


MODELOS DE LINGUAGEM DE GRANDE ESCALA PARA EXTRAÇÃO DE INFORMAÇÕES ESTRUTURADAS DE DOCUMENTOS FINANCEIROS: UMA AVALIAÇÃO COMPARATIVA¹

 <https://doi.org/10.22533/at.ed.394122608018>

Matheus Souza Rosa

<https://lattes.cnpq.br/4881180920145342>

Fábio Marques da Cruz

¹ Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA).

RESUMO: Este trabalho apresenta um benchmark experimental para avaliação comparativa de Modelos de Linguagem de Grande Porte (LLMs) aplicados à tarefa de extração automática de informações a partir de documentos corporativos em formato PDF. O objetivo principal consiste em identificar qual modelo apresenta maior eficiência e confiabilidade na conversão de informações textuais não estruturadas em uma estrutura de dados padronizada no formato JSON, adequada ao armazenamento em banco de dados. O estudo foi conduzido utilizando um corpus composto por 50 comunicados ao mercado e fatos relevantes emitidos por empresas brasileiras de diferentes setores, incluindo instituições financeiras e empresas do setor de commodities. Para viabilizar a avaliação, foi desenvolvido um pipeline automatizado responsável pela extração do texto dos documentos, envio das informações aos modelos de linguagem avaliados e análise das respostas geradas. A avaliação considerou métricas estruturais e semânticas, incluindo validade do JSON gerado, completude dos campos obrigatórios e similaridade semântica entre o título extraído e o conteúdo do documento. Foram avaliados diferentes modelos disponíveis por meio de APIs com planos gratuitos, incluindo GPT-OSS-120B, DeepSeek-V3.1-671B, Llama-3.3-70B, GLM-4.6, Qwen3-235B-A22B e Gemini-2.5-Flash. Os resultados

demonstram que os modelos GPT-OSS-120B, DeepSeek-V3.1-671B e Gemini-2.5-Flash apresentaram desempenho superior e alta consistência estrutural, com geração confiável de JSON válido em todas as avaliações. O modelo GPT-OSS-120B obteve o maior score geral no benchmark e foi selecionado para a implementação do sistema final de extração automatizada. Os resultados evidenciam a importância de avaliar simultaneamente a qualidade semântica e a consistência estrutural em sistemas baseados em LLMs voltados à automação de pipelines de ingestão de dados.

PALAVRAS-CHAVE: Large Language Models; Extração de Informação; Processamento de Documentos PDF; Benchmark; Inteligência Artificial.

INTRODUÇÃO

A crescente digitalização das comunicações corporativas e financeiras tem resultado em um aumento significativo na quantidade de documentos disponibilizados em formato digital por empresas de capital aberto. Entre esses documentos destacam-se os comunicados ao mercado e os fatos relevantes, que representam instrumentos formais de divulgação de informações capazes de impactar decisões de investidores, analistas e órgãos reguladores. Em geral, tais documentos são disponibilizados ao público em formato PDF, um padrão amplamente utilizado para distribuição de conteúdos digitais devido à sua capacidade de preservar a formatação original do documento independentemente da plataforma utilizada para sua visualização.

Apesar de sua ampla adoção, o formato PDF apresenta limitações importantes quando considerado sob a perspectiva do processamento automatizado de informações. Diferentemente de formatos estruturados, como JSON ou XML, os documentos PDF são concebidos principalmente para leitura humana, o que dificulta sua interpretação por sistemas computacionais. Essa característica torna a tarefa de extração automática de dados particularmente desafiadora, sobretudo em cenários que envolvem documentos com estruturas complexas e heterogêneas.

Conforme destacado por Upadhyay et al. (2025), documentos corporativos frequentemente apresentam informações qualitativas concentradas em trechos de texto narrativo, enquanto dados quantitativos são majoritariamente representados em tabelas e gráficos. Dessa forma, a compreensão completa de um documento exige a capacidade de integrar informações provenientes de diferentes representações estruturais, o que representa um desafio significativo para métodos tradicionais de extração de informação.

O Processamento de Linguagem Natural (PLN) plné uma área da inteligência artificial dedicada ao desenvolvimento de técnicas que permitem que computadores compreendam e processem a linguagem humana. Nesse contexto, o objetivo do

processamento de linguagem natural é permitir que computadores realizem tarefas envolvendo linguagem humana, como compreensão, tradução e geração de texto (JURAFSKY; MARTIN, 2007, p. 1).

Nos últimos anos, os avanços nesse campo têm sido impulsionados pelo desenvolvimento dos chamados Large Language Models (LLMs). Esses modelos são treinados em grandes volumes de dados textuais e apresentam capacidade avançada de compreensão contextual, geração de linguagem natural e interpretação semântica de textos complexos. Em função dessas características, os LLMs têm sido explorados em diversas aplicações relacionadas à análise automatizada de documentos, incluindo sumarização de textos, classificação de conteúdo e extração estruturada de informações.

No contexto da análise de documentos corporativos, os modelos de linguagem de grande escala apresentam potencial significativo para superar algumas das limitações observadas em abordagens tradicionais de extração de informações (BROWN et al., 2020). Ao invés de depender exclusivamente de regras heurísticas ou de estruturas rigidamente definidas, os LLMs podem interpretar o contexto semântico do documento e produzir representações estruturadas das informações identificadas (JURAFSKY; MARTIN, 2007), permitindo a conversão de conteúdos originalmente não estruturados em formatos adequados para análise computacional.

Entretanto, apesar do potencial desses modelos, ainda existem desafios relacionados à consistência estrutural das saídas geradas e à confiabilidade das informações extraídas (ANDERSEN et al., 2025). Diferentes modelos podem apresentar variações significativas na forma como interpretam documentos complexos, o que torna necessário avaliar de maneira sistemática seu desempenho em tarefas específicas de extração de informação.

Nessa vertente, torna-se relevante a realização de estudos comparativos que permitam avaliar o desempenho de diferentes modelos de linguagem em cenários aplicados de processamento de documentos. A identificação de modelos capazes de produzir saídas semanticamente consistentes e estruturalmente padronizadas representa um passo importante para a construção de sistemas automatizados de extração e organização de dados corporativos.

Diante desse cenário, este capítulo descreve a metodologia experimental desenvolvida e aplicada para avaliação comparativa (benchmark) de Large Language Models (LLMs), com foco na tarefa de extração automática de informações a partir de comunicados ao mercado e fatos relevantes emitidos por empresas de capital aberto, disponibilizados em formato PDF. O objetivo central do benchmark é identificar o modelo que apresenta maior eficácia e consistência na conversão de dados textuais e

tabulares, predominantemente não estruturados, em um formato JSON padronizado, confiável e adequado ao armazenamento em banco de dados.

Portanto, a seleção de um modelo com bom desempenho constitui um passo fundamental para a etapa subsequente deste projeto, que consiste no desenvolvimento de um agente automatizado de extração de dados em PDFs. Esse agente será responsável por processar grandes volumes de documentos corporativos e registrar, de forma estruturada, as informações extraídas em um banco de dados, servindo como base para aplicações posteriores relacionadas à análise de dados, monitoramento de eventos corporativos e suporte à tomada de decisão.

Assim, a qualidade do modelo escolhido impacta diretamente a confiabilidade, a escalabilidade e a segurança operacional de todo o pipeline de ingestão de dados. A adoção de um modelo com desempenho consistente é essencial para garantir que as informações extraídas dos documentos sejam armazenadas de forma padronizada e possam ser utilizadas com segurança em aplicações analíticas e sistemas automatizados.

Dessa forma, este capítulo estabelece de maneira objetiva e mensurável os critérios de avaliação, o ambiente experimental e os procedimentos adotados para identificar o modelo de linguagem mais adequado à implementação do agente de extração de dados. A definição de uma metodologia clara e reprodutível permite avaliar o desempenho dos modelos analisados de maneira sistemática, assegurando precisão, padronização e confiabilidade no processo de estruturação e persistência das informações extraídas.

REVISÃO DA LITERATURA

Extração de Informação em Documentos Textuais

A extração automática de informações constitui um dos pilares clássicos do Processamento de Linguagem Natural (PLN). Essa tarefa consiste em identificar, interpretar e converter dados relevantes de textos não estruturados em representações semanticamente organizadas. Conforme preconizado por Jurafsky e Martin (2007), a extração de informações transformam o fluxo informacional bruto em dados estruturados, viabilizando sua integração em sistemas computacionais e bancos de dados.

Historicamente, as metodologias de extração baseavam-se em heurísticas rígidas, regras linguísticas manuais ou modelos estatísticos superficiais. Tais abordagens, contudo, demonstram fragilidade frente à heterogeneidade de documentos corporativos e financeiros, que frequentemente amalgamam narrativas textuais,

tabelas complexas e elementos gráficos (GRISHMAN, 2019). Subjacente a essa limitação, a necessidade de sistemas dotados de maior plasticidade e capacidade de generalização impulsionou a transição para métodos baseados em aprendizado profundo (Deep Learning) e, mais recentemente, em modelos de linguagem de grande escala.

Modelos de Linguagem de Grande Escala

A evolução contemporânea da inteligência artificial culminou no desenvolvimento dos Modelos de Linguagem de Grande Escala (Large Language Models – LLMs). Diferente de arquiteturas predecessoras, os LLMs modernos fundamentam-se majoritariamente na arquitetura Transformer, que introduziu mecanismos de atenção (self-attention) capazes de modelar dependências contextuais de longo alcance (VASWANI et al., 2017). Essa inovação permitiu que os modelos discernissem nuances semânticas em documentos extensos, superando as limitações de recorrência e convolução de modelos anteriores.

Para além da arquitetura densa tradicional, a literatura recente destaca a eficiência de modelos baseados em Mixture of Experts (MoE), técnica que permite escalar o número de parâmetros sem um aumento proporcional no custo computacional de inferência. Essa escalabilidade é crucial para a compreensão de domínios especializados, uma vez que o aprendizado contínuo em vastos volumes de dados permite que o agente identifique padrões intrincados em ambientes inicialmente desconhecidos (RUSSELL; NORVIG, 1995).

Concomitante ao aumento de parâmetros, o refinamento desses modelos possibilitou uma capacidade de generalização que permite a execução de tarefas complexas sem a necessidade de ajustes finos (fine-tuning) extensivos para cada novo domínio.

Avaliação de Desempenho de Modelos de Linguagem

A mensuração da eficácia de um LLM transcende a simples análise de fluência textual. Segundo Jurafsky e Martin (2007), métricas de avaliação robustas são imperativas para determinar a adequação de um modelo a tarefas críticas. No cenário de geração de dados estruturados, a avaliação deve ser bifocal: deve-se aferir tanto a qualidade semântica da resposta quanto sua conformidade estrutural.

A exigência de saídas em formatos rígidos, como objetos JSON (JavaScript Object Notation), introduz um novo desafio métrico. A integridade do sistema automatizado depende de a resposta respeitar estritamente a sintaxe esperada e a completude dos campos obrigatórios. Assim, a literatura sugere a combinação de métricas de

consistência sintática com avaliações de similaridade semântica, assegurando que o modelo não apenas “compreenda” o conteúdo, mas seja capaz de encapsulá-lo em uma moldura estrutural apta para a ingestão direta em pipelines de dados.

Aplicações de Inteligência Artificial na Extração de Dados

A aplicação de inteligência artificial em tarefas de extração de dados tem sido amplamente explorada em diferentes áreas do conhecimento. Estudos recentes investigam o uso de modelos de linguagem para automatizar a análise de documentos científicos, relatórios corporativos e bases textuais extensas.

Um exemplo relevante é apresentado por Andersen et al. (2025), que investigaram o uso de ferramentas de inteligência artificial como revisores auxiliares na etapa de extração de dados em revisões sistemáticas. No estudo, modelos de IA como ChatGPT e Elicit foram avaliados em comparação com dados extraídos por revisores humanos, considerados como padrão-ouro. Os resultados indicaram que os sistemas apresentaram níveis elevados de desempenho, com métricas de precisão próximas de 90%. Entretanto, também foram identificados alguns casos de geração de informações incorretas ou inexistentes no texto original, fenômeno descrito pelos autores como confabulação¹. Esses achados reforçam o potencial das ferramentas de IA para auxiliar na extração de dados, ao mesmo tempo em que evidenciam a necessidade de validação humana em aplicações científicas críticas.

No contexto de documentos corporativos, a extração automatizada de informações apresenta desafios adicionais devido à heterogeneidade estrutural dos arquivos. Relatórios empresariais frequentemente combinam textos descritivos, tabelas e dados quantitativos distribuídos em diferentes seções do documento, dificultando a aplicação de métodos tradicionais de processamento textual.

Nesse cenário, o uso de modelos de linguagem de grande escala surge como uma alternativa promissora para transformar documentos não estruturados em representações estruturadas, permitindo a integração dessas informações em pipelines automatizados de análise de dados.

Síntese da Literatura e Lacuna de Pesquisa

A literatura revisada demonstra avanços significativos no uso de modelos de linguagem para tarefas de análise e extração de informações em documentos textuais. Entretanto, apesar do progresso recente, ainda existem desafios relacionados à confiabilidade e à consistência estrutural das respostas geradas por esses sistemas.

1. No contexto de modelos de linguagem, confabulação refere-se à geração de informações plausíveis, porém incorretas ou não presentes nos dados de origem, fenômeno também conhecido na literatura como hallucination em sistemas de inteligência artificial.

Em particular, a geração de dados estruturados a partir de documentos complexos exige não apenas compreensão semântica do conteúdo, mas também a capacidade de produzir saídas em formatos específicos, como estruturas JSON válidas e completas. A ausência dessa conformidade estrutural pode comprometer a integração dos resultados em sistemas automatizados.

Diante desse contexto, torna-se relevante investigar o desempenho comparativo de diferentes modelos de linguagem na tarefa de extração estruturada de informações a partir de documentos corporativos. Assim, o presente estudo busca avaliar empiricamente o desempenho de diversos modelos de linguagem de grande escala quanto à sua capacidade de gerar respostas semanticamente adequadas e estruturalmente consistentes, contribuindo para o desenvolvimento de pipelines automatizados de extração de dados baseados em inteligência artificial.

METODOLOGIA

Definição da Estrutura de Saída

Primeiramente, foi definido como requisito fundamental que os Modelos de Linguagem avaliados gerem como saída um objeto JSON (JavaScript Object Notation) com estrutura previamente especificada.

A escolha do JSON deve-se à sua simplicidade, legibilidade e ampla compatibilidade com sistemas de armazenamento e pipelines de processamento de dados. A padronização da saída é essencial para garantir consistência entre os resultados e permitir a automatização do registro das informações extraídas em banco de dados. A estrutura JSON definida neste estudo é composta por quatro campos obrigatórios: 'titulo', 'data_publicacao', 'tematica' e 'conteudo'.

O campo 'titulo' corresponde ao título formal do comunicado, que deve ser identificado e extraído com precisão a partir do documento. O campo 'data_publicacao' representa a data de emissão do comunicado, exigindo não apenas sua localização no texto, mas também sua normalização para o formato padrão DD-MM-AAAA. O campo 'tematica' demanda uma análise semântica do conteúdo, com o objetivo de produzir uma descrição concisa do tema central do comunicado, funcionando como uma etapa de inferência e sumarização. Por fim, o campo 'conteudo' deve conter o texto integral do documento, preservando sua completude informacional.

A capacidade do Modelo de Linguagem de preencher essa estrutura JSON de forma correta, consistente e semanticamente adequada, a partir de um PDF de

entrada, constitui o principal critério técnico de avaliação adotado no benchmark apresentado neste capítulo.

Corpus de Dados

Para avaliar sistematicamente o desempenho dos modelos de linguagem (LLMs) na tarefa de extração de informações de documentos PDF, foi desenvolvido um ambiente de benchmark automatizado. Esta arquitetura é composta por dois elementos centrais: um corpus de teste diversificado e um pipeline de software para orquestrar a execução e avaliação dos testes. O corpus utilizado na etapa de benchmark dos modelos foi selecionado cuidadosamente para refletir os desafios da extração de informações não estruturadas e testar a capacidade de generalização dos modelos.

Nesta etapa foram realizados testes com 50 documentos em formato PDF, emitidos por empresas brasileiras de setores variados. Os documentos utilizados na etapa de benchmark foram retirados da base de dados criada anteriormente, composta por comunicados ao mercado e fatos relevantes publicados em anos anteriores. Entre as fontes selecionadas estão instituições financeiras como Bradesco e Itaú, e empresas do setor de commodities como Vale e Petrobras. A escolha desses documentos teve como objetivo garantir uma ampla diversidade de layouts, formatos de tabelas, estilos de formatação e vocabulários específicos de cada setor. Essa heterogeneidade é fundamental para avaliar a robustez dos modelos e sua habilidade de extrair dados de forma precisa, independentemente da estrutura visual do documento.

Arquitetura do Pipeline de Benchmark

Para garantir a execução eficiente e reprodutível dos testes, foi implementado um pipeline automatizado. Esse fluxo de trabalho gerencia desde o processamento inicial dos documentos até a análise final dos resultados, utilizando um conjunto de módulos de software desenvolvidos especificamente para este projeto, conforme ilustrado na Figura 1.

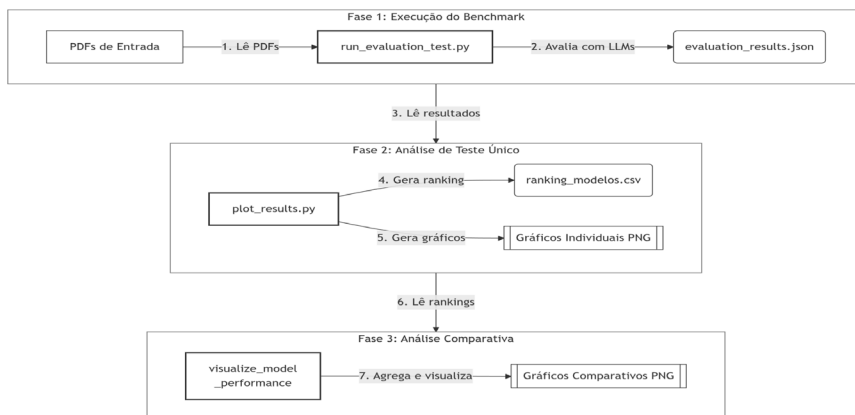


Figura 1- Fluxo do Processo de Ingestão dos PDFs até a Visualização dos Resultados

Fonte: Autoria própria (2025)

Os principais componentes do pipeline de benchmark foram desenvolvidos de forma modular, permitindo a automação completa do processo de extração, avaliação e consolidação dos resultados.

O passo anterior é o módulo de processamento de PDF (`pdf_processor.py`), responsável pela extração do texto bruto dos documentos. Sua implementação adota uma estratégia de fallback com o objetivo de minimizar falhas e interrupções durante o processamento.

Inicialmente, o módulo tenta extrair o conteúdo utilizando a biblioteca `pdfplumber`, reconhecida por sua maior precisão na interpretação de layouts complexos. Caso esse método falhe ou não retorne um texto válido, o sistema recorre secundariamente à biblioteca `pypdf` como alternativa.

Essa abordagem em duas etapas aumenta a confiabilidade do processo de extração, garantindo maior chance de extração de texto mesmo diante de documentos com diferentes estruturas, codificações e estilos de formatação. A orquestração do benchmark é realizada pelo script `run_evaluation_test.py`, que atua como o componente central de controle do fluxo de execução. Esse script é responsável por configurar a lista de modelos avaliados, definir um prompt template padronizado e iterar automaticamente sobre todos os arquivos PDF presentes no diretório de entrada. Para cada documento, o conteúdo textual extraído é submetido de forma idêntica a todos os modelos da lista, garantindo condições de avaliação equivalentes.

A execução do script gera como saída um arquivo estruturado denominado `evaluation_results.json`, que contém as respostas produzidas por cada modelo de linguagem para todos os documentos do corpus, bem como as métricas preliminares associadas a cada execução. Esse arquivo constitui a base de dados utilizada nas etapas posteriores de análise.

Conforme ilustrado na Figura 1, após a conclusão da execução do benchmark (Fase 1), os resultados passam por uma segunda etapa de processamento analítico, denominada Fase 2 - Análise de Teste Único. Nesta etapa, os resultados previamente armazenados são processados pelo script `plot_results.py`, responsável por consolidar os dados de desempenho de cada modelo para um conjunto específico de documentos. O script realiza a leitura do arquivo `evaluation_results.json`, calcula métricas agregadas e gera um arquivo de saída denominado `ranking_models.csv`, no qual os modelos avaliados são ordenados de acordo com suas respectivas pontuações de desempenho.

Além da geração do ranking, o script também produz gráficos individuais de desempenho, exportados em formato PNG. Esses gráficos permitem visualizar de maneira intuitiva o comportamento de cada modelo em relação às métricas definidas, facilitando a identificação de padrões de desempenho, variações entre modelos e possíveis inconsistências nos resultados.

Após a geração dos rankings individuais, os resultados seguem para a Fase 3 - Análise Comparativa, que corresponde à etapa final do pipeline analítico. Nesta fase, o script `visualize_model_performance.py` realiza a agregação dos rankings gerados para diferentes conjuntos de dados ou execuções experimentais. A partir dessa agregação, são produzidas visualizações comparativas que permitem analisar o desempenho relativo dos modelos de linguagem avaliados.

Os gráficos comparativos resultantes possibilitam identificar tendências gerais de desempenho, estabilidade entre diferentes conjuntos de documentos e eventuais discrepâncias entre modelos. Essas visualizações constituem um recurso importante para a interpretação dos resultados do benchmark, uma vez que sintetizam de forma visual as métricas quantitativas obtidas durante os experimentos.

A organização do pipeline em três fases — execução do benchmark, análise de teste único e análise comparativa — permite estruturar o processo experimental de forma modular e reproduzível. Essa abordagem facilita tanto a replicação dos experimentos quanto a extensão futura do sistema para novos modelos de linguagem, novos conjuntos de documentos ou novas métricas de avaliação.

Modelos de Linguagem Avaliados

A seleção dos modelos avaliados neste benchmark concentrou-se em Modelos de Linguagem acessíveis por meio de APIs que oferecem planos gratuitos, permitindo analisar não apenas o desempenho técnico, mas também a viabilidade prática de implementação.

Foram selecionados modelos de diferentes com o objetivo de garantir fornecedores diversidade arquitetural e metodológica no benchmark.

Modelo	Parâmetros	Arquitetura
DeepSeek-V3.1	671B	MoE
Gemini-2.5-Flash	N/D	Proprietário
GLM-4.6	355B	MoE
GPT-OSS-120B	116.8B	MoE
Llama-3.3-70B	70B	Transformer
Qwen3-235B	235B	MoE

Tabela 1 – Modelos Avaliados

Fonte: Autoria própria (2026).

Entre os modelos avaliados está o GPT-OSS:120b, um modelo open-weight disponibilizado pela OpenAI sob licença Apache 2.0, composto por aproximadamente 116,8 bilhões de parâmetros totais, dos quais cerca de 5,1 bilhões são ativados por token² em uma arquitetura Mixture of Experts (MoE)³. O modelo adota estratégias avançadas de eficiência computacional, como quantização dos especialistas MoE para o formato MXFP4⁴, reduzindo significativamente o consumo de memória e viabilizando sua execução em ambientes computacionais mais restritos, além de oferecer suporte nativo à geração de saídas estruturadas e a fluxos agentivos baseados em ferramentas (AGARWAL et al., 2025).

O DeepSeek-V3.1-671b representa um dos modelos open-source de maior escala atualmente disponíveis, contando com aproximadamente 671 bilhões de parâmetros totais e cerca de 37 bilhões de parâmetros ativos por token. Sua arquitetura Mixture of Experts (MoE) incorpora mecanismos como Multi-Head Latent Attention (MLA).

2. Em processamento de linguagem natural, um token corresponde a uma unidade básica de texto utilizada pelo modelo, que pode representar uma palavra, parte de uma palavra, caractere ou símbolo, dependendo do método de tokenização empregado.

3. Mixture of Experts (MoE) é uma arquitetura de redes neurais que utiliza múltiplos submodelos especializados, chamados de "experts", ativados seletivamente durante o processamento.

4. MXFP4 refere-se a um formato de representação numérica de baixa precisão utilizado em computação de alto desempenho e em modelos de inteligência artificial, no qual os valores são armazenados com menor número de bits do que formatos tradicionais de ponto flutuante.

O modelo foi pré-treinado com aproximadamente 14,8 trilhões de tokens e avaliado extensivamente em benchmarks de conhecimento, raciocínio e geração, apresentando desempenho competitivo em relação a modelos proprietários de ponta, o que o torna especialmente relevante para tarefas complexas de extração e compreensão semântica de documentos extensos (DEEPSEEK-AI et al., 2024).

O Gemini-2.5-flash, desenvolvido pela Google e disponibilizado por meio da API Gemini, corresponde à versão otimizada da família Gemini 2.5, focada em alta eficiência, baixa latência e custo computacional reduzido. Apesar de apresentar menor escala em comparação aos modelos de grande porte avaliados, o modelo foi projetado para oferecer respostas rápidas e consistentes, além de suporte multimodal, configurando-se como uma alternativa relevante para cenários de produção que demandam desempenho e escalabilidade (COMANICI et al., 2025).

O GLM-4.6, desenvolvido pela Z.ai (Zhipu AI), adota uma arquitetura Mixture of Experts com cerca de 355 bilhões de parâmetros totais e foi projetado para suportar janelas de contexto extremamente longas, podendo alcançar aproximadamente 200 mil tokens. Essa capacidade é particularmente relevante para o cenário investigado neste trabalho, uma vez que comunicados ao mercado e fatos relevantes frequentemente apresentam textos extensos e estruturas complexas, exigindo do modelo a manutenção de coerência e consistência ao longo de grandes volumes de conteúdo (Z.AI, 2024).

O GPT-OSS-120B é um modelo open-weight disponibilizado pela OpenAI sob licença Apache 2.0, composto por aproximadamente 116,8 bilhões de parâmetros totais, dos quais cerca de 5,1 bilhões são ativados por token em uma arquitetura Mixture of Experts (MoE). O modelo adota estratégias avançadas de eficiência computacional, como quantização dos especialistas MoE para o formato MXFP4, reduzindo significativamente o consumo de memória e viabilizando sua execução em ambientes computacionais mais restritos, além de oferecer suporte nativo à geração de saídas estruturadas e a fluxos agentivos baseados em ferramentas (AGARWAL et al., 2025).

O Llama-3.3-70b-versatile, desenvolvido pela Meta, é uma variante da família LLaMA com aproximadamente 70 bilhões de parâmetros, projetada para oferecer elevada versatilidade em tarefas de compreensão e geração de linguagem natural. O modelo é amplamente adotado na comunidade open-source e disponibilizado por meio de plataformas como o Ollama, destacando-se pelo equilíbrio entre capacidade expressiva e viabilidade computacional, características que o tornam adequado para aplicações de extração de informações e processamento de documentos (META, 2024).

O Qwen3-235b-a22b, da Alibaba, é uma variante de grande escala da família Qwen3, composta por aproximadamente 235 bilhões de parâmetros totais, com

cerca de 22 bilhões ativados por token em sua arquitetura Mixture of Experts (MoE). O modelo foi concebido para oferecer forte desempenho em tarefas de compreensão semântica, geração estruturada e processamento de longos contextos, aspectos diretamente relacionados aos requisitos do benchmark proposto neste estudo (YANG et al., 2025).

Sistema de Métricas e Avaliação

Para aferir a performance dos Modelos de Linguagem na tarefa de extração de dados a partir de documentos PDF, foi desenvolvido um sistema de métricas capaz de avaliar tanto a conformidade estrutural das respostas geradas quanto a qualidade semântica do conteúdo extraído. O conjunto de métricas foi organizado em três categorias principais, métricas de conformidade estrutural, métricas de qualidade semântica e métrica consolidada de desempenho.

As métricas de conformidade estrutural avaliam se a resposta gerada pelo modelo adere corretamente ao formato JSON especificado. A métrica `json_valido` verifica se a saída pode ser interpretada como um objeto JSON válido, enquanto a métrica `campos_completos` avalia a presença dos campos obrigatórios definidos na especificação. Essas métricas são agregadas na métrica `consistencia`, que expressa a qualidade estrutural geral da saída em uma escala contínua.

Métrica	Tipo	Descrição
<code>json_valido</code>	Binária (0 ou 1)	Verifica se a resposta gerada pelo modelo pode ser interpretada corretamente como um objeto JSON válido.
<code>campos_completos</code>	Binária (0 ou 1)	Avalia se todos os campos obrigatórios definidos na especificação do JSON (<code>titulo</code> , <code>data_publicacao</code> , <code>tematica</code> , <code>conteudo</code>) estão presentes na resposta.
<code>consistencia</code>	Contínua (0 até 1)	Métrica agregada que sintetiza o grau de conformidade estrutural da saída, considerando a validade do JSON e a presença dos campos obrigatórios.

Tabela 2 – Métricas de Conformidade Estrutural

Fonte: Autoria própria (2026).

Essa abordagem está alinhada com o princípio defendido por Honovich et al. (2022), segundo o qual avaliações de consistência em sistemas de geração de linguagem devem considerar representações verificáveis e estruturadas. As métricas de qualidade semântica concentram-se na avaliação do campo 'título' em relação ao conteúdo completo do comunicado. Para isso, são gerados embeddings⁵ utilizando o modelo all-MiniLM-L6-v2, permitindo calcular a similaridade de cosseno entre o título e o conteúdo.

Além disso, foram implementadas métricas adicionais relacionadas à clareza e concisão do título. Por fim, para sintetizar os diferentes aspectos avaliados, foi definida uma métrica denominada score geral, permitindo ranquear os modelos de forma objetiva e reproduzível, calculada pela seguinte expressão: $\text{score geral} = (\text{consistencia} \times 0.4) + (\text{relevancia} / 5 \times 0.3) + (\text{clareza} / 5 \times 0.1) + (\text{concisao} / 5 \times 0.2)$

A definição dos pesos utilizados no cálculo do score geral foi baseada na importância relativa de cada métrica para a tarefa de extração estruturada de informações. A métrica consistência estrutural recebeu o maior peso (0,4), uma vez que a correta formatação da saída em JSON e a presença dos campos obrigatórios constituem requisitos fundamentais para a integração automatizada das informações extraídas em bancos de dados e pipelines de processamento. Em cenários de ingestão automática de dados, respostas estruturalmente inválidas tornam-se inutilizáveis independentemente de sua qualidade semântica, justificando sua maior contribuição para a pontuação final.

A métrica relevância semântica recebeu peso 0,3, refletindo a importância do alinhamento entre o título gerado e o conteúdo do documento. Essa métrica avalia a capacidade do modelo de identificar corretamente o tema central do comunicado, aspecto essencial para aplicações de indexação, categorização e recuperação de informações.

As métricas concisão (0,2) e clareza (0,1) receberam pesos menores por apresentarem aspectos relacionados à qualidade comunicacional do título, mas que possuem impacto relativamente menor sobre a funcionalidade do sistema de extração. A concisão contribui para a geração de títulos informativos e diretos, enquanto a clareza está associada à legibilidade e à facilidade de interpretação do conteúdo gerado. Dessa forma, a ponderação adotada busca equilibrar critérios estruturais e semânticos, priorizando a integridade do formato de saída sem negligenciar a qualidade informacional do conteúdo extraído.

5. Embeddings são representações vetoriais de palavras, tokens ou outros elementos de dados em um espaço numérico multidimensional, nas quais itens semanticamente semelhantes tendem a apresentar representações próximas entre si.

RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos a partir da avaliação experimental dos modelos de linguagem selecionados. O objetivo principal dessa etapa foi analisar o desempenho dos modelos na tarefa de extração estruturada de informações a partir de documentos textuais, considerando critérios de qualidade semântica e conformidade estrutural da saída gerada.

Para isso, cada modelo foi submetido ao mesmo conjunto de documentos e instruções de extração, produzindo respostas no formato JSON estruturado, conforme definido na metodologia experimental. As respostas geradas foram então avaliadas utilizando o sistema de métricas descrito anteriormente, que considera aspectos como consistência estrutural, relevância, clareza e concisão das informações extraídas.

Os resultados obtidos foram consolidados a partir dos arquivos de avaliação gerados automaticamente durante a execução do pipeline experimental. Em seguida, scripts de análise foram utilizados para agregar os dados e produzir representações tabulares e gráficas, permitindo uma comparação direta entre os modelos avaliados.

A análise apresentada nesta seção busca identificar não apenas os modelos com melhor desempenho médio, mas também avaliar a estabilidade e confiabilidade estrutural das respostas, fatores essenciais para aplicações práticas que dependem da integração automatizada dos dados extraídos em pipelines de processamento.

Desempenho Geral dos Modelos

Os resultados individuais de cada avaliação foram armazenados em arquivos `evaluation_results.json`, organizados nos respectivos diretórios de saída de cada modelo. Em uma etapa posterior, os scripts `plot_results.py` e `visualize_model_performance.py` foram utilizados para consolidar os dados, gerar rankings comparativos e produzir as visualizações necessárias para a análise.

A Gráfico 1 apresenta o `score_geral` médio obtido por cada modelo para os diferentes conjuntos de dados analisados. A representação gráfica foi utilizada com o objetivo de facilitar a visualização comparativa do desempenho dos modelos entre as diferentes fontes de dados, permitindo identificar de forma mais clara padrões de desempenho e variações entre os resultados.

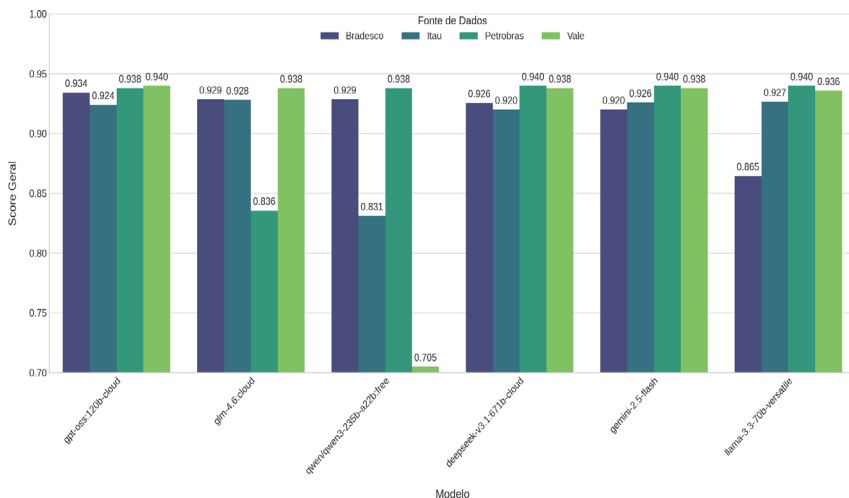


Gráfico 1 – Score Geral por Modelo e Fonte de Dados

Fonte: Autoria própria (2025)

A partir do Gráfico 1, observa-se que os modelos gpt-oss:120b, deepseek-v3.1:671b e gemini-2.5-flash apresentaram os melhores desempenhos globais, com pontuações elevadas e comportamento relativamente estável entre os quatro datasets analisados. O modelo gpt-oss:120b destacou-se ao alcançar o maior score geral total (3,736), mantendo valores consistentes para todas as empresas avaliadas. De forma bastante próxima, os modelos deepseek-v3.1:671b e gemini-2.5-flash obtiveram desempenhos praticamente idênticos, ambos com score geral total de 3,724, consolidando-se como alternativas altamente competitivas.

Esses resultados são consistentes com a literatura recente sobre modelos de linguagem de grande escala, que aponta que modelos com grande número de parâmetros e arquiteturas escaláveis tendem a apresentar melhor capacidade de generalização em diferentes tarefas de processamento de linguagem natural (BOMMASANI et al., 2021). De acordo com os autores, os chamados foundation models possuem maior capacidade de adaptação a diferentes domínios textuais quando comparados a modelos menores ou menos especializados.

Em contrapartida, modelos como qwen3-235b-a22b e glm-4.6 apresentaram maior variabilidade de desempenho entre os datasets. O modelo qwen3-235b-a22b, por exemplo, apresentou quedas expressivas no conjunto de dados da Vale, enquanto o glm-4.6 obteve desempenho inferior no dataset da Petrobras, indicando menor adaptabilidade frente a diferentes layouts e estilos textuais.

A confiabilidade estrutural das respostas foi analisada separadamente por meio do total de erros de consistência, definidos como respostas que não resultaram em um JSON válido ou que apresentaram ausência de campos obrigatórios.

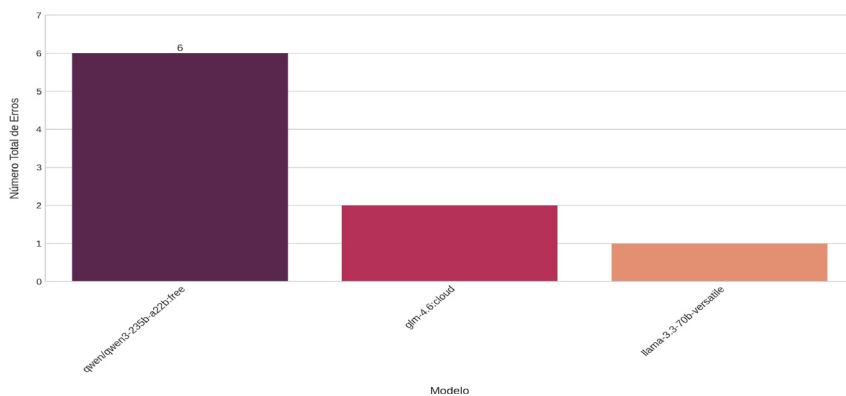


Gráfico 2 – Total de Erros de Consistência por Modelo

Fonte: Autoria própria (2025)

Análise de Confabulação e Fidelidade na Extração de Dados

Além da análise quantitativa apresentada anteriormente, foi realizada uma avaliação qualitativa do comportamento dos modelos com foco na fidelidade documental e no fenômeno de confabulação durante o processo de extração.

Observou-se que a confabulação não se manifestou predominantemente por meio da invenção de dados numéricos, mas sim através de uma reinterpretação contextual de campos descritivos. Modelos como Llama-3.3-70b e DeepSeek-v3.1 apresentaram tendência à “normalização” do conteúdo original. No campo destinado ao conteúdo, o Llama-3.3 frequentemente removeu metadados técnicos e quebras de linha específicas do documento original, convertendo o texto em uma estrutura mais fluida e contínua. Embora essa transformação favoreça a legibilidade, ela caracteriza uma alteração estrutural do documento fonte, o que pode ser indesejável em cenários que exigem preservação literal dos dados extraídos.

No que diz respeito à geração de títulos e temas, o comportamento de confabulação interpretativa foi mais evidente. O modelo DeepSeek-v3.1, em diversos casos, substituiu o título oficial do documento por sínteses derivadas do conteúdo interno, produzindo formulações que não correspondiam exatamente ao título original. Embora essa abordagem demonstre elevada capacidade de compreensão semântica, ela pode introduzir desvios em relação à fonte primária, comprometendo a fidelidade documental necessária em aplicações regulatórias.

Por outro lado, modelos como Gemini-2.5-flash e GPT-OSS-120B apresentaram comportamento mais conservador, priorizando a extração literal do conteúdo e preservando a formatação original, incluindo caracteres especiais e estrutura textual. Esse comportamento refletiu-se em maiores níveis de consistência estrutural e menor incidência de alterações interpretativas no conteúdo extraído.

A análise conjunta dos resultados indica que altos valores de consistência estrutural não garantem, necessariamente, fidelidade documental absoluta. Modelos podem gerar saídas estruturalmente válidas e semanticamente coerentes, mas ainda assim introduzir modificações interpretativas sutis. Dessa forma, a avaliação qualitativa complementa as métricas quantitativas, evidenciando que a escolha do modelo deve considerar não apenas o score geral, mas também o equilíbrio entre capacidade de síntese e preservação literal do conteúdo.

Em aplicações envolvendo documentos regulatórios, a preservação da integridade textual mostrou-se particularmente relevante, uma vez que alterações interpretativas podem comprometer a rastreabilidade e a segurança jurídica dos dados extraídos. Esses resultados reforçam a importância da consistência estrutural e da fidelidade documental como critérios decisivos na seleção do modelo para implementação do agente proposto neste trabalho.

Conclusão e Seleção do Modelo

Observa-se que os modelos gpt-oss:120b, deepseek-v3.1:671b e gemini-2.5-flash não registraram nenhum erro de consistência ao longo de todas as avaliações, alcançando 100% de sucesso na geração de estruturas JSON válidas. O llama-3.3-70b, apesar de apresentar bons valores médios de score_geral, registrou um único erro de consistência, posicionando-se ligeiramente abaixo do grupo líder. Já os modelos glm-4.6 e qwen3-235b-a22b demonstraram menor confiabilidade operacional, acumulando, respectivamente, 3 e 6 erros de consistência.

A análise conjunta dos resultados evidencia que, embora métricas de qualidade semântica sejam fundamentais, a consistência estrutural constitui um critério decisivo para a automação confiável de pipelines de extração de dados. Em aplicações práticas, respostas que não atendem ao formato esperado tornam-se inviáveis para integração automatizada, independentemente da qualidade informacional presente no conteúdo gerado.

Esse resultado está alinhado com estudos recentes que destacam a importância da geração estruturada de dados por modelos de linguagem em aplicações de automação e integração com sistemas computacionais (AGARWAL et al., 2025). Segundo os autores, modelos projetados para oferecer suporte nativo à geração de saídas estruturadas apresentam vantagens significativas em cenários que exigem integração com ferramentas externas e pipelines automatizados de processamento de dados.

Além dos resultados quantitativos, a análise qualitativa apresentada na seção anterior também evidenciou diferenças relevantes entre os modelos no que se refere à fidelidade documental e ao fenômeno de confabulação. Observou-se que alguns modelos, apesar de apresentarem bom desempenho médio, introduziram normalizações e reinterpretações do conteúdo original, alterando elementos estruturais dos documentos extraídos. Em contrapartida, modelos como gpt-oss:120b e gemini-2.5-flash demonstraram comportamento mais conservador, priorizando a preservação literal da informação e mantendo a estrutura textual original.

Nesse contexto, o modelo gpt-oss:120b emergiu como o competidor mais forte do benchmark, combinando o maior score_geral total com uma taxa de 100% de geração de JSONs válidos. Além do desempenho quantitativo, o modelo também apresentou comportamento consistente em termos de fidelidade documental, com menor incidência de confabulação e maior preservação da estrutura original dos textos analisados. Com base nessas evidências, o gpt-oss:120b foi selecionado para a implementação final do agente de extração de dados. A escolha fundamenta-se em seu equilíbrio entre conformidade estrutural, qualidade semântica, fidelidade documental e estabilidade de desempenho entre diferentes fontes de dados, características essenciais para garantir confiabilidade, escalabilidade e segurança operacional.

CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo avaliar o desempenho de diferentes modelos de linguagem de grande escala na tarefa de extração estruturada de informações a partir de documentos textuais, utilizando um pipeline experimental baseado na geração de saídas no formato JSON. Para isso, foi desenvolvido um ambiente de avaliação que permitiu comparar sistematicamente diferentes modelos de linguagem, considerando métricas relacionadas à qualidade semântica das respostas e à consistência estrutural das saídas geradas.

Os experimentos realizados permitiram analisar o comportamento de seis modelos de linguagem distintos, incluindo arquiteturas baseadas em Transformers densos e Mixture of Experts (MoE). A avaliação foi conduzida a partir de múltiplos conjuntos de dados, representando documentos provenientes de diferentes fontes e estilos textuais, o que possibilitou investigar a capacidade dos modelos de generalizar a tarefa de extração de informações em cenários variados.

Os resultados obtidos indicam que modelos com arquiteturas mais recentes e maior capacidade de representação tendem a apresentar melhor desempenho na tarefa analisada. Em particular, o modelo gpt-oss:120b destacou-se ao obter o maior score_geral agregado entre os modelos avaliados, além de apresentar

100% de consistência estrutural na geração de respostas em formato JSON. Esse comportamento demonstra elevada robustez para aplicações que dependem da integração automatizada de dados extraídos em sistemas computacionais.

A análise comparativa também evidenciou que, embora diversos modelos apresentem desempenho competitivo em termos de qualidade semântica, a consistência estrutural das respostas constitui um fator crítico para aplicações práticas de automação. Modelos que geram respostas estruturalmente inválidas podem comprometer pipelines de processamento de dados, mesmo quando as informações extraídas são semanticamente corretas.

Dessa forma, a avaliação realizada neste trabalho contribui para a compreensão do comportamento de modelos de linguagem em tarefas de extração estruturada de dados, destacando a importância de considerar simultaneamente métricas de qualidade informacional e confiabilidade estrutural na escolha de modelos para aplicações reais.

Como principal resultado prático do estudo, o modelo gpt-oss:120b foi selecionado para a implementação do agente automatizado de extração de dados proposto neste trabalho, devido ao seu desempenho superior e à sua elevada estabilidade operacional.

Como perspectivas para trabalhos futuros, destacam-se a ampliação do conjunto de dados utilizado nas avaliações, a inclusão de novos modelos de linguagem que venham a ser disponibilizados na literatura, bem como a investigação de técnicas adicionais de ajuste de prompts e engenharia de contexto para melhorar ainda mais a qualidade das respostas geradas. Além disso, futuras pesquisas podem explorar a integração desses modelos com sistemas de recuperação de informações (RAG) e outras abordagens híbridas que combinem modelos de linguagem com bases de conhecimento externas.

Por fim, espera-se que os resultados apresentados contribuam para o avanço das aplicações de modelos de linguagem em tarefas de processamento automatizado de documentos, oferecendo subsídios para o desenvolvimento de sistemas mais robustos, confiáveis e escaláveis.

REFERÊNCIAS

AGARWAL, S. et al. gpt-oss-120b & gpt-oss-20b Model Card. 2025. arXiv:2508.10925 [cs.CL]. Disponível em: <https://arxiv.org/abs/2508.10925>. Acesso em: 12 dez. 2025.

ANDERSEN, Helms et al. Using Artificial Intelligence Tools as Second Reviewers for Data Extraction in Systematic Reviews. *Cochrane Evidence Synthesis and Methods*, 2025. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cesm.70036>.

BOMMASANI, Rishi et al. On the Opportunities and Risks of Foundation Models. Stanford: Stanford University, 2021. Disponível em: <https://arxiv.org/abs/2108.07258>. Acesso em: 14 dez. 2025.

BROWN, Tom et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems (NeurIPS), 2020. Disponível em: <https://arxiv.org/pdf/2005.14165>.

COMANICI, G. et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. 2025. arXiv:2507.06261 [cs.CL]. Disponível em: <https://arxiv.org/abs/2507.06261>. Acesso em: 12 dez. 2025.

DEEPSEEK-AI; LIU, A.; FENG, B.; XUE, B. et al. DeepSeek-V3 Technical Report. 2024. arXiv:2412.19437 [cs.CL]. Disponível em: <https://arxiv.org/abs/2412.19437>. Acesso em: 12 dez. 2025.

GRISHMAN, Ralph. Twenty-five years of information extraction. *Natural Language Engineering*, Cambridge, v. 25, n. 6, p. 677–692, 2019. Disponível em: <https://doi.org/10.1017/S1351324919000512>. Acesso em: 15 mar. 2026.

HONOVICH, Or; CHOSHEN, Leshem; AHARONI, Roei; NEEMAN, Elad; SZPEKTOR, Idan; GOLDBERG, Yoav. TRUE: Re-evaluating Factual Consistency Evaluation. arXiv preprint, arXiv:2204.04991, 2022. Disponível em: <https://arxiv.org/abs/2204.04991>. Acesso em: 14 dez. 2025.

JURAFSKY, Dan; MARTIN, James H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 3. ed. draft. Stanford: Stanford University, 2007. Disponível em: [https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20\(2007\).pdf](https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20(2007).pdf). Acesso em: 14 dez. 2025.

META. LLaMA3.3-70b. Ollama Library. Disponível em: <https://ollama.com/library/llama3.3:70b>. Acesso em: 12 dez. 2025.

RUSSELL, Stuart; NORVIG, Peter. Artificial Intelligence: A Modern Approach. 4. ed. Pearson, 1995. Disponível em: https://www.academia.edu/download/125698862/artificial_intelligence_modern_approach.9780131038059.25368.pdf.

UPADHYAY, Shivani; ATAHEY, Messiah; MURTAZA, Syed Shariyar; NIE, Yifan; LIN, Jimmy. On the Comprehensibility of Multi-structured Financial Documents using LLMs and Pre-processing Tools. Waterloo: University of Waterloo; Toronto: Manulife, 2025. Disponível em: <https://arxiv.org/pdf/2506.05182>. Acesso em: 14 dez. 2025.

VASWANI, Ashish et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS), 2017. Disponível em: <https://arxiv.org/abs/1706.03762>.

YANG, A. et al. Qwen3 Technical Report. 2025. arXiv:2505.09388 [cs.CL]. Disponível em: <https://arxiv.org/abs/2505.09388>. Acesso em: 12 dez. 2025.

Z.AI. GLM-4.6v Blog Post. 2025. Z.ai Blog. Disponível em: <https://z.ai/blog/glm-4.6v>. Acesso em: 12 dez. 2025.