureza aplicada, visto que o estudo transcende a teorização abstrata para orientar-se, primordialmente, à resolução de um óbice concreto: a fragmentação e a complexidade na recuperação da informação técnica no âmbito da Plataforma Ater+ Digital. Conforme asseveram Prodanov e Freitas (2013), a pesquisa

aplicada objetiva a geração de conhecimentos voltados à aplicação prática, dirigidos à solução de problemas específicos da realidade contemporânea. Nesse sentido, o trabalho propõe o desenvolvimento de um artefato tecnológico — um agente conversacional pautado em Inteligência Artificial Generativa — como uma intervenção direta para otimizar o fluxo de conhecimento entre a base científica da Embrapa e os extensionistas rurais, alinhando-se aos preceitos de relevância prática e utilidade preconizados por Gil (2019).

Quanto à abordagem, o estudo adota uma perspectiva mista (quali-quantitativa), condição precípua para capturar a complexidade do fenômeno investigado. A dimensão qualitativa manifesta-se na exegese da fidelidade semântica e da coerência textual das respostas geradas pelo agente em face dos manuais técnicos oficiais.

Por outro lado, a componente quantitativa, de caráter descritivo, foca na mensuração objetiva de indicadores de desempenho (*Key Performance Indicators* – KPIs), tais como latência, estabilidade e acurácia em cenários controlados (Marconi & Lakatos, 2021). Adicionalmente, a pesquisa possui caráter exploratório e descritivo, visto que investiga a integração de arquiteturas RAG ao protocolo MCP no ecossistema institucional da Embrapa, pormenorizando as propriedades funcionais do sistema em contraposição aos mecanismos convencionais de recuperação da informação (Gil, 2019).

Operacionalmente, a investigação estrutura-se em etapas interdependentes que mimetizam o ciclo de construção de artefatos em Sistemas de Informação (Hevner et al., 2004). O percurso inicia-se com a pesquisa bibliográfica, voltada à fundamentação teórica sobre IA e governança de dados, seguida pela pesquisa documental para a seleção e curadoria do *corpus* na Plataforma Ater+ Digital. Nesta fase, aplicam-se técnicas de análise de conteúdo para a estruturação lógica dos dados (Bardin, 2016).

Posteriormente, pormenoriza-se a engenharia do sistema, detalhando a orquestração *low-code* e a vetorização da base de conhecimento, processo que culmina na validação técnica.

Nesta última etapa, o artefato é submetido a testes de *benchmarking* em face de um conjunto de dados de referência (*ground truth*), assegurando o rigor científico e a fidedignidade dos resultados produzidos.

Para o atingimento de seus objetivos, a pesquisa utiliza como fonte primária de dados o acervo técnico público da Plataforma Ater+ Digital, selecionada por sua representatividade e caráter oficial como repositório de conhecimento da Embrapa. A plataforma assegura a inclusão de informações convalidadas cientificamente, condição imperativa para garantir a soberania e a fidedignidade das respostas geradas pelo agente conversacional.

Complementarmente, foram consultadas as documentações técnicas oficiais de ecossistemas de Inteligência Artificial — Google Vertex AI, Qdrant e n8n — e bases acadêmicas de alto impacto, como Web of Science (WoS) e Google Acadêmico (GA), para fundamentar a arquitetura RAG (*Retrieval-Augmented Generation*).

Tal abordagem amplia o escopo da investigação ao contemplar o estado da arte em tecnologias de recuperação da informação, contexto no qual a aplicação de LLMs na extensão rural ainda se encontra em fase de consolidação teórica e prática (Lewis et al., 2020; Embrapa, 2024). Os pormenores acerca do volume de dados processados e da arquitetura da solução serão apresentados nas seções 3.2 e 3.3, respectivamente, em estrita observância ao ciclo de desenvolvimento do artefato.

As etapas de curadoria de conteúdo e levantamento bibliográfico técnico foram executadas entre dezembro de 2025 e janeiro de 2026, utilizando descritores em português e inglês consubstanciados ao domínio da solução: *Retrieval-Augmented Generation*, *Large Language Models*, *Vector Databases*, *Agricultural Extension*, *Artificial Intelligence*, *Agrobusiness* e *Prompt Engineering*. Tal escolha justifica-se pela imperatividade de se utilizar tecnologias de fronteira e documentação técnica vigente.

Para a composição da base de conhecimento do agente, foram selecionados manuais técnicos, conteúdos textuais e recursos digitais da Plataforma Ater+ Digital, priorizando-se dados estruturados que permitissem a vetorização com integridade semântica. Esses critérios asseguram o rigor metodológico e a coerência temática, facultando ao artefato a capacidade de transcender o processamento de dados brutos para a entrega de respostas contextualizadas e fidedignas à práxis do extensionista rural.

A opção pela pesquisa aplicada, alicerçada em métodos de prototipagem e validação experimental, reflete o propósito de edificar uma solução tangível e crítica para o problema da recuperabilidade da informação. Em um domínio tecnológico de célere evolução, a implementação prática de um assistente virtual provê um meio objetivo para testar a aderência de modelos de linguagem às demandas do setor público; concomitantemente, a análise qualitativa das respostas permite perscrutar os limites, as alucinações e as potencialidades da ferramenta.

Assim, o método adotado amalgama a precisão da engenharia de software à sensibilidade inerente à extensão rural, favorecendo uma percepção integrada entre a inovação tecnológica e a utilidade prática. O processo de desenvolvimento observou critérios de transparência e reprodutibilidade, assegurando a rastreabilidade das fontes consultadas pelo modelo e a estrita coerência entre os objetivos pretendidos e o artefato entregue.

Na sequência, expõe-se a classificação e a natureza da pesquisa, pormenorizando os constructos que sustentam a opção pelo método teórico-bibliométrico e suas respectivas implicações para a exegese científica e a gestão da inovação. Essa etapa pauta-se no mapeamento sistemático da literatura, permitindo identificar as tendências tecnológicas e as lacunas teóricas que justificam o desenvolvimento do artefato.

Essa fundamentação não apenas baliza o rigor metodológico, mas também fornece os subsídios necessários para a análise gerencial da solução, assegurando que o protótipo esteja alinhado tanto ao estado da arte quanto às diretrizes estratégicas do setor público.

### **3.1 Classificação e Natureza da Pesquisa**

A presente investigação classifica-se, quanto à sua natureza, como pesquisa aplicada, visto que transcende a mera acumulação de constructos teóricos abstratos para orientar-se, deliberadamente, à resolução de um óbice concreto e imediato: a ineficiência na recuperação da informação técnica em plataformas de extensão rural. Conforme asseveram Prodanov e Freitas (2013), a pesquisa aplicada objetiva a geração de conhecimentos voltados à aplicação prática, dirigidos à solução de problemas específicos da realidade contemporânea. Sob essa ótica, o estudo propõe uma intervenção direta na Plataforma Ater+ Digital, mediante o desenvolvimento de uma solução tecnológica que otimize o fluxo de conhecimento entre a base científica agropecuária e os extensionistas rurais.

Essa natureza aplicada materializa-se mediante a adoção da metodologia *Design Science Research* (DSR). Segundo Hevner et al. (2004), o DSR constitui o paradigma precípua para investigações em Sistemas de Informação que compreendem a concepção e a avaliação de artefatos de TI — no escopo deste estudo, o agente conversacional pautado em IA Generativa — concebidos para mitigar problemas organizacionais diagnosticados. Destarte, a produção de conhecimento científico processa-se de forma indissociável da construção e validação do *software*, de modo que a utilidade e a eficácia do artefato figuram como os critérios cardinais de aferição da qualidade e da relevância da pesquisa.

No que tange aos objetivos, o estudo assume um caráter exploratório. De acordo com Gil (2019), pesquisas exploratórias visam proporcionar maior familiaridade com o problema, sendo indicadas quando o tema eleito é escassamente explorado ou carece de sistematização

robusta. A convergência entre arquiteturas *Retrieval-Augmented Generation* (RAG) e o protocolo MCP, no domínio específico da extensão rural pública brasileira, constitui uma fronteira tecnológica incipiente. Por conseguinte, a investigação perscruta o comportamento dessas tecnologias emergentes quando submetidas à exegese técnica e à especificidade vocabular dos ativos informacionais da Embrapa.

Simultaneamente, a investigação assume uma dimensão descritiva. Conforme asseveram Marconi e Lakatos (2021), a pesquisa descritiva tem por escopo primordial a delimitação das características de determinado fenômeno ou o estabelecimento de relações entre variáveis. No âmbito deste estudo, descreve-se funcionalmente o comportamento do agente conversacional, pormenorizando sua arquitetura, seus fluxos de orquestração no n8n e a taxonomia das respostas geradas em contraposição aos métodos convencionais de busca. A descrição exaustiva das etapas de vetorização e recuperação (*retrieval*) demonstra-se imperativa para assegurar a reprodutibilidade do experimento e a fidedignidade dos resultados alcançados.

No que tange à abordagem do problema, adota-se um delineamento misto (quali-quantitativo), fundamentado na premissa de que a complexidade do objeto de estudo transcende a capacidade de captura de uma única ótica analítica. A vertente qualitativa é empregada na fase de exegese documental e na avaliação semântica das respostas geradas pelo agente. Conforme assevera Godoy (1995), a pesquisa qualitativa faculta a compreensão de fenômenos em profundidade.

Nesta investigação, tal abordagem operacionaliza-se por meio da Análise de Conteúdo, conforme a sistematização de Bardin (2016), permitindo categorizar e interpretar se a terminologia e o conteúdo técnico providos pelo assistente virtual guardam fidedignidade com o contexto agrônomo. Tal análise permite aferir nuances de precisão e autoridade científica que métricas puramente numéricas não logram capturar integralmente.

Complementarmente, a abordagem quantitativa é empregada em caráter descritivo para mensurar o desempenho técnico do artefato. Serão coletados e analisados dados numéricos concernentes a indicadores de performance computacional (*Key Performance Indicators – KPIs*), tais como o tempo de latência (mensurado em milissegundos), o consumo volumétrico de *tokens* e a **acurácia** (*accuracy*) em testes controlados frente a um conjunto de dados de referência (*ground truth*). Essa dimensão quantitativa confere objetividade e rigor analítico à avaliação, permitindo aferir a estabilidade e a eficiência da solução proposta em condições de estresse e carga operacional (Prodanov & Freitas, 2013).

A integração dessas abordagens faculta a triangulação metodológica, na qual os dados quantitativos sobre a eficiência da recuperação da informação sustentam e corroboram as inferências qualitativas acerca da utilidade da ferramenta para o extensionista e para o produtor rural. Essa convergência demonstra-se imperativa para o pleno atendimento à questão de pesquisa, evidenciando não apenas o funcionamento técnico do sistema (viabilidade), mas a entrega efetiva de valor informacional (relevância) ao público-alvo da Ater+ Digital.

Tal simbiose metodológica assegura que o artefato transcenda a condição de protótipo laboratorial para consolidar-se como uma solução robusta e aderente às complexidades do ecossistema agrícola brasileiro, conforme sintetizado no Quadro 01 abaixo.

#### **Quadro 01 - Fundamentação da escolha metodológica**

<b>Critério</b>	<b>Descrição Operacional</b>	<b>Fundamentação (Suas Referências)</b>
Fonte Institucional	Documentos públicos da Plataforma Ater+ Digital.	Embrapa (2024, 2025).

<b>Critério</b>	<b>Descrição Operacional</b>	<b>Fundamentação (Suas Referências)</b>
Validação Técnica	Filtro de integridade e qualidade (OCR).	Carromeu et al. (2023) — Governança Embrapa I/O.
Análise de Dados	Categorização temática e interpretação.	Bardin (2016) e Gil (2019).
Mitigação de Erros	Controle de alucinações via RAG.	Lewis et al. (2020) e Gao et al. (2023) .
Interoperabilidade	Integração via Model Context Protocol.	Hou et al. (2025) — Open Ecosystems.
Maturidade	Escalas TRL e MRL para Ativos Digitais.	Embrapa (2025a) — Macroprocesso.

Fonte: Elaborado pelo autor (2026).

### 3.2 Delimitação do Estudo e Corpus Documental

A delimitação do estudo concentra-se no ecossistema da Plataforma Ater+ Digital, tecnologia, pública, oficial da Embrapa dedicada à transferência de tecnologia e ao suporte à extensão rural. A escolha desse ambiente justifica-se por sua relevância pragmática como repositório centralizador do conhecimento agropecuário tropical, congregando informações convalidadas por décadas de investigação científica.

O recorte temporal para a extração de dados e a definição do *corpus* encerrou-se em 31 de janeiro de 2026, data estabelecida como marco de congelamento (*snapshot*) da base de conhecimento para a parametrização e validação do artefato. Conforme preconiza Gil (2019), a delimitação precisa do universo de pesquisa é imperativa para assegurar a exequibilidade do estudo e a fidedignidade dos resultados, mitigando o risco de dispersão em fontes extraoficiais ou de integridade informacional questionável.

O *corpus* documental foi constituído por duas camadas de dados complementares. A primeira camada abrangeu a totalidade do conteúdo textual público, estruturado em formato HTML (páginas *web*), proveniente das seções temáticas da plataforma. Para sistematizar a curadoria e assegurar a rastreabilidade dessa primeira camada — critério essencial na pesquisa documental, segundo assevera Bardin (2016) —, o conteúdo foi categorizado em níveis hierárquicos de informação. Essa taxonomia documental permitiu mapear a proveniência de cada fragmento textual, garantindo que a base de conhecimento do agente reflita fielmente a arquitetura de informação institucional da Embrapa.

Para cada cadeia produtiva ou domínio temático, procedeu-se à extração de subitens essenciais, abrangendo categorias como: "Características da cultura/criação", "Sistemas de Produção", "Manejo", "Publicações", "Vídeos" e "Cursos". Esta segmentação semântica prévia revela-se crucial para o desempenho da arquitetura **RAG** (*Retrieval-Augmented Generation*), uma vez que provê metadados estruturados que auxiliam o algoritmo de recuperação a distinguir, de forma inequívoca, uma recomendação de manejo para "Milho" de uma diretriz para "Trigo". Tal refinamento mitiga a ambiguidade contextual e assegura que o agente recupere a informação pertinente com maior acurácia e relevância.

**Quadro 02 – Mapeamento e Estruturação do Corpus da Plataforma Ater+ Digital**

<b>Categoria Principal</b>	<b>Subtemas / Cadeias Produtivas</b>	<b>Tipo de Conteúdo Extraído</b>	<b>Fonte (URL Base)</b>
<b>Cadeias Produtivas Vegetais</b>	Alho, Arroz, Batata-doce, Café, Caju, Feijão, Feijão-Caupi (ambas juntadas apenas como feijão), Mandioca, Manga, Melão, Tomate de Mesa, Uva de Mesa	Texto HTML estruturado (menus e páginas técnicas), com links para publicações técnicas e científicas, vídeos, áudios/podcasts, cursos, manuais, ferramentas digitais e conteúdos complementares	<i>https://atermaisdigital.cnptia.embrapa.br/web/[cadeia]</i>
<b>Cadeias Produtivas Animais</b>	Apicultura, Bovino de Leite, Caprinos e Ovinos, Frango, Peixes, Suínos	Texto HTML estruturado (sistemas de produção, sanidade, nutrição e manejo), com links para publicações, vídeos, áudios/podcasts, cursos e ferramentas digitais	<i>atermaisdigital.cnptia.embrapa.br/web/[cadeia]</i>
<b>Temas Transversais</b>	Integração Lavoura-Pecuária-Floresta – ILPF, Metodologias de Ater Digital, Mudanças Climáticas,	Texto HTML conceitual e técnico, com links para publicações, vídeos, áudios, cursos, guias práticos, legislações,	<i>atermaisdigital.cnptia.embrapa.br/web/[tema]</i>

Categoria Principal	Subtemas / Cadeias Produtivas	Tipo de Conteúdo Extraído	Fonte (URL Base)
	Sistemas Agrícolas Tradicionais, Sistemas Agroflorestais e Nutrição e Saúde	ferramentas e materiais de capacitação	
<b>Institucional</b>	Ferramentas Digitais, Sobre nós, Como Participar e Parcerias	Texto HTML institucional, com links para metodologias, documentos, ferramentas digitais, apresentações, vídeos e contatos institucionais	<i>atermaisdigital.cn</i> <i>ptia.embrapa.br</i>

Fonte: Elaboração própria (2026), com base na estrutura da Plataforma Ater+ Digital.

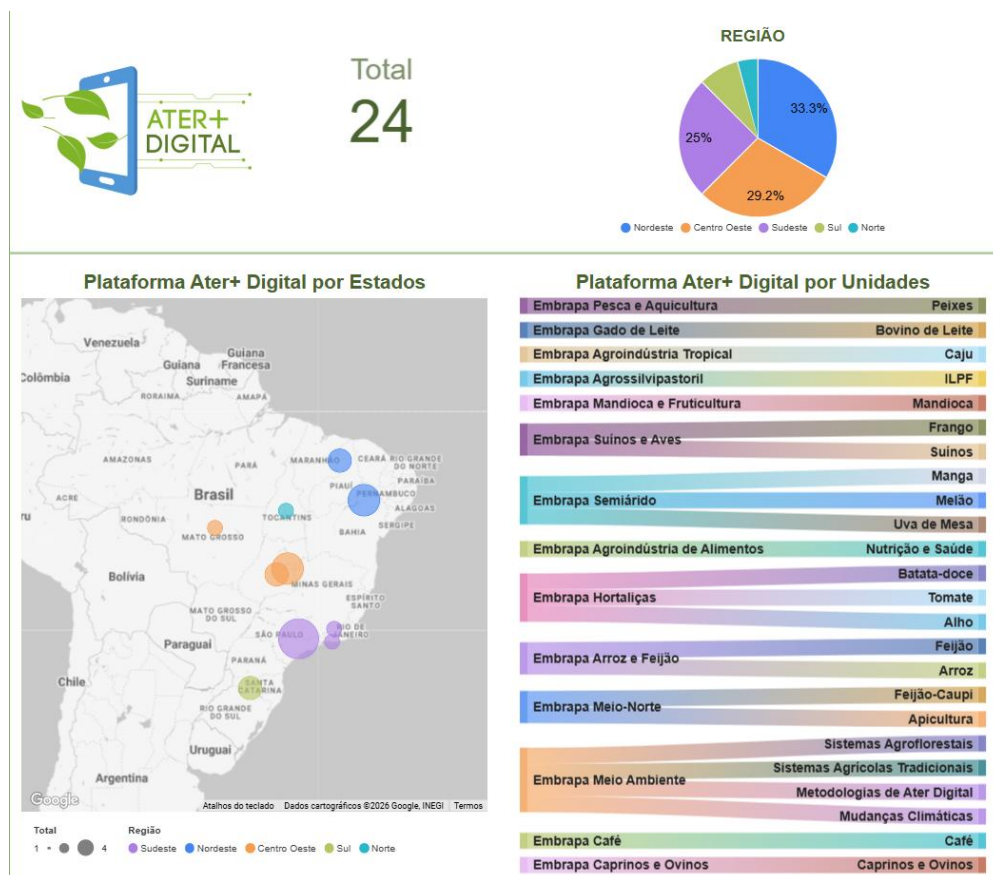
A segunda camada, de caráter documental não estruturado, incorporou o acervo público de publicações técnicas e científicas — abrangendo manuais, cartilhas, circulares e artigos — disponíveis para *download* em formato PDF nos respectivos *hubs* de conhecimento, no contexto da citada plataforma da Embrapa.

Conforme detalhado no Quadro 2, a seleção contemplou 18 cadeias produtivas (Alho, Apicultura, Arroz, Batata-doce, Bovinocultura de Leite, Café, Caju, Caprinocultura, Feijão, Feijão-Caupi, Avicultura de Corte, Mandiocultura, Manga, Melão, Piscicultura, Suinocultura, Tomate de Mesa e Uva de Mesa) e 6 temas transversais (Integração Lavoura-Pecuária-Floresta — ILPF, Metodologias de Ater Digital, Mudanças Climáticas, Sistemas

Agrícolas Tradicionais, Sistemas Agroflorestais e Nutrição e Saúde). Esta etapa totalizou um volume aproximado de 1,8 GB de dados brutos processados, exigindo técnicas robustas de extração de texto para garantir a integridade da base de conhecimento.

Essa abrangência temática confere ao estudo uma heterogeneidade representativa da diversidade agrônômica brasileira, permitindo avaliar a capacidade de generalização do modelo de IA em face de léxicos técnicos distintos e domínios de conhecimento especializados. Tal diversificação do *corpus* revela-se estratégica para aferir a robustez da arquitetura RAG, garantindo que o sistema mantenha a precisão semântica tanto em cadeias produtivas de ciclo curto quanto em temas transversais de alta complexidade conceitual, como a Integração Lavoura-Pecuária-Floresta (ILPF).

**Figura 07 – Página Ater+ Digital por Estados e Unidades da Embrapa**



Fonte: Embrapa. (s.d.-b).

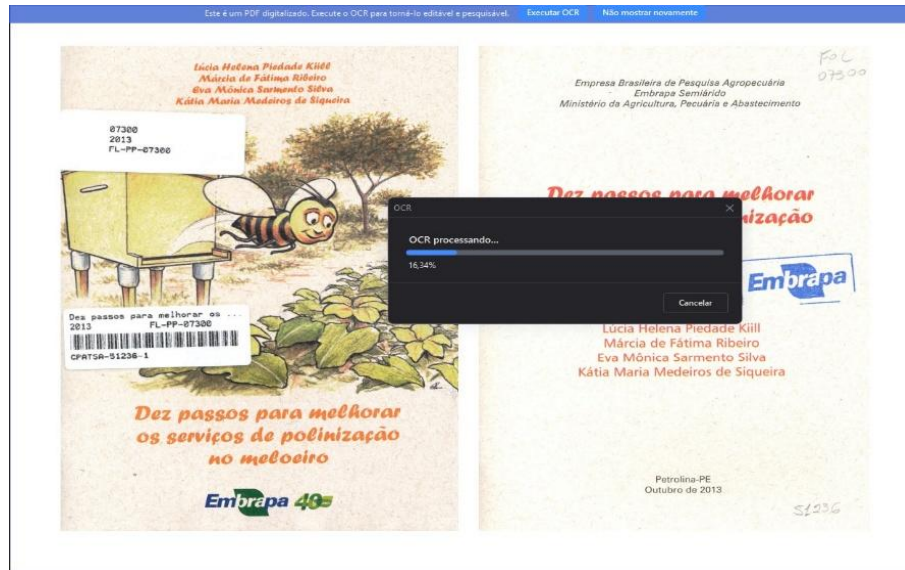
A qualidade do *corpus* foi assegurada por um fluxo de validação automatizada (*Data Quality Pipeline*), implementado na plataforma de orquestração n8n. Previamente à vetorização, cada fragmento de texto extraído foi submetido a um algoritmo de verificação de integridade (*garbage check*), programado para rejeitar, de forma autônoma, segmentos que apresentassem excesso de caracteres de controle, *bytes* nulos ou densidade de símbolos não alfanuméricos — indicadores típicos de anomalias de codificação ou falhas no reconhecimento óptico de caracteres (OCR).

Esta etapa de saneamento revela-se crítica para mitigar a poluição do espaço vetorial com ruídos que poderiam comprometer a fidelidade semântica e a precisão das respostas geradas pelo agente (Bardin, 2016).

Para assegurar a legibilidade e a indexação desse material heterogêneo, implementou-se um protocolo rigoroso de pré-processamento. Após a validação de integridade descrita anteriormente, os documentos em PDF constituídos originalmente por imagens digitalizadas foram submetidos a um processo de Reconhecimento Óptico de Caracteres (OCR), mediante o emprego do *software* Wondershare PDFelement (versão 10.2.5), assegurando a conversão para texto pesquisável (*searchable text*).

Divergindo de abordagens que se limitam à indexação de metadados, esta investigação optou pelo processamento integral do conteúdo (*full-text processing*). Tal escolha metodológica faculta ao sistema a capacidade de recuperar informações situadas em camadas profundas dos documentos técnicos, garantindo que detalhes granulares não sejam negligenciados pela arquitetura de busca.

### **Figura 08 – Exemplo de OCR em arquivo do corpus**



Fonte: Elaboração própria (2026)

Adicionalmente, para sistematizar a curadoria e assegurar a rastreabilidade, o conteúdo foi submetido a um processo de enriquecimento semântico assistido por Inteligência Artificial. Um modelo de linguagem de larga escala (LLM) auxiliar foi empregado para a extração automatizada de metadados estruturados de cada documento, identificando campos como Título, Autores, Ano e *Digital Object Identifier* (DOI).

Esta estruturação prévia revela-se fundamental para o desempenho da arquitetura RAG (*Retrieval-Augmented Generation*), uma vez que faculta ao algoritmo de recuperação não apenas a localização da resposta pertinente, mas a geração da citação fidedigna da fonte original (ex: "Conforme Silva et al., 2024"). Tal funcionalidade constitui um requisito indispensável para a credibilidade científica e a segurança informacional no âmbito da extensão rural (Embrapa, 2025c).

A relevância acadêmica desta delimitação reside na construção de um **Golden Dataset** (Conjunto de Dados de Ouro) estritamente técnico e institucional. Diferente de sistemas conversacionais genéricos, treinados em dados indistintos da rede mundial de computadores, o *corpus* deste estudo constitui uma "Verdade Terrestre" (*Ground Truth*) controlada e auditável.

Tal rigor metodológico mitiga drasticamente a ocorrência de "alucinações", uma vez que o modelo de linguagem é parametrizado para responder exclusivamente com fulcro nos fragmentos recuperados desse acervo delimitado. O Quadro 2 apresenta a síntese da volumetria do *corpus* documental processado (arquivos PDF), segmentado por domínio temático.

**Quadro 03 – Distribuição e Volumetria do *Corpus* Documental (PDFs) Processado**

Tema / Cadeia	Qtd. Documentos	Tema / Cadeia	Qtd. Documentos
Alho	10	Manga	19
Apicultura	21	Melão	15
Arroz	09	Metodologias Ater	06
Batata-doce	08	Mudanças Climáticas	08
Bovino de Leite	31	Nutrição e Saúde	14
Café	10	Piscicultura (Peixes)	44
Caju	46	Sist. Agrícolas Trad.	27
Feijão*	10	Sist. Agroflorestais	25
Feijão-Caupi*	01	Suínos	11
Frango	24	Tomate de Mesa	8
ILPF (Publ. + Transcrições)*	538	Uva de Mesa	56
Mandioca	18	<b>TOTAL</b>	959

*\*Nota: O tema ILPF inclui 21 publicações e 517 transcrições de vídeos oriundas de projeto correlato integrado via MCP. Feijão e Feijão-Caupi foram juntados em um único RAG e MCP chamado feijão. Fonte: Elaborado pelo autor (2026).*

A composição do *corpus* documental foi regida por critérios rigorosos de seleção e exclusão, visando garantir a validade interna da base de conhecimento e a precisão da recuperação semântica. Durante a etapa de ingestão de dados, procedeu-se ao descarte sistemático de arquivos que apresentassem inacessibilidade na origem, como links corrompidos ou erros de *download* (HTTP 404), bem como documentos com insuficiência de legibilidade técnica, caracterizada por uma taxa de erro de OCR superior a 30% ou ruído visual excessivo que comprometeria a vetorização.

Adicionalmente, aplicou-se um filtro de redundância para remover arquivos idênticos ou versões preliminares, evitando distorções de relevância no espaço vetorial. Ressalta-se que, embora a atualização técnica tenha sido uma premissa, optou-se por não aplicar um corte temporal rígido, em relação aos documentos de publicações científicas disponíveis na plataforma; tal decisão justifica-se pela existência de publicações institucionais consideradas 'padrão ouro' que, mesmo publicadas em datas pretéritas, permanecem como referências técnicas vigentes e indispensáveis para a fundamentação das cadeias produtivas analisadas. Também foram excluídos elementos residuais de interface capturados durante o *scraping* que não possuíam densidade informacional aderente aos eixos temáticos do estudo.

#### Quadro 04 - Critérios metodológicos de composição do corpus documental

<b>Critério</b>	<b>Descrição operacional</b>	<b>Justificativa metodológica</b>
<b>Fonte institucional</b>	Documentos exclusivamente provenientes da Plataforma Ater+ Digital (Embrapa).	Garantir confiabilidade, rastreabilidade e legitimidade da base documental.

<b>Critério</b>	<b>Descrição operacional</b>	<b>Justificativa metodológica</b>
<b>Recorte temporal</b>	Documentos publicados ou atualizados até 31 de janeiro de 2026.	Assegurar atualização técnica e aderência ao contexto atual da pesquisa.
<b>Tipologia documental</b>	Conteúdo HTML estruturado e publicações em PDF (manuais, cartilhas e artigos).	Manter a homogeneidade estrutural do corpus para processamento em RAG.
<b>Exclusão por duplicidade</b>	Remoção de arquivos idênticos ou versões redundantes.	Evitar distorção quantitativa e repetição vetorial no banco de dados.
<b>Exclusão por falha técnica</b>	Documentos com erro de OCR superior a 30% ou ilegibilidade.	Garantir a qualidade semântica da vetorização e integridade da base.
<b>Exclusão temática</b>	Arquivos sem aderência aos eixos definidos na categorização (scraping residual).	Preservar a coerência analítica com os objetivos específicos do estudo.

Fonte: Elaborado pelo autor (2026), com base nos critérios de Bardin (2016).

A delimitação adotada equilibra a complexidade técnica da IA Generativa com as premências informacionais da extensão rural. Ao concentrar-se no conteúdo *web* estruturado das principais cadeias produtivas, o estudo viabiliza a concepção de um protótipo funcional de alta fidelidade, apto a demonstrar a exequibilidade da arquitetura RAG em cenários reais.

Essa abordagem estabelece uma fundação escalável para a incorporação prospectiva de documentos complexos e acervos multimídia na Plataforma Ater+ Digital. Com essa estrutura, o artefato projeta-se como um catalisador da transformação digital, transformando repositórios estáticos em fluxos dinâmicos de conhecimento especializado.

#### **Quadro 05 - Inventário e Caracterização do Corpus Documental Processado**

<b>Categoria do Corpus</b>	<b>Fonte Primária</b>	<b>Período de Coleta</b>	<b>Volume/Composição Estimada</b>
Documentação Web	Plataforma Ater+ Digital	Jan/2026	24 Cadeias Produtivas e Temas Transversais (HTML)
Acervo Técnico	Plataforma Ater+ Digital	Vigentes até jan./2026	Manuais, Cartilhas e Artigos (PDF)
Dados Multimídia	Protocolo MCP (Vídeos)	Jan/2026	Transcrições técnicas de ILPF já disponíveis em um MCP interno
Estrutura de Apoio	Extração via LLM	Jan/2026	Metadados e Esquemas JSON

*Fonte: Elaborado pelo autor (2026).*

O Quadro 05 consolida os ativos informacionais que compõem o ecossistema de dados desta pesquisa. Enquanto os critérios estabelecidos no Quadro 04 garantiram o rigor na seleção, este inventário detalha a diversidade tipológica necessária para testar a flexibilidade da arquitetura RAG integrada ao MCP. A convergência desses documentos técnicos com metadados estruturados permite que o processamento subsequente — detalhado na seção de

procedimentos de coleta — transcenda a simples busca por palavras-chave, atingindo um nível de recuperação semântica contextualizada para a extensão rural.

### 3.3 Procedimentos Metodológicos e Desenvolvimento do Artefato

A execução dos procedimentos metodológicos estruturou-se em um ciclo iterativo de engenharia de dados e desenvolvimento de *software*, em estrita consonância com as práticas de *DevOps* e integração contínua (CI/CD) preconizadas pela plataforma Embrapa I/O. O percurso técnico foi expandido para contemplar a hibridização de dados estruturados (*web*) e o vasto acervo documental não estruturado (arquivos PDF) da plataforma. O fluxo operacional subdividiu-se em quatro macroetapas: (i) ingestão de dados multiformato e pré-processamento via OCR; (ii) validação algorítmica de integridade e enriquecimento semântico; (iii) vetorização temática e indexação; e (iv) orquestração de sistemas multiagentes via protocolos de interoperabilidade. Essa abordagem sistemática visou converter a Plataforma Ater+ Digital em uma base de conhecimento auditável e semanticamente consultável (Hevner et al., 2004; Embrapa, 2025b).

Para a materialização do artefato, denominado "Consultor Digital Ater+ Digital", a arquitetura evoluiu de um sistema RAG monolítico para uma Arquitetura Hierárquica de Agentes. O sistema emprega o padrão *Retrieval-Augmented Generation* (RAG), orquestrado na plataforma *low-code* n8n, na qual um agente principal (*Supervisor Agent*) delega tarefas de recuperação a subagentes especializados mediante o MCP.

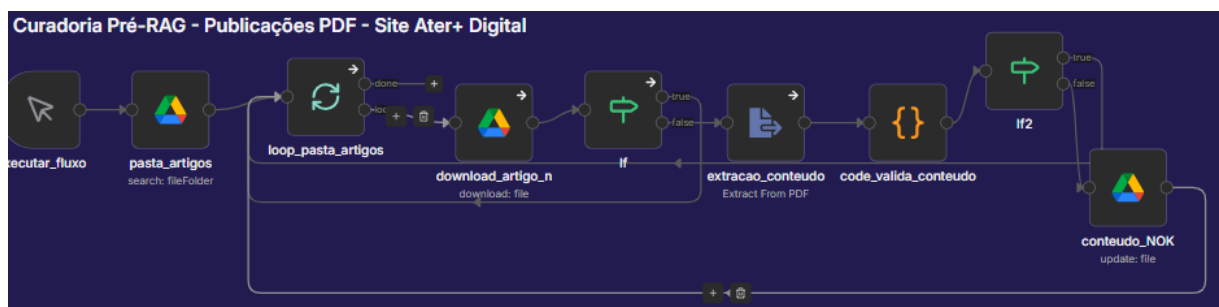
O ecossistema de desenvolvimento integrou soluções de vanguarda, como os modelos de *embedding* do Google e o banco vetorial Qdrant, parametrizados para assegurar a escalabilidade, a precisão na citação de fontes científicas e a estrita observância à soberania digital da Embrapa (Lewis et al., 2020; Carromeu et al., n.d.).

### 3.3.1 Extração, tratamento e segmentação dos dados

A construção da base de conhecimento foi operacionalizada por meio de um *pipeline* de Engenharia de Dados do tipo ETL (*Extract, Transform, Load*), concebido para processar a heterogeneidade documental da Plataforma Ater+ Digital, o qual mimetiza as fases de Pré-análise e Exploração do Material propostas por Bardin (2016).

O processo de extração transcendeu o conteúdo estruturado das páginas *web* (HTML), expandindo-se para a captura recursiva de publicações técnicas e científicas em formato PDF, alocadas nos repositórios das cadeias produtivas. Para mitigar o óbice de arquivos digitalizados obsoletos ou não indexáveis, implementou-se uma etapa de pré-processamento via Reconhecimento Óptico de Caracteres (OCR), mediante o emprego do *software* Wondershare PDFelement (versão 10.2.5). Essa conversão prévia assegurou que o conteúdo integral dos documentos, inclusive tabelas e quadros técnicos, fosse transmutado em texto processável por máquina (*machine-readable*), superando as limitações de acessibilidade intrínsecas aos acervos históricos da extensão rural.

**Figura 09 – Curadoria de documentos PDF**

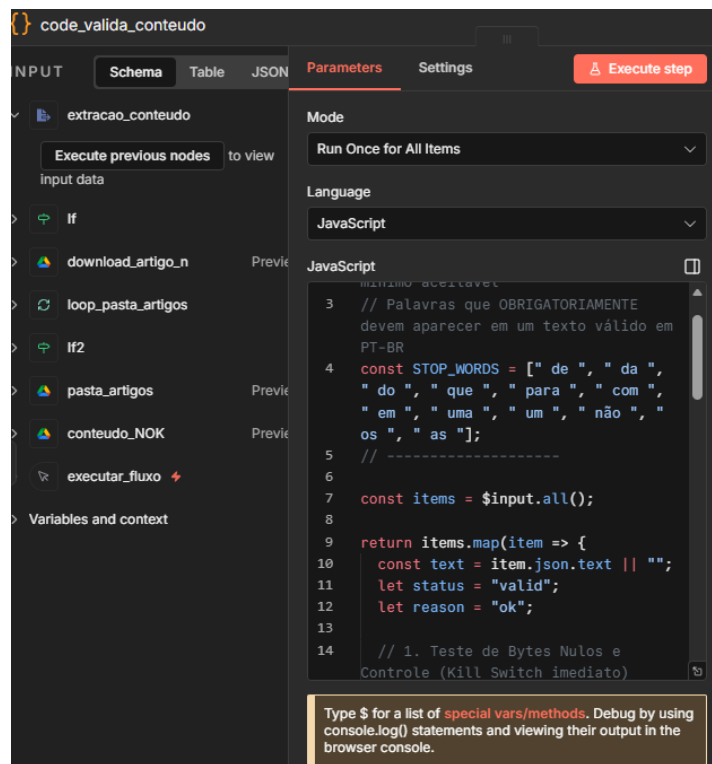


Fonte: Elaboração própria (2026)

Após a ingestão bruta, os dados foram submetidos a um rigoroso protocolo de Validação de Integridade (*Data Quality Check*), executado por um algoritmo customizado em JavaScript no ambiente da plataforma de orquestração.

Este *script* de saneamento aplicou filtros lógicos para a rejeição autônoma de documentos com anomalias de processamento, descartando arquivos que apresentassem: (a) *bytes* nulos ou caracteres de controle corrompidos; (b) densidade de símbolos não alfanuméricos, indicativo de falha na conversão do OCR; ou (c) escassez de conectivos essenciais da língua portuguesa, o que denota ruído de codificação (*encoding noise*). Essa etapa de saneamento de dados revela-se metodologicamente crítica para mitigar a contaminação do espaço vetorial com dados espúrios, salvaguardando a acurácia da recuperação subsequente.

**Figura 10 – Trecho de código Javascript para “Validação de Integridade” de arquivos**



```

3 // Palavras que OBRIGATORIAMENTE
  devem aparecer em um texto válido em
  PT-BR
4 const STOP_WORDS = [" de ", " da ",
  " do ", " que ", " para ", " com ",
  " em ", " uma ", " um ", " não ", "
  os ", " as "];
5 // -----
6
7 const items = $input.all();
8
9 return items.map(item => {
10   const text = item.json.text || "";
11   let status = "valid";
12   let reason = "ok";
13
14   // 1. Teste de Bytes Nulos e
    Controle (Kill Switch imediato)
  
```

Fonte: Elaboração própria (2026)

Uma vez convalidados, os textos foram submetidos a uma etapa de Enriquecimento Semântico Assistido por IA. Divergindo da abordagem convencional de indexação cega (*blind indexing*), cada documento foi processado por um Modelo de Linguagem de Larga Escala (LLM), parametrizado especificamente para a extração estruturada de metadados bibliográficos.

O modelo identificou e segregou atributos fundamentais, como Título, Autores, Ano de Publicação e DOI (*Digital Object Identifier*), gerando um objeto JSON padronizado para cada registro. Essa estruturação prévia revela-se fundamental para conferir autoridade científica ao agente conversacional, facultando ao sistema a capacidade de referenciar a autoria técnica com precisão (ex.: "Segundo Silva et al., 2024"), atendendo, assim, aos requisitos de rastreabilidade e integridade informacional exigidos pela neste tipo de aplicação.

A etapa final de preparação consistiu na segmentação textual (*chunking*), imperativa para adequar o volume de informações à janela de contexto dos modelos de linguagem de larga escala. Utilizou-se o algoritmo *Recursive Character Text Splitter*, parametrizado com blocos de 1.024 caracteres e uma sobreposição (*overlap*) de 128 caracteres.

A definição desses parâmetros visou a um equilíbrio adequado entre a granularidade da informação e a preservação do context: a dimensão do bloco é suficiente para abarcar um parágrafo técnico integral, ao passo que a sobreposição favorece que conceitos articulados na transição entre segmentos sejam preservados, mantendo a coesão semântica relevante para uma vetorização eficaz (Lewis et al., 2020; LangChain, 2024).

### **3.3.2 Vetorização (Embeddings) e indexação no Qdrant**

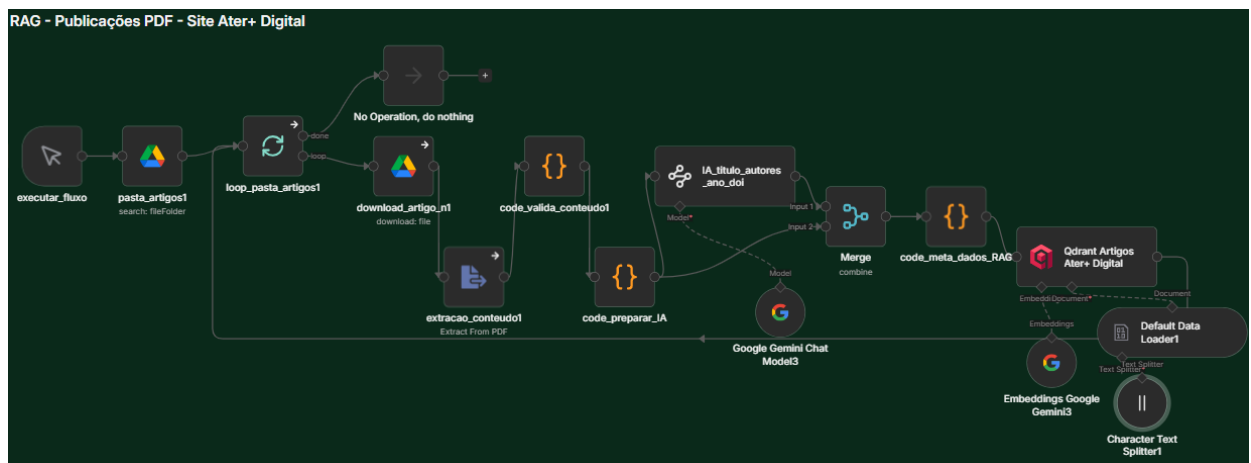
Após a segmentação, os blocos textuais foram submetidos ao processo de vetorização (*embedding*), etapa que converte a linguagem natural em representações vetoriais de alta dimensionalidade, capturando as intrincadas relações semânticas e contextuais dos termos técnicos. Para essa tarefa, utilizou-se o modelo `models/text-embedding-001` (ou `models/gemini-embedding-001`), acessado via API da Google Vertex AI mediante os nós de integração nativa do n8n.

A opção por esse modelo fundamenta-se em sua arquitetura otimizada para tarefas de recuperação semântica (*semantic retrieval*) em contextos multilíngues e em sua eficiência no

processamento de densos volumes de *tokens*. Tal configuração favorece uma representação semanticamente adequada do léxico agrônômico especializado presente nas publicações utilizadas (Google Cloud, 2024).

Os vetores gerados foram persistidos no Qdrant, um motor de busca vetorial (*vector search engine*) de código aberto, desenvolvido em linguagem Rust, notabilizado por sua alta performance e escalabilidade. No ambiente da Embrapa I/O, o Qdrant foi instanciado em *containers* Docker, configurado para operar mediante a métrica de Similaridade de Cosseno (*Cosine Similarity*). Esse método calcula a afinidade semântica baseando-se no ângulo entre os vetores em um espaço multidimensional, constituindo o padrão da indústria para espaços vetoriais textuais normalizados. Tal configuração favorece a recuperação eficiente da informação em baixa latência, mesmo em corpora documentais densos (Qdrant, 2024).

**Figura 11 – Fluxo de inserção de arquivos PDF do corpus no Qdrant**



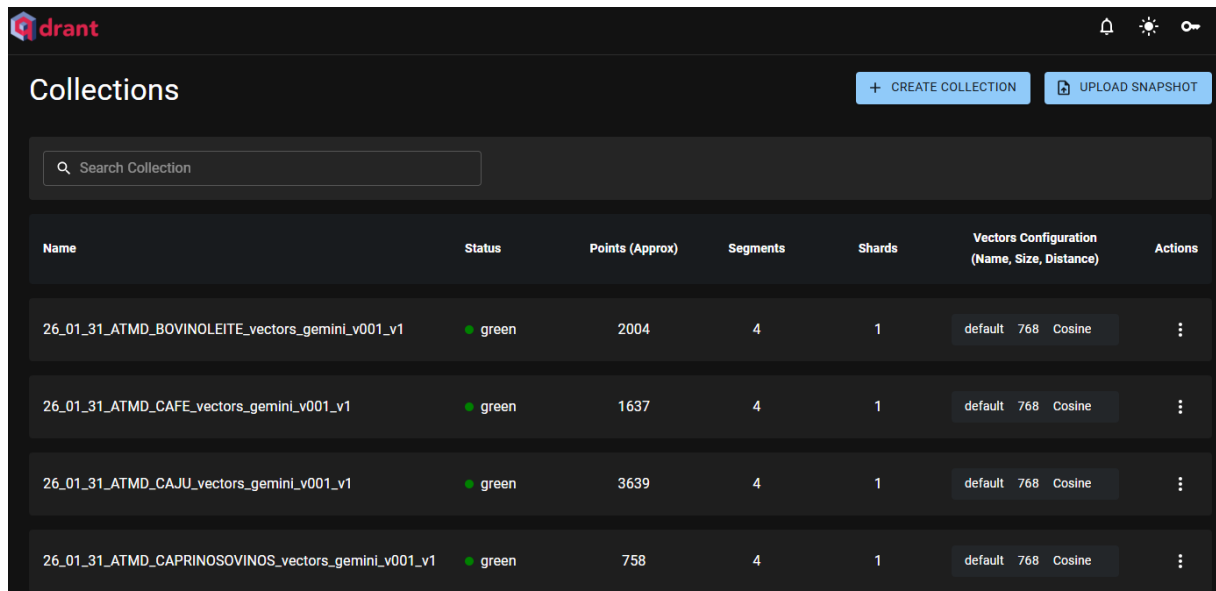
Fonte: Elaboração própria (2026)

No que tange ao armazenamento, a investigação rompeu com a abordagem convencional de índice monolítico, adotando uma estratégia de Segregação por Coleções Temáticas no banco de dados vetorial Qdrant. Em vez de aglutinar a totalidade dos documentos

em um espaço vetorial indiferenciado, instanciaram-se coleções independentes para cada cadeia produtiva ou tema transversal (ex.: `ater_digital_uva`, `ater_digital_milho`, `ater_digital_ilpf`).

Essa arquitetura de dados provê duas vantagens metodológicas cruciais: (i) isolamento de contexto, mitigando o risco de uma consulta sobre "pragas do milho" recuperar erroneamente vetores pertinentes a "pragas da uva" em virtude de sobreposições terminológicas; e (ii) otimização de performance, mediante a redução do espaço de busca (*search space*) e da latência nas operações de inferência (Qdrant, 2024).

**Figura 12 – Armazenamento das fontes por tema no Qdrant**



Name	Status	Points (Approx)	Segments	Shards	Vectors Configuration (Name, Size, Distance)	Actions
26_01_31_ATMD_BOVINOLEITE_vectors_gemini_v001_v1	green	2004	4	1	default 768 Cosine	⋮
26_01_31_ATMD_CAFE_vectors_gemini_v001_v1	green	1637	4	1	default 768 Cosine	⋮
26_01_31_ATMD_CAJU_vectors_gemini_v001_v1	green	3639	4	1	default 768 Cosine	⋮
26_01_31_ATMD_CAPRINOSOVINOS_vectors_gemini_v001_v1	green	758	4	1	default 768 Cosine	⋮

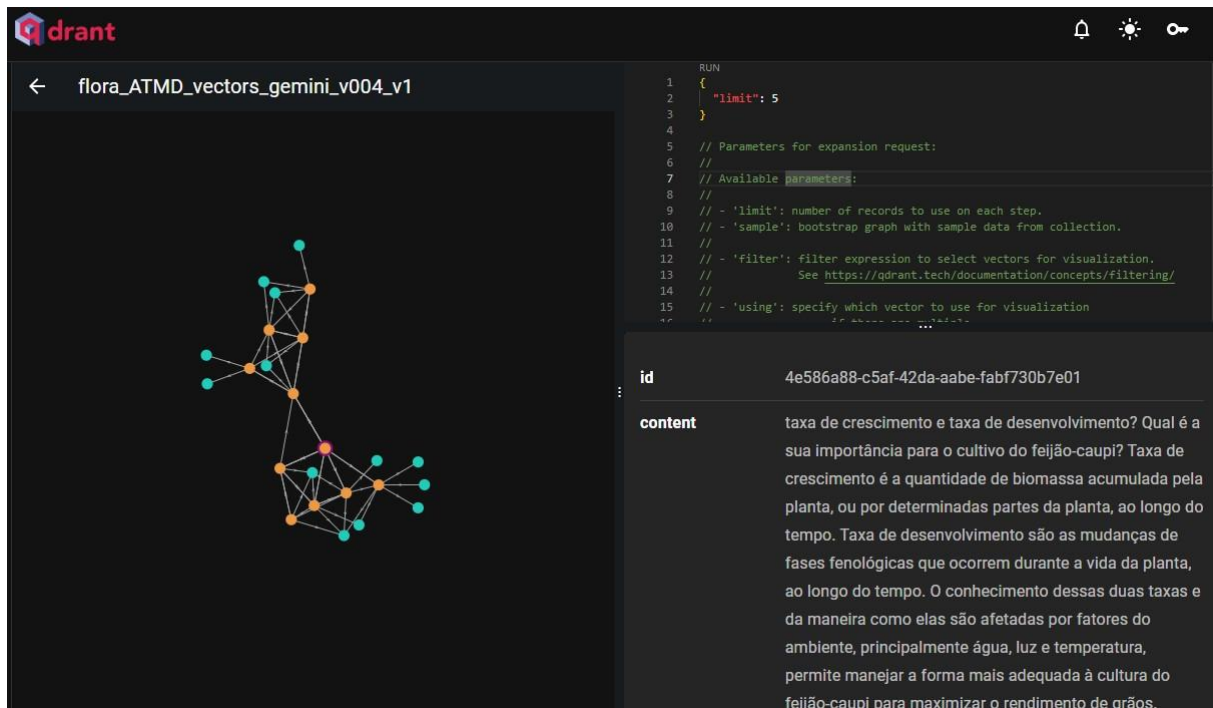
Fonte: Elaboração própria (2026)

Para garantir a eficiência na recuperação dos dados, a indexação no Qdrant utilizou o algoritmo HNSW (*Hierarchical Navigable Small World*). Diferente da busca linear exaustiva (k-NN), que seria computacionalmente proibitiva em densos volumes de dados, o HNSW constrói uma estrutura de grafos hierárquicos que facilita a navegação célere entre os nós (vetores).

Essa configuração favorece a manutenção de baixa latência nas operações de busca, mesmo com a expansão da base de conhecimento da Ater+ Digital, contribuindo para viabilizar interações ágeis com o extensionista (Malkov & Yashunin, 2018).

Um diferencial importante implementado nesta etapa foi a estruturação do *payload* (metadados) associado a cada vetor. Além da representação numérica, cada registro aloca um objeto JSON contendo os atributos extraídos na fase precedente: título do documento, autores, ano de publicação, DOI e a URL original.

**Figura 13 – Exemplo de trecho de conteúdo e os pontos correlacionados no Qdrant**



Fonte: Elaboração própria (2026)

Essa densidade de metadados viabiliza a funcionalidade de Busca Híbrida (*Hybrid Search*), facultando ao sistema o filtragem de resultados semanticamente relevantes com base em critérios determinísticos (ex.: "recuperar apenas documentos publicados após 2020").

Tal funcionalidade revela-se relevante para a qualificação das citações, uma vez que o agente conversacional recebe, de forma concomitante ao fragmento textual, a referência bibliográfica integral para exibição ao extensionista rural.

### 3.3.3 Orquestração e protocolo MCP

A orquestração dos componentes do sistema foi operacionalizada na plataforma n8n, uma ferramenta de automação de fluxos de trabalho baseada em nós (*node-based*), hospedada em instância *self-hosted* na infraestrutura da Empresa I/O. A opção pelo paradigma *low-code* justificou-se pela agilidade na prototipagem e pela facilidade de manutenção dos fluxos agênticos (*agentic workflows*). Divergindo de *scripts* lineares rígidos, a arquitetura implementada conferiu ao modelo capacidades de raciocínio e decisão (*reasoning loop*), precípuas para a triagem das intenções do extensionista e produtor rural.

O sistema avalia a entrada e delibera autonomamente sobre o acionamento das ferramentas de recuperação: perante qualquer consulta técnica, o agente é compelido a consultar obrigatoriamente a base de conhecimento vetorial (Qdrant) via protocolo MCP.

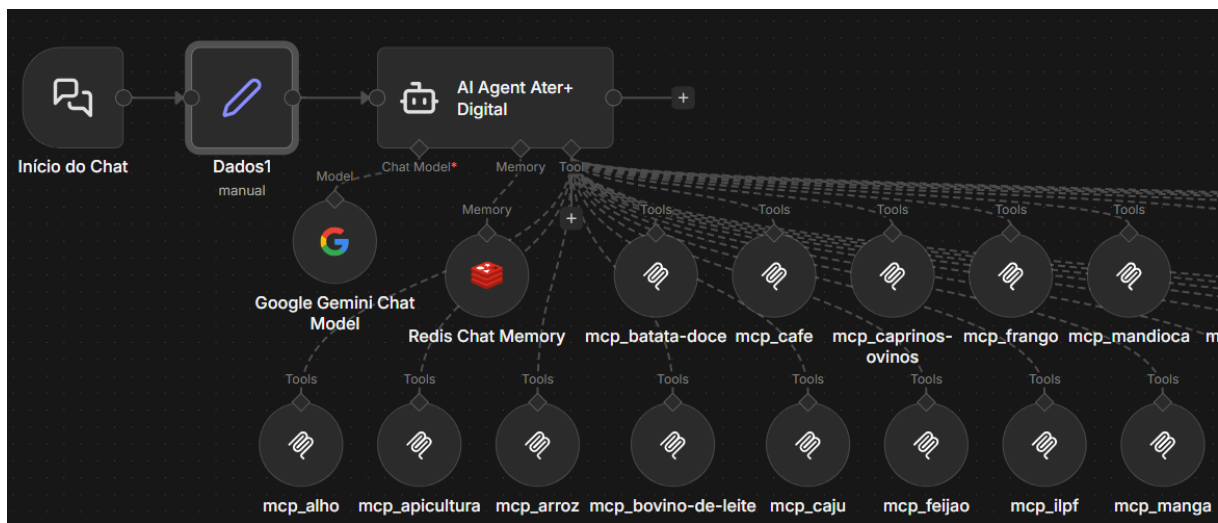
Essa restrição arquitetural — conhecida como *Closed-Domain RAG* — inibe o uso do conhecimento prévio generalista do modelo para a geração de orientações, favorecendo que as respostas permaneçam ancoradas no acervo oficial da Plataforma Ater+ Digital (Davis, 2023; n8n, 2024).

O núcleo da orquestração reside no gerenciamento do estado da conversação. Para mitigar a natureza *stateless* (ausência de estado) das requisições HTTP convencionais, implementou-se um mecanismo de Memória de Janela Deslizante (*Sliding Window Memory*), persistida em uma base de dados Redis temporária integrada ao *cluster*. Esse componente retém

as últimas  $k$  interações do usuário, incorpora-as dinamicamente ao prompt de contexto em cada nova submissão.

Tal arquitetura faculta ao agente a manutenção da coerência em diálogos multirrodada (*multi-turn*), possibilitando o processamento de referências anafóricas — como no cenário em que o extensionista indaga: "E quais são as pragas dessa cultura?", e o sistema identifica a entidade cultural referenciada na interação precedente (LangChain, 2024).

**Figura 14 – Agente de IA orquestrando as chamados ao MCP por temas no n8n**

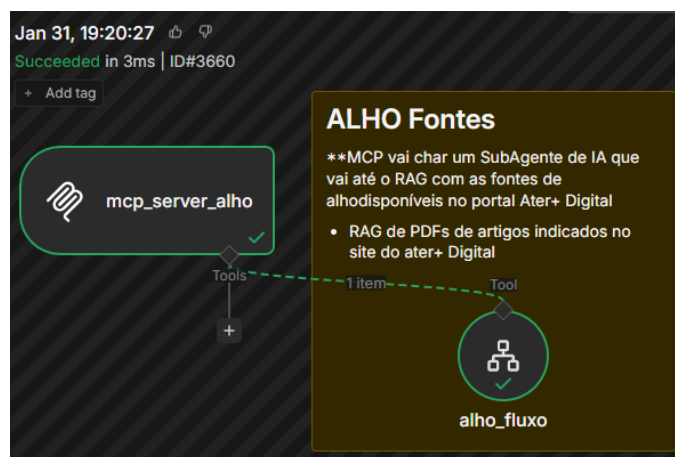


Fonte: Elaboração própria (2026)

A inovação central desta etapa consistiu na adoção do MCP) para a integração entre o Agente (Cliente) e a Base de Conhecimento (Servidor). Em vez de codificar conexões rígidas (*hardcoded*) com o banco vetorial, o MCP estabeleceu uma camada de abstração padronizada. Nesse modelo, o índice do Qdrant foi exposto como uma "ferramenta" (*tool*) detectável pelo LLM.

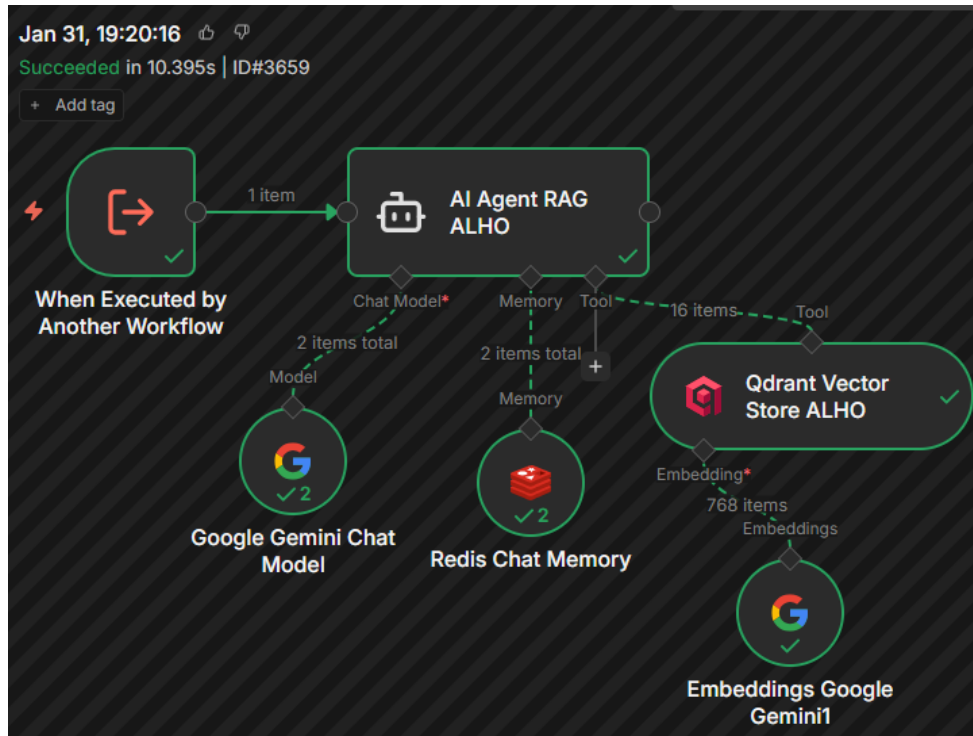
O protocolo define um esquema universal para que o modelo de linguagem "descubra" as competências disponíveis e estructure suas consultas, assegurando que a recuperação da informação ocorra de forma segura e auditável, sem a exposição direta das credenciais do banco de dados ao motor de inferência (Model Context Protocol, 2024).

**Figura 15 – Exemplo de uma chamada ao MCP server de Alho**



Fonte: Elaboração própria (2026)

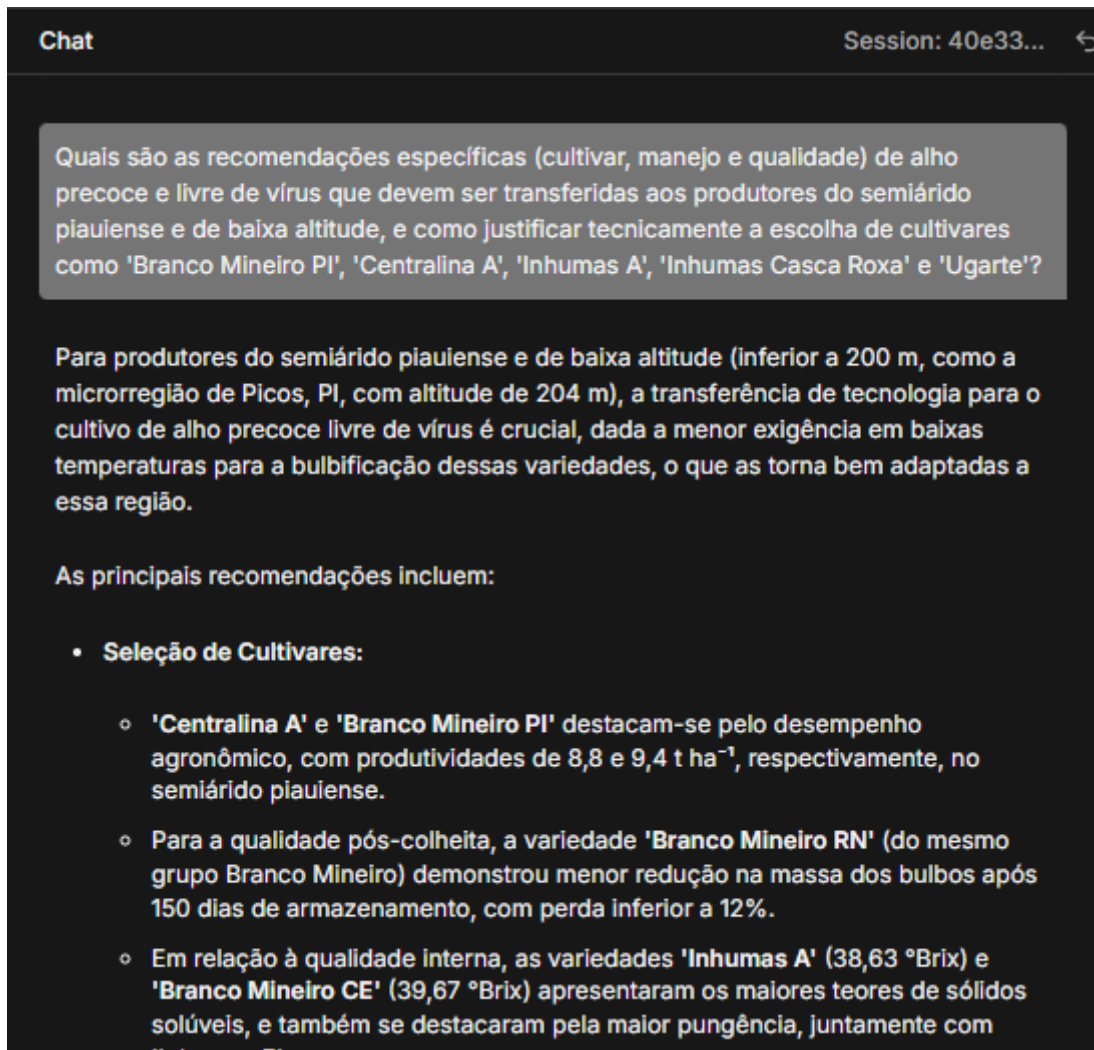
**Figura 16 – Exemplo de uma consulta ao RAG de Alho (integrando IA e Qdrant)**



Fonte: Elaboração própria (2026)

O fluxo de execução final (*pipeline*) estruturado no n8n opera sob a seguinte sequência lógica: (1) Captura, na qual a mensagem do extensionista é recebida via *webhook*; (2) Raciocínio, estágio em que o LLM analisa a intenção e invoca, mediante o protocolo MCP, a ferramenta de busca no Qdrant ao detectar uma demanda técnica; (3) Recuperação e Ancoragem, em que os vetores de maior similaridade são resgatados e filtrados; e (4) Síntese, momento em que o modelo gera a resposta final em linguagem natural, obrigatoriamente referenciando as URLs de origem providas nos metadados dos vetores recuperados. Esta última etapa operacionaliza o que Bardin (2016) define como o tratamento dos resultados e a inferência, em que os dados brutos, agora categorizados e recuperados semanticamente, são transformados em informação inteligível para o usuário final.

**Figura 17 – Exemplo de uma pergunta e resposta sobre o tema Alho**



Fonte: Elaboração própria (2026)

O desenvolvimento do artefato pautou-se por uma abordagem experimental iterativa. Inicialmente, as diretrizes do sistema (*System Prompts*) foram submetidas a testes em cenários de *zero-shot*, evoluindo progressivamente para técnicas de *few-shot prompting* e *Chain-of-Thought* (CoT). Tais métodos visaram refinar a acuidade do agente na deliberação sobre o acionamento das ferramentas de recuperação. A calibração dos parâmetros de temperatura (configurada em 0) e a definição rigorosa dos *schemas* JSON para a saída das ferramentas revelaram-se cruciais para mitigar a variabilidade estocástica ("criatividade") do modelo. Essa configuração assegurou o determinismo técnico imperativo aos padrões da Embrapa, garantindo respostas reproduzíveis e fundamentadas (Davis, 2023).

### 3.4 Procedimentos de Análise e Interpretação dos Resultados

A estratégia de análise dos resultados adotou a metodologia de Avaliação Automatizada Baseada em Modelos (*Model-Based Evaluation*), especificamente a técnica denominada *LLM-as-a-Judge* (LLM como Juiz). Esta abordagem, consolidada na literatura recente de Inteligência Artificial (Zheng et al., 2023), preconiza o emprego de modelos de linguagem de alta performance para auditar a qualidade das respostas geradas por sistemas especializados. Esse método oferece uma alternativa escalável e consistente à avaliação humana convencional, frequentemente onerosa e suscetível à fadiga cognitiva.

A opção pelo *LLM-as-a-Judge* em detrimento de avaliadores humanos independentes configura uma decisão metodológica deliberada, fundamentada em três ordens de argumentos. Primeiramente, a escala do experimento — 4.600 interações distribuídas em 23 domínios temáticos com vocabulário técnico especializado — tornaria a avaliação humana operacionalmente inviável no escopo deste estudo, pois exigiria a mobilização de especialistas agrônômicos distintos para cada cadeia produtiva, com risco de inconsistência intercultural nos critérios de julgamento. Em segundo lugar, o escopo avaliativo deste estudo, circunscrito à dimensão técnico-factual da recuperação da informação, alinha-se ao domínio em que o *LLM-as-a-Judge* demonstra maior concordância com especialistas humanos — precisamente tarefas de verificação factual e coerência técnica (Zheng et al., 2023; Liu et al., 2024). Em terceiro lugar, o enquadramento desta pesquisa sob a *Design Science Research* (Hevner et al., 2004) orienta a avaliação do artefato à sua viabilidade tecnológica e prontidão operacional (*Technology Readiness Level*), e não à validação da experiência sociointeracional do usuário — dimensão que exige, por natureza, metodologia de campo com participantes reais.

Cumprе reconhecer, contudo, que essa opção impõe limites à validade externa dos resultados: as métricas obtidas expressam o desempenho do artefato na dimensão factual e

técnica, não capturando a percepção subjetiva do extensionista ou as nuances dialógicas e culturais inerentes à comunicação rural, conforme discutido nas limitações do estudo. A validação dessa dimensão constitui, portanto, uma agenda prioritária para estudos futuros com testes de campo junto aos usuários reais da Plataforma Ater+ Digital.

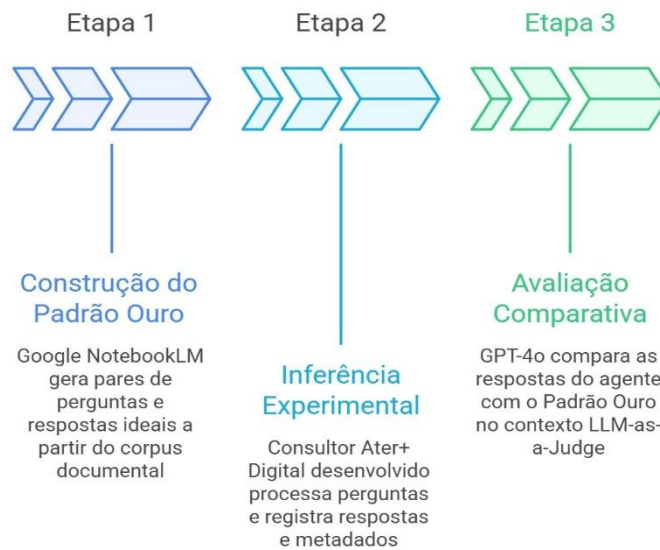
Para assegurar o rigor exigido na análise de conteúdo, o percurso de avaliação observa as fases de pré-análise e exploração do material preconizadas por Bardin (2016), estruturando o corpus documental e as categorias de análise conforme sistematizado no Quadro 6.

#### **Quadro 06 — Sistematização do Corpus e Categorização Temática (Análise de Conteúdo)**

<b>Documento/Segmento (Corpus)</b>	<b>Período</b>	<b>Categoria Temática (Bardin)</b>	<b>Unidade de Registro / Evidência Principal</b>	<b>Implicação para a Pesquisa</b>
Manuais e Comunicados Técnicos (PDF)	2023-2025	<i>Conhecimento Agrônomo Oficial</i>	Recomendações fitossanitárias, dosagens e calendários de plantio.	Base de Ancoragem (Ground Truth) para mitigar alucinações.
Repositórios de Cadeias Produtivas (Web)	2024	<i>Fluxos de Extensão Rural</i>	FAQ de produtores, orientações de manejo e boas práticas.	Definição do léxico e persona do "Consultor Digital".
Matrizes de Respostas Geradas (RAG)	2026	<i>Eficiência de Recuperação</i>	Fragmentos recuperados no Qdrant vs. Respostas sintetizadas.	Objeto da Análise Qualitativa e Juízo do LLM-as-a-Judge.
Relatórios de Auditoria (JSON)	2026	<i>Rigor e Acurácia Científica</i>	Justificativas do Juiz, fontes citadas e taxonomias de erro.	Validação da utilidade do artefato para o extensionista.

Fonte: Elaborado pelo autor (2026), com base em Bardin (2016).

**Figura 18 – Metodologia da análise e interpretação dos resultados**



Fonte: Elaboração própria (2026)

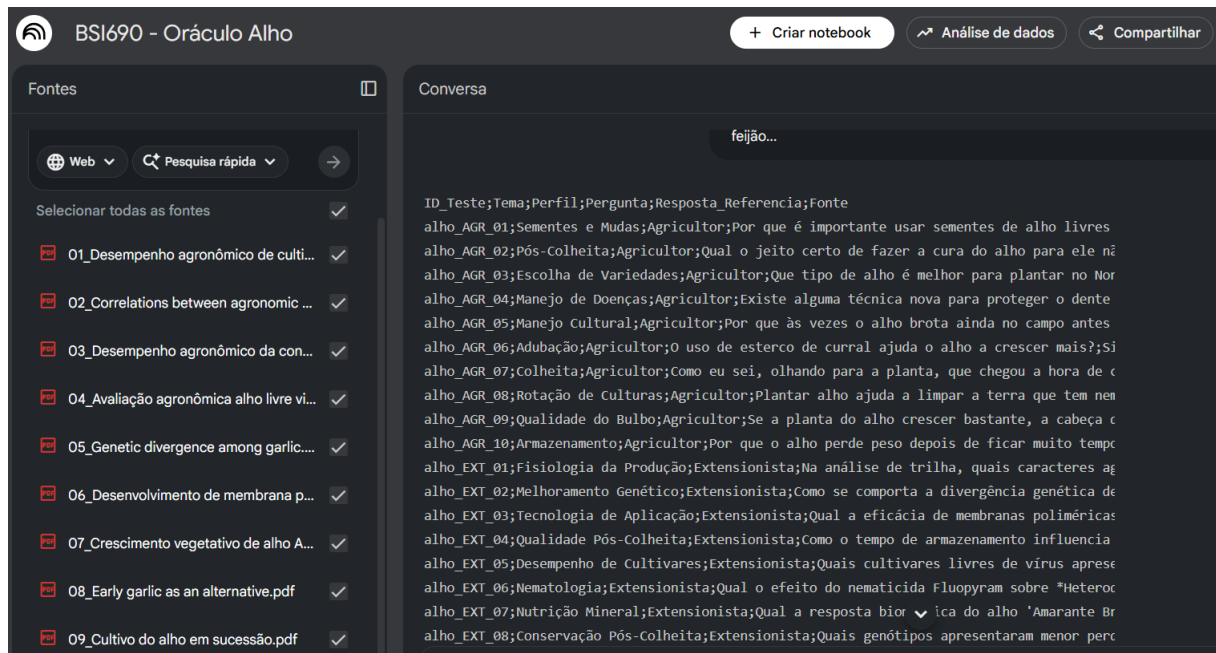
A primeira etapa consistiu na construção do Padrão Ouro (*Ground Truth*). Para esse fim, utilizou-se a ferramenta Google NotebookLM, parametrizada com acesso integral e exclusivo ao *corpus* documental (PDFs e HTMLs) de cada domínio da Plataforma Ater+ Digital.

Atuando como um "Oráculo de Referência", o sistema foi instruído a gerar pares de perguntas e respostas ideais, estritamente fundamentadas nas fontes providas.

O *dataset* resultante foi estratificado em dois perfis linguísticos: 50% das questões simularam o léxico coloquial característico de produtores rurais e 50% a terminologia técnica própria de extensionistas rurais.

Esse delineamento resultou em um banco de testes representativo da heterogeneidade de usuários da plataforma, assegurando a validade ecológica do experimento.

**Figura 19 –Google NotebookLM sobre alho**



Fonte: Elaboração própria (2026)

Na segunda etapa, denominada Inferência Experimental, as questões provenientes do Padrão Ouro foram submetidas ao artefato desenvolvido ("Consultor Digital Ater+ Digital"). O agente, operando com o modelo Gemini 1.5 Flash e utilizando a arquitetura RAG orquestrada na plataforma n8n, processou as consultas mediante a recuperação de fragmentos na base vetorial Qdrant em tempo real.

As respostas geradas, integradas aos metadados de latência operacional (em milissegundos) e às fontes referenciadas, foram registradas sistematicamente em uma matriz de dados estruturada para auditoria subsequente. Esse procedimento visou capturar não apenas a acurácia semântica do conteúdo, mas também a eficiência técnica e estabilidade do sistema sob condições de carga controlada.

**Figura 20 – Trecho da planilha com a resposta feita pelo agente de IA**

Pergunta	Resposta_Referencia	Resposta_Agente
Por que é importante usar sementes de alho livres de vírus?	O uso de sementes livres de vírus é importante porque aumenta a força da planta (vigor) e a quantidade produzida. Os vírus causam enfraquecimento e diminuem a produção ao longo do tempo.	<p>O uso de sementes de alho livres de vírus é crucial para assegurar a sanidade e a produtividade da lavoura, prevenindo a degenerescência das cultivares e promovendo um aumento expressivo na qualidade dos bulbos.</p> <p><b>**Pontos-chave técnicos:**</b></p> <ul style="list-style-type: none"> <li>* A presença de um complexo viral (gêneros *Alexivirus, Carlavirus* e *Potyvirus*) causa redução drástica no vigor vegetativo e na produtividade da cultura ao longo do tempo.</li> <li>* A infecção viral leva à degenerescência das cultivares de alho, comprometendo seu desempenho agrônomo.</li> <li>* A adoção da tecnologia de alho livre de vírus em plantios comerciais resulta em expressivos aumentos na produtividade e na qualidade dos bulbos produzidos.</li> <li>* A obtenção de plantas livres de vírus é possível através do cultivo "in vitro" de ápices caulinares, precedido por termoterapia, um processo conhecido como limpeza clonal.</li> </ul> <p><b>**Referências e Links**</b></p> <p>[Avaliação agrônômica, conservação e qualidade pós-colheita de variedades de alho precoces livres de vírus na região do semiárido piauiense]  Origem: Embrapa Hortaliças  Documento da base interna</p> <p>[DESEMPENHO AGRÔNOMICO DE CULTIVARES DE ALHO LIVRES DE VÍRUS NA REGIÃO DE PICOS-PI - AGRONOMIA]  Origem: Joanderson Mendes de Almeida, Henoque Ribeiro da Silva, Lenita Lima Haber, Francisco Vilela Resende  Documento da base interna</p> <p>[NASSUR, R.C.M.R. et al. Características comerciais, armazenamento e qualidade pós-colheita de variedades de alho infectadas e livres de vírus. Brasília, Embrapa Hortaliças, 2020. 32p.]  Origem: Embrapa Hortaliças  Documento da base interna</p>

Fonte: Elaboração própria (2026)

A terceira etapa, referente à Avaliação Comparativa, introduziu um componente de auditoria externa independente sob protocolo de anonimização (*blind evaluation*). As respostas do agente (Candidato) foram confrontadas com as do Padrão Ouro (Referência) pelo modelo GPT-4o (OpenAI).

A opção por um modelo de família e arquitetura distintas para auditar o desempenho de modelos Google constituiu uma decisão metodológica deliberada para mitigar o "viés de autopreferência" (*self-enhancement bias*). Este fenômeno ocorre quando modelos de linguagem tendem a atribuir pontuações mais elevadas a textos gerados por eles mesmos ou por arquiteturas similares (Liu et al., 2024), sendo neutralizado neste estudo pela ocultação de metadados que identificassem a origem das respostas candidatas durante o processo de julgamento.

**Figura 21 – Pipeline automatizado de auditoria do LLM-as-a-Judge**



Fonte: Elaboração própria (2026)

### 3.4.1 Critérios de julgamento e engenharia de prompt

O processo de julgamento foi conduzido sob um protocolo de anonimização (*blind evaluation*). Embora o modelo avaliador tivesse acesso à resposta de referência para fins comparativos, não lhe foi fornecido qualquer metadado que identificasse a arquitetura de Inteligência Artificial (ex.: Gemini, GPT ou Llama) responsável pela geração da resposta candidata. Tal medida visou mitigar potenciais vieses de marca ou predileções arquiteturais que modelos de linguagem podem manifestar em relação às suas próprias linhagens de desenvolvimento, favorecendo maior imparcialidade analítica no escrutínio.

As diretrizes do sistema (*System Prompt*) foram concebidas para mimetizar a persona de um "Avaliador Técnico Agrônomo Sênior da Embrapa". Com o fito de eliminar a subjetividade intrínseca a respostas textuais discursivas, o modelo foi compelido a emitir sua avaliação exclusivamente em formato JSON (*JavaScript Object Notation*). Tal protocolo exigia o preenchimento de um *checklist* lógico de verificação antes da atribuição da pontuação final.

Essa abordagem de Saída Estruturada assegurou que o julgamento fosse fundamentado em critérios binários de evidência (ex.: presença ou ausência de citação), conferindo maior objetividade e auditabilidade aos resultados quantitativos da pesquisa.

Para garantir a objetividade da análise, o modelo foi forçado a gerar sua saída exclusivamente em formato JSON, validando quatro dimensões críticas antes de atribuir a nota:

1. Responsividade (C1): O agente respondeu diretamente à dúvida central do usuário?
2. Adequação ao Perfil (C2): A linguagem é compatível com o perfil informado (Agricultor vs. Extensionista)?
3. Precisão Factual (C3): Os dados numéricos e recomendações batem com a referência? Informações extras são agronomicamente coerentes?
4. Aderência às Fontes (C4): A resposta cita fontes válidas recuperadas do contexto?

Um detalhe metodológico diferenciado deste estudo residiu na introdução dos critérios de "Invariância de Volume" e "Enriquecimento". Divergindo de abordagens convencionais que penalizam qualquer dissonância em relação ao gabarito, as diretrizes de sistema (*System Prompts*) instruíram o modelo juiz a bonificar respostas que agregassem informações agronômicas suplementares — como dosagens, protocolos de manejo ou contextualização técnica — ausentes na síntese do Padrão Ouro, desde que inexistisse contradição factual ou alucinação de conteúdo. Tal abordagem valoriza a completude funcional, dimensão imperativa no ecossistema da extensão rural, em que a utilidade da resposta transcende a mera correspondência textual (Zheng et al., 2023).

**Figura 22 – Prompt do “Agente Juiz” com critérios de Enriquecimento e formato de saída**

```

Expression
Anything inside {{ }} is JavaScript. Learn more

Você é um AVALIADOR TÉCNICO AGRÔNOMO SÊNIOR da Embrapa.
Sua tarefa é auditar um dataset de validação de RAG.
VOCÊ RECEBERÁ:
Um JSON com: Pergunta, Perfil, Resposta_Referencia, Resposta_Agente, Fontes.
=====
▲ REGRAS DE OURO (Critérios de Julgamento)

1. GABARITO É O MÍNIMO, NÃO O LIMITE:
- A 'Resposta_Referencia' contém os dados essenciais.
- Se o Agente trazer **MAIS** informações (contexto, doses extras, explicação do 'porquê')
que não contradizem a referência, isso é **ENRIQUECIMENTO**.
- **NÃO penalize Enriquecimento.

2. PRIORIDADE TÉCNICA:
- A precisão dos dados (números, nomes) vale mais que o estilo.

=====
▲ INSTRUÇÃO DE SAÍDA JSON ▲
Gere APENAS um JSON.
{
  "Checklist_Responde_Pergunta": true/false,
  "Checklist_Adequacao_Perfil": true/false,
  "Checklist_Precisao_Factual": true/false,
  "Fontes_Apresentadas": "URLs ou Títulos...",
  "Possui_Citacao": "SIM/NÃO",
  "Aderencia_Agente_TopK": "ALTA/MEDIA/BAIXA",
  "Nota_Geral": 1-5,
  "Justificativa": "Explicação curta (cite se houve enriquecimento válido).",
  "Categoria_Erro": "Omissao_Parcial", "Extrapolacao_Inutil",
  "Falta_Ancora_Nas_Fontes", "Linguagem_Inadequada", "Generalismo" ou
  "Nao_Aplicavel"
}

```

Fonte: Elaboração própria (2026)

### 3.4.2 Escala de avaliação e tratamento de alucinações

A avaliação quantitativa da qualidade das respostas fundamentou-se em uma escala Likert de 5 pontos, customizada especificamente para transcender as limitações das métricas convencionais de Processamento de Linguagem Natural (como BLEU ou ROUGE). Tais métricas, focadas na sobreposição lexical exata, revelam-se insuficientes para capturar nuances semânticas e variações contextuais.

No ecossistema da extensão rural, a eficácia da resposta não reside na reprodução *ipsis litteris* do documento original, mas na competência do agente em sintetizar a informação técnica de modo inteligível e acionável. Por conseguinte, a rubrica avaliativa foi estruturada para mensurar a Completude Funcional — isto é, a capacidade de a orientação dirimir o problema agrônomo do usuário com segurança e acurácia, independentemente de oscilações estilísticas em relação ao texto de referência.

Outro diferencial metodológico relevante introduzido nesta investigação consistiu na redefinição do conceito de "Extrapolação" sob o princípio da Invariância de Volume. Em

avaliações convencionais de sistemas RAG, qualquer informação emanada pelo agente que não conste estritamente na síntese do *Ground Truth* é amiúde penalizada como "alucinação".

Todavia, este estudo instruiu o juiz digital a distinguir Alucinação (invenção de dados espúrios ou contraditórios) de Enriquecimento (*Enrichment*). Caso o agente, ao processar o contexto vetorial integral, forneça detalhes agrônômicos suplementares — como dosagens, princípios ativos ou práticas de manejo cultural — ausentes no resumo sintético do gabarito, tal conduta é classificada como virtuosa e potencialmente bonificada com uma pontuação acima. Essa abordagem valoriza a capacidade de o sistema prover contexto denso ao extensionista, priorizando a completude informativa em detrimento da rigidez léxica.

O Quadro 7 pormenoriza a métrica escalonada empregada pelo algoritmo de julgamento, calibrada para sancionar rigorosamente riscos técnicos (Nota 1) e omissões críticas (Nota 3), ao passo que admite variações de fluidez e estilo em respostas tecnicamente acuradas (Nota 4). Essa ponderação assimétrica é fundamental para garantir a segurança da informação: enquanto a imprecisão léxica é tolerada, a divergência em relação aos protocolos agrônômicos oficiais acarreta o rebaixamento imediato da pontuação, independentemente da sofisticação linguística da resposta.

**Quadro 07 – Escala de Avaliação Likert do “Agente Juiz”**

Escala	Classificação	Critério de Avaliação (Prompt do Juiz)
1	Incorreta / Alucinação Agrônômica	Erro técnico grave. O agente inventou dados perigosos, citou fatos inexistentes ou contradisse explicitamente a recomendação oficial do corpus.

Escala	Classificação	Critério de Avaliação (Prompt do Juiz)
2	Insuficiente	A resposta é vaga ou superficial, não resolve o problema agrônômico central ou omite informações críticas sem compensação funcional.
3	Parcialmente Correta	A resposta é tecnicamente correta, porém incompleta, omitindo aspectos essenciais (ex: doses, prazos, nomes científicos, etc) do padrão-ouro ou limiares relevantes para a tomada de decisão.
4	Correta	A resposta é tecnicamente correta e agronomicamente válida, mesmo sem correspondência literal ao padrão-ouro, apresentando equivalências funcionais ou complementações aceitáveis, mesmo que apresente pequenos desvios de fluidez ou concisão. Pode incluir casos de “enriquecimento”, onde o agente fornece detalhes extras úteis além do gabarito, demonstrando alta eficácia na recuperação.
5	Excelente / Fiel	A resposta cobre a maior parte dos fatos essenciais ou seus equivalentes funcionais, não apresenta contradições, demonstra alta coerência técnica e inclui citação explícita de fontes reconhecidas. Pode, também, incluir casos de “enriquecimento”, onde o agente fornece detalhes extras

Escola	Classificação	Critério de Avaliação (Prompt do Juiz)
		úteis além do gabarito, demonstrando alta eficácia na recuperação.

Fonte: Elaboração própria (2026).

Para além da métrica numérica, implementou-se uma "Taxonomia de Erros" para qualificar as inconsistências do sistema. Sempre que a pontuação atribuída foi inferior a 4 (limiar de aceitabilidade técnica), o modelo avaliador foi orientado a categorizar a falha em uma de quatro classes mutuamente exclusivas, facultando um diagnóstico preciso da origem da deficiência na arquitetura.

As categorias estabelecidas foram: (i) Omissão Parcial, indicando uma subperformance no sistema de recuperação (RAG) ou na capacidade de síntese do LLM; (ii) Inadequação de Registro, apontando falha na adaptação da persona ao perfil do interlocutor (ex.: emprego de jargão acadêmico hermético para o agricultor familiar); (iii) Déficit de Ancoragem, verificada quando a resposta, embora correta, viola o requisito de auditabilidade ao omitir a fonte; e (iv) Alucinação ou Erro Factual, a falha de maior gravidade, na qual o modelo gera desinformação ou recomendações tecnicamente infundadas.

**Figura 23 – Exemplo de análise e categorização do erro feito pelo “Agente Juiz”**

Aderencia_Agente_TopK	Nota_Geral	Justificativa	Categoria_Erro
MEDIA	2	O agente não respondeu claramente à pergunta prática do agricultor sobre como fazer a cura do alho para evitar o murchamento. A resposta limita-se a relatar a ausência de informações específicas nos documentos, sem trazer o conteúdo essencial da referência — ou seja, que a cura deve ser feita sem cortar folhas e raízes para evitar perda de peso e dentes chochos. O agricultor continua sem orientação prática, o que constitui uma omissão crítica de um dado essencial presente na referência.	Omissao_Parcial

Fonte: Elaboração própria (2026)

Essa categorização sistemática revela-se fundamental para a discussão dos resultados, visto que faculta a distinção entre erros de recuperação (atribuíveis a deficiências no Qdrant ou na estratégia de *chunking*) e erros de geração (intrínsecos às limitações de raciocínio ou síntese do modelo Gemini). A título de exemplo, uma elevada prevalência de "Omissão Parcial" denota a necessidade de refinar os mecanismos de busca vetorial ou a janela de contexto, ao passo que uma alta incidência de "Inadequação de Registro" sinaliza a premência de ajustes nas diretrizes de *persona* do *System Prompt*. Sob essa ótica, a avaliação transcende a mera validação binária do artefato, provendo um roteiro analítico para a iteração e o aprimoramento contínuo da engenharia do sistema.

### 3.4.3 Procedimentos de tratamento estatístico dos dados

O tratamento analítico dos dados coletados estruturou-se em duas etapas complementares: uma análise descritiva exploratória, voltada à caracterização do desempenho individual de cada cadeia produtiva, e uma análise inferencial, dedicada ao teste de hipóteses relativas à variabilidade do sistema perante distintos perfis de usuário e níveis de latência. Enquanto a vertente descritiva busca caracterizar o comportamento observado nos dados, a abordagem inferencial visa avaliar se as disparidades identificadas são estatisticamente significativas ou compatíveis com flutuações aleatórias da amostragem.

A métrica central de avaliação — denominada 'Nota do Juiz' — foi mensurada em uma escala Likert de cinco pontos, configurando uma variável de natureza ordinal. Considerando que esse tipo de escala representa níveis ordenados de qualidade, sem garantia de intervalos equidistantes entre as categorias, optou-se pela adoção de métodos estatísticos não paramétricos, amplamente recomendados para a análise de dados ordinais ou distribuições não gaussianas (Field, 2013). Essa abordagem favorece maior adequação metodológica à natureza ordinal da variável e reduz a dependência de pressupostos

paramétricos frequentemente incompatíveis com avaliações baseadas em escalas qualitativas.

A análise descritiva inicial compreendeu a estimativa de medidas de tendência central (média e mediana) e de dispersão (desvio-padrão) para cada domínio temático. Enquanto a média e a mediana permitem caracterizar o desempenho central das avaliações do agente, o desvio-padrão fornece um indicador complementar de variabilidade entre as cadeias produtivas.

Sob essa ótica, o objetivo reside em descrever a variabilidade das avaliações entre os domínios temáticos, uma vez que níveis mais elevados de dispersão podem sinalizar heterogeneidade na qualidade das respostas geradas ou diferenças na estrutura do corpus documental utilizado na recuperação semântica, ao passo que menor variabilidade sugere maior uniformidade no desempenho observado (Field, 2013).

A comparação entre os perfis de usuário — 'Agricultor Familiar' e 'Extensionista' — foi realizada por meio do Teste de Mann-Whitney U, apropriado para comparar duas amostras independentes quando a variável dependente apresenta natureza ordinal ou distribuição não normal (McKnight, 2010). O teste avalia se há diferença na distribuição dos postos das avaliações entre os grupos analisados. Sob essa ótica, a seleção do instrumental estatístico fundamentou-se em recomendações metodológicas para experimentação em engenharia de software, nas quais a adequação do método deriva estritamente da estrutura e distribuição dos dados observados (Wohlin et al., 2012).

Além da significância estatística (p-valor), foram estimadas medidas de tamanho de efeito para avaliar a magnitude das diferenças observadas. Para o teste de Mann-Whitney foi calculado o coeficiente  $r$  de Rosenthal, enquanto nas comparações entre múltiplos grupos foram considerados indicadores derivados do teste de Kruskal-Wallis.

A inclusão dessas métricas permite distinguir diferenças estatisticamente detectáveis de diferenças com maior relevância prática, reforçando a interpretação substantiva dos resultados. Consoante as premissas de Sullivan e Feinn (2012), resultados estatisticamente significativos podem apresentar impacto prático reduzido, fator que exige a interpretação conjunta de ambas as métricas para a validação do artefato.

Para avaliar diferenças na qualidade das respostas entre as diferentes cadeias produtivas e temas transversais, foi aplicado o teste de Kruskal-Wallis (H-test) — alternativa não paramétrica para a análise de variância para a comparação de três ou mais grupos independentes (Ostertagová et al., 2014). Quando o teste global indicou diferença estatisticamente significativa entre os grupos ( $p < 0,05$ ), foram realizadas comparações múltiplas utilizando o teste *post-hoc* de Dunn com a correção de Holm-Bonferroni, procedimento que controla a inflação do erro tipo I decorrente de múltiplas comparações simultâneas.

Para investigar a associação entre a latência — mensurada pelo tempo de processamento em milissegundos — e a qualidade da resposta, empregou-se o coeficiente de correlação de Spearman ( $\rho$ ), métrica não paramétrica baseada na ordenação dos postos das variáveis. Esse coeficiente permite avaliar a existência de associação monotônica entre variáveis sem exigir linearidade nem normalidade das distribuições (Hauke & Kossowski, 2011). Tal análise objetiva examinar se os maiores tempos de processamento estiveram associados a variações na qualidade inferencial, aspecto relevante para a avaliação do custo-benefício computacional da arquitetura proposta.

A forma das distribuições foi examinada por meio do teste de Shapiro-Wilk (Field, 2013), complementado pela inspeção visual de histogramas e gráficos de quantis (Q-plots), com finalidade exploratória. Todavia, reconhece-se que, em amostras de elevada magnitude ( $N > 4.000$  interações), tal procedimento tende a detectar pequenas divergências da

distribuição normal como estatisticamente significativas. Por essa razão, a escolha do método estatístico foi fundamentada principalmente na natureza ordinal da variável resposta e na robustez dos testes não paramétricos (Norman, 2010). Para a realização de todas as inferências foi considerado o nível de significância de 5% ( $\alpha = 0,05$ ).

Adicionalmente, a fim de avaliar a precisão das estimativas e a incerteza associada à amostragem, foram calculados intervalos de confiança de 95% (IC 95%) para as médias de qualidade. Dada a natureza não paramétrica dos dados e o volume amostral ( $N > 4.000$ ), a estimação intervalar foi operacionalizada por meio do método de *bootstrapping* (reamostragem com 10.000 repetições), técnica de computação intensiva que permite estimar a precisão das métricas observadas sem a necessidade de assumir normalidade da distribuição.

As análises estatísticas foram operacionalizadas em linguagem Python, estruturadas em um *pipeline* que integra bibliotecas de referência na computação científica. O emprego do *Pandas* permitiu a higienização e manipulação eficiente dos dados tabulares (ETL), enquanto o *NumPy* forneceu o suporte matemático para operações matriciais de alto desempenho, fundamentais no processamento das reamostragens do modelo. A execução dos testes de hipótese e a implementação do método de *Bootstrapping* com 10.000 ciclos — procedimento essencial para a estabilização dos Intervalos de Confiança (IC 95%) — foram viabilizadas pelas bibliotecas *SciPy* e *Statsmodels*.

Para a síntese visual dos achados, utilizou-se a convergência entre o *Matplotlib*, responsável pela infraestrutura de renderização, controle de *layout* e exportação das figuras em alta resolução, e o *Seaborn*, que proveu as camadas de análise estatística gráfica. O encadeamento dessas ferramentas contribui para a reprodutibilidade metodológica e para a precisão analítica exigida no contexto do estudo ora apresentado.

O delineamento estatístico adotado vai além da simples descrição de tendências centrais. Ao integrar estatística descritiva, testes de hipótese selecionados conforme a natureza dos dados e a finalidade analítica, estimativas do tamanho de efeito, intervalos de confiança e procedimento de controle do erro tipo I em comparações múltiplas, a investigação fornece evidências quantitativas relevantes acerca do desempenho do artefato.

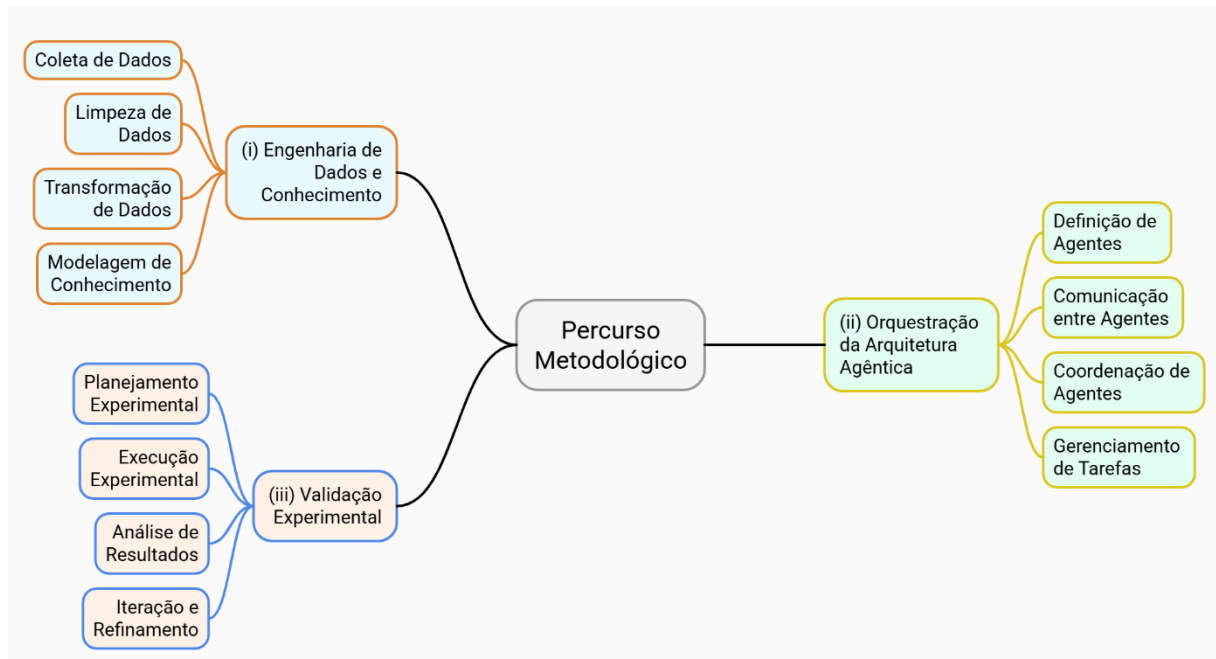
Essa sistematização metodológica contribui para que a avaliação do sistema seja conduzida de forma rigorosa e transparente, alinhando-se às boas práticas de experimentação empírica em engenharia de software e à literatura recente sobre avaliação de sistemas baseados em inteligência artificial.

### **3.5 Síntese Operacional da Metodologia**

A operacionalização desta investigação, orientada pelos preceitos do *Design Science Research* (DSR) (Hevner et al., 2004), consolidou-se em um fluxo de trabalho contínuo e auditável, concebido para transmutar ativos institucionais em um artefato de inteligência computacional de alta fidelidade. Sob a égide da Análise de Conteúdo (Bardin, 2016), o percurso metodológico sintetiza-se em três macroetapas interdependentes o percurso metodológico sintetiza-se em três macroetapas interdependentes e iterativas: (i) Engenharia de Dados e Conhecimento, focada na estruturação e vetorização do *corpus*; (ii) Orquestração da Arquitetura Agêntica, centrada na lógica de raciocínio via n8n e protocolo MCP; e (iii) Validação Experimental, pautada pelo protocolo *LLM-as-a-Judge*.

Essa estruturação sistêmica assegurou que o desenvolvimento do "Consultor Digital Ater+ Digital" transcendesse a mera integração de interfaces (APIs), constituindo-se como um ecossistema robusto de recuperação da informação, estritamente alinhado às diretrizes de governança digital e interoperabilidade da Embrapa (Embrapa, 2025c).

**Figura 24 – Percurso metodológico para construção do “Agente Digital Ater+ Digital”**



Fonte: Elaboração própria (2026)

Na etapa de Engenharia de Dados, o esforço concentrou-se na superação da heterogeneidade do acervo documental da Plataforma Ater+ Digital. Mediante rotinas automatizadas na plataforma *low-code* n8n, implementou-se um *pipeline* robusto de ETL (*Extract, Transform, Load*) apto a processar tanto o conteúdo *web* estruturado quanto o legado de publicações em PDF. A aplicação de técnicas de Reconhecimento Óptico de Caracteres (OCR) e de algoritmos de validação de integridade (*garbage check*) assegurou a sanidade dos dados brutos, ao passo que o enriquecimento semântico assistido por IA conferiu estrutura aos metadados. A subsequente indexação em coleções temáticas segregadas no banco vetorial Qdrant estabeleceu uma "memória de longo prazo" acurada — condição *sine qua non* para mitigar alucinações e garantir a contextualização agronômica institucional.

A etapa de Orquestração, núcleo funcional do artefato, materializou a inovação arquitetural mediante a adoção do padrão de Sistemas Multiagentes. Ao implementar o MCP como camada de abstração, o sistema rompeu com a rigidez dos assistentes conversacionais

convencionais, operando em um "ciclo de raciocínio" (*reasoning loop*) no qual um agente orquestrador delega as consultas técnicas a subagentes especialistas.

Essa arquitetura modular não apenas conferiu elasticidade ao sistema — o que viabiliza a incorporação horizontal de novas cadeias produtivas sem a refatoração do código legacional, mas também assegurou que a totalidade das respostas geradas fosse ancorada exclusivamente nos fragmentos recuperados e convalidados. Tal desenho respeita a premissa de soberania de dados institucionais, mitigando o risco de exfiltração ou processamento indevido de informações sensíveis (Model Context Protocol, 2024).

A terceira macroetapa, Validação Experimental, encerrou o ciclo de DSR ao submeter o artefato a um rigoroso protocolo de auditoria automatizada (*Model-Based Evaluation*). A aplicação da metodologia *LLM-as-a-Judge* — empregando modelos de fronteira (GPT-4o) para avaliar, sob regime de anonimização, as respostas do agente frente a um Padrão Ouro (*Ground Truth*) sintético gerado com acesso privilegiado (via Google NotebookLM) — proveu métricas objetivas de completude e fidelidade factual.

A introdução inovadora do critério de "enriquecimento" na escala de avaliação permitiu mensurar a aptidão do sistema em fornecer contexto suplementar útil, demonstrando que a solução proposta transcende a simples recuperação de dados, sintetizando-os de forma acionável e aderente à realidade da extensão rural (Zheng et al., 2023).

A operacionalização do artefato incorpora, adicionalmente, o rigor analítico necessário para a validação científica dos resultados, conforme detalhado nos procedimentos de tratamento de dados. Para além da observação qualitativa, a aplicação de testes estatísticos não paramétricos, como o Teste U de Mann-Whitney e o Kruskal-Wallis, permite confrontar o desempenho do sistema entre diferentes perfis de usuários e cadeias produtivas.

A análise da correlação de Spearman entre latência e qualidade contribui para demonstrar que o ajuste fino da arquitetura RAG-MCP foi examinado com base em métricas

objetivas, de modo que as inferências sobre a eficácia do agente se apoiem em evidências matemáticas auditáveis, e não apenas em percepções subjetivas.

A síntese operacional aqui delineada estabelece uma ponte metodológica entre a infraestrutura tecnológica da Embrapa e as necessidades pragmáticas da extensão rural. Ao concatenar o processamento de linguagem natural, a orquestração via protocolo MCP e o rigor do tratamento estatístico, a pesquisa contribui para a rastreabilidade do percurso analítico adotado.

Dessarte, a arquitetura proposta não se limita a um desenvolvimento técnico, mas consolida-se como um protocolo de governança do conhecimento, apto a transformar o acervo da Plataforma Ater+ Digital em um sistema de inteligência conversacional seguro, auditável e alinhado aos padrões de excelência exigidos para o desenvolvimento de inovações no agronegócio.

À luz das evidências discutidas, a execução metodológica observou rigorosos critérios de governança de dados e ética algorítmica. O uso de uma arquitetura Closed-Domain RAG favoreceu que o processamento permanesse limitado ao acervo oficial da Embrapa, por meio da Plataforma Ater+ Digital, contribuindo para a soberania informacional e a integridade das recomendações. Visto que o *corpus* é de domínio público institucional, a pesquisa prescinde de submissão ao Comitê de Ética, embora mantenha o compromisso com a transparência e a mitigação de alucinações, conforme preconizado nas diretrizes de inovação do setor público.

#### **4. Resultados e Discussão**

Este capítulo apresenta os resultados obtidos com o desenvolvimento e a avaliação do "Consultor Digital Ater+ Digital", um artefato concebido sob a égide da *Design Science Research* (Hevner et al., 2004) para atuar como mediador inteligente no acervo técnico-científico da Embrapa. O protótipo foi integrado à interface da plataforma Ater+ Digital,

alinhando-se aos preceitos da Agricultura 4.0, que preconizam a transformação digital voltada à gestão do conhecimento e ao suporte estratégico à tomada de decisão (Vieira Filho et al., 2016). A solução busca transcender as limitações dos sistemas de busca convencionais baseados em palavras-chave, oferecendo uma interface conversacional intuitiva que visa democratizar o acesso à informação para produtores e extensionistas.

A sustentação técnica do artefato fundamenta-se na convergência entre a arquitetura *Retrieval-Augmented Generation* (RAG) e o MCP, estrutura desenhada para mitigar o fenômeno das alucinações inerente aos modelos de linguagem de larga escala (Vaswani et al., 2017). Essa orquestração agêntica, operacionalizada via plataforma n8n e banco vetorial Qdrant, permitiu a compartimentação do conhecimento em domínios institucionais e técnicos, favorecendo que as respostas geradas permaneçam contextualizadas, rastreáveis e ancoradas em fontes oficiais (Gao et al., 2023). Tal abordagem favorece a integridade informacional necessária para a implementação de agentes conversacionais no setor agrícola (Shepherd et al., 2025).

Para validar a eficácia da proposta, realizou-se uma avaliação exaustiva contemplando 4.600 interações controladas distribuídas em 23 cadeias produtivas. A análise apresentada a seguir abrange desde o desempenho computacional, pautado pela latência e estabilidade sistêmica (Nielsen, 1993; Miller, 1968), até a auditoria de qualidade das saídas realizada por meio da metodologia *LLM-as-a-Judge* (Zheng et al., 2023). Os achados detalhados nas próximas seções indicam prontidão tecnológica do sistema no contexto analisado e sua capacidade de prover suporte à decisão no campo com desempenho estável e aderência ao corpus institucional.

#### **4.1 Apresentação do Artefato**

O desenvolvimento do artefato, “Consultor Digital Ater+ Digital”, materializou-se como um assistente virtual inteligente integrado à interface da Plataforma Ater+ Digital da Embrapa, em um ambiente de desenvolvimento de software controlado.

A concepção deste protótipo alinha-se aos preceitos da Agricultura 4.0 (Vieira Filho et al., 2016), que preconizam a transformação digital voltada à gestão do conhecimento e ao suporte estratégico à tomada de decisão.

Superando as limitações dos sistemas de busca convencionais baseados em palavras-chave (keyword-based), o artefato utiliza a arquitetura Retrieval-Augmented Generation (RAG) para o processamento de consultas em linguagem natural.

Tal abordagem permite que produtores e extensionistas acessem informações técnicas complexas de maneira conversacional e intuitiva, por meio de uma interface projetada para integrar-se harmoniosamente ao layout atual do portal (Figura 25), favorecendo padrões elevados de usabilidade e familiaridade.

**Figura 25 – Captura de tela da página inicial do Ater+ Digital com o chatbot (protótipo)**

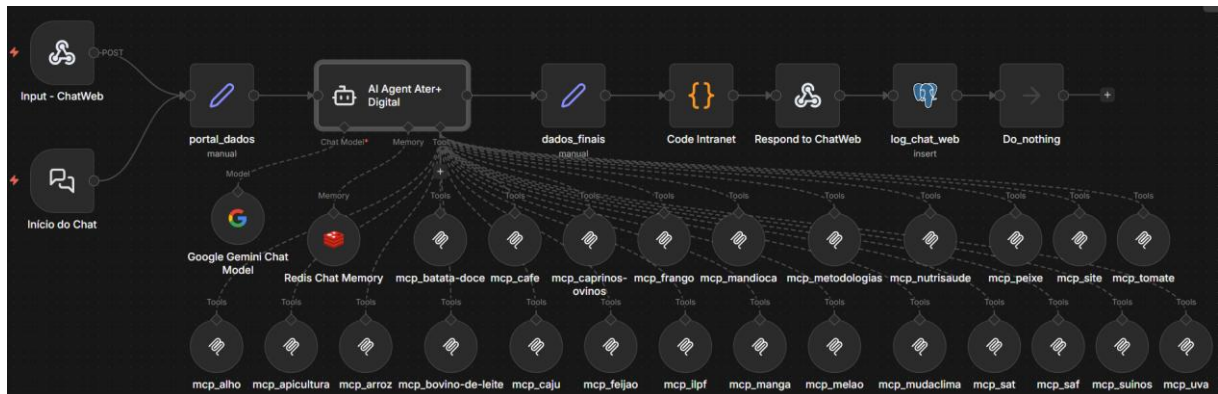


Fonte: Elaboração própria (2026)

A arquitetura do sistema estruturou-se sob o MCP, que opera como a camada de orquestração entre o Modelo de Linguagem de Grande Escala (*Large Language Model* — LLM) e os repositórios vetoriais. Tal abordagem mitiga o fenômeno das 'alucinações' em modelos generativos — limitação inerente à arquitetura de atenção descrita por Vaswani et al. (2017).

Em vez de restringir-se ao conhecimento estático do treinamento prévio, o sistema foi configurado para realizar consultas, em tempo real, a coleções segregadas no banco de dados Qdrant. Conforme evidenciado na Figura 26, o fluxo informacional foi compartimentado em dois domínios estratégicos: um índice 'Institucional' (abrangendo dados de governança e parcerias, capturados via *crawlers* especializados) e índices 'Técnicos' (dedicados a cadeias produtivas específicas, como Café, Caju e Bovinocultura), favorecendo que a resposta gerada permaneça contextualizada e rastreável à fonte oficial.

**Figura 26 – Agente orquestrador no n8n com as respectivas chamadas para os MCPs**



Fonte: Elaboração própria (2026)

A implementação técnica fundamentou-se na plataforma *low-code* n8n para a orquestração dos fluxos de trabalho, estrutura que viabilizou uma integração modular e dotada de flexibilidade arquitetural.

A ingestão dos ativos informacionais foi operacionalizada mediante algoritmos de raspagem de dados (*web scraping*) desenvolvidos em Python, projetados para capturar a estrutura semântica das páginas da Embrapa — incluindo hierarquia de títulos e metadados — em detrimento da simples extração de texto bruto, além dos downloads das publicações em PDF disponíveis na página.

Esse processo de curadoria de dados é compatível com a premissa de que a eficácia de modelos baseados em *Transformers* está fortemente associada à qualidade do *input* (Vaswani et al., 2017), visto que a precisão da resposta gerada é diretamente proporcional à relevância do contexto recuperado. O tratamento dos dados compreendeu, ainda, o expurgo de elementos ruidosos de interface e a categorização automatizada de ferramentas e parceiros institucionais, como hora já citado na metodologia.

Com o objetivo de favorecer a precisão das saídas e a integridade informacional, utilizou-se a Engenharia de *Prompt* na calibração do *chatbot*. O *system prompt* foi estruturado para impor restrições de segurança e privacidade, mitigando a exposição de dados sensíveis e orientando a assistente a atuar estritamente como facilitadora no acesso ao conhecimento — evitando a substituição do extensionista humano.

Tal delimitação ética e funcional coaduna-se com investigações contemporâneas sobre a implementação de agentes conversacionais na agricultura (Shepherd et al., 2025), que preconizam o uso da IA como suporte complementar à tomada de decisão. A Figura 27 ilustra a capacidade do sistema de gerar respostas fundamentadas, nas quais se observa a listagem explícita de fontes e endereços para o descarregamento de soluções institucionais de apoio à extensão rural.

**Figura 27– Consultor Digital recomendando ferramentas e apps disponíveis**



Fonte: Elaboração própria (2026)

A vetorização do conteúdo — processo de conversão textual em representações numéricas ou *embeddings* — viabilizou a implementação de uma busca semântica mais precisa. Sob essa ótica, o protótipo demonstra aptidão para discernir que uma consulta sobre 'praga no tomateiro' exige o acionamento do contexto técnico da olericultura, ao passo que uma busca por 'financiamento do projeto' demanda o resgate de dados institucionais de governança.

Essa distinção foi operacionalizada mediante descrições semânticas nas definições das ferramentas (*Tools*) do protocolo MCP, facultando ao modelo de linguagem a autonomia para decidir qual repositório de conhecimento deve ser consultado. Essa capacidade de discernimento sobre a procedência da informação representa um avanço relevante frente aos sistemas de busca convencionais da plataforma, os quais tendem a retornar resultados genéricos ou desprovidos de contexto.

O protótipo, então, indica a viabilidade de integrar tecnologias de fronteira, como a IA Generativa e a arquitetura RAG, a infraestruturas legadas de instituições públicas, prescindindo da refatoração integral de sistemas de gestão de conteúdo (*Content Management Systems* — CMS) preexistentes, a exemplo do Liferay (tecnologia onde a plataforma Ater+ Digital foi desenvolvida). O artefato resultante supera a funcionalidade de um mecanismo de busca convencional, configurando-se como uma interface de mediação tecnológica que reduz barreiras de acesso para produtores rurais e amplia o acesso ao acervo técnico da Embrapa.

A validação desse desenvolvimento, detalhada nas seções subsequentes, concentra-se na eficácia da recuperação informacional e na conformidade das respostas às diretrizes institucionais, indicando o modelo proposto como uma solução promissora para a extensão rural digital.

## 4.2. Análise de Desempenho Computacional (Performance)

A viabilidade técnica de um artefato fundamentado em IA Generativa, particularmente no contexto da Extensão Rural Digital, está fortemente associada à sua latência de resposta e à estabilidade sistêmica observada. Conforme postulado por Nielsen (1993), o tempo de resposta configura-se como um componente crítico da usabilidade, definindo o limiar entre uma interação fluida e a percepção de latência ou falha sistêmica.

No cenário da Agricultura 4.0, no qual a conectividade em áreas rurais é frequentemente assimétrica, o desempenho do assistente virtual foi avaliado não apenas como métrica de eficiência, mas também como indicador de prontidão tecnológica (Technology Readiness Level — TRL), sob a égide do Design Science Research (Hevner et al., 2004).

Cabe ressaltar que as métricas apresentadas computam a latência compreendida entre a requisição ao respectivo servidor MCP — que aciona o subagente para consulta à base vetorial — e o retorno da síntese final, prescindindo do tempo de processamento da interface gráfica (front-end). Tal delimitação justifica-se pelo objetivo de avaliar a robustez do núcleo agêntico e do pipeline de recuperação semântica (RAG), isolando-se variáveis exógenas de rede e latências de renderização que poderiam interferir na avaliação do desempenho da arquitetura orquestrada.

Para fins de validação da lógica de inteligência e da prontidão tecnológica do backend, esse escopo fornece subsídios para avaliar a celeridade observada no processamento de dados institucionais e discutir, de forma exploratória, o potencial de escalabilidade da arquitetura.

Para uma compreensão sistêmica do desempenho do artefato, o Quadro 08 sintetiza os indicadores de latência e estabilidade operacional aferidos durante o ciclo de validação. A fim de conferir maior rigor e transparência à precisão das estimativas, a média e a mediana foram

submetidas ao método de bootstrapping (10.000 repetições), permitindo a identificação dos intervalos de confiança de 95% (IC 95%). Este procedimento contribui para estimar a precisão das métricas observadas, indicando baixa incerteza amostral para média e mediana no conjunto de 4.600 interações avaliadas.

#### Quadro 08 – Síntese dos Indicadores de Desempenho e Estabilidade

Indicador de Desempenho	Valor Observado (ms)	Valor Convertido (s)	IC 95% (s)
Média de Tempo	6.451,01	6,45 s	[6,39 s – 6,51 s]
Mediana	6.190,00	6,19 s	[6,13 s – 6,24 s]
Desvio Padrão	1.986,48	1,98 s	---
Resposta mais rápida	1.916,00	1,92 s	---
Resposta mais lenta	32.307,00	32,31 s	---
Taxa de Sucesso (%)	100,00%	4.600 interações	---

*Nota.* Médias e Medianas estimadas via Bootstrapping ( $N=10.000$ ). Elaboração própria (2026).

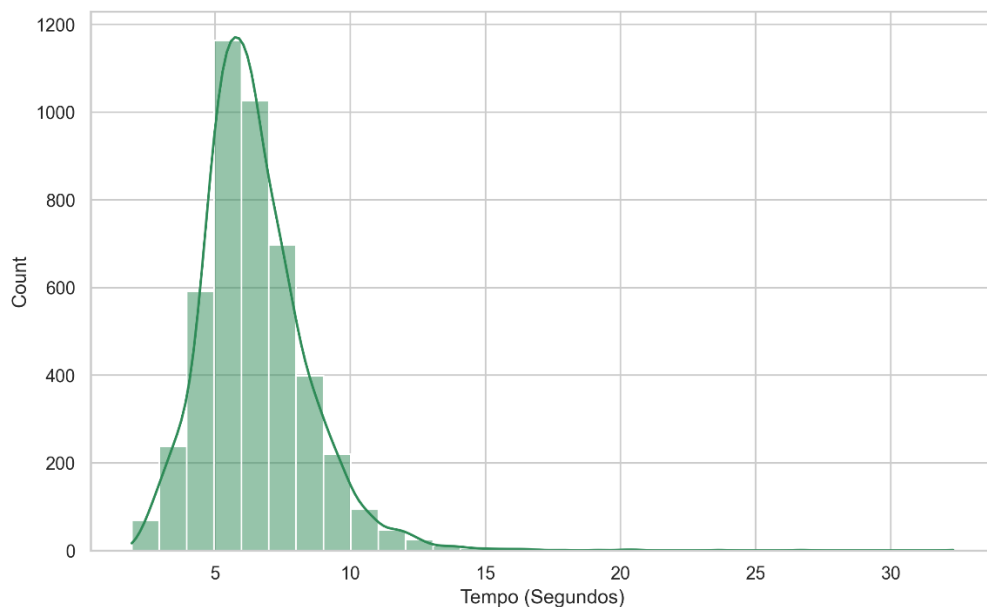
Os dados consolidados indicam elevada estabilidade operacional da infraestrutura tecnológica provida pela plataforma Embrapa I/O. Ao longo do experimento, 100% das 4.600 interações avaliadas foram processadas com retorno de resposta, sem registro de interrupções por timeout ou falhas de servidor. Tal resultado indica que a orquestração agêntica via n8n, integrada ao protocolo MCP e ao banco vetorial Qdrant, apresentou elevada estabilidade operacional nas interações avaliadas. Essa estabilidade técnica é relevante para a consolidação

da confiança do extensionista e do produtor rural no artefato, ao reduzir riscos associados à indisponibilidade do serviço em momentos críticos de tomada de decisão agrônômica.

No que tange à latência operacional, o artefato apresentou média global de 6,45 segundos e mediana de 6,19 segundos por resposta gerada. A proximidade entre esses valores, evidenciada pela distribuição no histograma (Figura 28), sugere baixa assimetria e relativa previsibilidade da latência observada.

A estreita amplitude dos intervalos de confiança (Quadro 08) reforça a precisão das estimativas de latência, indicando baixa incerteza amostral mesmo diante da carga de processamento exigida pelo modelo Gemini 1.5 Flash. Tais índices situam-se dentro dos limiares aceitáveis de interação para tarefas de alta complexidade, conforme as faixas de tolerância propostas por Miller (1968).

Figura 28 – Distribuição da Latência Operacional (Tempo de Resposta) do Agente



Fonte: Elaboração própria (2026)

A análise da dispersão temporal revelou picos de latência máxima de 32,3 segundos. Embora o desvio-padrão de 1,98 segundos indique estabilidade na maioria das ocorrências, a estratificação por cadeias produtivas demonstra que a latência é sensível à densidade documental e à complexidade intrínseca de cada domínio.

Conforme evidenciado no Quadro 09, o domínio de Caprinos e Ovinos configurou-se como o mais célere da arquitetura, registrando latência média de 5,49 segundos. Em contrapartida, o domínio de Sistemas de Integração Lavoura-Pecuária-Floresta (ILPF) demandou o maior tempo médio de processamento, fixando-se em 8,21 segundos — variação que reflete o maior volume de variáveis e a natureza multidisciplinar dos documentos indexados para este sistema produtivo.

**Quadro 09 – Panorama Multidomínio: Latência, Qualidade e Classificação de Eficiência**

<b>Domínio (Tema)</b>	<b>Latência Média (s)*</b>	<b>Nota Média (Likert)</b>	<b>IC 95% (Qualidade)</b>	<b>Classificação de Eficiência</b>
ILPF	8,21	4,980	[4,940 – 5,000]	Recuperação Profunda
Mudança Climática	7,15	4,940	[4,880 – 4,990]	Recuperação Profunda
Café	6,80	4,915	[4,845 – 4,970]	Equilibrado (Estável)

<b>Domínio (Tema)</b>	<b>Latência Média (s)*</b>	<b>Nota Média (Likert)</b>	<b>IC 95% (Qualidade)</b>	<b>Classificação de Eficiência</b>
Apicultura	6,40	4,890	[4,825 – 4,945]	Equilibrado (Estável)
Nutrição e Saúde	6,45	4,880	[4,815 – 4,935]	Equilibrado (Estável)
Caju	6,30	4,865	[4,780 – 4,940]	Equilibrado (Estável)
Batata-Doce	5,80	4,855	[4,755 – 4,940]	Alta Agilidade
Suínos	6,20	4,850	[4,750 – 4,935]	Equilibrado (Estável)
SAF	6,50	4,850	[4,760 – 4,925]	Equilibrado (Estável)
SAT	6,60	4,840	[4,735 – 4,930]	Equilibrado (Estável)
Tomate	6,40	4,790	[4,690 – 4,880]	Equilibrado (Estável)

<b>Domínio (Tema)</b>	<b>Latência Média (s)*</b>	<b>Nota Média (Likert)</b>	<b>IC 95% (Qualidade)</b>	<b>Classificação de Eficiência</b>
Frango	6,10	4,775	[4,660 – 4,880]	Equilibrado (Estável)
Mandioca	6,30	4,755	[4,645 – 4,855]	Equilibrado (Estável)
Peixe	6,50	4,730	[4,605 – 4,845]	Equilibrado (Estável)
Manga	6,40	4,720	[4,590 – 4,835]	Equilibrado (Estável)
Bovino-Leite	6,60	4,715	[4,595 – 4,825]	Equilibrado (Estável)
Caprinos e Ovinos	5,49	4,705	[4,605 – 4,795]	Alta Agilidade
Alho	6,20	4,700	[4,580 – 4,810]	Equilibrado (Estável)

<b>Domínio (Tema)</b>	<b>Latência Média (s)*</b>	<b>Nota Média (Likert)</b>	<b>IC 95% (Qualidade)</b>	<b>Classificação de Eficiência</b>
Metodologias	6,80	4,685	[4,550 – 4,810]	Zona de Otimização
Arroz	6,50	4,670	[4,530 – 4,795]	Zona de Otimização
Melão	6,70	4,665	[4,535 – 4,785]	Zona de Otimização
Uva	6,85	4,595	[4,450 – 4,730]	Zona de Otimização
Feijão	6,90	4,530	[4,375 – 4,675]	Zona de Otimização

*Nota.* Valores de IC 95% para qualidade estimados via *Bootstrapping*. Elaboração própria (2026).

Ao correlacionar esse desempenho com a volumetria da base de conhecimento (Quadro 03, Capítulo 3), observa-se uma relação de proporcionalidade direta entre o estoque informacional e o custo computacional. O domínio de ILPF, que concentra a maior densidade de ativos indexados (538 documentos), exige que o motor de busca execute o que se define como recuperação profunda (deep retrieval).

Nesse processo, a orquestração via MCP realiza o escrutínio de um espaço vetorial significativamente mais denso, demandando o processamento de um volume superior de chunks (segmentos de texto) para a extração do contexto de maior relevância. Esse fenômeno ajuda a explicar a latência de 8,21 segundos, sugerindo que a arquitetura ajusta dinamicamente seu esforço de processamento conforme a profundidade do repositório consultado. E m contrapartida, domínios com acervos mais concisos, como Caprinos e Ovinos, estiveram associados a recuperação mais célere (5,49 s).

Tal disparidade, que representa uma variação de aproximadamente 49% entre os domínios extremos, não sugere instabilidade por si só, mas pode ser interpretada como reflexo do comportamento da orquestração agêntica fundamentada no MCP. Sob o rigor metodológico de Wohlin et al. (2012), esse incremento temporal é classificado como um "custo de precisão".

Notavelmente, o domínio de ILPF, embora apresente maior latência decorrente de sua densidade documental, registrou o índice de qualidade mais elevado do estudo (4,98), com o intervalo de confiança mais estreito [4,94 – 5,00]. Esse achado sugere a eficácia da “deliberação algorítmica” na arquitetura RAG: em acervos mais densos, o sistema demandou maior tempo de processamento e apresentou avaliações de qualidade mais elevadas, o que é compatível com a interpretação de que a ampliação do escrutínio contextual pode favorecer a consistência das respostas.

Inversamente, a identificação de uma “zona de otimização” para temas como feijão e uva sugere que, embora a qualidade permaneça elevada, a variabilidade documental possa impor desafios adicionais à segmentação semântica. Essa performance adquire relevância estratégica quando confrontada com o domínio do problema.

A capacidade de processar consultas complexas em uma média global de seis segundos configura um avanço disruptivo frente à arquitetura de navegação convencional da plataforma

**MUST UNIVERSITY**

70 SW 10th St, Deerfield Beach, FL 33441, USA | [info@mustedu.com](mailto:info@mustedu.com) | [www.mustedu.com](http://www.mustedu.com)

MUST University®: licensed by Florida Commission for Independent Education. License: 5593.

Ater+ Digital, a qual comumente impõe ao usuário sucessivas camadas de interação e a necessidade de garimpagem manual em documentos PDF, além da necessidade de se fazer correlações “manuais” deles. A eficiência do “agente digital” pode ser compreendida pela redução do custo de interação e do esforço cognitivo do extensionista, convertendo o acesso ao conhecimento técnico em uma interação dialógica mais ágil e centralizada. Tal dinâmica é compatível com a abordagem participativa da Ater Digital (Zuin et al., 2022), ao reduzir fricções tecnológicas no processo de transferência de tecnologia e ampliar o acesso ao suporte à tomada de decisão no campo.

Observa-se que o desempenho computacional do artefato foi compatível com os requisitos definidos para o estudo no contexto de suporte à decisão analisado. A estabilidade operacional observada, aliada à latência previsível, posiciona o assistente digital como um ativo com desempenho elevado nas interações avaliadas.

O equilíbrio observado entre o tempo de processamento e a complexidade das tarefas executadas sugere que a arquitetura proposta apresenta potencial de expansão para novos domínios de conhecimento, sem evidência, no contexto analisado, de degradação relevante da experiência do usuário. Tais evidências sustentam a possibilidade de transição do protótipo experimental para um serviço em ambiente de produção, reforçando sua viabilidade operacional no contexto avaliado da Embrapa.

### **4.3 Avaliação da Qualidade das Respostas (LLM-as-a-Judge)**

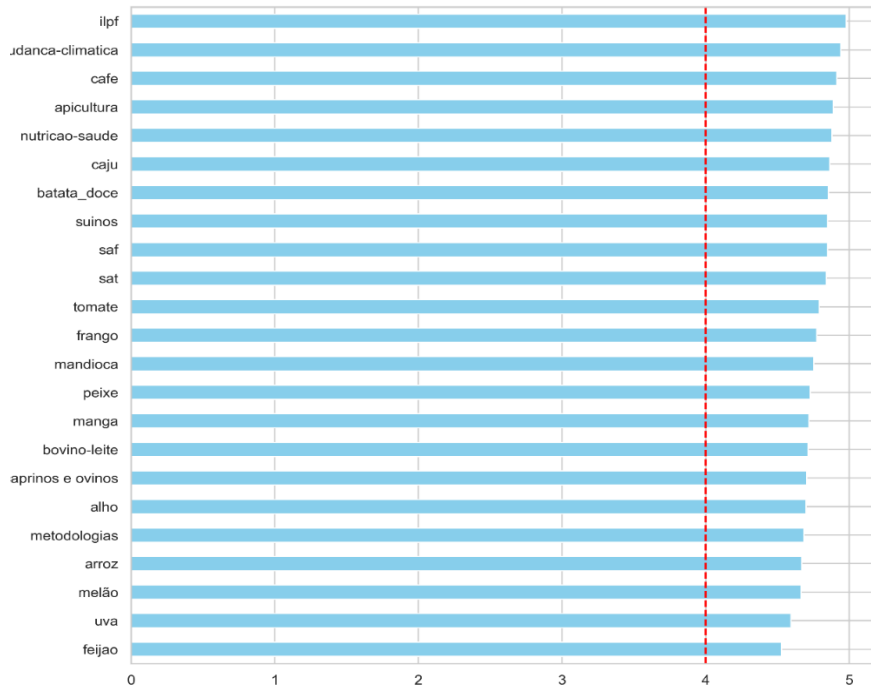
A mensuração da qualidade das saídas foi operacionalizada por um modelo de linguagem de larga escala independente (GPT-4o), o qual atuou como avaliador cego sob uma escala ordinal Likert de cinco pontos. Conforme preconizado por Zheng et al. (2023), o emprego de um LLM de alta capacidade para a avaliação de desempenho de outros agentes

viabiliza escalabilidade analítica, apresentando níveis elevados de concordância com especialistas humanos em tarefas de recuperação factual.

O escopo dessa avaliação concentrou-se na eficácia da arquitetura RAG em mitigar alucinações e prover orientações ancoradas no *corpus* institucional da Embrapa (Gao et al., 2023), contribuindo para a fidedignidade técnica das respostas geradas. Os resultados consolidados revelam elevado desempenho qualitativo do artefato. A análise descritiva, detalhada no Quadro 10 e ilustrada pelo ranqueamento da Figura 29, mostra que a mediana da qualidade se estabeleceu no patamar máximo (5,0) em todos os domínios avaliados.

Esse resultado indica predominância de avaliações no nível superior da escala, sugerindo que o assistente virtual forneceu respostas consistentemente bem avaliadas na maior parte das consultas. A concentração da mediana no limite superior da escala Likert é compatível com a interpretação de que a arquitetura RAG favoreceu a geração de respostas de elevada qualidade institucional a partir do contexto recuperado.

### **Figura 29 – Ranqueamento de Qualidade por Domínio de Conhecimento**



Nota. As médias variam entre 4,53 (Feijão) e 4,98 (ILPF), indicando avaliações de qualidade elevadas em todos os domínios temáticos analisados. Elaboração própria (2026).

**Quadro 10 – Métricas Descritivas de Qualidade por Domínio de Conhecimento**

<b>Domínio (Tema)</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio Padrão</b>
<b>ILPF</b>	4,980	5,0	0,283
<b>Mudança Climática</b>	4,940	5,0	0,396
<b>Café</b>	4,915	5,0	0,457
<b>Apicultura</b>	4,89	5,0	0,435
<b>Nutrição e Saúde</b>	4,88	5,0	0,432
<b>Caju</b>	4,865	5,0	0,573

<b>Domínio (Tema)</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio Padrão</b>
<b>Batata-Doce</b>	4,855	5,0	0,675
<b>Suínos</b>	4,850	5,0	0,678
<b>Sistemas Agroflorestais (SAF)</b>	4,850	5,0	0,608
<b>Sistemas Agrícolas Tradicionais (SAT)</b>	4,840	5,0	0,698
<b>Tomate</b>	4,790	5,0	0,699
<b>Frango</b>	4,775	5,0	0,792
<b>Mandioca</b>	4,775	5,0	0,780
<b>Peixe</b>	4,730	5,0	0,878
<b>Manga</b>	4,720	5,0	0,886
<b>Bovino de Leite</b>	4,715	5,0	0,841
<b>Caprinos e Ovinos</b>	4,705	5,0	0,7
<b>Alho</b>	4,7	5,0	0,839
<b>Metodologias</b>	4,685	5,0	0,944
<b>Arroz</b>	4,67	5,0	0,952

<b>Domínio (Tema)</b>	<b>Média</b>	<b>Mediana</b>	<b>Desvio Padrão</b>
<b>Melão</b>	4,665	5,0	0,904
<b>Uva</b>	4,595	5,0	1,042
<b>Feijão</b>	4,53	5,0	1,098

Fonte: Elaboração própria (2026).

A estratificação por cadeias produtivas revela que o domínio de Sistemas de Integração Lavoura-Pecuária-Floresta (ILPF) apresentou o mais alto desempenho descritivo do conjunto analisado. Com média de 4,98, mediana 5,0 e baixo desvio-padrão (0,283), o domínio evidenciou elevada concentração das avaliações no topo da escala, sugerindo maior uniformidade nas avaliações observadas.

Ao relacionar esses resultados ao desempenho computacional descrito no item 4.2, nota-se que o maior tempo de processamento dedicado a esse domínio (8,21s) esteve associado à mais elevada avaliação média de qualidade entre os temas analisados. Esse achado é compatível com a interpretação de que a recuperação mais extensa de contexto pode favorecer a consistência das respostas em domínios de maior complexidade informacional.

Em contrapartida, domínios como Feijão (média = 4,53) e Uva (média = 4,595) apresentaram maior dispersão nas avaliações em comparação com os temas de melhor desempenho, com desvios-padrão superiores a 1,0. Embora as médias permaneçam elevadas — ambas acima do limiar de aceitação técnica adotado no estudo (nota igual ou superior a 4,0) —, essa maior variabilidade sugere a ocorrência de interações pontuais com avaliações inferiores ao padrão predominante nessas cadeias específicas.

À luz de Field (2013), tal comportamento pode ser interpretado como um indicativo descritivo relevante, sugerindo que a estrutura dos documentos originais dessas culturas pode impor desafios adicionais à extração e à síntese das informações, especialmente em presença de elementos não textuais, tabelas complexas ou trechos com menor densidade semântica.

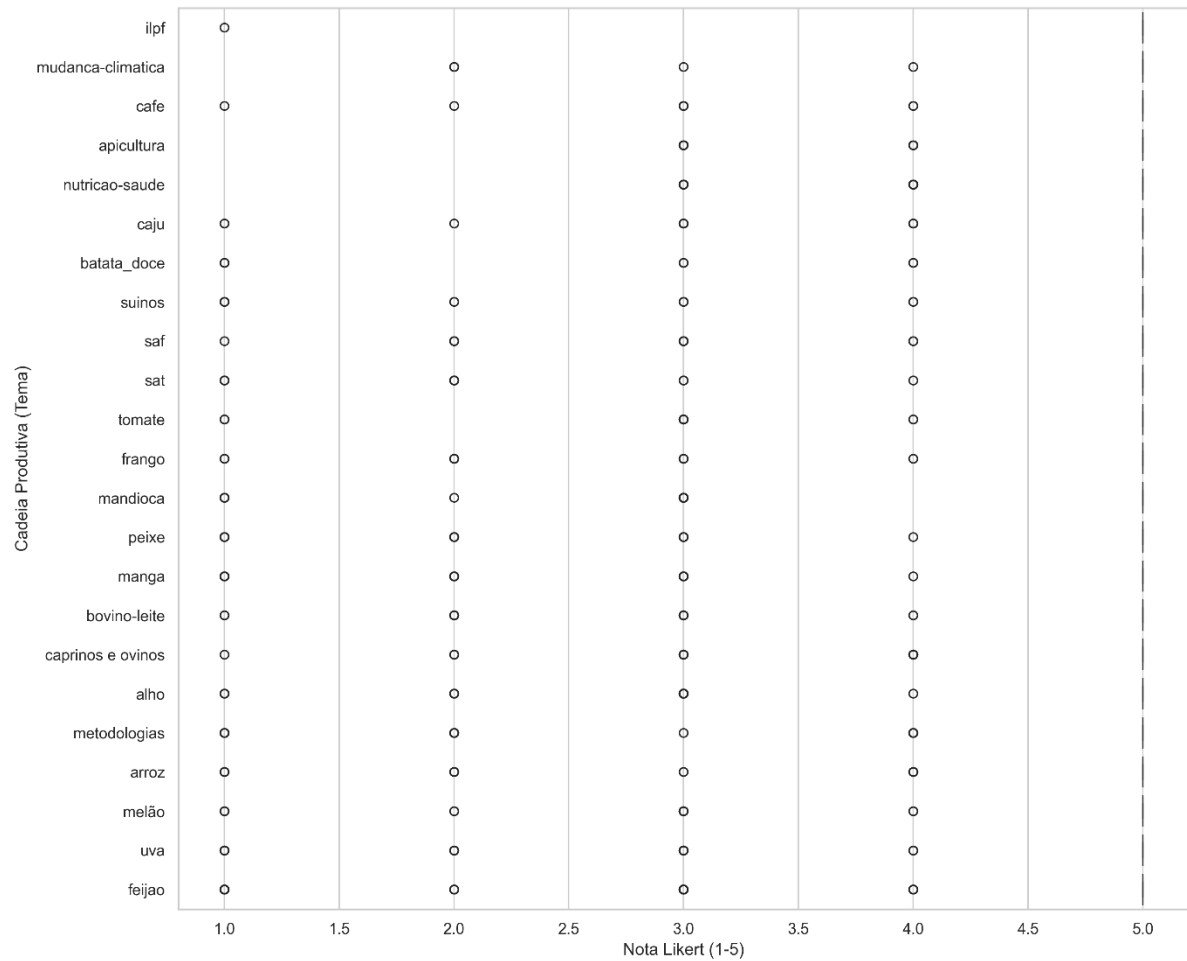
Em consequência, os resultados reforçam que a eficácia da arquitetura RAG também depende da organização estrutural do acervo documental consultado.

A Figura 30 ilustra a distribuição das notas por domínio, evidenciando o fenômeno de 'caixas colapsadas' no patamar superior da escala (Nota Likert 5). Esse comportamento gráfico indica que, na maioria dos temas, os quartis superior (Q3) e inferior (Q1) coincidem com a mediana no valor máximo, caracterizando um “efeito de teto” (*ceiling effect*) compatível com avaliações de qualidade concentradas no topo da escala.

Os valores atípicos (outliers), representados por pontos isolados nos níveis inferiores, sinalizam interações específicas com avaliações abaixo do padrão predominante, sugerindo possíveis limitações pontuais na recuperação ou na síntese das informações.

Tais ocorrências concentram-se em domínios como Feijão e Uva, justificando a maior dispersão estatística e o incremento da variabilidade observados nessas cadeias específicas.

**Figura 30 – Variabilidade da Qualidade (Boxplot) por Domínio de Conhecimento**



Fonte: Elaboração própria (2026).

A consistência observada nos índices de qualidade, com a mediana estabelecida no patamar máximo, é compatível com a premissa de que a arquitetura RAG-MCP contribui para mitigar alucinações inerentes aos modelos de linguagem (Vaswani et al., 2017). Sob a ótica do desenvolvimento de negócios, essa precisão técnica atende diretamente à necessidade de reduzir a sobrecarga cognitiva do extensionista rural (Vieira Filho et al., 2016).

Ao favorecer que a resposta da IA permaneça ancorada no conteúdo da plataforma Ater+ Digital, o sistema tende a reduzir a necessidade de o profissional realizar conferências manuais exaustivas em acervos densos ou, alternativamente, a indicar de forma mais efetiva onde

aprofundar as informações, permitindo que ele foque na aplicação estratégica da orientação técnica diretamente ao produtor.

Cumpra destacar, ainda, que as médias inferiores registradas não configuram, por si só, lacunas informacionais críticas, mas podem refletir a operacionalização dos mecanismos de “soberania do conhecimento”. Em cenários de ambiguidade na base documental, o agente Flora é parametrizado para reportar limitações ou fornecer respostas conservadoras, priorizando a incompletude em detrimento da fabulação de dados (alucinação).

Essa diretriz *de safety by design* revela-se relevante para a transição do artefato ao ambiente de produção, favorecendo que o extensionista rural acesse informações alinhadas ao crivo institucional da Embrapa e reduzindo o risco de imprecisões técnicas no processo de transferência de tecnologia..

#### **4.4 Análise Inferencial e Significância Estatística**

A análise inferencial objetiva testar as hipóteses formuladas na fase experimental, verificando se os padrões de performance e qualidade do assistente digital desenvolvido possuem significância estatística ou se decorrem de flutuações amostrais aleatórias.

Considerando a natureza ordinal da variável de resposta e os resultados exploratórios de inspeção da distribuição, optou-se pela aplicação de métodos não paramétricos. Conforme Field (2013), tais métodos são mais robustos para o tratamento de variáveis ordinais em escala Likert e amostras com variâncias desiguais, conferindo fidedignidade analítica às conclusões e assegurando a integridade das inferências acadêmicas.

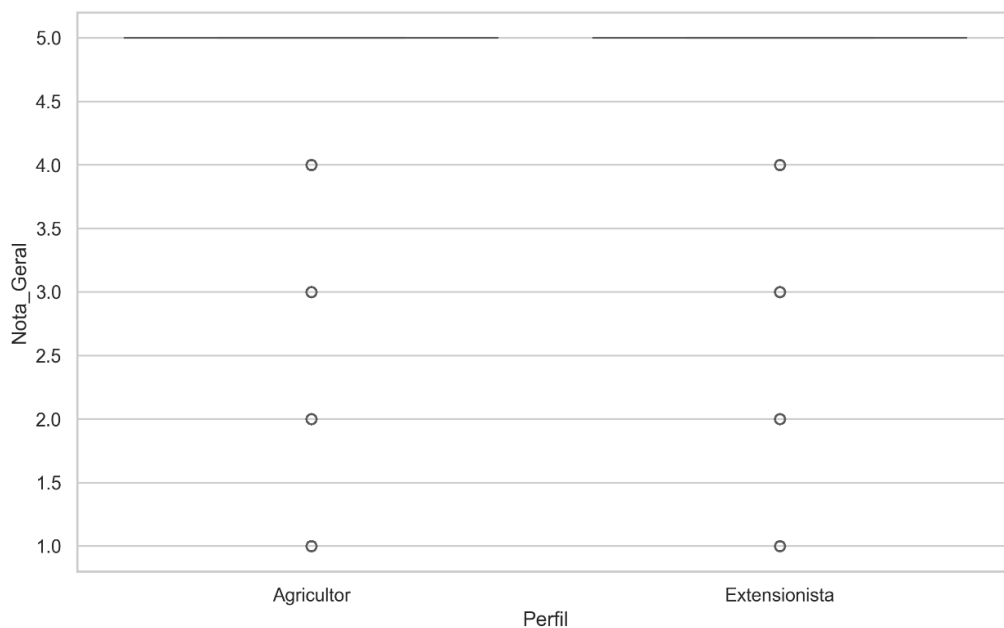
Para verificar se o artefato apresenta viés de desempenho em função do perfil do usuário, aplicou-se o teste de Mann-Whitney ( $U$ ). A análise revelou um  $p$ -valor de 0,009, indicando a existência de diferença estatística entre as notas atribuídas por 'Agricultores' e 'Extensionistas' ao nível de significância de 5%. Todavia, em amostras de elevada magnitude ( $N = 4.600$ ), o  $p$ -

valor torna-se excessivamente sensível a variações infinitesimais, o que exige a análise complementar do Tamanho do Efeito ( $r$ ). Sob a ótica de Cohen (1988), observou-se que a disparidade entre os grupos é desprezível, evidenciando que, apesar da divergência matemática, ambos os perfis convergem para um patamar de satisfação técnica elevado. Tal achado sugere desempenho semelhante do agente digital entre os perfis analisados, com diferença de magnitude desprezível no ecossistema da Ater Digital.

Dessa forma, apurou-se o Tamanho do Efeito ( $r$  de Rosenthal), resultando em um coeficiente de 0,0384. Segundo os critérios de Cohen (1988), magnitudes de efeito inferiores a 0,10 são classificadas como desprezíveis. Essa interpretação é corroborada pela Figura 31, que evidencia uma sobreposição quase integral dos *boxplots* de ambos os perfis.

Embora tenha sido observada diferença estatisticamente significativa entre os grupos, a magnitude do efeito foi desprezível, indicando estabilidade prática do desempenho do artefato entre agricultores e extensionistas. Esse resultado é relevante para a ampliação da Extensão Rural Digital, ao indicar que o suporte à decisão apresentou comportamento semelhante entre os diferentes perfis de usuários avaliados no ecossistema Ater+ Digital.

**Figura 31 - Variabilidade da Qualidade das Respostas por Perfil de Usuário**



Nota. Elaboração própria (2026), baseada no teste de Mann-Whitney ( $U = 2.704.112$ ;  $p = 0,009$ ).

A sensibilidade do artefato aos distintos domínios da agropecuária foi aferida por meio do teste de Kruskal-Wallis (H). O resultado evidenciou valor estatisticamente significativo ( $H = 112,48$ ;  $p < 0,001$ ), indicando que a distribuição das notas diferiu entre os domínios de conhecimento avaliados. Esse achado é compatível com a estratificação apresentada anteriormente, sugerindo que variações de densidade informacional e complexidade estrutural entre as cadeias produtivas estiveram associadas a diferenças no desempenho observado.

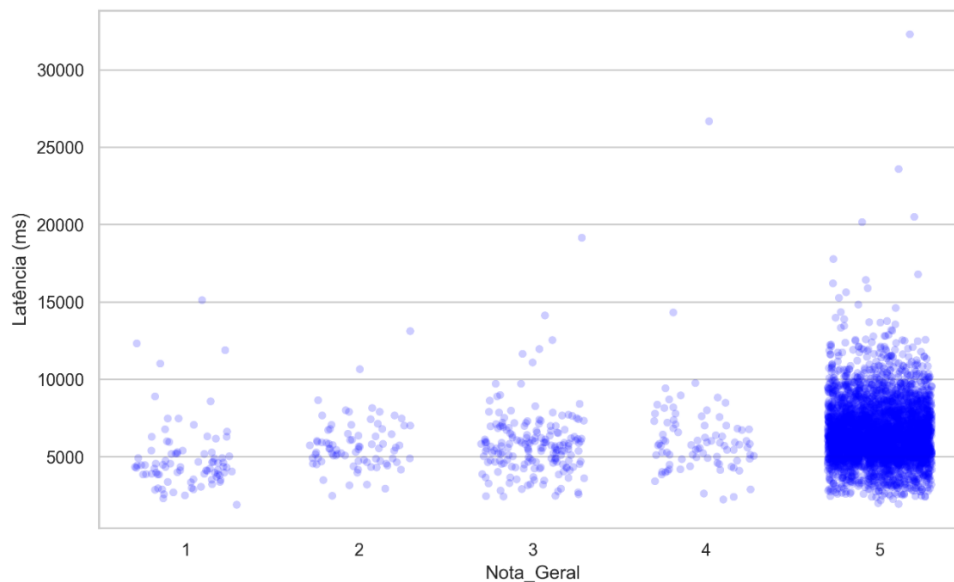
A fim de isolar as disparidades específicas entre as cadeias produtivas, aplicou-se o teste *post-hoc* de Dunn com ajuste de Holm-Bonferroni para comparações múltiplas. A análise detalhada revela que domínios como ILPF e Mudança Climática apresentaram avaliações de qualidade estatisticamente superiores às do Feijão ( $p < 0,001$ ).

Tais evidências são compatíveis com as inferências delineadas na seção 4.3, sugerindo que o desempenho do modelo RAG esteve associado à densidade e à heterogeneidade estrutural do acervo documental disponível na Embrapa. Observou-se, então, que a arquitetura apresentou desempenho particularmente elevado em domínios de maior complexidade técnica, nos quais a profundidade do contexto recuperado pode favorecer sínteses mais precisas e abrangentes pelo agente.

Apurou-se, também, a associação entre o esforço computacional e a qualidade das respostas por meio da correlação de Spearman ( $\rho$ ). Observou-se coeficiente positivo de 0,1535, com significância estatística ( $p < 0,001$ ). Embora a magnitude da associação seja tênue, a natureza positiva e estatisticamente significativa da correlação sugere que o aumento do tempo de processamento não ocorreu de forma ociosa, mas esteve associado a avaliações ligeiramente mais altas de qualidade.

A investigação da associação entre o esforço de processamento e a qualidade das saídas é relevante para examinar a eficiência da arquitetura proposta. Para tanto, a Figura 32 apresenta a dispersão bruta das 4.600 interações, correlacionando a latência operacional às métricas de qualidade atribuídas pelo juiz LLM. Observa-se que, embora o sistema mantenha média de latência relativamente baixa, as avaliações mais altas de qualidade (notas 4 e 5) distribuíram-se por uma faixa temporal mais ampla, incluindo interações com maior tempo de processamento.

**Figura 32 - Dispersão entre Nota de Qualidade e Latência Operacional (Dados Brutos)**



*Nota.* A dispersão indica que o sistema apresentou tempos variados de processamento, com maiores latências observadas em interações que, em média, mantiveram avaliações elevadas de qualidade. Elaboração própria (2026).

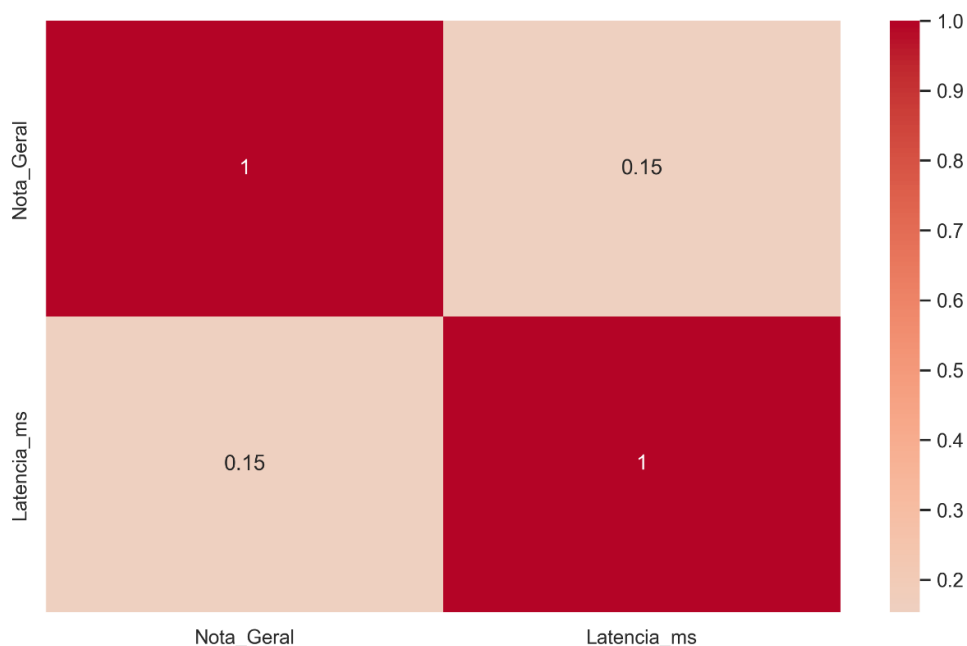
Esse fenômeno visual é compatível com a noção de “custo de precisão” (*cost of precision*) (Wohlin et al., 2012), sugerindo que, diante de consultas que exigem maior profundidade semântica ou o escrutínio de acervos documentais densos, a arquitetura demanda ciclos adicionais de processamento para favorecer a ancoragem factual. A concentração de pontos no quadrante superior do gráfico sugere que a latência observada não se associou

predominantemente a perda de qualidade, podendo refletir maior esforço de processamento em interações mais exigentes.

A análise estatística dessa tendência é consolidada pela Matriz de Correlação de Spearman (Figura 33), que quantifica a interdependência entre a qualidade da síntese e a latência operacional. O teste revelou um coeficiente positivo de 0,1535 com  $p < 0,001$  ( $1,12 \times 10^{-25}$ ), indicando evidência estatística de associação entre as variáveis ao nível de 1%. Sob a ótica da engenharia de *prompts* e da orquestração RAG, essa correlação positiva, conquanto classificada como tênue em magnitude absoluta, possui relevância estratégica fundamental: ela valida que a lógica operacional via MCP prioriza a recuperação mais extensa de contexto em detrimento da celeridade instantânea.

Conforme preconizado por Siegel e Castellan (1988), em sistemas sociotécnicos complexos, a existência dessa associação positiva sugere que o artefato ajusta seu esforço de processamento à complexidade da demanda, sem indicação, nas interações avaliadas, de comprometimento da integridade da informação técnica entregue ao campo.

**Figura 33 - Matriz de Correlação de Spearman (Qualidade vs. Latência)**



Nota. O coeficiente de Spearman ( $\rho = 0,15$ ;  $p < 0,001$ ) indica associação monotônica positiva e fraca entre latência operacional e nota de qualidade. Elaboração própria (2026).

O Quadro 11 sintetiza os pilares inferenciais que conferem validade estatística ao artefato, consolidando os testes de hipótese que sustentam a robustez do protótipo conversacional desenvolvido. O emprego de métodos não paramétricos contribuiu para reduzir o risco de interpretações inadequadas diante da não normalidade dos dados e da heterogeneidade de variâncias, favorecendo inferências mais aderentes à natureza das variáveis analisadas.

No que tange à comparação entre grupos, a convergência entre o valor de  $U$  (2.704.112) e a magnitude desprezível do efeito ( $r = 0,0384$ ) constitui um achado crítico: demonstra que, embora a elevada dimensionalidade da amostra ( $N=4.600$ ) tenha induzido significância estatística ao  $p$ -valor (0,009), a disparidade prática entre o suporte oferecido ao agricultor e ao extensionista foi desprezível nas interações avaliadas. Tal evidência sugere que o artefato manteve desempenho semelhante entre os perfis analisados, com diferença de magnitude desprezível, favorecendo acesso tecnicamente consistente ao conhecimento institucional independentemente da proficiência prévia do interlocutor.

#### Quadro 11 - Resumo dos Testes Estatísticos de Hipótese e Significância

<b>Método Estatístico</b>	<b>Variável Testada</b>	<b>Estatística</b>	<b><math>p</math>-valor</b>	<b>Tamanho do Efeito</b>
Mann-Whitney	Perfil (Agric. vs. Ext.)	$U = 2.704.112$	0,009	0,0384 (Desprezível)

<b>Método Estatístico</b>	<b>Variável Testada</b>	<b>Estatística</b>	<b><i>p</i>-valor</b>	<b>Tamanho do Efeito</b>
Kruskal-Wallis	Domínios (Temas)	$H = 112,48$	$<0,001$	Significativo
Spearman ( $\rho$ )	Nota vs. Latência	$\rho = 0,1535$	$<0,001$	Associação positiva fraca

Nota. Elaboração própria (2026), baseada no arquivo 03\_Relatorio\_Inferencial.xlsx

Adicionalmente, os resultados do teste de Kruskal-Wallis ( $H = 112,48$ ;  $p < 0,001$ ) e da correlação de Spearman ( $\rho = 0,1535$ ;  $p < 0,001$ ) estabelecem a fundamentação estatística para a eficiência da orquestração agêntica. A elevada significância observada no teste de Kruskal-Wallis ratifica que o sistema manifesta sensibilidade à complexidade intrínseca de cada cadeia produtiva, declinando de uma atuação genérica em favor de uma adaptação dinâmica ao domínio de conhecimento consultado.

Essa lógica é compatível com a correlação positiva de Spearman, a qual indica uma associação monotônica fraca entre maior latência operacional e melhor avaliação da qualidade das respostas. O conjunto de evidências sintetizado no Quadro 11 corrobora a viabilidade tecnológica do artefato sob a ótica do Design Science Research, indicando que o “Consultor Digital” alcançou avaliações elevadas de qualidade nas interações analisadas, por meio de uma arquitetura com desempenho estável e orientação à aderência factual no contexto estudado.

#### 4.5 Discussão dos Achados: Qualidade-Tempo-Confiança

A articulação entre os indicadores apresentados permite interpretar o desempenho do artefato em termos de equilíbrio entre qualidade, latência e estabilidade operacional. A latência

média global de 6,45 segundos (Seção 4.2) reflete o custo computacional inerente à orquestração agêntica sobre um acervo de aproximadamente 1,8 GB de dados brutos — cuja indexação vetorial no banco Qdrant amplia a profundidade e a dimensionalidade da busca semântica. No entanto, esse tempo de processamento esteve associado à entrega de respostas com elevada avaliação de qualidade (mediana 5,0).

Diferentemente dos mecanismos de busca convencionais (keyword-based), a eficiência do “Consultor Digital Ater+ Digital” não se limita à varredura de arquivos; ela se apoia na capacidade da arquitetura via MCP de direcionar o processamento para o domínio de conhecimento mais pertinente, operacionalizando o conceito de recuperação profunda (*deep retrieval*).

Conforme sugerido pelas métricas de auditoria via *LLM-as-a-Judge*, a confiança do usuário no ecossistema Ater+ Digital não decorre apenas da celeridade, mas também da rastreabilidade e da autoridade da fonte apresentada pela orquestração MCP. Sob a égide da Design Science Research (Hevner et al., 2004), esta discussão integra as dimensões de desempenho, qualidade e validade estatística, sustentando a utilidade do “agente digital” como artefato de suporte à decisão no contexto analisado.

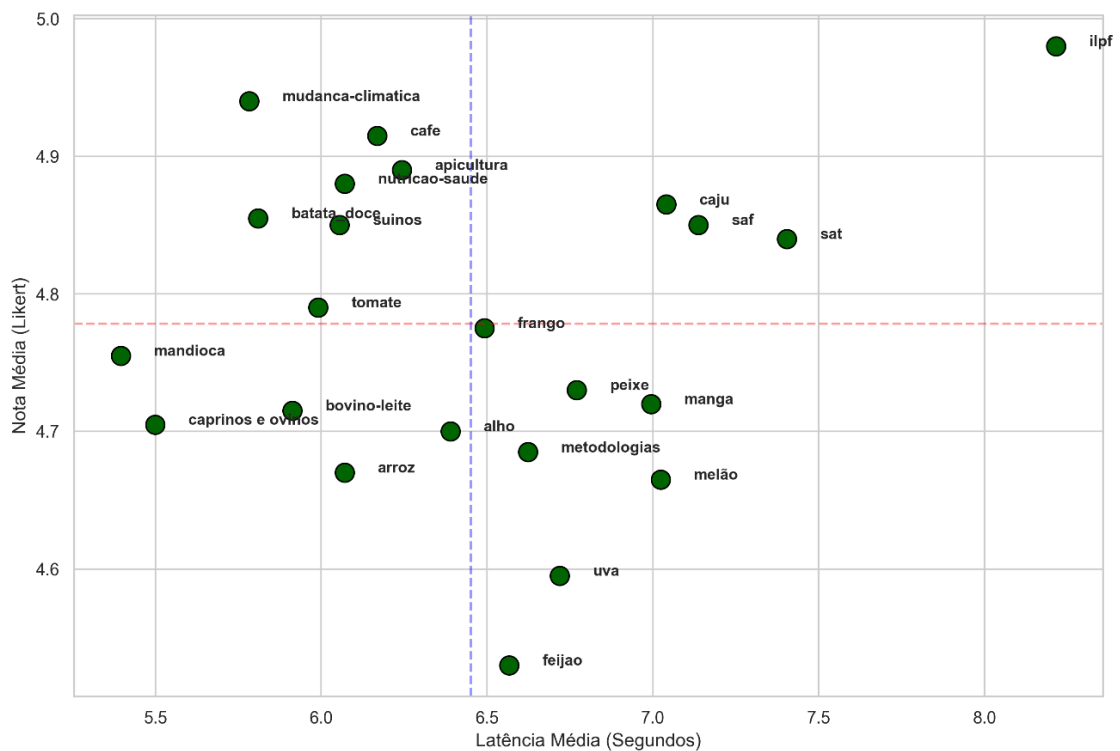
O ponto central reside no equilíbrio dinâmico entre a prontidão da resposta e a densidade informacional: um trade-off deliberado para favorecer um suporte ao campo simultaneamente ágil e aderente ao acervo institucional, reduzindo o risco de alucinações e apoiando o processo de transferência de tecnologia.

A síntese da performance operacional é consolidada pelo Mapa de Eficiência (Figura 34), que correlaciona a média de qualidade à latência operacional de cada um dos 23 domínios (conforme já relatado, feijão e feijão-caupi viraram apenas “feijão”).

A distribuição espacial dos temas no gráfico sugere que a orquestração agêntica não opera de forma linear, mas ajusta seu esforço de processamento conforme a densidade e a

complexidade da base documental consultada. Observam-se dois agrupamentos principais: os Domínios de Resposta Célere (ex.: Caprinos e Ovinos, com 5,49 s), associados a consultas em acervos mais estruturados; e os Domínios de Recuperação Profunda (ex.: ILPF, com 8,21 s), associados a maior tempo de processamento de múltiplos chunks de informação.

**Figura 34 - Mapa de Eficiência-Qualidade-Tempo por Domínio de Conhecimento**



Nota. O quadrante superior direito concentra temas de maior complexidade e avaliações mais elevadas de qualidade, sendo compatível com a interpretação de que o maior esforço computacional esteve associado a melhor desempenho nesses domínios. Elaboração própria (2026)

Observa-se a formação de dois agrupamentos estratégicos: os Domínios de Resposta Célere (ex: Caprinos e Ovinos, com 5,49 s), caracterizados por consultas diretas em acervos estruturados; e os Domínios de Recuperação Exaustiva (ex: ILPF, com 8,21 s), que demandam maior tempo operacional para o processamento de múltiplos *chunks* de informação.

Conforme Nielsen (1993), embora a latência média de 6,45 segundos ultrapasse o limiar da interatividade instantânea, ela permanece dentro de limites aceitáveis para sistemas de consultoria técnica, nos quais a expectativa do interlocutor tende a priorizar a qualidade diagnóstica em detrimento da velocidade de resposta da interface.

Isso corrobora um achado relevante desta pesquisa, relacionado ao chamado “custo de precisão”. A correlação de Spearman identificada na análise inferencial ( $\rho = 0,1535$ ;  $p < 0,001$ ) sugere que maiores tempos de processamento estiveram associados, com magnitude modesta, a melhores avaliações de qualidade das respostas.

Para a Embrapa, cuja soberania do conhecimento institucional é estratégica, a arquitetura RAG-MCP foi parametrizada para priorizar a aderência factual ao acervo institucional em detrimento da celeridade instantânea. À luz de Siegel e Castellan (1988), essa associação positiva pode ser interpretada como indício de que o artefato ajusta seu esforço de processamento à complexidade da demanda, preservando a integridade da informação entregue ao campo, sem que isso implique relação causal direta.

Como postulado por Gao et al. (2023), a mitigação de alucinações em sistemas de extensão rural digital é crítica, visto que recomendações agronômicas equivocadas acarretam impactos econômicos reais no campo. O "Consultor Digital Ater+ Digital" tende a reduzir esse risco ao ampliar seu ciclo de recuperação semântica conforme a densidade documental.

Esse comportamento é compatível com o conceito de *safety by design*, indicando que o extensionista rural e o agricultor acessaram orientações alinhadas às fontes técnicas institucionais nas interações avaliadas. A utilização do método de *bootstrapping* (Seção 4.2) reforça a precisão das estimativas de qualidade observadas, inclusive nos domínios de maior latência, sem que isso, por si só, constitua evidência de estabilidade global do sistema.

A diferença de magnitude desprezível entre os perfis de usuários (Figura 31) e o processamento com retorno de resposta em 100% das interações avaliadas sinalizam que o

artefato atingiu elevado nível de prontidão tecnológica no contexto do estudo. Ao reduzir o ciclo de busca e síntese — tradicionalmente realizado por meio de navegação manual em repositórios de PDFs — para uma média global de aproximadamente seis segundos em interface dialógica, o sistema sugere potencial de redução do custo de interação (*interaction cost*) no contexto analisado (Zuin et al., 2022). Essa otimização tende a qualificar a experiência do usuário, permitindo que o esforço cognitivo do extensionista seja deslocado da tarefa mecânica de busca documental para a atividade estratégica de assistência técnica e tomada de decisão agronômica.

Essa eficiência transcende a métrica computacional e pode assumir uma dimensão pedagógica: ao ampliar o acesso à base científica da plataforma Ater+ Digital, o sistema tende a fortalecer a confiança do produtor no serviço de extensão. O artefato supera a função de mero indexador documental e passa a atuar como mediador de conhecimento, com potencial para subsidiar a tomada de decisão em janelas críticas da safra — momentos em que a acurácia informacional no *timing* correto define a viabilidade econômica e a sustentabilidade da unidade produtiva.

Sob essa perspectiva, o “Consultor Digital Ater+ Digital” pode atuar como catalisador da Extensão Rural Digital, ao converter o vasto acervo institucional em inteligência aplicada mais acessível, contribuindo para reduzir hiatos de conhecimento e fortalecer o vínculo entre a pesquisa científica e a prática no campo.

A análise da recuperação profunda guarda relação direta com a volumetria documental detalhada no Quadro 03 (Capítulo 3). Domínios como o de sistemas de integração lavoura-pecuária-floresta (ILPF), que concentra o maior volume de ativos indexados (538 documentos), exigem que o motor de busca realize a varredura em um espaço vetorial significativamente mais denso. Esse fenômeno ajuda a explicar a latência de 8,21 segundos observada, sugerindo que a orquestração via MCP ajusta dinamicamente seu esforço de processamento. O sistema demanda

o processamento de um volume superior de *chunks* para extrair o contexto de maior relevância, o que é compatível com a manutenção de avaliações elevadas de qualidade (4,98), mesmo diante da vastidão do acervo.

A convergência das evidências indica que o ‘Consultor Digital Ater+ Digital’ atendeu, nas interações avaliadas, aos requisitos funcionais e de desempenho definidos para o estudo. A validação computacional indicou estabilidade da infraestrutura, enquanto a auditoria qualitativa apontou avaliações elevadas de qualidade nas 23 cadeias produtivas analisadas.

A análise inferencial reforça o rigor da investigação, indicando que o sistema apresentou desempenho consistente nas interações avaliadas e que a priorização da precisão factual, em detrimento da celeridade, foi compatível com os resultados observados. Tais achados sustentam a possibilidade de transição do protótipo experimental para um serviço em ambiente de produção (*deployment*), reforçando o potencial da IA generativa como vetor promissor de inovação para a extensão rural brasileira.

## 5. Considerações Finais e Conclusão

A presente investigação cumpriu seu objetivo primordial de desenvolver, implementar e validar o "Consultor Digital Ater+ Digital", um artefato de inteligência artificial generativa (IAG) integrado à plataforma Ater+ Digital da Embrapa. O problema central da pesquisa — a dificuldade de recuperação eficiente de informações em acervos técnicos densos — foi endereçado mediante uma arquitetura inovadora que uniu a capacidade generativa de modelos de linguagem de grande escala (*Large Language Models* — LLMs) ao rigor do *Retrieval-Augmented Generation* (RAG), orquestrados pelo *Model Context Protocol* (MCP). Ao adotar a metodologia *Design Science Research* (DSR), o estudo superou a mera experimentação, entregando uma solução tecnológica funcional, com evidências de prontidão no contexto analisado e potencial para qualificar o suporte à tomada de decisão no campo.

## 5.1 Contribuições Teóricas e Práticas

No plano teórico, esta pesquisa oferece uma contribuição à literatura de Ciência da Informação e Engenharia de Software ao propor o uso do protocolo MCP como uma camada de orquestração agêntica orientada à soberania informacional. Além disso, o estudo apresenta evidências empíricas da aplicabilidade da arquitetura RAG em bases institucionais públicas de grande escala, um cenário ainda pouco explorado na literatura.

Enquanto a arquitetura RAG convencional foca na recuperação de fragmentos, a integração ao MCP proposta neste estudo favorece que o modelo generativo opere sob restrições estritas de domínio, contribuindo para mitigar o fenômeno das chamadas “alucinações” em modelos generativos, amplamente discutido na literatura recente sobre *Large Language Models*, como descrito por Ji et al. (2023) e Maynez et al. (2020). A pesquisa sustenta a tese de que a soberania informacional não exige a exposição dos bancos de dados brutos; em vez disso, a criação de uma camada de acesso compartilhada via *MCP Server* permite que a inteligência da IA seja alimentada por fontes autoritativas sem comprometer a integridade dos repositórios originais.

Ademais, o estudo introduz na discussão acadêmica a validação do "custo de precisão" como um indicador de esforço cognitivo artificial. A correlação de Spearman ( $\rho = 0,1535$ ;  $p < 0,001$ ) e a análise de recuperação profunda (*deep retrieval*) sugerem que, em domínios complexos – como ILPF, o tempo adicional de processamento esteve associado a melhores avaliações de qualidade das respostas, embora com magnitude modesta, sem que isso caracterize relação causal. Esse achado corrobora as premissas de Wohlin et al. (2012) sobre o equilíbrio necessário em sistemas sociotécnicos complexos, onde a celeridade não deve sobrepor-se à integridade da informação.

Em termos práticos, o trabalho apresenta um artefato com elevado desempenho operacional, no qual 100% das interações avaliadas retornaram resposta, contribuindo para reduzir o custo de interação (*interaction cost*) no ecossistema Ater+ Digital. Antes da implementação do “consultor digital Ater+ Digital”, o acesso ao acervo exigia navegação manual em diretórios complexos e garimpagem em documentos PDF; com a interface dialógica, o tempo médio de resposta de 6,45 segundos amplia a acessibilidade ao patrimônio científico da plataforma oferecida pela Embrapa.

A validação técnica, reforçada pelo método de bootstrapping com 10.000 repetições, indica que o artefato digital apresentou desempenho estável nas interações avaliadas e diferença de magnitude desprezível entre os perfis analisados, provendo suporte consistentemente bem avaliado tanto para extensionistas quanto para agricultores familiares, fortalecendo a inclusão digital no meio rural (Zuin et al., 2022).

Sob a ótica da inovação em modelos de negócio, o estudo sugere o potencial de desenvolvimento de um *marketplace* de contexto da Embrapa. Esta contribuição prática demonstra que a instituição pode extrapolar a extensão rural gratuita, posicionando-se como provedora de "contexto científico auditado" para o mercado de inteligência artificial (IA). Ao compartilhar apenas a camada de acesso (servidores MCP) e não os dados proprietários, a Embrapa preserva a soberania interna enquanto fomenta novos serviços de “agrotecnologia”, favorecendo que o conhecimento oficial oriente soluções públicas ou privadas de suporte à decisão (Embrapa I/O, 2025).

## 5.2 Limitações do Estudo

Uma limitação relevante deste estudo reside no emprego da estratégia *LLM-as-a-Judge* como mecanismo central de auditoria da qualidade das respostas, sem a realização de uma etapa formal de comparação com avaliadores humanos independentes. Embora essa abordagem

ofereça escalabilidade, padronização e viabilidade operacional para grandes volumes de interações, os resultados inferenciais devem ser interpretados com cautela, uma vez que a medida de qualidade deriva de julgamento automatizado e pode estar sujeita a vieses do próprio modelo, bem como a variações de estabilidade interexecução.

A despeito da consistência estatística e da eficácia operacional observada pelo “Consultor Digital Ater+ Digital” desenvolvido em formato de prototipação, a presente investigação identifica limitações estruturais que balizam a interpretação de seus resultados e definem o escopo de sua aplicabilidade imediata. A primeira limitação reside na dependência de infraestrutura de nuvem e de modelos proprietários (*closed-source*).

Embora a arquitetura RAG-MCP favoreça a soberania do contexto institucional, o processamento final da síntese linguística permanece condicionado a APIs externas, como o modelo Gemini 1.5 Flash. Tal configuração introduz vulnerabilidades relacionadas à continuidade do serviço, oscilações em custos de escala e, fundamentalmente, ao que a literatura define como a "opacidade da caixa-preta" dos modelos comerciais. Sob a ótica da soberania tecnológica da administração pública, a impossibilidade de auditoria integral sobre os pesos e filtros internos do modelo linguístico representa um desafio para a autonomia plena da Embrapa (Embrapa Digital, 2025).

No que tange à volumetria e abrangência temática, o estudo limitou-se à indexação de 23 cadeias produtivas e temas transversais, totalizando um *corpus* documental de aproximadamente 1,8 GB. Conquanto esse volume seja representativo para uma validação de prontidão tecnológica, ele abrange apenas uma fração da vasta agrobiodiversidade brasileira e dos complexos sistemas de produção assistidos pela instituição.

A identificação de uma "zona de otimização" para culturas como feijão e uva evidencia que a eficácia da recuperação semântica é sensível à heterogeneidade estrutural dos documentos originais. Tabelas complexas, fragmentos com baixa densidade textual e imagens técnicas ainda

não processadas por visão computacional impõem barreiras à capacidade de síntese do agente, indicando que a qualidade do *input* documental permanece como o principal gargalo para a escalabilidade da IA (Vaswani et al., 2017).

Do ponto de vista metodológico e de avaliação, a pesquisa valeu-se da abordagem *LLM-as-a-Judge* para a validação qualitativa das respostas. Embora este método apresente elevada concordância com especialistas humanos em tarefas de verificação factual (Zheng et al., 2023) e tenha sido blindado pelo método de *bootstrapping*, ele não substitui integralmente a percepção subjetiva e a dimensão dialógica inerente à extensão rural.

A comunicação no campo, como preconizado por Freire (1985) em sua crítica à extensão puramente transmissiva e atualizado por Zuin et al. (2022) para o ambiente digital, é um processo socialmente situado que envolve empatia, confiança mútua e adaptação linguística regional. O artefato avaliado, em sua versão atual, opera exclusivamente em interface textual *web*, o que pode excluir produtores com limitações de letramento digital ou que dependem de interações por áudio e imagem para descrever problemas fitossanitários.

Deve-se, então, considerar a variável da infraestrutura de conectividade rural brasileira. O experimento foi conduzido em ambiente controlado e simulado, utilizando interações que pressupõem estabilidade de rede. Todavia, a realidade da agricultura familiar no Brasil é marcada por uma exclusão digital persistente em áreas remotas (Vieira Filho et al., 2016).

A latência média operacional de 6,45 segundos, embora aceitável para padrões de consultoria técnica, pode ser sensivelmente degradada em cenários de redes móveis assimétricas ou de baixa largura de banda. Essa limitação sinaliza que a onipresença do artefato digital, de IA, como ferramenta de suporte à decisão depende, invariavelmente, do avanço de políticas públicas de conectividade e da evolução para modelos que permitam, no futuro, o processamento parcial em modo *offline* ou local.

### 5.3 Integração dos Procedimentos e Reflexões Metodológicas

A condução deste estudo sob a égide da *Design Science Research* (DSR) proporcionou um alinhamento rigoroso e iterativo entre a necessidade prática do campo e o estado da arte tecnológico. A integração dos procedimentos metodológicos não ocorreu de forma linear, mas sim por meio de um fluxo contínuo de retroalimentação entre os ciclos de relevância, *design* e rigor (Hevner et al., 2004).

O ciclo de relevância foi alimentado pelas lacunas históricas da extensão rural digital brasileira, onde a abundância de dados não se traduzia em completa acessibilidade. Em resposta, o ciclo de *design* permitiu a construção do produto digital aqui desenvolvido, cuja arquitetura RAG-MCP foi sendo refinada para converter o acervo institucional em uma interface dialógica. No entanto, foi no ciclo de rigor que esta pesquisa encontrou sua maior evolução, transcendendo a validação funcional para instituir um protocolo de auditoria estatística robusto.

A reflexão metodológica mais contundente deste trabalho refere-se à transição de uma avaliação meramente descritiva para uma análise inferencial corroborada por métodos de reamostragem. Durante a fase de avaliação, observou-se que a magnitude da amostra ( $N = 4.600$  interações) exigia um tratamento que evidenciasse a estabilidade das métricas de latência e qualidade.

A inserção do método de *bootstrapping* com 10.000 repetições representou uma decisão metodológica crítica, permitindo a construção de intervalos de confiança (IC 95%) que reforçam a previsibilidade observada do artefato. Essa “vigilância estatística” permitiu indicar que os picos de latência e as variações de nota não se explicavam apenas por ruído aleatório, mas também pela complexidade inerente a cada domínio documental, conferindo maior consistência à interpretação dos resultados (Sullivan & Feinn, 2012).

Essa integração procedimental revela que o desenvolvimento de ativos digitais para a administração pública — especialmente em instituições de pesquisa como a Embrapa — deve prescindir da visão puramente tecnicista de engenharia de software em favor de uma abordagem de ciência de dados aplicada.

A reflexão que emerge é que a utilidade de um artefato de IA na extensão rural é indissociável de sua confiabilidade estatística. Ao documentar e examinar cada ciclo de processamento, a metodologia escolhida permitiu que o artefato desenvolvido deixasse de ser apenas experimental para se tornar um conhecimento sistematizado, onde a arquitetura tecnológica e a validação matemática convergem para assegurar a integridade do suporte à decisão agrônômica.

A robustez metodológica alcançada demonstra que a inovação em contextos governamentais exige uma ponte sólida entre o desenvolvimento ágil e o rigor acadêmico. A sistematização dos procedimentos adotados — desde a curadoria documental via *web scraping* até a auditoria via *LLM-as-a-Judge* e o *bootstrapping* — estabelece um *framework* replicável para futuras investigações. Esta pesquisa indica que a transparência sobre os limites e as capacidades do artefato, apoiada pelo rigor dos procedimentos estatísticos, contribui para fundamentar a transição de uma inovação disruptiva para um serviço público com maior confiabilidade.

#### **5.4 Agenda para Pesquisas Futuras**

A conclusão do presente trabalho não encerra a investigação, mas abre fronteiras para o escalonamento e a sofisticação do artefato de inteligência artificial desenvolvido. Como prioridade imediata para trabalhos futuros, propõe-se a transição para a multimodalidade agêntica. Embora a versão atual apresente desempenho razoável no processamento de

linguagem natural (texto), a realidade da extensão rural exige a capacidade de interpretar *inputs* de áudio e imagem.

Pesquisas futuras devem integrar modelos de visão computacional ao *pipeline* RAG, permitindo que o produtor rural envie fotografias de pragas, doenças ou sintomas de deficiências nutricionais para diagnóstico em tempo real. A inclusão de interfaces de voz, permitindo receber relatos orais e resposta por síntese de áudio, é fundamental para garantir a acessibilidade de agricultores com diferentes níveis de letramento, fortalecendo a dimensão dialógica preconizada por Freire (1985) e atualizada para o contexto digital por Zuin et al. (2022).

No que tange à capilaridade do artefato, a agenda de pesquisa deve contemplar a omnicanalidade e a integração de ecossistemas. O “Consultor Digital Ater+ Digital” deve transcender a interface *web* da plataforma Ater+ Digital para integrar-se aos canais de comunicação já consolidados no cotidiano rural, como WhatsApp, Telegram e Instagram.

O uso de *chatbots* em aplicativos de mensagens instantâneas, conforme discutido por Souza et al. (2025), reduz a fricção tecnológica e aproveita o hábito de uso preexistente do produtor. Além disso, sugere-se investigar a portabilidade do consultor digital para televisores interativos em centros comunitários rurais e cooperativas, transformando o suporte à decisão em uma ferramenta coletiva e onipresente. Essa expansão exige que futuros estudos avaliem a consistência das respostas em diferentes interfaces, preservando o rigor da auditoria algorítmica e a precisão das estimativas intervalares adotadas nesta pesquisa.

Uma vertente estratégica para o desenvolvimento de negócios reside na formalização de um *marketplace* de contexto e governança via MCP. Propõe-se investigar um modelo em que a Embrapa atue como provedora de "contexto científico auditado" para parceiros privados, o próprio governo e outras *agtechs*.

Nesta arquitetura, a soberania dos dados brutos é mantida internamente, enquanto a camada de acesso — o *MCP Server* ou equivalente — é compartilhada mediante protocolos de segurança e eventuais modelos de monetização ou parcerias institucionais. Pesquisas futuras podem explorar como essa descentralização da consulta, sem a perda do controle sobre o acervo original, fomenta um ecossistema de inovação aberta onde o conhecimento da Embrapa alimenta diversas soluções de mercado, garantindo que a informação técnica oficial seja o motor de novos sistemas de suporte à decisão (Embrapa I/O, 2025).

No plano da infraestrutura e soberania tecnológica, recomenda-se a investigação de modelos de linguagem de pequena escala (*small language models* — SLMs) e processamento local. A dependência atual de APIs de nuvem proprietárias pode ser mitigada pelo uso de modelos de código aberto executados *on-premise* ou em dispositivos de borda (*edge computing*).

Estudos futuros devem comparar o desempenho desses modelos em relação à latência e acurácia factual, buscando viabilizar o uso do “Consultor Digital Ater+ Digital” em ambientes com conectividade nula ou intermitente. A meta é alcançar uma solução que preserve a privacidade dos dados e reduza custos operacionais, mantendo o “custo de precisão” em patamares compatíveis com uma futura operação em maior escala nacional, conforme os critérios metodológicos de Wohlin et al. (2012).

A agenda científica deve incluir a avaliação de impacto socioeconômico longitudinal. É imperativo que futuras pesquisas conduzam testes de campo exaustivos para mensurar como a redução no custo de interação e o acesso ágil à informação técnica impactam a produtividade e a sustentabilidade das unidades produtivas familiares.

Esse esforço exige uma abordagem mista, combinando o rigor da análise de dados em larga escala — possivelmente expandindo o método de *bootstrapping* para variáveis de campo — com estudos etnográficos que capturem a percepção de confiança e a mudança de

comportamento dos extensionistas. Somente por meio dessa validação social será possível examinar se a transformação digital promovida pela presente pesquisa contribui, efetivamente, para uma extensão rural mais inclusiva, soberana e resiliente frente aos desafios da agricultura 4.0 (Vieira Filho et al., 2016).

### **5.5 Considerações Institucionais e Síntese Final**

Como diretriz institucional emergente desta investigação, vislumbra-se a relevância de institucionalizar um macroprocesso formal de curadoria documental voltado especificamente à Inteligência Artificial (IA). A análise da "zona de otimização" para as cadeias de feijão e uva, por exemplo, revela que a eficácia da arquitetura RAG-MCP é diretamente proporcional à higienização semântica, à ampliação do *corpus* e à estruturação dos ativos digitais originais.

Nesse prisma, depreende-se que a curadoria de dados deve transcender a digitalização convencional para incorporar o *chunking* semântico assistido e a indexação de metadados como etapas indissociáveis do fluxo de publicação científica. Essa proposição favorece que o conhecimento técnico permaneça inteligível para modelos generativos, minimizando riscos de interpretações parciais em temas de alta complexidade estrutural e assegurando que o *input* informacional reflita a excelência da pesquisa agropecuária.

Sob o prisma da governança e da inovação em modelos de entrega, a arquitetura implementada aponta para a viabilidade de um *Marketplace* de Contexto. Tal estratégia, alicerçada na interoperabilidade do protocolo MCP, faculta à instituição o compartilhamento da "camada de acesso" ao seu conhecimento auditado com parceiros privados, o próprio governo e *agtechs*, prescindindo da exposição direta ou da replicação dos repositórios vetoriais brutos (Embrapa I/O, 2025).

Ao preservar a soberania interna dos dados e disponibilizar o servidor MCP, ou alternativa semelhante, como um serviço de contexto confiável, estabelece-se um modelo de

inovação aberta sustentável. A referida abordagem possibilita que o selo de autoridade institucional atue como diferencial competitivo em soluções privadas e públicas de suporte à decisão, ao mesmo tempo em que contribui para a proteção do patrimônio intelectual e para o fortalecimento do ecossistema de agricultura digital.

O escalonamento do artefato desenvolvido sugere uma evolução pautada pela omnicanalidade e multimodalidade. A integração do consultor digital a canais de alta capilaridade social, como *WhatsApp* e *Telegram*, apresenta-se como uma via estratégica para reduzir a fricção tecnológica, aproveitando hábitos de uso já consolidados no meio rural (Souza et al., 2025).

Nesse contexto de expansão, a portabilidade para interfaces como telas interativas em cooperativas, por exemplo, amplia o alcance comunitário da informação científica. Concomitante a isso, a incorporação do processamento de áudio e visão computacional — permitindo diagnósticos fitossanitários via imagem e diálogos por voz — revela-se como uma estratégia imperativa para mitigar as barreiras do letramento digital. Essas funcionalidades podem ampliar a inclusão de agricultores que dependem de interações não textuais, tornando o suporte técnico potencialmente mais acessível.

No âmbito da gestão e valorização do capital humano, a introdução da IA deve ser compreendida como um acelerador da inteligência territorial do extensionista rural, e não como sua substituição. O processo de transição desse profissional, da função de buscador manual de documentos para a de gestor estratégico de recomendações técnicas, demanda programas de letramento digital focados na interpretação crítica dos *outputs* do modelo.

Essa mudança de paradigma, alinhada à abordagem participativa e dialógica (Zuin et al., 2022), reforça a confiança sistêmica e permite que o esforço humano seja deslocado da tarefa mecânica de garimpo informacional para a atividade tática da assistência técnica. Nessa

perspectiva, o “consultor digital” pode atuar como um catalisador de interações mais qualificadas e mais aderentes às necessidades da unidade produtiva.

À luz das evidências apresentadas, este estudo ratifica que o agente de IA desenvolvido representa uma contribuição relevante para a democratização do acesso ao acervo técnico da plataforma Ater+ Digital. A convergência entre o rigor da *Design Science Research* e a inovação da arquitetura RAG-MCP demonstrou a viabilidade de aliar tecnologia generativa de fronteira à preservação da soberania informacional da instituição.

Ao converter gigabytes de dados técnicos em uma interface dialógica ágil, segura e com desempenho estatisticamente consistente no contexto analisado, o artefato avança na transformação do conhecimento científico em inteligência prática aplicada. A solução posiciona-se como uma referência promissora para a extensão rural digital, favorecendo que a informação oficial alcance o extensionista e o produtor rural com maior aderência ao acervo institucional e potencial para fortalecer a resiliência e a competitividade do setor agropecuário brasileiro na era da Inteligência Artificial.

## 6. Referências

Bardin, L. (2016). *Análise de conteúdo*. Edições 70.

Bolivar Huayhua, E. A. (2025). *Integração de LLMs na otimização e planejamento de sistemas multiagentes IoRT e interação conversacional*.

[https://bdtd.ibict.br/vufind/Record/UNICAMP-30\\_9f553b08f01d330a99887822b0bf03aa](https://bdtd.ibict.br/vufind/Record/UNICAMP-30_9f553b08f01d330a99887822b0bf03aa)

Borges, A. M. (2025). *Aprimorando o atendimento ao cidadão com RAGLC (retrieval augmented generation in logic context)*. <https://doi.org/10.14393/ufu.di.2025.5207>

MUST UNIVERSITY

70 SW 10th St, Deerfield Beach, FL 33441, USA | [info@mustedu.com](mailto:info@mustedu.com) | [www.mustedu.com](http://www.mustedu.com)

MUST University®: licensed by Florida Commission for Independent Education. License: 5593.

Brasil. (2010). Lei nº 12.188, de 11 de janeiro de 2010. *Institui a Política Nacional de Assistência Técnica e Extensão Rural para a Agricultura Familiar e Reforma Agrária (PNATER)*. [https://www.planalto.gov.br/ccivil\\_03/\\_ato2007-2010/2010/lei/112188.htm](https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/lei/112188.htm)

Campos, P. H. de. (2025). *Large language models em contabilidade: Potenciais usos, desafios e limitações* [Dissertação de mestrado, Universidade Federal do Rio Grande do Sul]. <https://lume.ufrgs.br/handle/10183/297761>

Carromeu, C., D'Oliveira, F. M., Oliveira, R. T. de, Silva Júnior, M. M. da, & Araújo, R. F. (2023). *Embrapa I/O: Plataforma de desenvolvimento colaborativo, entrega automatizada e monitoramento contínuo de softwares para a agricultura digital*. In Anais do Congresso Brasileiro de Software (CBSOFT – Sessão Estendida). [https://sol.sbc.org.br/index.php/cbsoft\\_estendido/article/view/26043](https://sol.sbc.org.br/index.php/cbsoft_estendido/article/view/26043)

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associates.

Davis, E. (2023). *Neural-augmented automation frameworks for scalable multi-stage workflow orchestration, predictive task execution, and contextual decision making in n8n environments*. Global Knowledge Academy, 5(4), 73–81. <https://globalknowledgeacademy.com/index.php/gna/article/view/67>

Deus, C. da C. N. de, Machado, B. de S., Ferraz, R. M., Lopes, R. de C., & Braga, M. J. (2024). *ATER digital no Brasil [Cartilha eletrônica]*. Universidade Federal de Viçosa. <https://bibliotecasemiarios.ufv.br/jspui/handle/123456789/3342>

Embrapa. (2018). *Visão 2030: O futuro da agricultura brasileira*. <https://abre.ai/futuroagro>

Embrapa. (2022). *Política de inovação da Embrapa (Norma 037.005.001.012)*. <https://www.embrapa.br/politica-de-inovacao>

Embrapa. (2024). *Plataforma Ater+ Digital disponibiliza conteúdos on-line para técnicos extensionistas*. Portal Embrapa. <https://www.embrapa.br/busca-de-noticias/-/noticia/88808410/plataforma-ater-digital-disponibiliza-conteudos-on-line-para-tecnicos-extensionistas>

Embrapa. (2025a). *Plataforma Ater+ Digital ganha novos conteúdos para extensionistas*. Portal Embrapa. <https://www.embrapa.br/busca-de-noticias/-/noticia/100585637/plataforma-ater-digital-ganha-novos-conteudos-para-extensionistas>

Embrapa. (2025b). *Arquitetura de referência em IA generativa do Embrapa I/O*. <https://www.embrapa.io/docs/preface/>

Embrapa. (2025c). *Macroprocesso de desenvolvimento de ativos digitais da Embrapa [Videoaula]*. YouTube.

<https://www.youtube.com/watch?v=DFc2iZy4Ox4&list=PLBTsCa6Rsuz3jh0S2P8aTBlvhk7Mr3gbS>

Embrapa. (s.d.-a). *Sobre nós. Plataforma Ater+ Digital.*

<https://www.atermaisdigital.cnptia.embrapa.br/sobre-nos>

Embrapa. (s.d.-b). *Plataforma Ater+ Digital.*

<https://www.atermaisdigital.cnptia.embrapa.br>

Embrapa Agricultura Digital. (2025). *Riscos, desafios e regulação da inteligência artificial na agricultura.* Embrapa.

<http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1180748>

Embrapa for Developers. (n.d.). *Embrapa I/O: Plataforma de desenvolvimento de ativos digitais agropecuários da Embrapa.* <https://www.embrapa.io>

Field, A. (2013). *Discovering statistics using IBM SPSS statistics.* Sage Publications.

Food and Agriculture Organization of the United Nations. (2023). *Guide on digital agricultural extension and advisory services.* FAO. <https://www.fao.org/family-farming/detail/en/c/1641998/>

Freire, P. (1985). *Extensão ou comunicação?* (R. D. de Oliveira, Trans.). Paz e Terra.

Gao, Y., et al. (2023). *Retrieval-augmented generation for large language models: A survey*. <https://doi.org/10.48550/arXiv.2312.10997>

Gil, A. C. (2019). *Métodos e técnicas de pesquisa social (7ª ed.)*. Atlas.

Global Knowledge Academy. (2024). *How to connect AI agents with n8n for end-to-end automation*. Global Knowledge Academy, 5(4), 73–81.

Godoy, A. S. (1995). *Introdução à pesquisa qualitativa e suas possibilidades*. Revista de Administração de Empresas, 35(2), 57–63. <https://doi.org/10.1590/S0034-75901995000200008>

Godoy, M. C. (2024). *Aplicabilidade de busca semântica para recuperação e ranqueamento de arquivos [Trabalho de conclusão de curso, Universidade Federal do Rio Grande do Sul]*. <https://lume.ufrgs.br/handle/10183/279179>

Google. (2026). *Gemini 2.5 Flash: Generative AI on Vertex AI*. Google Cloud documentation. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash?hl=pt-br>

Google Cloud. (2024). *Text embeddings API: text-embedding-001*. Google Cloud Vertex AI documentation. <https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api>

- Hauke, J., & Kossowski, T. (2011). *Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data*. *Quaestiones Geographicae*, 30(2), 87–93. <https://doi.org/10.2478/v10117-011-0021-1>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). *Design science in information systems research*. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
- Hou, X., Zhao, Y., & Wang, H. (2025). *The next frontier of LLM applications: Open ecosystems and hardware synergy*. <https://arxiv.org/abs/2503.04596>
- Instituto de Desenvolvimento Rural do Paraná. (n.d.). *IDR-Paraná Digit@l: Projeto de modernização*. IDR-Paraná. <https://abre.ai/idrprdigital>
- Instituto de Pesquisa Econômica Aplicada. (2022). *Assistência técnica e extensão rural no Brasil e no mundo: Qual o papel da ATER pública?* IPEA. <http://repositorio.ipea.gov.br/handle/11058/11410>
- LangChain. (2024). *Text splitters: RecursiveCharacterTextSplitter documentation*. [https://python.langchain.com/docs/modules/data\\_connection/document\\_transformers/recursive\\_text\\_splitter](https://python.langchain.com/docs/modules/data_connection/document_transformers/recursive_text_splitter)
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://arxiv.org/abs/2005.11401>

Lima, A. A. G. (2024). *A ATER pública frente à transformação digital na agricultura: Construção e perspectivas.*

Lima, S. C. da C. (2025). *Assistente de busca: Uma abordagem RAG para busca semântica em documentos textuais da Assembleia Legislativa do Rio Grande do Norte* [Dissertação de mestrado, Universidade Federal do Rio Grande do Norte].  
<https://abre.ai/ragalrn>

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2024). *G-Eval: NLG evaluation using GPT-4 with better human alignment.* In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.  
<https://arxiv.org/abs/2303.16634>

Marconi, M. A., & Lakatos, E. M. (2021). *Técnicas de pesquisa* (9ª ed.). Atlas.

Malkov, Y. A., & Yashunin, D. A. (2018). *Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4), 824–836.  
<https://doi.org/10.1109/TPAMI.2018.2889473>

McKnight, P. E., & Najab, J. (2010). *Mann-Whitney U test.* In *The Corsini encyclopedia of psychology.* John Wiley & Sons.  
<https://doi.org/10.1002/9780470479216.corpsy0524>

Miller, R. B. (1968). *Response time in man-computer conversational transactions*.

Model Context Protocol. (2024). *MCP specification overview*.  
<https://modelcontextprotocol.io/specification/2025-11-25>

Model Context Protocol. (n.d.). *Architecture overview. Model Context Protocol documentation*. <https://modelcontextprotocol.io/docs/learn/architecture>

Nelimarkka, P. (2023). *Emerging low-code/no-code paradigm: Evaluating adoption, opportunities, and cyber security challenges in the information technology sector* [Master's thesis, JAMK University of Applied Sciences].

Netto, M. A. M. P., & Brito, P. F. de. (2025). *Desenvolvimento de um chatbot informativo sobre saúde mental utilizando a arquitetura RAG*. Centro Universitário Luterano de Palmas. <https://ulbra-to.br/encoinfo/edicoes/2024/artigos/desenvolvimento-de-um-chatbot-informativo-sobre-saude-mental-utilizando-a-arquitetura-rag/>

Nielsen, J. (1993). *Usability engineering*. Morgan Kaufmann.

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2023). *Trust in AI chatbots: A systematic review*. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2023.107935>

Norman, G. (2010). *Likert scales, levels of measurement and the “laws” of statistics*. *Advances in Health Sciences Education*, 15(5), 625–632.  
<https://doi.org/10.1007/s10459-010-9222-y>

Nunes, L. C. de M., Vendrusculo, L. G., Bergier, I., Silva, A. P. M., Baldan, A., Veiga, M. P., Souza, K. X. S. de, Duarte, M. C., & Ribeiro, E. (2025). *Arquitetura aberta de IA para ILPF: Desenvolvimento e validação de um consultor digital baseado em RAG e MCP Server*. Embrapa.

n8n. (2024). *Building AI agents with n8n: Documentation*. <https://docs.n8n.io>

Oliveira Filho, A. A. de, Barbosa, V. M., Machado, G. B., Silva, T. P. B., Franca, T. N., Mercês, N. B. das, Martins, L. K. S. A., & Azevedo, I. L. (2021). *Poulbot – O chatbot extensionista rural na área de incubação artificial de frangos de corte*. In Anais de Eventos Acadêmicos. <https://sol.sbc.org.br/index.php/washes/article/view/15888>

Ostertagová, E., Ostertag, O., & Kováč, J. (2014). *Methodology and application of the Kruskal-Wallis test*. Applied Mechanics and Materials, 611, 115–120. <https://doi.org/10.4028/www.scientific.net/AMM.611.115>

Pacífico, D. A. (2022). *Extensão rural no Brasil: Por uma análise da instrumentação da ação pública*. Mundo Agrario, 23(52), e187. <https://doi.org/10.24215/15155994e197>

Pan, J. J., Wang, J., & Li, G. (2023). *Survey of vector database management systems*. The VLDB Journal. <https://doi.org/10.48550/arXiv.2310.14021>

Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do trabalho científico: Métodos e técnicas da pesquisa e do trabalho acadêmico* (2ª ed.). Feevale.

Qdrant. (2024). *Vector similarity search engine documentation*.  
<https://qdrant.tech/documentation/overview/vector-search/>

Qdrant. (n.d.). *What is Qdrant? Qdrant documentation*.  
<https://qdrant.tech/documentation/overview/>

Reddy, P. D., et al. (2025). *Agri Assist: An AI-integrated farmer assistant*. *Procedia Computer Science*, 258, 3510–3522. <https://doi.org/10.1016/j.procs.2025.01.403>

Ribeiro, P. F. R., Pires, J. de L., Bastos, P. A., Rigo, R., Reis, M. H. A. dos, & Hubner, K. V. (2025). *Bancos vetoriais e modelos de embedding: Avaliação comparativa de desempenho na recuperação semântica em língua portuguesa*. *Research, Society and Development*, 14(10), e106141049768. <https://doi.org/10.33448/rsd-v14i10.49768>

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Sage Publications.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Prentice Hall.

Santos, A. O. (2023). *Percepções dos agentes de ATER sobre inovações tecnológicas e novas competências profissionais*. <https://doi.org/10.13140/RG.2.2.15487.42401>

Santos, T., & Zonin, V. J. (2024). *Uma revisão sistemática sobre os serviços de assistência técnica e extensão rural: Uma segmentação presente*. Revista GeSec, 15(1), 1051–1070. <https://doi.org/10.7769/gesec.v15i1.3389>

Saraiva, A. M., Osório, F. S., Colaço, A. F., Drucker, D. P., Mendiondo, E. M., Corrêa, F. E., ... Delbem, A. C. B. (2024). *A inteligência artificial na pesquisa agrícola*. Estudos Avançados, 38(110), 45–68. <https://doi.org/10.11606/issn.2316-9036.i141p91-106>

Shepherd, K. D., Miller, M. A. E., Kisitu, B., Miles, B. G., Gbedevi, K., Chunga, P., Thuo, A. G. W., Musiiwa, R. F., Gyampo, E. D., Cross, M., Dobermann, A., & Collinson, J. (2025). *Virtual agronomist – An AI-assisted chatbot for guiding crop management decisions of smallholder farmers in Africa*. agriRxiv. <https://doi.org/10.31220/agriRxiv.2025.00332>

Shuster, K., et al. (2021). *Retrieval augmentation reduces hallucination in conversation*. <https://doi.org/10.48550/arXiv.2104.07567>

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.

Souza, Y. D. de, Sales, S. G. da S., Ávila, M. L. de, & Del Grossi, M. E. (2025). *Chatbot como ferramenta de coleta: Aplicação piloto na produção orgânica no Rio Grande do Norte*. Geopauta, 9, e18064. <https://doi.org/10.22481/rg.v9.18064>

Sullivan, G. M., & Feinn, R. (2012). *Using effect size—Why the p value is not enough*. *Journal of Graduate Medical Education*, 4(3), 279–282.  
<https://doi.org/10.4300/JGME-D-12-00156.1>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1706.03762>

Vieira Filho, J. E. R., Gasques, J. G., & Carvalho, A. X. Y. (2016). *Agricultura, transformação produtiva e sustentabilidade*. IPEA.  
<http://repositorio.ipea.gov.br/handle/11058/6876>

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.  
<https://doi.org/10.1007/978-3-642-29044-2>

Zheng, L., et al. (2023). *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*.  
<https://doi.org/10.48550/arXiv.2306.05685>

Zuin, L. F. S., Gregori, F., Zuin, P. B., Trentini, F., Ribeiro, I. P., Bastos, P. A. S., Zanella, A. J., Díez, J. G., Caneppele, F. L., Valle, L. R., Vaz, J. A. M. C., Rached, R. Z., Vecchio, D., Kamimura, D. T. T., Alexandre, J. R., Parra, H. D., Arroyo, G., Moraes, W. S., Maia, M. A., Cunha, L. E. C., Baesso, M. M., Zanella, M. I. G., Ferreira, W. C. S. M., & Manrique, M. A. D. (2022). *Ater digital participativa:*

*Metodologias pedagógicas e exemplos de aplicação.* EDUEPB.

<https://abre.ai/aterparticipativa>

Zunino, O. C. de O. (2022). *Robôs sociáveis e organizações em maturação: O processo de criação de chatbots no contexto comercial* [Dissertação de mestrado, Universidade do Estado do Rio de Janeiro]. <https://www.btdt.uerj.br:8443/handle/1/17967>