IAGRO: UM CHATBOT BASEADO EM GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG) PARA RECOMENDAÇÃO DE CULTURAS AGRÍCOLAS A PARTIR DE DADOS DE SOLO E CLIMA

Universidade Presbiteriana Mackenzie
Pró-Reitoria de Pesquisa e Pós-Graduação - Coordenadoria de Fomento à Pesquisa

Beatriz Pimenta de Camargo[1]
Apoio: PIVIC Mackenzie

Gustavo Scalabrini Sampaio[1]

[1] Faculdade de Computação e Informática (FCI)

Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

<beatriz.camargo@mackenzista.com.br>[1]
<gustavo.sampaio@mackenzie.br>[2]

2026

Resumo

O agronegócio ocupa posição central na economia brasileira, respondendo por parcela expressiva do Produto Interno Bruto e da geração de empregos. Apesar disso, o acesso a informações técnicas que orientem a escolha da cultura mais adequada às condições de solo e clima ainda representa um desafio, sobretudo para pequenos e médios produtores, uma vez que tais informações costumam estar dispersas em bases de dados de difícil consulta e exigem conhecimento especializado para sua interpretação. Neste contexto, os Modelos de Linguagem de Grande Escala (LLMs) apresentam-se como alternativa promissora para a consulta em linguagem natural, embora estejam sujeitos a alucinações e limitados ao conhecimento adquirido em treinamento. Neste trabalho, propõe-se o desenvolvimento do IAgro, um sistema conversacional construído sobre a arquitetura de Geração Aumentada por Recuperação (*Retrieval-Augmented Generation* – RAG), capaz de responder em linguagem natural a perguntas sobre recomendação de culturas a partir de parâmetros de solo e clima. Para tal, utilizou-se o conjunto de dados público *Crop Recommendation Dataset*, composto por 2.200 registros distribuídos entre 22 culturas agrícolas e descritos por sete atributos numéricos. Os registros foram agrupados por cultura e convertidos em documentos textuais padronizados, vetorizados pelo modelo de *embeddings text-embedding-3-large* e indexados no banco vetorial Pinecone, com orquestração realizada pela biblioteca LangChain. A validação foi conduzida por meio de um conjunto de perguntas que simulam consultas reais de produtores, comparando-se as respostas geradas com os valores obtidos diretamente da base. Os resultados evidenciam que o ajuste da granularidade dos documentos e da cobertura da recuperação foi determinante para a correção das respostas em perguntas comparativas, demonstrando a viabilidade da arquitetura RAG aplicada a dados agrícolas estruturados.

Palavras-chave: Agronegócio, Geração Aumentada por Recuperação, Modelos de Linguagem de Grande Escala, Busca Semântica.

Abstract

Agribusiness plays a central role in the Brazilian economy, accounting for a significant share of the Gross Domestic Product and of job creation. Nevertheless, access to technical information that guides the choice of the most suitable crop for given soil and climate conditions remains a challenge, especially for small and medium-sized farmers, since such information is usually scattered across databases that are difficult to query and require specialized knowledge to interpret. In this context, Large Language Models (LLMs) emerge as a promising alternative for natural language querying, although they are subject to hallucinations and limited to the knowledge acquired during training. In this study, the development of IAgro is proposed, a conversational system built upon the Retrieval-Augmented Generation (RAG) architecture, able to answer questions in natural language about crop recommendation based on soil and climate parameters. For this purpose, the public Crop Recommendation Dataset was used,

comprising 2,200 records distributed across 22 agricultural crops and described by seven numerical attributes. The records were grouped by crop and converted into standardized textual documents, vectorized by the text-embedding-3-large embedding model and indexed in the Pinecone vector database, with orchestration performed by the LangChain library. Validation was conducted through a set of questions simulating real farmer queries, comparing the generated answers with values obtained directly from the database. The results show that adjusting document granularity and retrieval coverage was decisive for the correctness of answers to comparative questions, demonstrating the feasibility of the RAG architecture applied to structured agricultural data.

Keywords: Agribusiness, Retrieval-Augmented Generation, Large Language Models, Semantic Search.

1 INTRODUÇÃO

Atualmente, o Brasil ocupa uma posição de destaque no ranking mundial dos países que mais investem no setor do agronegócio. Segundo a Agrow Negócios (AGROW NEGÓCIOS, 2023), o Brasil está em 5º lugar entre os países que mais investem recursos nessa indústria; além disso, é um dos países que mais produz alimentos, como arroz, trigo e milho, ocupando a 4ª posição do ranking mundial. Pela grande diversidade de vegetação, clima, solo e por sua ampla faixa territorial, o país tem a possibilidade de inúmeros cultivos. De acordo com dados do CEPEA (CEPEA, 2025), o agronegócio representou 23,3% do Produto Interno Bruto (PIB) nacional de 2024, sendo também um dos setores que mais geram empregos (EMBRAPA, 2020).

Devido à relevância do agronegócio, diversas iniciativas de modernização do setor, por meio da integração tecnológica, como a aplicação de inteligência artificial (IA), proporcionam inúmeras possibilidades de pesquisa e inovação. Porém, muito se discute sobre a forma que esse setor irá receber essa modernização, sendo necessário que não haja perda da qualidade do cultivo, algum fator que gere risco para a plantação ou algo que seja prejudicial para o meio ambiente e, assim, para os consumidores finais. Além disso, o agronegócio é regido por diversas leis agrárias e ambientais, as quais têm o objetivo de preservar o meio ambiente e desenvolver uma produção responsável e sustentável (EMBRAPA, 2020). Dessa forma, é necessário que a inovação tecnológica esteja alinhada com essas legislações, o que reforça a importância de soluções que aproximem o produtor da tecnologia sem comprometer as boas práticas produtivas.

Apesar desse potencial, o acesso a informações técnicas que orientem a escolha da cultura mais adequada às condições de solo e clima ainda representa um desafio, sobretudo para pequenos e médios produtores. Tais informações, quando disponíveis, costumam estar dispersas em bases de dados técnicas de difícil consulta para o produtor leigo, exigindo conhecimento especializado para sua interpretação (EMBRAPA, 2019). Essa barreira de acesso constitui a lacuna que este trabalho busca preencher, ao propor uma forma de consulta simples e em linguagem natural sobre dados agrícolas.

Diante desse cenário, este trabalho tem por objetivo desenvolver uma solução tecnológica que auxilie pequenos e médios produtores na tomada de decisão quanto ao seu cultivo. A proposta consiste na criação de um *chatbot*, denominado IAgro, capaz de responder, em linguagem natural, a perguntas relacionadas à recomendação de culturas a partir de parâmetros de solo e clima. Para isso, a ferramenta foi construída sobre a arquitetura de Geração Aumentada por Recuperação (*Retrieval-Augmented Generation* – RAG), que combina a recuperação de informações de uma base de dados com a geração de respostas por um modelo de linguagem de grande escala (LLM). Para atingir esse objetivo geral, foram definidos os seguintes objetivos específicos: (i) estruturar uma base de dados de culturas em formato adequado para busca semântica; (ii) aplicar técnicas de vetorização (*embeddings*) sobre os dados, armazenando-os em um banco vetorial; (iii) implementar um mecanismo de recuperação semântica que selecione as informações relevantes para cada consulta; e (iv) integrar essa recuperação a um modelo de linguagem, de modo a gerar

respostas em linguagem natural fundamentadas nos dados recuperados.

Como contribuições, este trabalho apresenta desdobramentos tanto de ordem prática quanto científica. Do ponto de vista prático, o IAgro oferece ao pequeno e médio produtor um canal de consulta em linguagem natural, dispensando conhecimento técnico especializado para interpretar bases de dados agronômicas e reduzindo o tempo e o esforço associados à busca por orientação técnica. Ao fundamentar as respostas em dados de solo e clima, a solução apoia a escolha de culturas mais adequadas às condições da propriedade, o que tende a favorecer o aproveitamento dos recursos disponíveis e a reduzir perdas decorrentes de decisões mal informadas, alinhando-se às práticas de produção responsável previstas na legislação ambiental. Do ponto de vista científico, o trabalho contribui ao demonstrar a viabilidade da arquitetura RAG aplicada a dados agrícolas estruturados, abordagem ainda pouco explorada nesse domínio quando comparada às soluções baseadas em visão computacional; ao documentar o impacto da granularidade dos documentos e da cobertura da recuperação sobre a qualidade das respostas, oferece subsídios metodológicos reaproveitáveis por outros trabalhos que utilizem RAG sobre bases tabulares. Por fim, o estudo fornece uma base replicável para extensões futuras, como a incorporação de fontes oficiais de dados e o desenvolvimento de interfaces voltadas ao produtor final.

O trabalho está organizado da seguinte forma: o Referencial Teórico (2) apresenta os conceitos de Modelos de Linguagem de Grande Escala, de representação vetorial de texto e de Geração Aumentada por Recuperação, além dos trabalhos relacionados. A seção de Metodologia (3) descreve a base de dados utilizada, a estruturação dos documentos, a vetorização e indexação, o mecanismo de recuperação e geração de respostas e o processo de teste e validação. A seção de Resultados e Discussão (4) apresenta e analisa os resultados obtidos. Por fim, a seção de Conclusão (5) resume os principais achados do estudo e aponta trabalhos futuros.

2 REFERENCIAL TEÓRICO

Esta seção apresenta os fundamentos teóricos que sustentam o desenvolvimento deste trabalho. São abordados os conceitos de Modelos de Linguagem de Grande Escala (LLMs), de representação vetorial de texto por meio de *embeddings* e busca semântica, e da arquitetura de Geração Aumentada por Recuperação (RAG), que articula os dois anteriores. Por fim, são apresentados os trabalhos relacionados que fundamentaram as decisões metodológicas adotadas.

2.1 *Large Language Models* (LLMs)

As LLMs representam um grande avanço no campo do Processamento de Linguagem Natural (PLN). São modelos de inteligência artificial treinados com grandes volumes de dados textuais, capazes de compreender, gerar e interagir em linguagem natural. Com bilhões de parâmetros, as LLMs identificam padrões complexos e desempenham tarefas como resposta a perguntas, tradução, resumo e criação de textos (RAIAAN et al., 2024). Neste projeto, um LLM

é empregado como componente final responsável por interpretar as informações recuperadas da base de dados e redigir a resposta ao usuário em linguagem natural.

2.2 *Embeddings* e Busca Semântica

Modelos computacionais não operam diretamente sobre texto, mas sobre números. A primeira etapa desse processo é a tokenização, na qual o texto é fragmentado em unidades menores chamadas *tokens*, que podem corresponder a palavras inteiras, a partes de palavras ou a sinais de pontuação, e cada *token* recebe um identificador numérico. Esses identificadores, porém, são apenas rótulos arbitrários e não carregam qualquer noção de significado: nada neles indica que “chuva” esteja mais próximo de “precipitação” do que de “trator”.

É nesse ponto que entram os *embeddings*. Um *embedding* é um vetor de números reais, com centenas ou milhares de dimensões, que representa o significado de um *token*, palavra ou trecho de texto. Esses vetores não são atribuídos manualmente: são aprendidos durante o treinamento do modelo a partir dos contextos em que cada termo aparece, seguindo a hipótese distributiva de que palavras que ocorrem em contextos semelhantes tendem a ter significados semelhantes. O resultado é um espaço vetorial semântico, no qual a posição de cada vetor reflete o sentido do texto que ele representa, de modo que termos com significados próximos ocupam regiões próximas.

Uma propriedade notável desse espaço é que ele preserva relações de sentido na forma de operações geométricas. O exemplo clássico da literatura mostra que, ao subtrair o vetor de “homem” do vetor de “rei” e somar o vetor de “mulher”, o vetor resultante fica próximo do vetor de “rainha”, indicando que a diferença entre os pares codifica a noção de gênero de maneira consistente (MIKOLOV; YIH; ZWEIG, 2013). De modo geral, palavras semanticamente relacionadas agrupam-se em regiões vizinhas do espaço, e as relações entre pares de termos aparecem como deslocamentos aproximadamente paralelos.

A proximidade entre vetores é medida por métricas de similaridade, sendo a similaridade por cosseno a mais utilizada. Ela avalia o ângulo entre dois vetores: quanto menor o ângulo, maior a similaridade semântica entre os textos representados, independentemente do comprimento dos vetores. É essa medida que viabiliza a busca semântica, na qual a pergunta do usuário é convertida em vetor pelo mesmo modelo de *embeddings* utilizado para indexar os documentos, e a recuperação seleciona aqueles cujos vetores estão mais próximos do vetor da pergunta.

A vantagem dessa estratégia sobre a busca léxica tradicional, baseada em correspondência de palavras-chave, é que a busca semântica recupera conteúdo relevante mesmo quando não há coincidência exata de termos. Uma consulta como “qual cultura precisa de bastante água?” pode recuperar um documento que menciona precipitação elevada, ainda que a palavra “água” não apareça nele. Essa técnica é a base do mecanismo de recuperação utilizado neste projeto: cada uma das culturas da base é representada por um vetor, e a consulta do produtor é comparada a esses vetores para localizar as informações pertinentes.

2.3 Geração Aumentada por Recuperação (RAG)

Apesar da capacidade das LLMs de gerar texto fluente, esses modelos apresentam limitações relevantes quando aplicados a domínios específicos. Seu conhecimento está restrito aos dados vistos durante o treinamento, o que impede o acesso a informações posteriores ao corte temporal do modelo ou a bases privadas de uma organização. Além disso, as LLMs estão sujeitas ao fenômeno das alucinações, no qual o modelo produz afirmações plausíveis, porém incorretas ou não fundamentadas, com aparente confiança (JI et al., 2023). Em um contexto de recomendação agrícola, no qual a decisão do produtor depende da exatidão dos valores informados, esse comportamento representa um risco concreto.

A Geração Aumentada por Recuperação (*Retrieval-Augmented Generation* – RAG), proposta por Lewis et al. (2020), foi concebida para mitigar essas limitações ao acoplar ao modelo de linguagem um mecanismo de acesso a uma fonte externa de conhecimento. Em vez de responder apenas a partir dos parâmetros aprendidos no treinamento, o modelo recebe, junto à pergunta, trechos recuperados de uma base própria e formula a resposta com base neles.

O funcionamento da arquitetura divide-se em duas fases. Na fase de indexação, realizada previamente, os documentos da base de conhecimento são convertidos em *embeddings* e armazenados em um banco de dados vetorial, que mantém tanto o vetor quanto o texto original de cada documento. Na fase de consulta, executada a cada pergunta, três etapas se sucedem: (i) recuperação, em que a pergunta é vetorizada pelo mesmo modelo de *embeddings* e um componente de recuperação (*retriever*) seleciona, por similaridade, os documentos mais pertinentes; (ii) aumento, em que os trechos recuperados são inseridos no *prompt* enviado ao modelo, compondo o contexto da resposta; e (iii) geração, em que a LLM redige a resposta final fundamentada nesse contexto (LEWIS et al., 2020; GAO et al., 2024).

Entre as vantagens dessa abordagem, destacam-se a atualização do conhecimento sem necessidade de retreinar ou ajustar o modelo — bastando alterar o conteúdo da base indexada —, a redução de alucinações pelo ancoramento das respostas em dados verificáveis e a rastreabilidade, uma vez que é possível identificar quais documentos originaram cada resposta (GAO et al., 2024). Comparada ao *fine-tuning*, alternativa usual para especializar um modelo em um domínio, a RAG apresenta custo computacional consideravelmente menor e maior flexibilidade diante de bases que mudam com frequência.

Cabe observar que a qualidade da resposta final é limitada pela qualidade da recuperação: se os documentos selecionados não contiverem a informação necessária, o modelo não terá como produzir uma resposta correta. Por esse motivo, parâmetros como a granularidade dos documentos indexados e a quantidade de trechos recuperados por consulta exercem influência direta sobre o desempenho do sistema (GAO et al., 2024) — aspecto que se mostrou determinante nos resultados obtidos neste trabalho, conforme discutido na Seção 4. No caso deste projeto, a base de conhecimento é composta pelos documentos descritivos das culturas agrícolas, e a arquitetura RAG é o que permite que as respostas do sistema reflitam exclusivamente os dados

de solo e clima dessa base.

2.4 Trabalhos Relacionados

Com o objetivo de compreender o estado da arte na aplicação de inteligência artificial ao agronegócio e de fundamentar as escolhas metodológicas deste projeto, foi conduzido um levantamento de trabalhos relacionados nas bases IEEE Xplore, MDPI, Springer, ScienceDirect, Frontiers e BIO Web of Conferences. Ao todo, treze trabalhos foram analisados, abrangendo desde soluções de visão computacional para diagnóstico de doenças em plantas até aplicações de modelos de linguagem e bases de conhecimento estruturadas no domínio agrícola. Dessa análise, três trabalhos mostraram-se especialmente relevantes para o presente projeto, seja pela proximidade da aplicação, seja pela abordagem técnica adotada, e são descritos a seguir.

Jain et al. (2022) propuseram um aplicativo móvel que utiliza algoritmos de detecção de objetos para diagnosticar doenças em folhas de arroz diretamente no dispositivo, sem necessidade de conexão com a internet. Foram comparadas as arquiteturas YOLOv3-*tiny* e YOLOv4-*tiny*, sendo que a YOLOv4-*tiny* apresentou desempenho superior, alcançando acurácia geral de 98,13%. O trabalho evidencia o potencial do uso de IA aplicada diretamente ao suporte ao produtor rural.

Krupitzer (2024) explora as potenciais aplicações da IA generativa no agronegócio e na indústria alimentícia, apontando aplicações como o diagnóstico de doenças, a previsão de colheitas e a extração de informações de grandes volumes de texto. O autor também destaca desafios como a integração de diferentes tipos de dados, a aplicabilidade dos modelos e a confiança dos usuários, aspectos relevantes para a adoção de ferramentas como a proposta neste projeto.

Zhao et al. (2024) propuseram uma solução que integra grafos de conhecimento agrícola (*Agricultural Knowledge Graphs*) e LLMs para a detecção eficiente de doenças em plantas. A abordagem, que utiliza LLMs para interpretar informações estruturadas e oferecer recomendações, alcançou precisão de 94%. Esse trabalho está particularmente relacionado ao presente projeto por combinar uma base de conhecimento estruturada com a capacidade interpretativa de uma LLM, princípio semelhante ao adotado na arquitetura RAG do IAgro.

3 METODOLOGIA

Neste projeto foi utilizada uma abordagem quantitativa aplicada, de caráter exploratório e tecnológico, na qual o foco está no desenvolvimento de uma solução computacional para aplicação prática no campo do agronegócio. O desenvolvimento foi conduzido em linguagem Python, em ambiente de *notebook* Jupyter, e organizado nas etapas descritas a seguir.

3.1 Coleta e Caracterização da Base de Dados

Como base de conhecimento do sistema foi utilizado o conjunto de dados público *Crop Recommendation Dataset* (INGLE, 2020), disponível na plataforma Kaggle. Trata-se de uma base sintética derivada de dados reais: seus registros foram gerados por técnicas de aumento de dados (*data augmentation*) a partir de conjuntos reais de precipitação, clima e uso de fertilizantes referentes à Índia, compilados pela *Indian Chamber of Food and Agriculture* (ICFA). Assim, embora os registros individuais não correspondam a medições de campo específicas, os intervalos de valores e as relações entre os atributos refletem condições agronômicas reais.

O conjunto é composto por 2.200 registros distribuídos entre 22 culturas agrícolas (como arroz, milho, algodão, café e melão, entre outras), com 100 registros por cultura, o que caracteriza uma base balanceada. Cada registro é descrito por sete atributos numéricos, são eles nitrogênio (N), fósforo (P) e potássio (K) presentes no solo, temperatura, umidade relativa, pH do solo e precipitação (chuva), além do rótulo indicando a cultura correspondente. A base foi carregada e manipulada com o auxílio da biblioteca *pandas*.

3.2 Estruturação dos Dados em Documentos

Como as etapas subsequentes de vetorização operam sobre texto, os dados tabulares foram convertidos em documentos textuais. Para isso, os registros foram agrupados por cultura e, para cada uma delas, calcularam-se os valores médios dos sete atributos numéricos, sintetizando em um único conjunto de valores as cem ocorrências existentes na base.

A geração dos documentos seguiu uma estrutura padronizada, na qual um modelo de texto fixo (*template*) é combinado com os valores numéricos correspondentes a cada cultura. Nesse modelo, trechos pré-definidos identificam o nome da cultura e rotulam cada atributo — nitrogênio, fósforo, potássio, temperatura, umidade, pH e precipitação — enquanto os campos variáveis são preenchidos com as médias calculadas. A aplicação desse mesmo modelo a todas as culturas garante que os documentos resultantes compartilhem vocabulário e organização idênticos, variando apenas nos valores, o que favorece a comparabilidade entre os vetores gerados na etapa seguinte. Dessa forma, cada uma das 22 culturas passou a ser representada por um documento textual descritivo, estrutura adequada à indexação semântica.

3.3 Vetorização e Indexação

Cada documento foi processado pelo modelo de *embeddings text-embedding-3-large*, da OpenAI (OPENAI, 2024), que converte o texto em um vetor numérico de alta dimensionalidade representando seu significado. Os vetores resultantes foram armazenados em um banco de dados vetorial, o Pinecone (PINECONE, 2024), responsável por indexar as representações e possibilitar buscas por similaridade de forma eficiente. Para cada documento, o banco armazena tanto o vetor utilizado na busca quanto o texto original, utilizado posteriormente na composição da resposta.

A orquestração entre a base de dados, o modelo de *embeddings* e o banco vetorial foi realizada por meio da biblioteca LangChain (LANGCHAIN, 2024).

3.4 Recuperação e Geração de Respostas

Na etapa de consulta, a pergunta do usuário é convertida em vetor pelo mesmo modelo de *embeddings*, garantindo que pergunta e documentos compartilhem o mesmo espaço vetorial. Um componente de recuperação (*retriever*) compara o vetor da pergunta com os vetores armazenados no Pinecone e seleciona os documentos mais semelhantes. Os documentos recuperados são então fornecidos como contexto a um modelo de linguagem (*ChatOpenAI*, baseado no GPT-3.5-turbo), por meio do componente *RetrievalQA* da biblioteca LangChain. O modelo lê o contexto e redige a resposta final em linguagem natural, fundamentada exclusivamente nas informações recuperadas.

3.5 Teste e Validação

O sistema foi submetido a testes práticos por meio de um conjunto de perguntas simulando consultas reais de produtores, tais como a identificação da cultura recomendada para solos com alto teor de nitrogênio, culturas adaptadas à alta umidade, culturas que se desenvolvem em pH ácido e culturas que demandam maior volume de chuva. A validação foi conduzida comparando as respostas geradas pelo sistema com os valores obtidos diretamente da base de dados, permitindo verificar a correção das respostas e identificar limitações da configuração adotada. A Figura 1 apresenta a visão geral da arquitetura implementada, desde o carregamento da base até o retorno da resposta ao usuário.

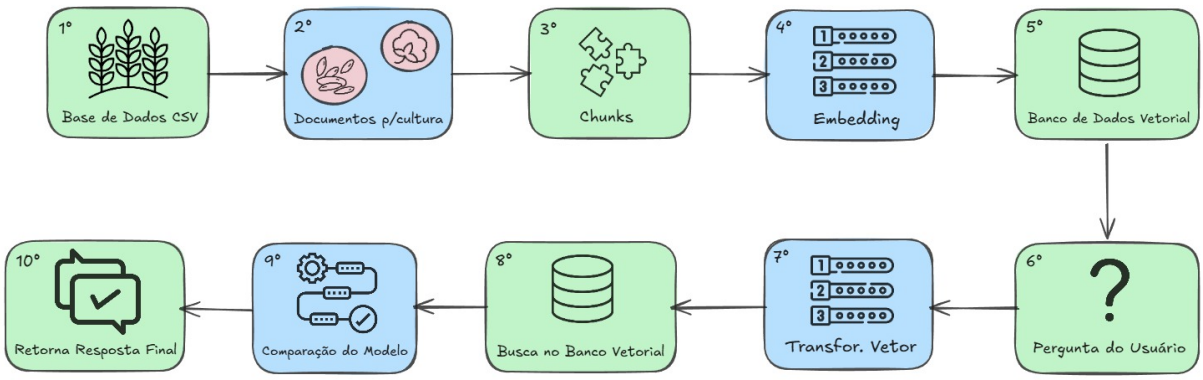


Figura 1 – Arquitetura do Modelo RAG IAgro.

4 RESULTADOS E DISCUSSÃO

4.1 Análise dos resultados

O projeto desenvolvido foi capaz de responder, em linguagem natural, às consultas formuladas sobre a recomendação de culturas, demonstrando a viabilidade da arquitetura RAG

para o domínio agrícola. Durante o processo de teste e validação, contudo, observou-se que a configuração inicial do sistema apresentava limitações em determinado tipo de pergunta, o que exigiu ajustes na metodologia de recuperação.

Na configuração inicial, cada registro da base era transformado em um documento individual (resultando em aproximadamente 2.200 documentos) e o mecanismo de recuperação selecionava apenas os três documentos mais similares a cada consulta. Essa configuração mostrou-se adequada para perguntas sobre características de uma cultura específica, como ilustram os Cenários 1, 2 e 3, mas apresentava imprecisão em perguntas comparativas, como "qual cultura precisa de mais chuva?", uma vez que o modelo tinha acesso a apenas três registros, e não ao conjunto completo de culturas necessário para a comparação.

Cenário 1

Pergunta: Qual cultura é recomendada para solo com alto nitrogênio?

Resposta: A cultura recomendada para solo com alto nitrogênio é a cultura de *muskmelon*, que possui 100 de nitrogênio.

Cenário 2

Pergunta: Quais culturas se adaptam a alta umidade?

Resposta: As culturas que se adaptam a alta umidade com base nos dados fornecidos são *papaya*, *muskmelon*, *cotton*, *apple*, *pomegranate*, *coconut*, *watermelon* e *mango*.

Cenário 3

Pergunta: Qual cultura cresce bem em pH ácido?

Resposta: A cultura que cresce bem em pH ácido é a cultura de *lentil*, com um pH de 5,8.

Para solucionar essa limitação, foram realizadas duas alterações. Primeiramente, a granularidade dos documentos foi ajustada: em vez de um documento por registro, passou-se a gerar um documento por cultura, com base nas médias de seus atributos, reduzindo a redundância e produzindo 22 documentos representativos. Em segundo lugar, a cobertura da recuperação foi ampliada, de modo que o mecanismo passasse a recuperar todas as culturas da base para cada consulta. Com isso, o modelo de linguagem passou a receber o conjunto completo de culturas antes de formular a resposta, o que corrigiu as respostas às perguntas comparativas.

Após os ajustes, o sistema passou a responder corretamente a esse tipo de consulta, conforme apresentado no Cenário 4. Ao ser questionado sobre qual cultura necessita de maior volume de chuva, o sistema indicou o arroz (*rice*), com aproximadamente 236 mm de precipitação. Essa resposta foi validada por meio da consulta direta à base de dados: ao ordenar as culturas de forma decrescente pela coluna de precipitação (*rainfall*), confirmou-se que o arroz apresenta os maiores valores registrados, ocupando as primeiras posições da ordenação. Dessa forma, a

resposta gerada pelo sistema mostrou-se coerente com os dados reais da base, evidenciando a correção do comportamento após os ajustes realizados.

Cenário 4

Pergunta: Qual é a cultura que mais precisa de água?

Resposta: A cultura que mais precisa de água é a de arroz.

4.2 Discussão

Cabe ressaltar uma análise crítica dos resultados. Entre os pontos positivos, destaca-se que o sistema cumpriu o objetivo proposto, respondendo corretamente às consultas formuladas sem exigir do usuário qualquer conhecimento sobre a estrutura da base ou sobre linguagens de consulta. As respostas mantiveram-se ancoradas nos dados recuperados e puderam ser verificadas por consulta direta à base, o que evidencia a eficácia da arquitetura RAG em manter a geração fundamentada e reduzir a ocorrência de informações não sustentadas pelos dados. Destaca-se também que a correção do comportamento em perguntas comparativas foi obtida por meio de ajustes na estruturação dos documentos e na cobertura da recuperação, sem necessidade de substituir o modelo de linguagem ou de realizar qualquer treinamento adicional, o que reforça a flexibilidade e o baixo custo de manutenção da abordagem.

Por outro lado, algumas limitações devem ser reconhecidas. A ampliação da recuperação para o conjunto completo de culturas é viável neste caso justamente porque a base é pequena e estruturada (22 culturas). Em bases maiores, essa estratégia não seria escalável, e a busca semântica voltaria a exercer papel de filtragem das informações mais relevantes. Além disso, observou-se que perguntas de natureza agregada (como mínimos, máximos e médias) são mais sensíveis à configuração da recuperação do que perguntas sobre características isoladas.

Outra limitação relevante diz respeito à ausência de uma interface gráfica para o usuário. Nesta versão, as consultas são realizadas por meio de *strings* definidas diretamente no código-fonte, em ambiente de *notebook*, o que restringe os testes ao conjunto de perguntas previamente escritas pelo desenvolvedor e impede, no estágio atual, a avaliação da ferramenta com produtores reais em condições de uso. Por fim, cabe registrar que a base utilizada foi construída a partir de dados do contexto agroclimático indiano, de modo que as recomendações geradas não devem ser transpostas diretamente para a realidade brasileira sem a devida adaptação da base de conhecimento.

5 CONCLUSÃO

Este trabalho apresentou o desenvolvimento do IAgro, um sistema conversacional voltado ao auxílio da tomada de decisão na agricultura, capaz de responder em linguagem natural a perguntas sobre a recomendação de culturas a partir de parâmetros de solo e clima. O sistema foi construído sobre a arquitetura de Geração Aumentada por Recuperação (RAG), integrando a

base de dados *Crop Recommendation*, modelos de *embeddings* e um modelo de linguagem, com o apoio do banco vetorial Pinecone e da biblioteca LangChain. Os objetivos específicos foram atendidos: a base foi estruturada em documentos adequados à busca semântica, os dados foram vetorizados e indexados, um mecanismo de recuperação semântica foi implementado e integrado a um modelo de linguagem para a geração das respostas.

O processo de teste e validação mostrou-se fundamental, pois permitiu identificar e corrigir uma limitação da configuração inicial relacionada a perguntas comparativas, resultando em respostas corretas e verificáveis contra a base de dados. Como principal contribuição, o trabalho demonstra a viabilidade da abordagem RAG para consultas agrícolas baseadas em dados estruturados, bem como a importância do ajuste da granularidade dos dados e da cobertura da recuperação para a qualidade das respostas.

Como trabalhos futuros, propõe-se a ampliação da base de conhecimento, possivelmente com a integração de fontes oficiais e adequadas ao contexto nacional, como as APIs da Embrapa (EMBRAPA, 2025); o desenvolvimento de uma interface gráfica que torne a ferramenta acessível ao produtor final e viabilize sua avaliação com usuários reais; e a investigação de técnicas complementares, como o uso de ferramentas de cálculo pelo modelo, de modo a aumentar a precisão em perguntas de natureza agregada.

6 REFERÊNCIAS

AGROW NEGÓCIOS. *Quais os países mais investem em agronegócio?* São Paulo: aGrow Negócios, 2023. AGrow Negócios. Disponível em: <INSERIR LINK>. Acesso em: 28 abr. 2025. Disponível em: <INSERIRLINK>.

CEPEA. *PIB do agronegócio brasileiro*. Piracicaba: CEPEA/ESALQ-USP, 2025. Centro de Estudos Avançados em Economia Aplicada (CEPEA). Disponível em: <INSERIR LINK>. Acesso em: 28 abr. 2025. Disponível em: <INSERIRLINK>.

EMBRAPA. *Agricultura digital no Brasil: tendências, desafios e oportunidades: resultados de pesquisa online*. Brasília: Embrapa, 2019. Empresa Brasileira de Pesquisa Agropecuária. Disponível em: <INSERIR LINK>. Acesso em: 28 abr. 2025. Disponível em: <INSERIRLINK>.

EMBRAPA. *A agricultura brasileira*. Brasília: Embrapa, 2020. Empresa Brasileira de Pesquisa Agropecuária. Disponível em: <INSERIR LINK>. Acesso em: 28 abr. 2025. Disponível em: <INSERIRLINK>.

EMBRAPA. *Agro API: Portal de APIs Agropecuárias*. Campinas: Embrapa Agricultura Digital, 2025. Empresa Brasileira de Pesquisa Agropecuária. Disponível em: <https://www.agroapi.cnptia.embrapa.br/portal/>. Acesso em: 28 abr. 2025. Disponível em: <https://www.agroapi.cnptia.embrapa.br/portal/>.

GAO, Y. et al. Retrieval-augmented generation for large language models: a survey. *arXiv preprint arXiv:2312.10997*, arXiv, Ithaca, USA, 2024.

INGLE, A. *Crop Recommendation Dataset: maximize agricultural yield by recommending appropriate crops*. San Francisco, USA: Kaggle, 2020. Kaggle. Disponível em:

<https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>. Acesso em: 23 ago. 2026. Disponível em: <https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>.

JAIN, S. et al. Automatic rice disease detection and assistance framework using deep learning and a chatbot. *Electronics*, MDPI AG, v. 11, n. 14, p. 2110, jul. 2022.

JI, Z. et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, Association for Computing Machinery (ACM), v. 55, n. 12, p. 1–38, 2023.

KRUPITZER, C. Generative artificial intelligence in the agri-food value chain: overview, potential, and research challenges. *Frontiers in Food Science and Technology*, Frontiers Media SA, v. 4, set. 2024.

LANGCHAIN. *LangChain: build context-aware reasoning applications*. San Francisco, USA: LangChain, 2024. LangChain. Disponível em: <https://python.langchain.com/>. Acesso em: 23 ago. 2026. Disponível em: <https://python.langchain.com/>.

LEWIS, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vancouver, Canada: Curran Associates, 2020. v. 33, p. 9459–9474.

MIKOLOV, T.; YIH, W.-t.; ZWEIG, G. Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Atlanta, USA: Association for Computational Linguistics, 2013. p. 746–751.

OPENAI. *New embedding models and API updates: text-embedding-3-large*. San Francisco, USA: OpenAI, 2024. OpenAI. Disponível em: <https://openai.com/index/new-embedding-models-and-api-updates/>. Acesso em: 23 ago. 2026. Disponível em: <https://openai.com/index/new-embedding-models-and-api-updates/>.

PINECONE. *Pinecone: the vector database for machine learning applications*. New York, USA: Pinecone Systems, 2024. Pinecone Systems. Disponível em: <https://www.pinecone.io/>. Acesso em: 23 ago. 2026. Disponível em: <https://www.pinecone.io/>.

RAIAAN, M. A. K. et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, Institute of Electrical and Electronics Engineers (IEEE), v. 12, p. 26839–26874, 2024.

ZHAO, X. et al. Implementation of large language models and agricultural knowledge graphs for efficient plant disease detection. *Agriculture*, MDPI AG, v. 14, n. 8, p. 1359, ago. 2024.