INSTITUTO FEDERAL
Pernambuco

Análise Comparativa de Modelos em Sistemas Multi-Agentes para Geração de Documentos Analíticos

Alexandre Henrique Silva do Nascimento
ahsn@discente.ifpe.edu.br
Flávio Rosendo da Silva Oliveira
flavio.oliveira@paulista.ifpe.edu.br

RESUMO

A evolução dos Grandes Modelos de Linguagem impulsionou avanços significativos na automação de tarefas intelectuais. Entretanto, a confiabilidade das respostas e a propensão a alucinações ainda representam limitações relevantes para seu uso em contextos profissionais. Este trabalho apresenta uma análise comparativa do desempenho de três modelos: GPT-4o-mini, Llama 3–8B e Llama 3–70B. A comparação de desempenho entre os modelos ocorreu em dois estudos de caso distintos: (i) a validação de um texto acadêmico com foco em detecção de inconsistências e alucinações, e (ii) a análise financeira comparativa de múltiplos balancetes. Para conduzir os experimentos, desenvolveu-se um sistema multi-agente, utilizando o framework CrewAI para orquestrar agentes autônomos responsáveis pela geração, revisão e validação dos documentos. A metodologia emprega um *pipeline* estruturado de verificação por meio de *checklists* preenchidos manualmente, permitindo comparar as saídas dos modelos com os documentos originais e mensurar sua precisão. Os resultados evidenciam diferenças significativas entre os três modelos quanto à fidelidade ao conteúdo, clareza textual, consistência analítica e taxa de alucinações, demonstrando que modelos proprietários e de código aberto apresentam comportamentos distintos dependendo da natureza da tarefa analisada.

Palavras-chave: inteligência artificial; sistemas multi-agentes; grandes modelos de linguagem; documentos analíticos.

ABSTRACT

The evolution of Large Language Models has driven significant advances in the automation of intellectual tasks. However, the reliability of their responses and their tendency to produce hallucinations remain relevant limitations for their use in professional contexts. This work presents a detailed comparative analysis of the performance of three models: GPT-4o-mini, Llama 3–8B, and Llama 3–70B. These models were applied to two distinct case studies: (i) the validation of an academic text with a focus on detecting inconsistencies and hallucinations, and (ii) the comparative financial analysis of multiple balance sheets. To conduct the experiments, a multi-agent system was developed using the CrewAI framework to orchestrate autonomous agents responsible for generating, reviewing, and validating documents. The methodology employs a structured verification pipeline based on manually

created checklists, allowing the comparison of model outputs against the original documents and enabling the assessment of their accuracy. The results reveal significant differences among the three models in terms of content fidelity, textual clarity, analytical consistency, and hallucination rates, demonstrating that proprietary and open-source models exhibit distinct behaviors depending on the nature of the task.

Keywords: artificial intelligence; multi-agent systems; large language models; analytical documents.

1 INTRODUÇÃO

A Inteligência Artificial Generativa vem passando por um avanço significativo impulsionado pelos Grandes Modelos de Linguagem (LLMs), como GPT, Llama e outros modelos abertos e proprietários. Esses modelos, capazes de compreender e gerar texto em linguagem natural, vêm moldando uma nova fase da computação, ampliando o suporte ao desenvolvimento de software, à análise de dados e à automação de tarefas complexas. O uso de LLMs vem se expandindo para domínios que exigem não só geração de texto, mas também raciocínio estruturado, verificações de consistência e análise crítica de informações.

Apesar desses avanços, a ocorrência de alucinações, definidas como a geração de informações plausíveis, porém factualmente incorretas, é considerada um dos maiores problemas atuais na utilização da Inteligência Artificial. Esse fenômeno compromete a confiabilidade de sistemas que utilizam IA para produzir documentação técnica, resumos analíticos ou interpretações de dados sensíveis, sendo amplamente discutido na literatura sobre geração de linguagem natural e LLMs (Ji et al., 2023).

Nesse cenário, sistemas multi-agentes surgem como uma alternativa promissora, permitindo que diferentes agentes especializados colaborem entre si para executar tarefas de forma mais estruturada e confiável. Em vez de depender de um único modelo, a combinação de agentes responsáveis por funções como planejamento, verificação factual, síntese e revisão possibilita mitigar erros e aumentar o rigor analítico da solução. Essa abordagem já vem sendo explorada em estudos sobre agentes autônomos baseados em LLMs e arquiteturas multiagentes, embora ainda apresente desafios relacionados à coordenação, à consistência e à validação cruzada das respostas (Wang et al., 2023; Guo et al., 2024).

Pesquisas recentes têm investigado de forma sistemática o uso de Grandes Modelos de Linguagem como componentes centrais de agentes autônomos capazes de planejar, raciocinar e executar tarefas complexas. Wang et al. (2023) apresentam uma análise abrangente sobre agentes baseados em LLMs, propondo uma arquitetura composta por módulos de perfil, memória, planejamento e ação, que permite estruturar sistemas capazes de interagir com ambientes e resolver tarefas de forma relativamente autônoma.

De forma complementar, Guo et al. (2024) investigam arquiteturas multiagentes baseadas em LLMs, destacando que a divisão de tarefas entre agentes especializados pode reduzir limitações observadas em modelos isolados, como erros de interpretação ou inconsistências analíticas. Os autores classificam diferentes formas de cooperação entre agentes, incluindo arquiteturas paralelas, sequenciais e hierárquicas, nas quais cada agente assume responsabilidades específicas dentro do sistema.

Outro avanço relevante nesse campo foi apresentado por Wu et al. (2023), que desenvolveram o framework AutoGen, uma plataforma voltada à construção de aplicações baseadas em múltiplos agentes que cooperam por meio de interações conversacionais estruturadas. Nesse modelo, diferentes agentes podem assumir papéis distintos, como planejamento, execução e revisão, interagindo iterativamente até alcançar uma solução satisfatória.

A organização deste trabalho está estruturada da seguinte forma: a Seção 2 apresenta o referencial teórico; a Seção 3 discute os trabalhos relacionados; a Seção 4 descreve a

metodologia adotada; a Seção 5 apresenta os estudos de caso e os resultados obtidos; e, por fim, a Seção 6 apresenta as considerações finais.

2 REFERENCIAL TEÓRICO

2.1 Agentes Inteligentes e *Large Language Models*

Os LLMs representam um marco na evolução da Inteligência Artificial, caracterizando-se por sua capacidade de compreender, processar e gerar linguagem natural em níveis elevados de desempenho em tarefas específicas. Esses modelos são treinados sobre grandes volumes de dados textuais e utilizam predominantemente a arquitetura Transformer, introduzida por Vaswani et al. (2017), que revolucionou o campo ao empregar mecanismos de atenção capazes de capturar relações complexas entre palavras, frases e contextos longos.

A expansão dos LLMs permitiu avanços significativos em áreas como tradução automática, sumarização, classificação de documentos, extração de informação, geração criativa de conteúdo e análise contextual de textos. Entretanto, limitações estruturais ainda persistem. Modelos de larga escala são conhecidos por gerar alucinações, isto é, informações aparentemente plausíveis, mas factualmente incorretas, além de erros numéricos, falhas na compreensão temporal e inconsistências na identificação de relações lógicas. Tais limitações tornam-se mais evidentes em tarefas que exigem precisão factual ou rigor metodológico, como relatórios financeiros, análises acadêmicas ou avaliações de dados sensíveis (Ji et al., 2023).

Assim, embora os LLMs sejam ferramentas poderosas, seu uso seguro e confiável depende de mecanismos complementares de validação, coordenação e controle.

A literatura aponta que a ausência de mecanismos de validação em LLMs pode gerar consequências críticas no mundo real. Em contextos práticos, como em análises financeiras, de saúde e jurídicas, a dependência exclusiva de um único modelo já resultou em respostas com dados contábeis distorcidos ou citações de jurisprudências inexistentes, o que pode induzir tomadores de decisão a erros graves. Como destacam Ji et al. (2023), esses casos reais demonstram que, embora os LLMs sejam ferramentas poderosas, seu uso seguro depende obrigatoriamente de arquiteturas que validem as informações geradas

2.2 Agentes Inteligentes e Sistemas Multi-Agentes

Em sistemas baseados em LLMs, agentes podem ser entendidos como componentes especializados responsáveis por funções distintas dentro de um *pipeline*, como planejamento, análise, verificação e síntese. Quando organizados de forma cooperativa, esses componentes permitem distribuir tarefas e estruturar melhor a resolução de problemas complexos (Wang et al., 2023; Guo et al., 2024).

No contexto de LLMs, essa organização favorece a criação de camadas adicionais de verificação, análise semântica e síntese final. Estudos recentes reforçam essa perspectiva ao demonstrar que a colaboração entre múltiplos agentes baseados em LLMs pode ampliar significativamente a capacidade de resolução de tarefas complexas. Tran et al. (2025) analisam diferentes estratégias de cooperação entre agentes, evidenciando que a divisão estruturada de responsabilidades contribui para reduzir erros e aumentar a confiabilidade das respostas geradas. De forma semelhante, Yan et al. (2025) destacam que a comunicação entre agentes constitui um elemento central nesses sistemas, permitindo coordenação eficiente e troca estruturada de informações durante a execução das tarefas.

3 TRABALHOS RELACIONADOS

A literatura documenta um salto significativo na evolução da geração de linguagem natural desde a introdução da arquitetura Transformer (Vaswani et al., 2017). Inicialmente, os Grandes Modelos de Linguagem eram utilizados de forma isolada, atuando como preditores

de texto em fluxo único. Trabalhos dessa fase inicial apontavam limitações severas de contexto, perda de coerência em documentos longos e alta propensão a alucinações. Mais recentemente, a pesquisa avançou para a construção de Sistemas Multi-Agentes (MAS). Obras atuais, como as de Guo et al. (2024) e Tran et al. (2025), evidenciam que a atual orquestração de LLMs em papéis específicos reduziu drasticamente essas falhas iniciais, espelhando a divisão de trabalho humano e viabilizando a resolução de tarefas muito mais complexas e confiáveis do que as abordagens do início da década.

Nos últimos anos, diversos estudos têm investigado o uso de arquiteturas baseadas em Grandes Modelos de Linguagem combinados com sistemas multi-agentes para execução de tarefas complexas. Trabalhos como os de Wang et al. (2023), Guo et al. (2024) e Wu et al. (2023) exploram como agentes especializados podem colaborar em *pipelines* estruturadas de planejamento, execução e verificação de resultados, ampliando a capacidade analítica dos modelos de linguagem. Essas abordagens têm sido aplicadas em diferentes domínios, incluindo geração de documentos, análise de dados e automação de processos.

Os trabalhos analisados relacionam-se diretamente com esta análise por abordarem a geração automática de documentos utilizando LLMs e enfrentarem desafios semelhantes aos observados neste estudo, como a necessidade de preservar a integridade informacional, reduzir alucinações e estruturar conteúdos de maneira coerente. O sistema AGIR, conforme discutido por Perrina et al. (2023), demonstra que a combinação de modelos de linguagem com estruturas semânticas e etapas intermediárias de verificação contribui para aumentar significativamente a fluência e a precisão de relatórios gerados automaticamente, fundamentando a abordagem aqui adotada ao avaliar textos produzidos em diferentes domínios, como revisão acadêmica e análise financeira. De forma complementar, o trabalho de Musumeci et al. (2024) apresenta uma arquitetura multi-agente baseada em LLMs que distribui etapas de geração de documentos entre agentes especializados, reduzindo erros e mitigando limitações inerentes aos modelos. Esses princípios também orientam a metodologia deste trabalho, que utiliza agentes distintos para verificação, correção e síntese dos relatórios gerados. Assim, ambos os estudos oferecem sustentação teórica e metodológica ao demonstrar que LLMs podem produzir relatórios úteis quando apoiados por mecanismos adicionais de controle e por *pipelines* estruturados, servindo como base para a investigação realizada nesta pesquisa, que compara diferentes modelos abertos e proprietários e examina o impacto de uma arquitetura multiagente na qualidade final dos documentos produzidos.

4 METODOLOGIA

Este trabalho adota uma abordagem experimental estruturada a partir de dois estudos de caso distintos, documentados na seção 5. O objetivo central é avaliar a aplicação de um sistema multi-agente baseado em LLMs em cenários de complexidade variada, examinando sua capacidade de gerar análises consistentes, precisas e comparáveis. No primeiro estudo de caso, o sistema foi utilizado para produzir um relatório de análise contendo sugestões de correção para um texto científico previamente elaborado, com foco na detecção de inconsistências gramaticais, necessidades de citação e problemas de clareza. No segundo estudo de caso, o sistema foi aplicado à geração de um relatório financeiro comparativo a partir de balancetes referentes ao período de junho a outubro de um determinado ano, envolvendo extração de dados, análise temporal e síntese. Para ambos os cenários, foram utilizados três modelos de linguagem distintos: GPT-4o-mini que é um modelo proprietário da OpenAI, Llama 3-8B que é um modelo *open source* de menor porte e Llama 3-70B que é um modelo *open source* de maior capacidade. A escolha desses três modelos permite uma análise comparativa sobre o desempenho desses modelos que possuem características diferentes, em especial seu poder de processamento representado na quantidade de parâmetros treinados. Os resultados gerados por cada modelo foram posteriormente avaliados por meio de métricas qualitativas aplicadas manualmente por um avaliador humano, permitindo comparar o desempenho dos modelos e identificar diferenças

significativas em sua capacidade analítica, fidelidade aos dados originais e propensão a alucinações.

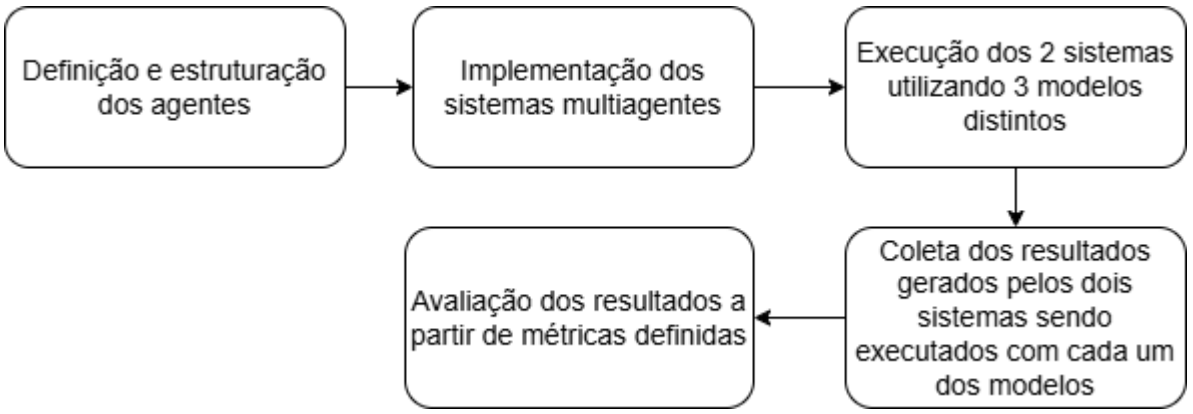
A metodologia seguiu um *pipeline* estruturado em cinco etapas principais, conforme ilustrado na Figura 1, que apresenta o fluxograma metodológico adotado neste trabalho e que será detalhado nas subseções a seguir.

4.1 Definição dos agentes

A primeira etapa da metodologia consistiu na definição dos agentes especializados que compõem o sistema multi-agente para cada estudo de caso. A escolha de três agentes por cenário foi fundamentada na necessidade de distribuir responsabilidades analíticas específicas, permitindo que cada agente se concentrasse em uma dimensão particular do problema, reduzindo a complexidade cognitiva de cada tarefa individual e aumentando a precisão das análises. Essa abordagem metodológica é corroborada por Guo et al. (2024) e Wang et al. (2023), que demonstram que a divisão estruturada de tarefas em sistemas multi-agentes mitiga significativamente o risco de alucinações e falhas de raciocínio observadas quando um único modelo tenta processar múltiplas dimensões simultaneamente.

Para o estudo de caso 1, focado na análise de textos científicos, foram definidos três agentes que cobrem as principais dimensões de qualidade de um documento acadêmico: (i) o Analista Gramatical, responsável pela correção técnica da língua; (ii) o Analista de Citações, encarregado do rigor metodológico e fundamentação bibliográfica; (iii) e o Analista de Clareza, dedicado à legibilidade e fluidez comunicativa. A escolha desses três agentes baseou-se na constatação de que a qualidade de um texto científico depende simultaneamente da correção formal, da fundamentação teórica e da clareza expositiva, dimensões que, embora inter-relacionadas, exigem competências analíticas distintas.

Figura 1 – Fluxograma metodológico realizado neste trabalho. Os dois sistemas mencionados se referem aos dois estudos de caso abordados na Seção 5.



Fonte: Elaborada pelos autores(2026)

No segundo estudo de caso, a arquitetura multi-agente foi organizada em três funções complementares. **(i)** O primeiro agente, denominado Extrator de Dados Financeiros, foi responsável por identificar e estruturar as informações contábeis presentes nos balancetes, incluindo receitas, despesas, resultados, margens e principais categorias de gastos. **(ii)** O segundo agente, denominado Analista Financeiro Comparativo, teve como função comparar os períodos analisados, identificar variações relevantes, recorrências e padrões observáveis

nos dados financeiros, mantendo a análise restrita às evidências disponíveis nos documentos. **(iii)** O terceiro agente, denominado Sintetizador Analítico, foi responsável por consolidar os dados extraídos e as análises realizadas em um relatório final claro, descritivo e fiel aos balancetes, destacando achados principais, pontos para investigação e limitações da análise, sem propor metas ou recomendações normativas sem base documental explícita.

Tabela 1 - Objetivo dos Agentes do estudo de caso 1

Agente	Objetivo
Analista Gramatical	Realiza uma análise no arquivo para conferir erros gramaticais
Analista de Clareza	Analisa a legibilidade e fluidez comunicativa
Analista de Citações	Analisa o rigor metodológico e a fundamentação

Fonte: Elaborada pelos autores(2026)

Tabela 2 - Objetivo dos Agentes do estudo de caso 2

Agente	Objetivo
Extrator de Dados Financeiros	Realiza a extração e a estruturação de dados financeiros dos balancetes
Analista Financeiro Comparativo	Compara os períodos analisados e identifica padrões, variações e recorrências nos dados financeiros
Sintetizador Analítico	Consolida os dados e as análises em relatório descritivo, claro e fiel às evidências observadas

Fonte: Elaborada pelos autores(2026)

4.2 Implementação do Sistema com CrewAI

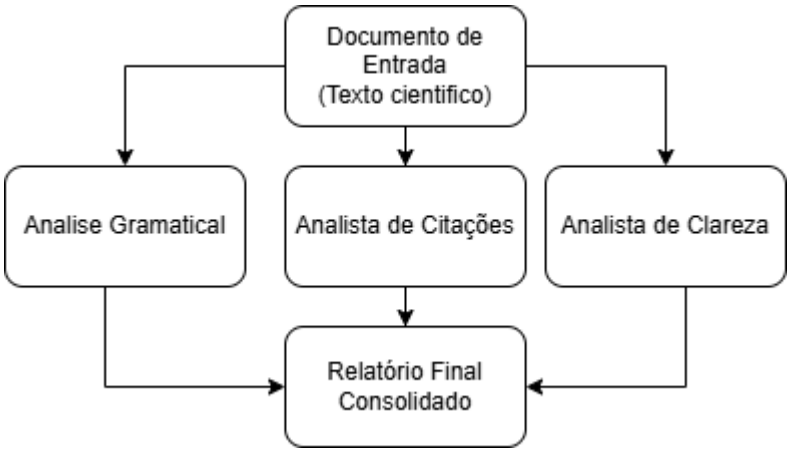
A segunda etapa metodológica envolveu a implementação técnica do sistema multi-agente utilizando o framework CrewAI, uma plataforma especializada na orquestração de agentes baseados em LLMs que permite definir papéis específicos, controlar fluxos de execução e gerenciar a troca de mensagens entre componentes (CREWAI, 2024). A escolha do CrewAI foi motivada por sua capacidade de coordenar múltiplos agentes de forma estruturada, oferecendo controle rigoroso sobre o escopo e o papel de cada agente, características essenciais documentadas por Wu et al. (2023) para o sucesso de aplicações multi-agentes.

Os dois estudos de caso implementados apresentam arquiteturas distintas quanto ao relacionamento entre agentes, conforme ilustrado na Figura 2 e Figura 3. No estudo de caso 1, conforme Figura 2, os três agentes operam de forma independente: cada agente recebe o mesmo documento de entrada, realiza sua análise especializada sem depender das saídas dos demais e produz um relatório específico. Ao final, as três análises independentes (i.e. gramatical, de citações e de clareza) são consolidadas em um relatório único que integra todas as dimensões avaliadas.

No estudo de caso 2, conforme Figura 3, os agentes operam de forma sequencial e interdependente: (i) o Extrator de Dados Financeiros processa os balancetes brutos e produz uma estrutura de dados padronizada; (ii) essa saída alimenta o Analista Financeiro Comparativo, que identifica padrões temporais e variações entre períodos; (iii) por fim, o Sintetizador Analítico recebe as análises técnicas e as traduz em formato executivo acionável. Vale ressaltar que, em análises com fluxo sequencial, o risco de falha sistêmica aumenta à medida que a quantidade de agentes também aumenta, visto que o sucesso do último agente (Sintetizador) depende intrinsecamente da qualidade das análises realizadas pelos agentes anteriores (Extrator e Analista).

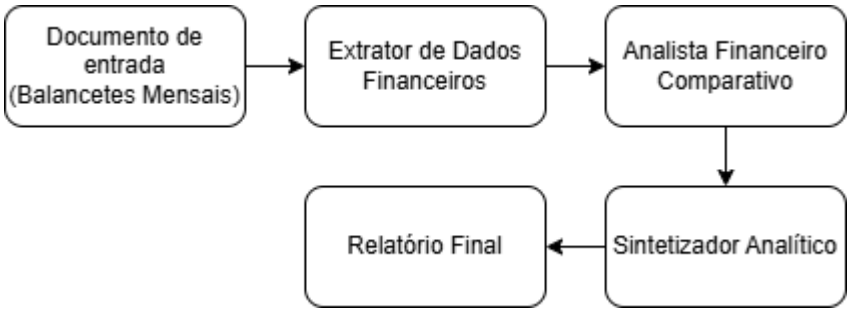
A implementação técnica envolveu a definição de *prompts* especializados para cada agente, especificando instruções claras, passos de execução estruturados e formatos de saída padronizados. Cada agente foi equipado com ferramentas customizadas para leitura de documentos, garantindo acesso uniforme aos dados de entrada. Os prompts foram projetados com restrições explícitas sobre o escopo de atuação de cada agente, evitando sobreposições funcionais e garantindo que cada componente se concentrasse exclusivamente em sua responsabilidade designada. Os códigos criados podem ser vistos através do link para o repositório Github contido no Apêndice A.

Figura 2 - Arquitetura de Agentes do Estudo de Caso 1



Fonte: Elaborada pelos autores(2026)

Figura 3 - Arquitetura de Agentes do Estudo de Caso 2



Fonte: Elaborada pelos autores(2026)

4.3 Execução dos Sistemas Utilizando Três Modelos de Linguagem Distintos

Após a definição dos agentes e a implementação dos dois sistemas multiagentes, procedeu-se à etapa de execução experimental, conforme indicado no fluxograma metodológico apresentado na Figura 1. Nesta fase, os dois estudos de caso desenvolvidos foram executados utilizando três modelos de linguagem distintos: GPT-4o-mini, Llama 3–8B e Llama 3–70B.

Cada sistema foi executado separadamente com cada um dos três modelos, resultando em seis execuções independentes: três correspondentes ao Estudo de Caso 1 e três ao Estudo de Caso 2. Essa estratégia permitiu isolar o desempenho de cada modelo sob as mesmas condições experimentais.

No Estudo de Caso 1, os três agentes especializados (i.e. Analista Gramatical, Analista de Citações e Analista de Clareza) atuaram de forma independente, cada um responsável por uma dimensão específica da qualidade textual. Cada agente recebe o mesmo documento de entrada e produz um relatório próprio. Após a execução das três análises, os resultados são consolidados em um relatório final, permitindo integrar as diferentes dimensões avaliadas. Já no Estudo de Caso 2, os agentes atuaram de maneira sequencial e interdependente, em um *pipeline* estruturado no qual a saída do agente extrator alimentava o agente analista, cuja saída, por sua vez, servia de base para o agente sintetizador analítico.

Todas as saídas geradas foram armazenadas integralmente para posterior análise comparativa e aplicação das métricas qualitativas descritas na Seção 4.4.

4.4 Coleta dos Resultados e Aplicação de Métricas Qualitativas

As duas etapas finais da metodologia envolveram a coleta sistemática dos relatórios gerados e a avaliação qualitativa de sua correção e qualidade. Após a execução de cada estudo de caso, foram coletados os relatórios produzidos por cada um dos três LLMs testados (i.e. GPT-4o-mini, Llama 3-8B e Llama 3-70B), resultando em seis conjuntos de documentos analíticos: três relatórios para o estudo de caso 1 e três relatórios para o estudo de caso 2.

A avaliação dos resultados foi realizada com base em checklists qualitativos, aplicados manualmente pelo próprio autor desta pesquisa, atuando como avaliador humano. Entre os riscos inerentes a essa metodologia, destaca-se o viés do avaliador único. Embora a identificação de erros gramaticais e a verificação de percentuais em demonstrativos financeiros sejam tarefas de natureza objetiva, um avaliador humano sem o conhecimento técnico aprofundado em cada uma dessas áreas específicas poderia, inadvertidamente, considerar como correta uma informação que, na realidade, foi alucinada pelo modelo.

No estudo de caso 1, o checklist contemplou os seguintes critérios:(i) Precisão dos problemas identificados; (ii) Localização exata; (iii) Identificação correta de afirmações; (iv) Contexto preservado; (v) Identificação de ambiguidades; (vi) Preservação do tom acadêmico; (vii) Estruturação adequada; (viii) Ausência de alucinações;

Para o estudo de caso 2, o checklist adotado compreendeu critérios distintos, voltados à avaliação da qualidade da extração e da análise dos dados financeiros. Os critérios para avaliação dos resultados são: (i) Preservação de informações-chave; (ii) Ausência de alucinações; (iii) Concisão adequada; (iv) Coerência interna; e (v) Identificação de padrões válidos; (vi) Conclusões fundamentadas; (vii) Aderência ao formato; (viii) Metadados corretos.

A metodologia de aferição manual foi escolhida devido à natureza qualitativa e contextual das dimensões avaliadas, que não podem ser capturadas adequadamente por métricas automatizadas convencionais.

Os resultados dessa avaliação qualitativa, detalhados na Seção 5, permitiram identificar diferenças significativas entre os três modelos quanto à sua capacidade de produzir análises precisas, consistentes e livres de alucinações, fornecendo *insights* sobre a adequação de cada modelo para aplicações profissionais em contextos de análise textual acadêmica e análise financeira.

Entre os riscos inerentes a essa metodologia, destaca-se o viés do avaliador único. Em contextos técnicos, um avaliador humano sem o conhecimento aprofundado específico poderia, inadvertidamente, considerar como correta uma informação gramatical ou financeira que, na realidade, foi alucinada pelo modelo.

5 - ESTUDOS DE CASO

Esta seção apresenta os dois estudos de caso conduzidos neste trabalho, cujo objetivo foi avaliar o desempenho de um sistema multi-agente baseado em LLMs em tarefas analíticas distintas. O primeiro estudo de caso (Seção 5.1) concentra-se na análise e revisão de um texto científico contendo erros propositalmente inseridos, visando avaliar a capacidade dos modelos em identificar inconsistências gramaticais, sugerir correções de clareza e validar necessidades de citação. O segundo estudo (Seção 5.2) aborda a análise financeira comparativa a partir de demonstrativos contábeis, examinando a precisão na extração de dados e a geração de resumos executivos. Os resultados gerados em ambos os cenários buscaram examinar a confiabilidade das análises produzidas.

5.1 - Estudo de Caso 1: Análise e revisão de textos científicos

O sistema multi-agente foi estruturado a partir de três agentes especializados, que atuam de forma independente, cada qual responsável por uma dimensão específica da qualidade textual: correção gramatical, análise de citações acadêmicas e clareza comunicativa.

O Agente Analista Gramatical foi concebido para atuar exclusivamente sobre os aspectos formais da língua portuguesa. Para a verificação ortográfica e gramatical, o agente não utilizou nenhuma fonte ou dicionário externo, baseando-se exclusivamente em seu próprio conhecimento pré-treinado acerca das normas padrão do idioma. Seu funcionamento baseia-se na leitura integral do documento e na identificação de desvios ortográficos, uso inadequado de acentuação, falhas de concordância verbal e nominal, problemas de regência, além de inadequações sintáticas e de pontuação. Esse agente não realiza avaliações de conteúdo, clareza ou fundamentação teórica, restringindo-se à dimensão gramatical. Como resultado de sua atuação, são produzidos registros estruturados que indicam a localização do erro no texto, sua descrição objetiva e uma proposta de correção que preserva o sentido original do enunciado.

O Agente Analista de Citações Acadêmicas foi projetado para avaliar a necessidade de fundamentação bibliográfica ao longo do texto científico. Sua atuação consiste na identificação de trechos que, segundo normas acadêmicas, exigem referência externa, tais como dados específicos, afirmações categóricas, menções a teorias ou conceitos formais e resultados atribuídos a pesquisas. Paralelamente, o agente é instruído a não sinalizar como problemáticos trechos que correspondam a conhecimento amplamente aceito, explicações metodológicas do próprio autor ou exemplos ilustrativos, evitando-se rigor excessivo. Para cada ocorrência identificada, o agente indica a localização do trecho, o trecho correspondente e a justificativa técnica para a necessidade de citação, além de sugerir o tipo de fonte apropriada a referenciar.

O Agente Analista de Clareza atua sobre aspectos relacionados à legibilidade e à fluidez do texto científico. Seu funcionamento consiste na detecção de problemas que dificultam a compreensão, tais como ambiguidades, períodos excessivamente longos, uso

inadequado de jargão técnico, falhas de coesão entre ideias e comprometimento do fluxo de leitura. Diferentemente do agente gramatical, este não avalia erros formais da língua nem a ausência de referências bibliográficas, concentrando-se exclusivamente na dimensão comunicativa do texto. Para cada ocorrência, são indicados o trecho correspondente à descrição objetiva do problema identificado e uma sugestão de reformulação que mantenha o tom acadêmico e o significado original do conteúdo.

O sumário das avaliações realizadas será descrito nas subseções 5.1.1 a 5.1.4. Os detalhes técnicos da implementação dos agentes encontram-se disponíveis no repositório público do projeto, conforme indicado no Apêndice A e o checklist completo utilizado na avaliação dos relatórios gerados neste estudo encontra-se disponível no Apêndice D.

5.1.1 - Qualidade da Análise Gramatical

A análise realizada pelos três modelos evidencia diferenças significativas de desempenho em termos de precisão, cobertura e qualidade técnica. O GPT-4o-mini apresentou o resultado mais consistente, pois identificou corretamente alguns erros gramaticais efetivamente presentes no texto, como “questãos”, “gerassão” e “rezolução”, além de reconhecer determinados problemas de clareza. O modelo também localizou com precisão os trechos citados, manteve o tom acadêmico nas sugestões de reformulação e estruturou o relatório em conformidade com a metodologia proposta. Ainda assim, sua avaliação não foi exaustiva, uma vez que deixou de identificar outras falhas gramaticais e de clareza igualmente presentes no texto. O Llama 3–70B, por sua vez, apresentou desempenho intermediário: embora tenha localizado trechos que de fato constam no documento e preservado a formalidade das sugestões, sua análise mostrou-se irregular, com diagnósticos apenas parcialmente corretos, omissão de outros problemas gramaticais e de clareza, além da indicação de citações desnecessárias. Já o Llama 3–8B obteve o desempenho mais limitado, pois não conseguiu identificar adequadamente os erros gramaticais presentes nem localizar com precisão os trechos problemáticos. Embora tenha reconhecido alguns problemas reais de clareza, sua análise foi incompleta e, em certos casos, inadequada ao contexto avaliado, inclusive com sugestões pouco pertinentes. Em conjunto, observa-se que os três modelos seguiram a estrutura técnica esperada, mas diferiram substancialmente quanto à qualidade da avaliação produzida: o GPT-4o-mini destacou-se pelo melhor equilíbrio entre precisão, fidelidade ao texto e organização do relatório; o Llama 3–70B apresentou resultados parciais e inconsistentes; e o Llama 3–8B evidenciou limitações analíticas mais acentuadas. A síntese desses resultados pode ser vista na Tabela 3.

Tabela 3 - resultados da análise sobre a qualidade da análise gramatical por cada modelo

Modelo	Resultado
gpt-4o-mini	Erros ('questãos', 'gerassão', 'rezolução') realmente existem no texto, mas deixou passar outros erros não tão aparentes
Llama 3–8B	Não conseguiu detectar os erros gramaticais contidos no texto
Llama 3–70B	Acertou no diagnóstico de um trecho detectado, mas errou em outro. Além de que deixou de detectar trechos com problemas gramaticais

Fonte: Elaborada pelos autores(2026)

5.1.2 - Validação das Necessidades de Citação

Na validação das necessidades de citação, os três modelos apresentaram desempenhos distintos quanto ao rigor analítico e à pertinência das recomendações produzidas. O GPT-4o-mini demonstrou um resultado equilibrado, ao reconhecer corretamente que o texto analisado não continha afirmações que demandassem comprovação externa, razão pela qual não indicou citações desnecessárias e preservou integralmente o contexto argumentativo. O Llama 3–8B também apresentou comportamento consistente nesse aspecto, ao não apontar a necessidade de referências em trechos que, de fato, não as exigiam. Em contraste, o Llama 3–70B foi o modelo que mais se afastou do resultado esperado, uma vez que classificou de forma equivocada alguns trechos como carentes de citação, gerando recomendações indevidas. Na tabela 4 se encontram resultados das análises feitas sobre a validação de necessidade de citação.

Tabela 4 - resultados da análise sobre a validação de necessidade de citação

Modelo	Resultado
gpt-4o-mini	Não identificou citações, pois não eram necessárias.
Llama 3–8B	Não identificou citações, pois não eram necessárias.
Llama 3–70B	As afirmações existem no texto, porém as recomendações de citação foram desnecessárias

Fonte: Elaborada pelos autores(2026)

5.1.3 - Avaliação de Clareza

Na avaliação da clareza, os três modelos demonstraram capacidade de identificar estruturas que comprometem a fluidez do texto, embora com níveis distintos de precisão e consistência. O GPT-4o-mini apresentou o desempenho mais robusto, ao reconhecer corretamente alguns períodos excessivamente longos, ambiguidades e problemas de coesão, além de oferecer sugestões de reformulação que preservaram tanto o sentido original quanto o tom acadêmico do texto. O Llama 3–70B, por sua vez, identificou adequadamente um trecho longo, mas deixou de reconhecer outros segmentos com problemas de clareza e fluidez, além de apresentar alguns diagnósticos equivocados. Já o Llama 3–8B foi capaz de detectar dois trechos potencialmente problemáticos, porém acertou em apenas um dos diagnósticos e se equivocou no outro, além de não identificar os demais problemas de clareza presentes no texto. Na tabela 5 se encontra uma avaliação dos 3 modelos em relação a avaliação de clareza.

Tabela 5 - resultados da análise sobre a avaliação de clareza

Modelo	Resultado
gpt-4o-mini	Conseguiu detectar alguns problemas de fluidez e clareza e fez o diagnóstico correto, mas outros não foram detectados. Sugestões mantêm sentido e tom corretamente
Llama 3–8B	Trechos com problemas de clareza e fluidez detectados de fato existem, mas não detectou os outros erros e, em um erro detectado, detectou um problema e fez uma sugestão de melhoria que não cabia no trecho detectado. Na outra localização de erro, ele acabou corrigindo automaticamente uma palavra, gramaticalmente errada, de um trecho que tinha problema de clareza, mas fez o diagnóstico correto. Sugestões formais, porém uma sugestão fora do trecho analisado.
Llama 3–70B	Localizou um trecho e fez o diagnóstico corretamente, mas os outros diagnósticos de clareza detectados foram equivocados. Além de que deixou de detectar outros trechos com problemas. Sugestões mantêm formalidade.

Fonte: Elaborada pelos autores(2026)

5.1.4 - Qualidade Técnica do Output

Na dimensão de qualidade técnica do output, os três modelos seguiram, de modo geral, a estrutura esperada para o relatório, embora tenham apresentado diferenças relevantes quanto à consistência e à precisão analítica. O GPT-4o-mini obteve o desempenho mais robusto, uma vez que organizou adequadamente o relatório nas seções previstas e não introduziu informações indevidas nem formulou diagnósticos equivocados. O Llama 3–70B também demonstrou boa aderência à estrutura proposta, porém apresentou alguns diagnósticos incorretos nas seções de análise gramatical, análise de clareza e validação da necessidade de citação. O Llama 3–8B, por sua vez, embora tenha seguido formalmente a estrutura do relatório, também incorreu em diagnósticos equivocados, especialmente na seção de análise de clareza. Assim, pode-se afirmar que os três modelos produziram relatórios tecnicamente válidos, mas o GPT-4o-mini destacou-se por sua maior precisão e fidelidade ao texto analisado, seguido pelo Llama 3–70B, enquanto o Llama 3–8B apresentou a menor consistência técnica entre os modelos avaliados. Na tabela 6 se encontra uma avaliação dos 3 modelos em relação a qualidade técnica do output.

Tabela 6 - resultados da análise sobre a qualidade técnica do output

Modelo	Resultado
gpt-4o-mini	O relatório segue corretamente a estrutura A–B–C definida na metodologia. Nenhum trecho inexistente foi citado.
Llama 3–8B	O relatório segue corretamente a estrutura A–B–C definida na metodologia. Diagnósticos com baixa qualidade na seção C.
Llama 3–70B	O relatório segue corretamente a estrutura A–B–C definida na metodologia. Análises com baixa qualidade na seção A, B e C

Fonte: Elaborada pelos autores(2026)

5.2 - Estudo de Caso 2: Análise Financeira Comparativa

O Estudo de Caso 2 teve como objetivo processar múltiplos demonstrativos contábeis e elaborar um relatório analítico. Para isso, foram utilizados diversos balancetes mensais sintéticos, disponíveis no repositório do GitHub, cujo link está no Apêndice A. Seus valores foram analisados, consolidados e comparados por meio de um conjunto de agentes desenvolvidos em Python. A orquestração desses agentes foi realizada com o framework CrewAI, que permitiu estruturar tarefas especializadas, como a leitura dos documentos, a identificação de padrões financeiros e a síntese dos resultados. Esse estudo buscou avaliar não apenas a capacidade dos modelos de interpretar dados numéricos e produzir análises coerentes, mas também sua precisão, consistência e propensão a alucinações ao lidar com informações financeiras.

O sistema foi estruturado a partir de três agentes especializados que atuam de forma sequencial e complementar:

- O Agente Extrator de Dados Financeiros é responsável pela extração e estruturação dos dados financeiros. Sua atuação envolve a identificação dos principais indicadores financeiros de cada período, que são: receita total, despesas totais, resultado líquido, top 5 maiores gastos com valores, período/mês de referência e margem percentual, bem como a correta associação desses valores ao respectivo mês e ano.
- O Agente Analista Financeiro Comparativo atua sobre os dados previamente estruturados pelo agente extrator. Sua função consiste em identificar padrões temporais e relações comparativas entre os períodos analisados, tais como o melhor e o pior desempenho financeiro, tendências gerais de evolução, variações percentuais e recorrência de categorias de gastos.
- O Agente Sintetizador Analítico é responsável por transformar as análises técnicas produzidas pelo agente comparativo em um relatório estruturado. Seu funcionamento consiste na organização das informações em um formato executivo, contemplando um resumo das principais tendências observadas, uma tabela comparativa entre os períodos e algumas observações.

Os detalhes técnicos da implementação dos agentes encontram-se disponíveis no repositório público do projeto, conforme indicado no Apêndice A e o checklist completo utilizado na avaliação dos relatórios gerados neste estudo encontra-se disponível no Apêndice E.

Ao avaliar as divergências e erros numéricos apresentados por alguns modelos neste cenário, é fundamental considerar o peso dessas falhas no contexto de aplicação. Um texto com um erro gramatical ou de coesão apresenta um prejuízo predominantemente formal. Por outro lado, em um cenário de análise executiva, um erro numérico em um balancete ou na margem de lucro traz consequências hipotéticas consideravelmente mais graves, uma vez que tais relatórios balizam as ações de tomadores de decisão em ambientes corporativos e

podem induzir a escolhas financeiras equivocadas.

5.2.1 - Avaliação da Fidelidade ao conteúdo original

Na dimensão de fidelidade ao conteúdo original, os três modelos apresentaram níveis distintos de rigor na preservação das informações extraídas dos balancetes financeiros. O GPT-4o-mini preservou adequadamente os principais dados dos documentos e apresentou boa organização analítica, mas registrou erros numéricos nos resultados referentes aos meses de junho e agosto, além de pequenas imprecisões no cálculo das margens de junho, agosto e setembro, o que compromete parcialmente a comparação entre os períodos. O Llama 3–70B também demonstrou boa preservação dos dados centrais dos balancetes e correta organização temporal das informações, sem inventar dados-base. Contudo, apresentou pequenas imprecisões nas margens de junho, julho e setembro e incluiu inferências quantitativas e gerenciais que não estavam explicitamente demonstradas nos documentos, como a estimativa de que a rubrica de salários corresponderia a cerca de 30% das despesas totais, o que reduz sua aderência estrita ao conteúdo-fonte. Já o Llama 3–8B preservou os principais pontos dos balancetes e manteve uma organização analítica satisfatória, sem inventar categorias ou alterar a estrutura documental. Ainda assim, apresentou pequenas imprecisões nas margens de junho, julho, agosto e setembro, além de erros numéricos nos resultados de junho, organização temporal inadequada e indicação incorreta de tendência de crescimento, comprometendo, em maior medida, a fidelidade da análise. Na tabela 7 se encontra uma avaliação dos 3 modelos em relação à fidelidade dos conteúdos que estão nos relatórios com os dados fornecidos.

Tabela 7 - resultados da análise sobre a Avaliação da Fidelidade ao conteúdo original

Modelo	Resultado
gpt-4o-mini	Preserva os principais pontos dos balancetes e apresenta boa organização analítica, mas contém erros numéricos nos resultados em junho e agosto, além de pequena imprecisão nas margens de junho, agosto e setembro.
Llama 3–8B	Preserva os principais pontos dos balancetes e apresenta boa organização analítica, mas contêm pequena imprecisão nas margens de junho, julho, agosto e setembro e erros numéricos nos resultados em junho.
Llama 3–70B	Preserva bem os principais dados dos balancetes e organiza corretamente os períodos, embora acrescente interpretações quantitativas não demonstradas explicitamente, como a estimativa de que salários representam cerca de 30% das despesas totais, além de que contém pequena imprecisão nas margens de junho, julho e setembro

Fonte: Elaborada pelos autores(2026)

5.2.2 - Avaliação da Qualidade dos Resumos Individuais

Na análise da qualidade dos resumos individuais, observa-se que os três modelos produziram sínteses coerentes, embora com diferenças relevantes quanto à concisão, à fluidez e à consistência analítica. O GPT-4o-mini apresentou um resumo conciso, com síntese clara e sem detalhamento excessivo. Além disso, demonstrou boa coerência discursiva e linguagem cautelosa. O Llama 3–70B também produziu um resumo coerente e bem estruturado, com boa organização das informações e uma análise analítica adequada,

embora tenha imprecisões nas margens percentuais contidas no resumo. O Llama 3–8B, por sua vez, apresentou um relatório igualmente conciso, com síntese clara e relativamente estável, além de boa coerência discursiva e linguagem cautelosa. Ainda assim, sua consistência analítica mostrou-se mais fragilizada em razão de erros ao detectar uma tendência inexistente de crescimento na receita total e no resultado liquido. Na tabela 8 se encontra uma avaliação dos 3 modelos em relação a fidelidade dos resumos que estão nos relatórios com os dados fornecidos.

Tabela 8 - resultados da análise sobre a Avaliação da Qualidade dos Resumos Individuais

Modelo	Resultado
gpt-4o-mini	O relatório apresenta boa concisão, com síntese clara e sem excesso de detalhamento. O relatório apresenta boa coerência discursiva e linguagem cautelosa. A coerência analítica também é assertiva
Llama 3–8B	O relatório é conciso e direto. O relatório apresenta boa coerência discursiva e linguagem cautelosa, mas a coerência analítica é parcialmente enfraquecida por erros na análise de tendência de crescimento de receita e de resultado.
Llama 3–70B	O relatório apresenta boa concisão, com síntese clara e estável. O relatório apresenta boa coerência discursiva e linguagem cautelosa. A coerência analítica é assertiva, mas tem um pequeno erro de arredondamento nas margens percentuais

Fonte: Elaborada pelos autores(2026)

5.2.3 - Avaliação da Qualidade da Análise Consolidada

Na análise consolidada, o GPT-4o-mini apresentou boa capacidade de identificar padrões nos dados, embora seu desempenho tenha sido parcialmente prejudicado por imprecisões numéricas. De modo geral, suas conclusões mostraram-se bem fundamentadas, ainda que, em alguns momentos, o modelo tenha incorrido em equívocos pontuais. O Llama 3–70B também foi capaz de reconhecer padrões compatíveis com os dados analisados, porém incluiu inferências quantitativas que não se sustentam integralmente nas evidências disponíveis, como a atribuição de aproximadamente 30% das despesas à rubrica de salários. Já o Llama 3–8B identificou apenas parcialmente os padrões presentes nos dados e, em alguns casos, superestimou oscilações pontuais ao interpretá-las como tendência de crescimento. Assim, embora suas conclusões apresentem elementos pertinentes, sua qualidade analítica mostrou-se apenas parcialmente satisfatória, em razão da presença de exageros e inconsistências. Na tabela 9 se encontra uma avaliação dos 3 modelos em relação a qualidade da análise feita nos relatórios.

Tabela 9 - resultados da análise sobre a Qualidade da Análise Consolidada

Modelo	Resultado
gpt-4o-mini	Fez uma boa análise de padrões, sendo parcialmente prejudicada por conta das imprecisões numéricas. As conclusões são bem fundamentadas, mas houve ocasiões onde cometeu equívoco.
Llama 3–8B	Identificou parcialmente os padrões dos dados, mas exagerou ao tratar oscilações como tendência de crescimento. A qualidade das conclusões foi relativamente boa, mas algumas contêm exagero e inconsistências.
Llama 3–70B	Fez boas análises de padrões, mas cometeu um erro, que foi a atribuição de aproximadamente 30% das despesas à rubrica de salários. As conclusões com base nos dados foram boas, mas cometeu um erro

Fonte: Elaborada pelos autores(2026)

5.2.4 - Avaliação da Estrutura e Organização do Output

Quanto à estrutura e à organização dos relatórios, o GPT-4o-mini aderiu adequadamente ao formato esperado de um relatório analítico, contemplando seções de resumo, consolidação, achados e limitações. Além disso, representou corretamente os períodos e as categorias analisadas, embora tenha apresentado inconsistências numéricas nos resultados de junho e agosto, bem como imprecisões nas margens de junho, agosto e setembro. O Llama 3–70B também seguiu de forma satisfatória a estrutura esperada, com boa organização das seções, apresentação consistente dos dados e correta ordenação temporal dos períodos. Contudo, houve equívocos em algumas inferências analíticas, especialmente na quantificação incorreta do peso da rubrica de salários no conjunto das despesas, além de pequenas imprecisões nas margens de junho, julho e setembro. O Llama 3–8B, por sua vez, também seguiu a estrutura geral esperada, mas apresentou leves erros numéricos, imprecisões nas margens percentuais e equívocos na identificação de tendências. De modo geral, os três modelos demonstraram capacidade de organizar os relatórios de forma tecnicamente adequada, embora com diferenças relevantes na precisão analítica e na confiabilidade dos resultados.Na tabela 10 se encontra uma avaliação dos 3 modelos em relação à estrutura do texto do relatório.

Tabela 10 - resultados da análise sobre a Avaliação da Estrutura e Organização do Output

Modelo	Resultado
gpt-4o-mini	Segue adequadamente a estrutura esperada de relatório analítico, com resumo, consolidação, achados e limitações. Os períodos e categorias estão corretamente representados, mas há inconsistências numéricas no resultado de junho e agosto e nas margens de junho, agosto e setembro.
Llama 3–8B	Segue a estrutura geral esperada. Há leves erros numéricos e em margens percentuais e também há equívocos na detecção de tendências.
Llama 3–70B	Segue adequadamente a estrutura esperada, com boa organização das seções e apresentação consistente dos dados. Os metadados temporais e a ordenação dos períodos estão corretos; Equívocos acontecem nas inferências analíticas e em uma quantificação específica incorreta sobre o peso da rubrica de salários nas despesas. E contém pequena imprecisão em margens de junho, julho e setembro

Fonte: Elaborada pelos autores(2026)

Os relatórios integrais gerados por cada modelo nos dois estudos de caso, bem como os checklists preenchidos durante a etapa de avaliação qualitativa, estão organizados nos Apêndices A a E.

6 - CONSIDERAÇÕES FINAIS

Este trabalho avaliou a aplicação de LLMs em uma arquitetura multi-agente baseada no framework CrewAI, considerando dois cenários distintos de análise documental: revisão de textos científicos e análise de documentos financeiros. Foram comparados os modelos GPT-4o-mini, Llama 3–70B e Llama 3–8B quanto à sua capacidade de executar tarefas analíticas por meio da especialização funcional de agentes.

Os resultados obtidos indicaram que todos os modelos foram capazes de desempenhar as tarefas propostas, porém com diferenças relevantes quanto à precisão e à consistência das análises. No estudo de revisão textual, o GPT-4o-mini apresentou maior confiabilidade na identificação de erros gramaticais e problemas de clareza, enquanto o Llama 3–70B demonstrou bom desempenho geral, porém com limitações na avaliação de necessidades de citação, avaliação de clareza e análise gramatical. O Llama 3–8B apresentou o pior desempenho, evidenciando limitações na análise de aspectos textuais mais complexos. No estudo de análise financeira, o GPT-4o-mini apresentou um melhor desempenho entre os modelos, embora com divergências numéricas pontuais e um pequeno equívoco nas suas análises. No mesmo cenário, o Llama 3–70B apresentou boa organização formal, porém com uma análise não integralmente sustentada pelos balancetes e algumas divergências numéricas pontuais, enquanto o Llama 3–8B demonstrou menor estabilidade na precisão numérica, na organização temporal dos períodos analisados e na detecção de tendências.

Uma análise comparativa dos riscos inerentes a cada arquitetura revela diferenças fundamentais entre os dois cenários avaliados. No Estudo de Caso 1, que empregou um fluxo de execução paralela e independente, o principal risco concentrava-se na limitação analítica individual de cada modelo. Em contrapartida, no Estudo de Caso 2, a arquitetura de fluxo sequencial elevou consideravelmente o risco sistêmico; a precisão do relatório do Sintetizador final estava inteiramente subordinada à qualidade da extração dos dados feita

pelos agentes anteriores. Isso evidencia que pipelines em cadeia são mais vulneráveis ao efeito cascata de alucinações e erros numéricos, um fator crítico em domínios corporativos.

Embora os resultados confirmem a viabilidade da abordagem multi-agente para a análise documental automatizada, o estudo apresenta limitações, especialmente quanto à ausência de métricas quantitativas padronizadas e à dependência de avaliação qualitativa. Trabalhos futuros poderão explorar a incorporação de métricas automáticas de validação, a aplicação da abordagem em outros domínios e o uso de agentes para fazer validações cruzadas com os dados gerados pelos agentes.

Embora os resultados confirmem a viabilidade da abordagem multi-agente para a análise documental automatizada, o estudo apresenta limitações, especialmente quanto à ausência de métricas quantitativas padronizadas e à dependência de avaliação qualitativa. Trabalhos futuros poderão explorar a incorporação de métricas automáticas de validação visando o aumento da possibilidade de se replicar o estudo, sendo isso muito importante para a continuidade do desenvolvimento científico, além de aplicar a abordagem em outros domínios e utilizar agentes para fazer validações cruzadas com os dados gerados.

REFERÊNCIAS

CREWAI. **CrewAI**. Disponível em: https://crewai.com/.

GUO, T. et al. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. In: **PROCEEDINGS OF THE THIRTY-THIRD INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE (IJCAI 2024)**. 2024. p. 8048–8057. Disponível em: https://www.ijcai.org/proceedings/2024/890. Acesso em: 12 maio 2026.

JI, Z. et al. Survey of Hallucination in Natural Language Generation. **ACM Computing Surveys**, v. 55, n. 12, 2023. Disponível em: https://dl.acm.org/doi/10.1145/3571730. Acesso em: 12 maio 2026.

MUSUMECI, E.; BRIENZA, M.; SURIANI, V.; NARDI, D.; BLOISI, D. D. LLM Based Multi-Agent Generation of Semi-structured Documents from Semantic Templates in the Public Administration Domain. In: **INTERNATIONAL CONFERENCE ON HUMAN-COMPUTER INTERACTION**. Cham: Springer, 2024. p. 98–117. Disponível em: https://arxiv.org/abs/2402.14871. Acesso em: 12 maio 2026.

OPENAI. **GPT-4 Technical Report**. 2023. Disponível em: https://cdn.openai.com/papers/gpt-4.pdf. Acesso em: 12 maio 2026.

PERRINA, F.; MARCHIORI, F.; CONTI, M.; VERDE, N. V. AGIR: Automating Cyber Threat Intelligence Reporting with Natural Language Generation. **arXiv preprint**, arXiv:2310.02655, 2023. Disponível em: https://arxiv.org/abs/2310.02655. Acesso em: 12 maio 2026.

TRAN, K.-T. et al. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. **arXiv preprint**, 2025. Disponível em: https://arxiv.org/abs/2501.06322. Acesso em: 12 maio 2026.

VASWANI, A. et al. Attention Is All You Need. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS**. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 12 maio 2026.

WANG, L. et al. A Survey on Large Language Model based Autonomous Agents. **arXiv preprint**, arXiv:2308.11432, 2023. Disponível em: https://arxiv.org/abs/2308.11432. Acesso em: 12 maio 2026.

WU, Q. et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. **arXiv preprint**, arXiv:2308.08155, 2023. Disponível em: https://arxiv.org/abs/2308.08155. Acesso em: 12 maio 2026.

YAN, B. et al. Beyond Self-Talk: A Communication-Centric Survey of LLM-Based Multi-Agent Systems. **arXiv preprint**, 2025. Disponível em: https://arxiv.org/abs/2502.14321. Acesso em: 12 maio 2026.

Apêndice A

Link para o GitHub. contendo os documentos gerados, além dos códigos feitos e os documentos usados para a análise: https://github.com/alexandre96dev/projeto-tcc

Relatório gerado pelo modelo gpt-4o-mini, no estudo de caso 1: https://github.com/alexandre96dev/projeto-tcc/blob/main/resultados_estudo_caso/relatorio_gpt_20251130_153513.md

Relatório gerado pelo modelo llama 3-8b, no estudo de caso 1: https://github.com/alexandre96dev/projeto-tcc/blob/main/resultados_estudo_caso/relatorio_llama_8b_20251130_153354.md

Relatório gerado pelo modelo Llama 3–70B, no estudo de caso 1: https://github.com/alexandre96dev/projeto-tcc/blob/main/resultados_estudo_caso/relatorio_llama_70b_20251130_153433.md

Relatório gerado pelo modelo gpt-4o-mini, no estudo de caso 2: https://github.com/alexandre96dev/projeto-tcc/blob/main/resultados_estudo_caso_2/analise_financeira_chatgpt_4o_mini_20260316_180126.md

Relatório gerado pelo modelo llama 3-8b, no estudo de caso 2:
https://github.com/alexandre96dev/projeto-tcc/blob/main/resultados_estudo_caso_2/analise_financeira_llama_8b_20260316_175948.md

Relatório gerado pelo modelo Llama 3–70B, no estudo de caso 2:
https://github.com/alexandre96dev/projeto-tcc/blob/main/resultados_estudo_caso_2/analise_financeira_llama_70b_20260316_180036.md

Apêndice B - Checklist de avaliação dos relatórios gerados no estudo de Caso 1:

Modelo	Categoria	Critério	Observação
ChatGPT 4o Mini	A. Qualidade da Análise Gramatical	Precisão dos problemas identificados	Erros ('questãos', 'gerassão', 'rezolução') realmente existem no texto, mas deixou passar outros erros não tão aparentes
ChatGPT 4o Mini	A. Qualidade da Análise Gramatical	Localização exata	Trechos citados aparecem literalmente no documento.
ChatGPT 4o Mini	B. Validação das Necessidades de Citação	Identificação correta de afirmações	Não identificou, corretamente, nenhuma necessidade de citação
ChatGPT 4o Mini	B. Validação das Necessidades de Citação	Contexto preservado	Não fez nenhuma indicação de citações que alterasse o contexto
ChatGPT 4o Mini	C. Avaliação de Clareza e Estilo	Identificação de ambiguidades	Conseguiu detectar alguns problemas de fluidez e clareza e fez o diagnóstico correto, mas outros não foram detectados
ChatGPT 4o Mini	C. Avaliação de Clareza e Estilo	Preservação do tom acadêmico	Sugestões mantêm tom formal adequado.
ChatGPT 4o Mini	D. Qualidade Técnica do Output	Estruturação adequada	O relatório segue corretamente a estrutura A–B–C definida na metodologia.
ChatGPT 4o Mini	D. Qualidade Técnica do Output	Ausência de alucinações	Nenhum trecho inexistente foi citado.
Llama 8B	A. Qualidade da Análise Gramatical	Precisão dos problemas identificados	Não conseguiu detectar os erros gramaticais contidos no texto
Llama 8B	A. Qualidade da Análise Gramatical	Localização exata	Não fez nenhuma localização de trechos
Llama 8B	B. Validação das Necessidades de Citação	Identificação correta de afirmações	Não identificou, corretamente, nenhuma necessidade de citação
Llama 8B	B. Validação das Necessidades de Citação	Contexto preservado	Não fez nenhuma indicação de citações que alterasse o contexto
Llama 8B	C. Avaliação de Clareza e Estilo	Identificação de ambiguidades	Trechos com problemas de clareza e fluidez detectados de fato existem, mas não detectou os outros erros e, em um erro detectado, detectou um problema e fez uma sugestão de melhoria que não cabia no trecho detectado. Na outra localização de erro, ele acabou corrigindo automaticamente uma palavra, gramaticalmente errada, de um trecho que tinha problema de clareza, mas fez o diagnóstico correto
Llama 8B	C. Avaliação de Clareza e Estilo	Preservação do tom acadêmico	Sugestões formais, porém uma sugestão fora do trecho analisado.
Llama 8B	D. Qualidade Técnica do Output	Estruturação adequada	O relatório segue corretamente a estrutura A–B–C definida na metodologia.
Llama 8B	D. Qualidade Técnica do Output	Ausência de alucinações	diagnósticos mal feitos na seção C
Llama 70B	A. Qualidade da Análise Gramatical	Precisão dos problemas identificados	Acertou no diagnóstico de um trecho detectado, mas errou em outro. Além de que deixou de detectar trechos com problemas gramaticais
Llama 70B	A. Qualidade da Análise Gramatical	Localização exata	Os trechos localizados nos diagnósticos feitos existem
Llama 70B	B. Validação das Necessidades de Citação	Identificação correta de afirmações	Detectou necessidade de citação, mas não precisa
Llama 70B	B. Validação das Necessidades de Citação	Contexto preservado	Exige citações desnecessárias
Llama 70B	C. Avaliação de Clareza e Estilo	Identificação de ambiguidades	Localizou um trecho e fez o diagnóstico corretamente, mas os outros diagnósticos de clareza detectados foram equivocados. Além de que deixou de detectar outros trechos com problemas
Llama 70B	C. Avaliação de Clareza e Estilo	Preservação do tom acadêmico	Sugestões mantêm formalidade.
Llama 70B	D. Qualidade Técnica do Output	Estruturação adequada	O relatório segue corretamente a estrutura A–B–C definida na metodologia.
Llama 70B	D. Qualidade Técnica do Output	Ausência de alucinações	Análises mal feitas na seção A, B e C

Apêndice C - Checklist de avaliação dos relatórios gerados no estudo de Caso 2:

Modelo	Categoria	Critério	Observação
ChatGPT 4o Mini	A. Fidelidade ao Conteúdo Original	Preservação de informações-chave	Preserva os principais pontos dos balancetes e apresenta boa organização analítica, mas contém erros numéricos nos resultados em junho e agosto, além de pequena imprecisão na margens de junho, agosto e setembro e , o que compromete parcialmente a comparação entre os períodos.
ChatGPT 4o Mini	A. Fidelidade ao Conteúdo Original	Ausência de alucinações	Não inventa categorias, períodos ou estrutura documental, mas apresenta incorreções parciais nos valores consolidados: resultado e margem em junho e agosto, além de pequena imprecisão na margem de setembro.
ChatGPT 4o Mini	B. Qualidade dos Resumos Individuais	Concisão adequada	O relatório apresenta boa concisão, com síntese clara e sem excesso de detalhamento.
ChatGPT 4o Mini	B. Qualidade dos Resumos Individuais	Coerência interna	O relatório apresenta boa coerência discursiva e linguagem cautelosa. A coerência analítica também é assertiva.
ChatGPT 4o Mini	C. Qualidade da Análise Consolidada	Identificação de padrões válidos	Fez uma boa análise de padrões, sendo parcialmente prejudicada por conta das imprecisões numéricas.
ChatGPT 4o Mini	C. Qualidade da Análise Consolidada	Conclusões fundamentadas	As conclusões são bem fundamentadas, mas houve ocasiões onde cometeu equívoco
ChatGPT 4o Mini	D. Estrutura e Organização	Aderência ao formato	Segue adequadamente a estrutura esperada de relatório analítico, com resumo, consolidação, achados e limitações.
ChatGPT 4o Mini	D. Estrutura e Organização	Metadados corretos	Os períodos e categorias estão corretamente representados, mas há inconsistências numéricas no resultado de junho e agosto e nas margens de junho, agosto e setembro.
Llama 8B	A. Fidelidade ao Conteúdo Original	Preservação de informações-chave	Preserva os principais pontos dos balancetes e apresenta boa organização analítica, mas contêm pequena imprecisão nas margens de junho, julho agosto e setembro e erros numéricos nos resultados em junho.
Llama 8B	A. Fidelidade ao Conteúdo Original	Ausência de alucinações	Não inventa categorias nem estrutura documental, mas apresenta margens imprecisas e organização temporal inadequada, além de tendência equivocada de crescimento
Llama 8B	B. Qualidade dos Resumos Individuais	Concisão adequada	O relatório é conciso e direto.
Llama 8B	B. Qualidade dos Resumos Individuais	Coerência interna	O relatório apresenta boa coerência discursiva e linguagem cautelosa, mas a coerência analítica é parcialmente enfraquecida por erros na análise de tendência de crescimento de receita e de resultado.
Llama 8B	C. Qualidade da Análise Consolidada	Identificação de padrões válidos	Identifica parcialmente os padrões dos dados, mas exagera ao tratar oscilações como tendência de crescimento.
Llama 8B	C. Qualidade da Análise Consolidada	Conclusões fundamentadas	A qualidade das conclusões está parcialmente boa, mas algumas contem exagero e inconsistências
Llama 8B	D. Estrutura e Organização	Aderência ao formato	Segue a estrutura geral esperada
Llama 8B	D. Estrutura e Organização	Metadados corretos	Há leves erros numéricos e em margens percentuais e também há equivocos na detecção de tendencias.
Llama 70B	A. Fidelidade ao Conteúdo Original	Preservação de informações-chave	Preserva bem os principais dados dos balancetes e organiza corretamente os períodos, embora acrescente interpretações quantitativas não demonstradas explicitamente, como a estimativa de que salários representam cerca de 30% das despesas totais,além de que contém pequena imprecisão na margens de junho, julho e setembro
Llama 70B	A. Fidelidade ao Conteúdo Original	Ausência de alucinações	Não inventa os dados-base e apresenta boa aderência documental, mas faz uma análise errada de que a rubrica de salários teria peso de aproximadamente 30% das despesas.
Llama 70B	B. Qualidade dos Resumos Individuais	Concisão adequada	O relatório apresenta boa concisão, com síntese clara e estável.
Llama 70B	B. Qualidade dos Resumos Individuais	Coerência interna	O relatório apresenta boa coerência discursiva e linguagem cautelosa. A coerência
Llama 70B	C. Qualidade da Análise Consolidada	Identificação de padrões válidos	Faz análises de padrões boas, mas comete erros, como a atribuição de
Llama 70B	C. Qualidade da Análise Consolidada	Conclusões fundamentadas	As conclusões com base nos dados são boas, mas comete erros.
Llama 70B	D. Estrutura e Organização	Aderência ao formato	Segue adequadamente a estrutura esperada, com boa organização das seções e
Llama 70B	D. Estrutura e Organização	Metadados corretos	Os metadados temporais e a ordenação dos períodos estão corretos; Equivocos