

SISTEMA INTELIGENTE PARA AVICULTURA BASEADO EM MODELOS DE LINGUAGEM DE GRANDE ESCALA

Caio Eduardo Hisamatsu Mizohata¹
Guilherme de Barros Pedro¹
Eloíza Martins Primo Capeloci²
Ronnie Shida Marinho³

RESUMO

Este trabalho apresenta o desenvolvimento e a análise da viabilidade de um sistema inteligente voltado a dar suporte informacional na área da avicultura, com o objetivo de transformar conteúdos técnicos complexos em respostas claras e objetivas. O sistema utiliza Modelos de Linguagem de Grande Escala (LLMs) para processar informações especializadas e gerar respostas acessíveis, contribuindo para a disseminação do conhecimento técnico no setor. Considerando a relevância da avicultura para a economia nacional, especialmente em polos produtivos como Bastos (SP), o estudo discute como a inteligência artificial pode aprimorar o acesso a informações sobre manejo, nutrição, sanidade e gestão. A metodologia envolveu o desenvolvimento de um protótipo funcional baseado em tecnologias de código aberto, implementado em Python e Django, integrando a arquitetura Retrieval-Augmented Generation (RAG) e o banco de dados vetorial ChromaDB para buscas contextuais. Foram realizados testes comparando os modelos Gemma:3.4b e MISTRAL:7b, analisando tempo de resposta, desempenho e qualidade textual. Os resultados indicam que a adoção de LLMs pode otimizar o processo de gestão de granjas, melhorando o acesso a informações técnicas e reduzindo o tempo de resposta em consultas relacionadas ao manejo.

Palavras-chave: avicultura; sistema inteligente; modelos de linguagem de grande escala; inteligência artificial.

1 INTRODUÇÃO

A avicultura, ramo da agropecuária voltado à criação de aves, configura-se como um dos principais pilares da economia brasileira. Destaca-se, nesse contexto, o município de Bastos, localizado no interior do Estado de São Paulo, reconhecido nacional e internacionalmente como um dos maiores polos exportadores do setor avícola. Essa relevância econômica impõe aos produtores locais a necessidade de manter elevados padrões de precisão e eficiência produtiva para sustentar a competitividade do mercado (Sapé Agro, 2025).

¹ Discente do curso de Tecnologia de Big Data no Agronegócio- FATEC Pompeia

² Docente do curso de Tecnologia de Big Data no Agronegócio- FATEC Pompeia

³ Docente do Departamento de Ciências da Natureza- UFSCar

Nesse cenário, a tomada de decisão torna-se um fator determinante para o desempenho das granjas, envolvendo preocupações relacionadas ao período de safra das matérias-primas, ao volume de produção e aos custos operacionais. As incertezas inerentes ao processo produtivo, aliadas à dificuldade de acesso rápido e sistematizado a informações técnicas, configuram-se como entraves significativos à eficiência e à efetividade operacional da atividade avícola (Caldas, 2018).

O avanço da inteligência artificial em diversos setores tem ampliado as possibilidades de apoio à tomada de decisão e de desenvolvimento de soluções capazes de reduzir as incertezas operacionais (*Baracho; Tolón, 2025*). No contexto da avicultura, essas tecnologias podem contribuir para mitigar dúvidas imediatas dos produtores, proporcionando acesso ágil a informações técnicas sobre manejo, sanidade e nutrição. Dessa forma, a incorporação de ferramentas baseadas em inteligência artificial apresenta-se como uma alternativa promissora para suprir lacunas informacionais e aprimorar a gestão produtiva no setor avícola.

A aplicação de tecnologias baseadas em Modelos de Linguagem de Grande Escala (LLMs) possibilita o desenvolvimento de sistemas capazes de compreender consultas em linguagem natural e processar grandes volumes de dados, fornecendo respostas personalizadas e contextualizadas (*Minaee et al., 2021*). Tais sistemas apresentam potencial para apoiar a tomada de decisão e reduzir entraves operacionais, ao facilitar o acesso rápido a informações relevantes sobre manejo, nutrição, sanidade e outros aspectos da produção avícola.

A relevância deste estudo está associada à necessidade de integrar soluções tecnológicas ao ambiente rural, conforme apontado por Albuquerque (2023), que destaca a inteligência artificial e a gestão digital como elementos essenciais para decisões mais ágeis e fundamentadas. Essa integração contribui diretamente para o aumento da competitividade e para a promoção da sustentabilidade na cadeia produtiva agropecuária.

Portanto, esta pesquisa tem como objetivo analisar a viabilidade e o desenvolvimento de um sistema personalizado que funcione como um assistente virtual de apoio informacional voltado à avicultura, com o propósito de transformar conhecimentos técnicos complexos em respostas objetivas, acessíveis e ágeis. Fundamentado em Modelos de Linguagem de Grande Escala (LLMs), o protótipo proposto visa fornecer informações e auxiliar na tomada de decisão dos produtores, contribuindo para maior eficiência e precisão nas práticas produtivas.

2 REFERENCIAL TEÓRICO

O referencial teórico é essencial para este trabalho, pois estabelece a junção de dois cenários, o da tecnologia com o Modelo de Linguagens Grandes (LLMs) com a área do campo com a Avicultura, cenários onde o protótipo estará atuando. Esta seção tratará das definições dos LLMs e posteriormente contextualizar com as necessidades do setor da Avicultura.

2.1 Modelos de Linguagem de Grande Escala (LLMs)

Os Modelos de Linguagem de Grande Escala (*Large Language Models – LLMs*) representam uma vertente avançada da inteligência artificial voltada ao processamento e à geração de linguagem natural. Esses modelos são treinados com volumes massivos de dados textuais provenientes de diversas fontes, como livros, artigos científicos, páginas da web e descrições de imagens, com o objetivo de identificar padrões linguísticos e estruturas semânticas complexas (*Microsoft Learn*, 2025).

Durante o processo de treinamento, o modelo aprende a predizer a próxima palavra em uma sequência textual, o que lhe permite produzir respostas coerentes, contextuais e semanticamente relevantes. Essa capacidade de aprendizado estatístico, em larga escala, possibilita aos LLMs executar múltiplas tarefas de Processamento de Linguagem Natural (PLN), incluindo tradução automática, sumarização de textos, geração de conteúdo, resposta a perguntas complexas e apoio a tarefas de programação (IBM Brasil, 2025).

A principal característica dessas arquiteturas reside na habilidade de generalização, que permite aplicar o conhecimento adquirido a novos contextos e domínios, tornando os LLMs ferramentas versáteis e de alto potencial para aplicações em diversos setores, inclusive no agronegócio e na avicultura, onde podem ser explorados para o apoio à tomada de decisão e à gestão de informações técnicas (IBM Brasil, 2025).

2.2 Avicultura

No Brasil, a avicultura desempenha um papel estratégico na economia nacional e no comércio internacional, sendo um dos setores mais dinâmicos do agronegócio e um importante gerador de emprego e renda. O país figura entre os maiores produtores e exportadores de carne e ovos do mundo (Agência Brasil, 2025), o que exige dos produtores processos produtivos cada vez mais precisos e eficientes.

Essa busca por eficiência está associada à transição para a chamada Avicultura 4.0, conceito que engloba a incorporação de tecnologias digitais, sensores, sistemas automatizados e ferramentas de análise de dados na gestão das granjas (Pizol, 2024). Nesse novo cenário, a geração de dados deixou de ser um desafio, o principal obstáculo passou a ser a gestão eficiente das informações e a capacidade de utilizá-las na tomada de decisões (Pegino; Ribeiro Neto; Urgia, 2021).

Devido à complexidade técnica envolvida em aspectos como planos de manejo, curvas de nutrição e controle sanitário, os produtores demandam acesso rápido, integrado e confiável a informações que sustentem decisões assertivas no dia a dia da produção. Nesse contexto, a integração entre Inteligência Artificial (IA) e Gestão Digital surge como um fator determinante para a competitividade e a sustentabilidade do agronegócio, conforme destacado por Albuquerque (2023).

Assim, a aplicação de Modelos de Linguagem de Grande Escala (LLMs) na avicultura apresenta-se como uma alternativa promissora para atender a essas necessidades, permitindo transformar grandes volumes de dados técnicos em informações acessíveis, contextualizadas e úteis ao processo decisório dos produtores.

3 MATERIAL E MÉTODOS

O desenvolvimento do protótipo foi realizado com base em ferramentas e tecnologias de código aberto voltadas à Inteligência Artificial, priorizando a transparência, a reproduzibilidade e o baixo custo computacional. O sistema foi implementado em Python 3.11.11, devido ao amplo suporte oferecido à área de *machine learning* e às bibliotecas especializadas disponíveis para o ecossistema da linguagem.

Para a camada de aplicação web, utilizou-se o framework Django 5.2.6, que possibilitou a implementação da arquitetura *Model-View-Template (MVT)*, o gerenciamento de requisições HTTP e a persistência de dados por meio do banco SQLite3. A interface do usuário foi desenvolvida com HTML5, CSS3 e JavaScript Vanilla, buscando simplicidade, responsividade e baixo tempo de carregamento.

A geração de respostas conversacionais baseou-se na utilização de Modelos de Linguagem de Grande Escala (LLMs) processados localmente através do ambiente Ollama 0.11.6. O sistema foi configurado em *containers* Docker, com suporte à aceleração gráfica via GPU NVIDIA (CUDA). Foram testados dois modelos distintos, Gemma 3.4B e MISTRAL 7B, escolhidos em função da disponibilidade de memória de vídeo e de seu desempenho reconhecido em tarefas de linguagem natural. A arquitetura adotada seguiu o paradigma *Retrieval-Augmented Generation (RAG)*, implementado com a biblioteca LangChain 0.2.16, a qual foi responsável pela orquestração do pipeline de recuperação e geração contextual de respostas.

Para o armazenamento e a consulta vetorial dos *embeddings*, empregou-se o banco ChromaDB 0.4.24, configurado para persistência local em disco. A geração dos *embeddings* textuais foi realizada por meio do modelo Nomic-Embed-Text, com vetores de 768 dimensões hospedados localmente, garantindo baixa latência, segurança e privacidade dos dados processados.

No tratamento de documentos, o sistema foi preparado para suportar diferentes formatos de entrada, incluindo PDF, Word, planilhas Excel e arquivos CSV, utilizando, respectivamente, as bibliotecas pdfplumber (v6.1.1), python-docx (v1.2.0), openpyxl (v3.1.5) e pandas (v2.3.3). Documentos extensos foram divididos em segmentos menores (*chunks*) por meio do método *RecursiveCharacterTextSplitter* da biblioteca LangChain, configurado com tamanho de 500 caracteres e sobreposição de 100, a fim de preservar a coesão semântica durante a indexação vetorial.

Para garantir desempenho e escalabilidade, foi implementado um sistema de cache em memória, configurado para armazenar até 500 entradas válidas por 10 minutos, reduzindo o tempo de resposta em consultas repetidas.

Operações tensoriais e de inferência foram otimizadas com o uso do PyTorch 2.8.0, enquanto o ONNX Runtime 1.20.1 foi empregado para acelerar a produção de *embeddings* em processadores convencionais (*CPU-based inference*). O ambiente

de desenvolvimento consistiu em máquinas locais com sistema operacional Windows 11, com gerenciamento de dependências Python por meio de ambientes virtuais (venv). A execução dos LLMs e do mecanismo RAG ocorreu em um *desktop* equipado com GPU NVIDIA RTX 2060 (6GB de VRAM).

O servidor Ollama foi configurado em ambiente *Docker Compose*, com passagem explícita de GPU e mapeamento da porta 11434, permitindo a comunicação direta com o servidor Django via protocolo HTTP.

O banco vetorial ChromaDB foi armazenado no diretório *db/chroma*, com sistema automático de validação e reparo de índices corrompidos, além de rotinas de backup incremental com carimbo temporal (*timestamp*). O registro de logs seguiu o padrão de arquivos rotativos, limitados a 5 MB por arquivo e três cópias de backup, implementados através do módulo nativo *logging* do Python. Essa configuração possibilitou rastreabilidade, diagnóstico e análise detalhada do comportamento do sistema durante a execução dos testes.

3.1 Métodos e Procedimentos Técnicos

O sistema foi estruturado em módulos distintos, cada um com funções específicas que garantem a validação, processamento, recuperação e geração de respostas a partir de documentos e perguntas textuais.

O módulo de documentos é responsável pela validação e processamento de arquivos. Ele aplica extratores específicos para cada tipo de arquivo, garantindo que o conteúdo seja interpretado corretamente. Para preservar o contexto semântico durante o processamento, utiliza-se o *RecursiveCharacterTextSplitter* do LangChain.

Já o módulo de geração de *embeddings* e indexação transforma cada fragmento de texto em vetores de 768 dimensões utilizando o modelo nomic-embed-text local. Esses vetores são armazenados no ChromaDB, acompanhados de metadados que permitem filtragem e rastreabilidade. O sistema assegura a persistência dos dados por meio de chamadas explícitas e validações de integridade.

Ainda, o módulo de recuperação contextual realiza buscas por similaridade de cosseno, utilizando como parâmetro k=4 e um limiar de 0,25. Para otimização, limita o contexto a 1.500 caracteres e emprega um sistema de cache de consultas, reduzindo o tempo de resposta para perguntas recorrentes.

No módulo de geração de respostas, o contexto recuperado é integrado ao LLM. O prompt considera instruções, histórico da conversa, contexto relevante e a pergunta do usuário. A geração de respostas utiliza o modelo ChatOllama com temperatura 0,7 e conta com um mecanismo de fallback automático de três níveis, garantindo resiliência frente a falhas.

O gerenciamento de serviços do sistema foi implementado seguindo o padrão *Singleton Thread-Safe* por meio do *ServiceManager*, evitando acessos concorrentes ao ChromaDB. Além disso, foi aplicado *Lazy Loading* nos componentes, melhorando a performance da aplicação, e foram incluídos métodos de limpeza para garantir a liberação ordenada de recursos.

3.2 Fluxo de Funcionamento do Sistema

O processamento de uploads inicia com a validação e extração de texto dos arquivos. O conteúdo extraído é enviado ao *RAGService* pela função `processar_documento_e_responder()`. Esse método divide o documento em fragmentos, gera os *embeddings* correspondentes e persiste os vetores no ChromaDB. Em seguida, busca o contexto relevante e gera uma resposta contextual, retornando um resumo estruturado com análise técnica.

O processamento de perguntas textuais sem anexos ocorre por meio do método `gerar_resposta_contextual()`. Inicialmente, verifica-se o cache, que apresentou uma taxa de acerto de 30% nos testes. Em caso de cache miss, é realizada uma busca vetorial filtrando por `session_id`, recuperando os documentos mais relevantes acima do limiar de similaridade. O conteúdo é então concatenado, formatado e utilizado na construção de um *prompt* especializado. A resposta gerada é armazenada em cache com TTL de 10 minutos e retornada em formato JSON, contendo metadados RAG (`used_rag`, `context_found`, `chunks_added`, `from_cache`) que permitem rastreabilidade e depuração.

Em relação à validação e testes, foram realizados testes funcionais manuais com documentos do setor avícola, incluindo planilhas financeiras, verificando a extração, indexação, recuperação contextual e geração de respostas técnicas estruturadas. Testes de carga identificaram limitações de hardware, com tempos de inferência entre 8 e 12 segundos para modelos de 7b-8b em GPUs de 6 GB,

orientando a escolha de modelos mais leves, como o Gemma 3:4b, para respostas rápidas.

Por fim, a arquitetura modular e o uso de *logs* detalhados permitiram rastrear o fluxo de dados, facilitando a identificação e correção de bugs críticos relacionados à persistência vetorial e à concorrência de threads.

4 RESULTADOS E DISCUSSÃO

A presente seção tem como objetivo apresentar e analisar os resultados obtidos a partir do uso de dados de produções obtidos do Instituto Brasileiro de Geografia e Estatística (IBGE, 2025), para a implementação e dos testes do protótipo desenvolvido. Foram comparados dois modelos de linguagem, MISTRAL:7b e Gemma:3.4b, considerando parâmetros como tempo de resposta, consumo de tokens, qualidade textual e adequação contextual. As figuras a seguir ilustram exemplos reais de interações com o sistema, demonstrando o comportamento de cada modelo frente a consultas distintas, bem como a eficiência operacional observada em cada caso. A análise é conduzida de forma conjunta com as imagens, buscando evidenciar como as diferenças de desempenho se traduzem em experiências práticas de uso.

A Figura 1 apresenta a primeira interação realizada com o assistente virtual, na qual foi feita uma pergunta sobre práticas de manejo. A resposta foi gerada pelo modelo MISTRAL:7b, evidenciando a capacidade do sistema em compreender comandos diretos e oferecer respostas contextualizadas.

Figura 1- Exemplo de interação com o assistente virtual.

Quais são as práticas de manejo recomendadas para reduzir mortalidade em frangos de corte durante as primeiras 3 semanas?

As práticas de manejo recomendadas para reduzir a mortalidade em frangos de corte durante as primeiras três semanas incluem:

1. **Higiene:** Mantenha a granja limpa e saudável, evitando a acumulação de excrementos e restos de alimentação.
2. **Ventilação:** Garanta uma boa ventilação para manter um ambiente fresco e seco.
3. **Alimentação:** Forneca uma dieta nutritiva adequada, com uma proporção correta de proteínas, gorduras, carboidratos e minerais.
4. **Vacinação:** Vacine as aves contra doenças comuns em frangos de corte, como colibacteíro, salmonelose e gastroenterite.
5. **Água:** Forneca uma fonte de água limpa e fresca, disponível a todo momento.
6. **Registro de saúde:** Registre todos os eventos relacionados à saúde das aves, como mortalidade, doenças e comportamento anormal.
7. **Controle de predadores:** Instale barras de proteção nas aberturas da granja para evitar que animais predadores entrem.
8. **Bem-estar animal:** Procure manter um ambiente confortável para as aves, com espaço suficiente para se movimentarem e se comportarem naturalmente.
9. **Observação diária:** Faça observações diárias nas aves para detectar sinais de doenças ou problemas precocemente.
10. **Consulta com profissionais:** Se houver dúvidas ou problemas, procure consultar um veterinário ou técnico especializado em avicultura.

 Pergunte alguma coisa...

Pressione Shift + Enter para quebrar linha, Enter para enviar • Anexe documentos para análise

Fonte: Elaborada pelos autores (2025)

A figura 1 ilustra a resposta do modelo MISTRAL ao ser questionado sobre práticas de manejo (em avicultura), que sumarizou as informações em 10 tópicos.

Na Figura 2, observa-se a geração automática de uma tabela a partir de um relatório técnico fornecido pelo usuário. O modelo MISTRAL demonstrou habilidade na interpretação de dados estruturados, transformando conteúdo textual em formato tabular comprehensível.

Figura 2- Relatório gerado pelo modelo MISTRAL a partir de uma planilha.

Faça um relatório dos dados fornecidos
ovos_202104.xlsx (31 KB)

Relatório de produção de ovos de galinha em 2020 e 2021, segundo o arquivo ovos_202104.xlsx:

- **Produção Total:** A produção anual de ovos de galinha foi de 3.976.053 milhões de dúzias em 2021, com uma variação de 0,22% em relação a 2020.
- **Primeiro Trimestre:** O primeiro trimestre de 2021 apresentou um aumento de 1,27% na produção em comparação ao mesmo período de 2020, totalizando 986.885 milhões de dúzias.
- **Segundo Trimestre:** O segundo trimestre de 2021 apresentou um aumento de 1,39% na produção em relação ao mesmo período de 2020, totalizando 990.849 milhões de dúzias.

- **Tabela:**

Mês	Produção (milhões de dúzias)	Efetivo (mil cabeças)	Varição %	Variação % do 2020
Total do ano	3.976.053,0	-	0,224721	-
Total do 1º Trimestre	986.885,0	-	1,265193	-
Janeiro	333.052,0	175.321,64	1,478367	1,792989
Fevereiro	310.536,0	175.382,24	-1,054973	1,501443
Março	343.296,0	177.186,80	3,244123	2,017387
Total do 2º Trimestre	990.849,0	-	1,387719	-

 Pergunte alguma coisa...

Pressione Shift + Enter para quebrar linha, Enter para enviar • Anexe documentos para análise

Fonte: Elaborada pelos autores (2025)

A figura 2 ilustra um resumo textual sobre a produção de ovos e uma tabela

que estrutura os dados analisados.

Já a Figura 3 ilustra uma resposta mais complexa, relacionada a melhorias futuras no relatório. Nesse caso, o modelo manteve a coerência textual e apresentou recomendações relevantes, destacando seu potencial para tarefas analíticas.

Figura 3- Geração de recomendações pelo modelo MISTRAL..

The screenshot shows a user interface for generating recommendations. At the top right, a purple button says "De acordo com os dados fornecidos, o que pode ser melhorado futuramente?". Below it, a text box contains a recommendation: "De acordo com os dados fornecidos, observa-se uma tendência geral de aumento na produção de ovos de galinha ao longo dos meses. Porém, em alguns meses, como fevereiro e março de 2021, a produção apresentou um declínio em comparação ao período correspondente de 2020. Para melhorar futuramente:" followed by a list of actions. A "Tabela de Produção de Ovos de Galinha:" is shown below, with a snippet of its content. At the bottom, there's a text input field "Pergunte alguma coisa..." and a note: "Pressione Shift + Enter para quebrar linha, Enter para enviar • Anexe documentos para análise".

Mês	Produção (milhões de dúzias)	Variação %	Variação % do 2020
Total do ano	3.976.053,0	0,224721	-
Total do 1º Trimestre	986.885,0	1,265193	-
Janeiro	333.052,0	1,478367	1,792989

Fonte: Elaborada pelos autores (2025).

A figura 3 ilustra a análise e as sugestões de melhorias futuras propostas pelo modelo com base nos dados do relatório.

Na sequência, as Figuras 4 a 6 demonstram o mesmo conjunto de testes realizados com o modelo Gemma:3.4b, o qual apresentou desempenho mais rápido nas mesmas consultas. A Figura 4 mostra a resposta do Gemma a uma pergunta sobre manejo, enquanto a Figura 5 evidencia a geração de uma tabela a partir de um relatório, e a Figura 6, por sua vez, apresenta a resposta do modelo sobre melhorias futuras.

Figura 4- Comparativo de detalhamento técnico do modelo Gemma.

Quais são as práticas de manejo recomendadas para reduzir mortalidade em frangos de corte durante as primeiras 3 semanas?

As primeiras três semanas são críticas para os frangos de corte, pois a mortalidade é maior devido a diversos fatores. Para reduzir a mortalidade nesse período, as seguintes práticas de manejo são altamente recomendadas:

- 1. Controle de Temperatura e Umidade:**
- Temperatura:** Manter a temperatura ambiente entre 32°C e 36°C, especialmente nos primeiros dias.
- Umidade:** Manter a umidade relativa entre 60% e 70% para evitar a desidratação.
- Legenda:** Controle da temperatura e umidade é crucial para a termorregulação dos frangos recém-nascidos.
- 2. Qualidade da Água e Alimentação:**
- Água:** Disponibilizar água limpa e fresca constantemente, com pH entre 6,5 e 7,5. A ingestão de água é fundamental para a termorregulação.
- Alimentação:** Fornecer ração específica para frangos iniciantes, com alto teor de proteína (18-20%) para estimular o crescimento.
- Legenda:** A água e a ração de alta qualidade são essenciais para o desenvolvimento e a termorregulação dos frangos.
- 3. Higiene e Limpeza:**
- Limpeza:** Limpar e desinfetar o galpão diariamente para evitar o acúmulo de microrganismos.
- Ventilação:** Garantir uma ventilação adequada para remover o excesso de umidade e gases.
- Legenda:** A higiene e a ventilação são cruciais para a prevenção de doenças.
- 4. Manejo do Grupo:**
- Densidade:** Evitar alta densidade de aves para facilitar a ventilação e reduzir o estresse.

 Pergunte alguma coisa... 

Pressione Shift + Enter para quebrar linha, Enter para enviar • Anexe documentos para análise

Fonte: Elaborada pelos autores (2025).

A figura 4 ilustra a resposta deste modelo ao ser questionado com a mesma pergunta feita ao modelo MISTRAL na figura 1. Nota-se que o Gemma sumarizou os tópicos de forma notavelmente mais técnica em comparação a resposta anterior.

Figura 5- Geração de um relatório estruturado pelo modelo Gemma.

Faça um relatório dos dados fornecidos
ovos_202104.xlsx (31 KB)

Relatório da Produção de Ovos de Galinha - 2020-2021 (Dados Preliminares)

Com base nos dados extraídos, a produção total de ovos de galinha no ano de 2021 foi de 3.976.053 mil dúzias, um aumento de 0,224721% em relação ao ano de 2020, que registrou 3.967.138 mil dúzias. A produção no primeiro trimestre de 2021 foi de 986.885 mil dúzias, com um aumento de 1,265193% em relação ao primeiro trimestre de 2020, que foi de 974.555 mil dúzias.

Mês	Produção de Ovos (Mil Dúzias)	Efetivo de Galinhas (Mil Cabeças)	Variação %
Total do Ano	3.976.053	-	0,224721
1º Trimestre	986.885	-	1,265193
Janeiro	333.052	175.321.638	1,478367
Fevereiro	310.536	175.382.243	-1,054973
Março	343.296	177.186.8	3,244123

Observação: Os dados referentes a 2021 são preliminares.

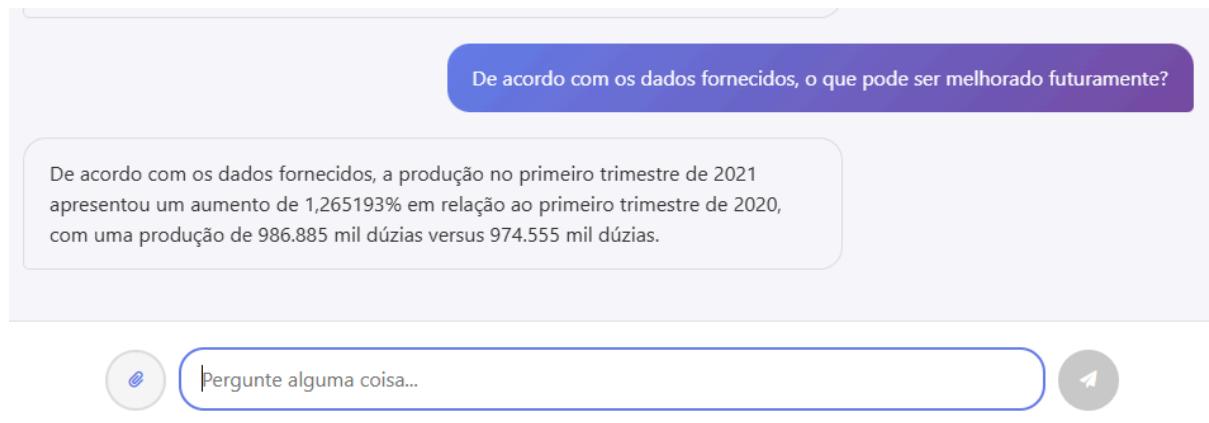
 Pergunte alguma coisa... 

Pressione Shift + Enter para quebrar linha, Enter para enviar • Anexe documentos para análise

Fonte: Elaborada pelos autores (2025).

A figura 5 ilustra um relatório gerado pelo modelo a partir do mesmo arquivo analisado pelo MISTRAL na figura 2. Nota-se que o relatório apresentou uma formatação distinta, incluindo um título, um resumo textual e uma nota de observação.

Figura 6- Resposta do Gemma a uma pergunta de acompanhamento sobre o relatório.

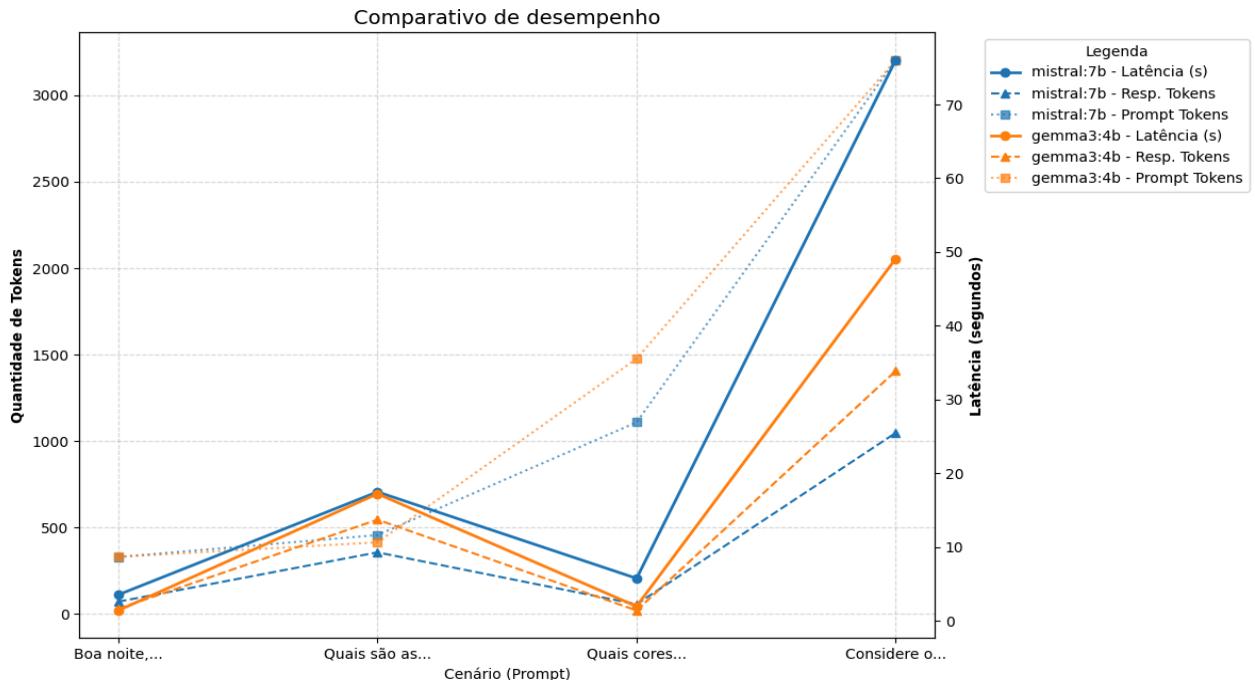


Fonte: Elaborada pelos autores (2025).

A figura 6 ilustra uma resposta do modelo Gemma ao ser questionado para futuras melhorias com base nos dados do relatório. O Gemma não foi capaz de sugerir ações estratégicas em comparação ao modelo MISTRAL, em vez disso, sua resposta consistiu em um único dado quantitativo do relatório.

A Figura 7 consolida a comparação entre os dois modelos, MISTRAL e Gemma, destacando os principais aspectos de desempenho observados durante os testes. O gráfico compara as métricas de contagem de tokens de entrada (Prompt Tokens), tokens de resposta (Response Tokens) e latência para quatro cenários de consulta distintos.

Figura 7- Comparativo de desempenho técnico entre os modelos MISTRAL e Gemma.



Fonte: Elaborada pelos autores (2025).

De acordo com os resultados, observou-se que o Gemma:3.4b apresentou menor consumo de tokens e menor latência em interações curtas, como saudações simples (“Boa noite, como está?”), o que o torna mais adequado para aplicações que exigem respostas rápidas e baixo custo computacional. Já o MISTRAL:7b, embora mais lento, demonstrou melhor coesão semântica e profundidade contextual em consultas complexas, como as que envolvem análise de cenários ou recomendações detalhadas, sugerindo uma compreensão mais refinada das informações apresentadas.

Essas diferenças de desempenho refletem diretamente o equilíbrio entre velocidade e qualidade textual. Enquanto o Gemma é otimizado para baixa latência e menor uso de recursos, o MISTRAL é voltado para respostas mais elaboradas e contextualmente precisas. O custo computacional do MISTRAL (modelo com 7 bilhões de parâmetros) é superior, porém suas respostas apresentam maior riqueza semântica, o que pode ser vantajoso em sistemas voltados à análise técnica e relatórios detalhados.

A escolha do modelo depende diretamente do objetivo da aplicação. O Gemma:3.4b é mais adequado para interações rápidas, assistência em tempo real e uso em dispositivos com hardware limitado. Já o MISTRAL:7b é mais indicado para

atividades que demandam maior profundidade analítica, como a elaboração de relatórios técnicos e a realização de diagnósticos complexos.

Os resultados indicam que ambos os modelos são viáveis para a implementação de um assistente virtual voltado à avicultura. O Gemma se destaca pela eficiência operacional e pela capacidade de escalar o sistema, enquanto o MISTRAL sobressai pela qualidade e riqueza textual das respostas.

4 CONSIDERAÇÕES FINAIS

O modelo MISTRAL:7b se sobressaiu como a alternativa mais indicada para aplicações diversas, graças à sua notável habilidade de gerar textos ricos, coerentes e com elevado nível de contextualização. Embora não apresente a mesma velocidade de processamento do Gemma:3.4b, essa diferença está diretamente associada às limitações do ambiente computacional utilizado durante os testes, que contava com hardware de médio desempenho e placas gráficas (*GPUs*) com capacidade restrita de memória. Essa condição influenciou o tempo de resposta do modelo, mas não comprometeu a qualidade de suas saídas.

Os resultados obtidos evidenciam que o Gemma:3.4b se destaca pela eficiência operacional, sendo mais adequado para cenários que exigem baixa latência e respostas rápidas, enquanto o MISTRAL:7b apresenta melhor desempenho em tarefas que demandam profundidade analítica, riqueza textual e maior compreensão semântica. Dessa forma, a escolha entre os dois modelos depende diretamente do objetivo da aplicação: velocidade e escalabilidade, no caso do Gemma; ou qualidade e detalhamento contextual, no caso do Mistral.

Como limitação principal, destaca-se o fato de que os testes foram conduzidos em ambiente controlado e com um conjunto reduzido de dados, o que impede generalizações mais amplas sobre o desempenho dos modelos em cenários reais de produção. Além disso, a utilização de *GPUs* com baixa memória de vídeo impactou diretamente o tempo de inferência e a estabilidade das execuções, limitando a exploração de modelos mais robustos.

Para trabalhos futuros, pretende-se realizar testes em larga escala, utilizando dados reais da avicultura como relatórios de produção, registros sanitários e parâmetros nutricionais de modo a avaliar a capacidade de generalização e

precisão das respostas em contextos práticos. Também pretende-se desenvolver módulos multimodais, capazes de processar simultaneamente texto, imagem e planilhas, bem como incorporar modelos de linguagem especializados no domínio agrícola, otimizando o desempenho e a aplicabilidade do sistema.

Enfim, o estudo demonstrou que o uso de Modelos de Linguagem de Grande Escala (LLMs) é viável e promissor para o desenvolvimento de assistentes virtuais voltados à avicultura, contribuindo para a transformação digital do setor e o acesso facilitado a informações técnicas, com potencial de impacto direto na eficiência produtiva e na tomada de decisão de produtores rurais.

REFERÊNCIAS

AGÊNCIA BRASIL. País tem recorde na produção de ovos, abate de bovinos, frango e porco. Agência Brasil, 18 mar. 2025. Disponível em: <https://agenciabrasil.ebc.com.br/economia/noticia/2025-03/pais-tem-recorde-na-producao-de-ovos-abate-de-bovinos-frango-e-porco>. Acesso em: 28 out. 2025.

ALBUQUERQUE, Nathan Costa. Aplicação da inteligência artificial no processo de gerenciamento no agronegócio. 2024. Trabalho de Conclusão de Curso (Bacharelado em Administração) – Universidade Estadual de Mato Grosso do Sul, Unidade Universitária de Ponta Porã. Disponível em: <<https://www.uems.br/anexos/download/23871>>. Acesso em: 4 out. 2025.

BARACHO, Marta dos Santos; TOLÓN, Yamilia Barrios. Inovações digitais na avicultura. In: CIÊNCIAS AGRÁRIAS: inovações, sustentabilidade e desafios para o século XXI. Cap. 4, p. 55-71. 2025. Disponível em: <https://www.editoracientifica.com.br/books/chapter/inovacoes-digitais-na-avicultura>. Acesso em: 28 out. 2025.

CALDAS, Edgard Onoda Luiz. Análise econômica e de eficiência produtiva da produção industrial de frangos de corte e perspectivas de integração em Minas Gerais. 2018. 171 f. Tese (Doutorado em Zootecnia) — Universidade Federal de Minas Gerais, Escola de Veterinária, Belo Horizonte, 2018. Disponível em: <https://repositorio.ufmg.br/items/06e31bbc-b181-4b6b-b36d-27943e3fa5ef>. Acesso em: 28 out. 2025.

IBM. **O que são LLMs** (Large Language Models). *IBM Brasil*, 2 nov. 2023. Disponível em: <<https://www.ibm.com/br-pt/think/topics/large-language-models>>. Acesso em: 5 out. 2025.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa trimestral da produção de ovos de galinha**. *IBGE*, [s.d.]. Disponível em: <<https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9216-pesquisa-trimestral-da-producao-de-ovos-de-galinha.html?edicao=43629&t=downloads>>. Arquivo consultado: *4_Producao_de_Ovos_de_Galinha > Resultados_completos*. Acesso em: 4 out. 2025.

MICROSOFT LEARN. **Entender LLMs**. *Microsoft Learn*, [s.d.]. Disponível em: <https://learn.microsoft.com/pt-br/training/modules/introduction-large-language-models/2-understand-large-language-models>. Acesso em: 5 out. 2025.

MINAEE, S. et al. **Large language models: A survey**. *arXiv preprint*, arXiv:2107.00824, 2021. Disponível em: <https://arxiv.org/abs/2107.00824>. Acesso em: 28 out. 2025.

PEGINO, Paulo Marcelo Ferrarese; RIBEIRO NETO, José Campos de Araújo; URPIA, Arthur Gualberto Bacelar da Cruz. **O ciclo da gestão do conhecimento na tomada de decisão sobre hedge cambial em uma cooperativa do agronegócio**. *Organizações & Sustentabilidade*, Londrina, v. 9, n. 1, p. 203–226, jan./dez. 2021. Disponível em: <<https://ojs.uel.br/revistas/uel/index.php/ros/article/download/46603/47803>>. Acesso em: 5 out. 2025.

PIZOL, José Vitor. **Avicultura 4.0: tecnologias e benefícios**. Nutrimosaic, 10 out. 2024. Disponível em: <https://nutrimosaic.com.br/avicultura-4-0-tecnologias-e-beneficios/>. Acesso em: 5 out. 2025.

SAPÉ AGRO. **A prova de que a avicultura é a força do Brasil**. Sapé Agro, 2025. Disponível em: <https://sapeagro.com.br/noticia/a-prova-de-que-a-avicultura-e-a-forca-do-brasil/36>. Acesso em: 28 out. 2025.