Capítulo 4

Avaliação de Grandes Modelos de Linguagem

Fundamentos, métodos tradicionais e desafios atuais

Livy Real
André Carvalho
Altigran Soares da Silva

Publicado em: 16/04/2026

4.1 A importância e os desafios intrínsecos da avaliação de LLMs

Talvez você já tenha ouvido falar do caso em que uma LLM (Large Language Model, Grande Modelo de Linguagem) vendeu um carro por 1 dólar[1]. Ou de quando um advogado estado-unidense, ex-advogado do próprio presidente, foi processado por inventar precedentes jurídicos[2]. Se não ouviu, certamente você já ficou preso em conversas intermináveis e inúteis de *chatbots* de diversas empresas. Todos esses casos aconteceram e acontecem todos os dias porque os responsáveis por tais sistemas, mesmo que inadvertidamente, desconsideram a necessidade de uma avaliação robusta de seus sistemas baseados em LLMs. A ampla divulgação comercial desses sistemas pode levar à percepção equivocada de que já atingimos a era da Inteligência Artificial Geral (AGI), o tão esperado momento em que as IAs terão 'inteligência' semelhante à humana e serão autodidatas. Isso ainda não corresponde à realidade. Hoje, em 2025, é possível afirmar que as saídas das LLMs sempre parecem boas, parecem verossímeis e factuais, mas com frequência não o são. Esse capítulo aborda a necessidade da avaliação de LLMs, considerando técnicas tradicionais e técnicas mais recentes, e vai defender a integração da avaliação das LLMs dentro do próprio ciclo de desenvolvimento de um sistema ou produto.

A avaliação de LLMs constitui uma questão fundamental no ciclo de vida de pesquisa e desenvolvimento desses modelos e de aplicações construídas com base neles, incluindo aí os agentes inteligentes. A sua importância transcende a mera verificação de desempenho, sendo crucial para o avanço da área a comparação entre diferentes arquiteturas, a identificação e mitigação de riscos e a garantia de aplicações práticas robustas e seguras (Meva; Kukadiya, 2025). A complexidade inerente à linguagem natural, combinada com a natureza gerativa e as capacidades emergentes dos LLMs, introduz desafios avaliativos que superam as abordagens tradicionais de Processamento de Linguagem Natural (PLN) (Minaee et al., 2025). A falta de interpretabilidade destes modelos ainda contribui à longa lista de novos desafios enfrentados na área atualmente.

LLMs, como modelos de linguagem estatísticos pré-treinados em larga escala, exibem "habilidades" interessantes que vão além da mera previsão de palavras. Incluem-se o aprendizado em contexto (in-context learning), em que os modelos assimilam novas tarefas a partir de poucos exemplos fornecidos no *prompt* , e o seguimento de instruções (*instruction following*), que permite que os modelos executem tarefas sem exemplos explícitos após o *instruction tuning*[3] (Minaee et al., 2025). Estas capacidades, embora revolucionárias, trazem novos desafios à avaliação, pois as métricas clássicas usadas para mensurar sistemas de PLN frequentemente não conseguem capturar a profundidade semântica, a criatividade ou a factualidade das respostas geradas. Como as LLMs são treinadas para a simples tarefa de predizer a próxima palavra de uma sequência, mas, a partir desse treinamento, surgem capacidades emergentes (habilidades dos modelos que não foram explicitamente treinadas), não sabemos de antemão quais habilidades devemos avaliar.

A avaliação do raciocínio passo a passo, por exemplo, é fundamental para compreender e aprimorar a capacidade dos LLMs em problemas complexos, exigindo critérios que vão além da simples precisão da

[1] https://the-decoder.com/people-buy-brand-new-chevrolets-for-1-from-a-chatgpt-chatbot/

[2] https://www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases

[3] *Instruction tuning* é o processo de re-treinar um modelo de linguagem já pré-treinado usando pares (instrução, resposta) para alinhar o modelo a seguir instruções (instruction following) (Zhang et al., 2025c).

resposta final, como factualidade, validade, coerência e utilidade. A necessidade de *frameworks* de avaliação mais sofisticados é cada vez mais evidente, especialmente à medida que os LLMs são integrados em domínios de alta criticidade, como a saúde, na qual segurança e eficácia são imperativas. A avaliação deve, portanto, ser multifacetada, abrangendo não apenas o desempenho técnico, mas também a segurança, a eficiência e o impacto social dos modelos (Bender et al., 2021).

A evolução do paradigma de avaliação é uma resposta direta à forma como os LLMs operam. Sistemas de PLN anteriores frequentemente se propunham a tarefas mais simples e as métricas tradicionais projetadas para tais tarefas frequentemente assumem a existência de uma única resposta correta (Minaee et al., 2025). *Benchmarks* estáticos e métricas rígidas guiaram por décadas o avanço das técnicas de PLN, mesmo que as limitações dessas metodologias fossem conhecidas e frequentemente apontadas (Bowman; Dahl, 2021; Kalouli et al., 2023; Ribeiro et al., 2020). No entanto, a natureza gerativa dos LLMs, que podem produzir diversas saídas válidas para a mesma entrada, torna o problema da avaliação mais evidente, na medida em que a própria definição de uma resposta "boa" ou "correta" se torna uma definição fluida e multidimensional (Meva; Kukadiya, 2025; Minaee et al., 2025). Esta observação implica uma transição: saímos de métricas puramente objetivas e baseadas em correspondência de strings e vamos em direção a abordagens de avaliação mais diversificadas, buscando diferentes relações semânticas e alinhadas com a percepção humana. O desafio não é apenas medir a correção, mas sim a qualidade em um espaço subjetivo, criativo e dependente do contexto, o que demanda uma evolução contínua dos *frameworks* de avaliação.

Adicionalmente, a avaliação não é um processo *post-hoc*, mas uma parte integral do ciclo de desenvolvimento dos LLMs. O aprendizado por reforço (Reinforcement Learning, RL) utiliza pontuações de avaliadores humanos para aprimorar o treinamento de um LLM. Isso estabelece um vínculo causal: métricas de avaliação falhas ou incompletas, especialmente aquelas que não se alinham com os valores humanos ou com as capacidades emergentes dos modelos, podem levar à otimização dos sistemas para a métrica em si, em vez de levá-los a novas versões realmente úteis e seguras. Assim, a qualidade e o alinhamento das pontuações de avaliação impactam diretamente a trajetória de aprendizado do modelo e seu desempenho subsequente, especialmente em esforços de alinhamento como o RLHF (Reinforcement Learning from Human Feedback).

4.1.1 Tipos de avaliação

A avaliação de LLMs pode ser categorizada de diversas maneiras para abordar os múltiplos aspectos de seu desempenho. Uma distinção fundamental é entre avaliação *intrínseca* e *extrínseca*[4]. A avaliação intrínseca foca na qualidade inerente do modelo, frequentemente medida por métricas que avaliam a probabilidade de sequências de texto, como a perplexidade, ou a sobreposição lexical com textos de referência. A avaliação extrínseca, por outro lado, mede o desempenho do LLM em uma tarefa finalística ou *downstream*[5] ou em um cenário de aplicação real, como a capacidade de um LLM de melhorar a experiência do usuário em um *chatbot* ou sua eficácia em tarefas de sumarização de notícias.

Outra categorização crucial diferencia a *avaliação humana* da *automática*. A avaliação automática tradicional emprega algoritmos e métricas computacionais para comparar a saída do modelo com referências ou para inferir a qualidade com base em características textuais. Este tipo inclui métricas estatísticas tradicionais (como BLEU e ROUGE). Atualmente, também há a avaliação automática baseada em modelos de linguagem, como a utilização do BERTScore e o paradigma LLM-as-a-Judge) (Gao et al., 2024). Vamos visitá-las nas próximas páginas.

Já a avaliação humana é amplamente considerada o "padrão ouro" especialmente para tarefas de Geração de Linguagem Natural (NLG) devido à sua capacidade de capturar nuances e critérios de qualidade subjetivos que as métricas automáticas não conseguem capturar (Dierk et al., 2025; Gao et al., 2024; Wang et al., 2025).

Há uma complementaridade e compromissos (*trade-offs*) inerentes entre a avaliação humana e a automática. Embora a avaliação humana seja o "padrão ouro", ela apresenta desafios significativos em termos de custo e escalabilidade, sendo um processo caro e demorado (Dierk et al., 2025; Gao et al., 2024). Em contraste, as métricas automáticas são eficientes e rápidas de calcular, mas muitas vezes carecem da capacidade de

[4] A Seção Paradigmas: Tipos de avaliação deste livro traz outras formas de categorizar possíveis avaliações de sistemas de PLN. Veja as Seções Avaliação intrínseca e Avaliação extrínseca para entender a aplicação de avaliação intrínseca e extrínseca a *datasets* específicos.

[5] Uma tarefa *downstream* é uma tarefa 'mais para o fim do fluxo'. Considere que um LLM é treinado para prever a próxima palavra (essa seria a *upstream task*, no início do fluxo). Os usos reais, análise de sentimento, classificação textual, agente conversacional, são as *downstream tasks*, as tarefas reais nas quais aquele modelo será empregado (Schnabel et al., 2015).

capturar a profundidade semântica e a coerência contextual que os humanos podem discernir (Minaee et al., 2025). Esta observação implica que um *framework* de avaliação holístico deve alavancar os pontos fortes de ambas as abordagens, em vez de depender exclusivamente de uma.

A combinação de verificações automáticas rápidas para qualidade básica e consistência com revisões humanas direcionadas para aspectos subjetivos, segurança e raciocínio complexo sugere um futuro no qual as abordagens de avaliação híbridas se tornarão o padrão. O trade-off entre custo/velocidade e profundidade/nuance possivelmente impulsionará as escolhas metodológicas na implantação prática de LLMs. A avaliação humana é essencial para validar textos gerados por máquinas e garantir o alinhamento com requisitos específicos do domínio e expectativas do usuário. A dificuldade intrínseca da avaliação de tarefas gerativas significa que a avaliação humana é sempre necessária. Além de não existir uma "resposta única correta" para muitas tarefas de geração de linguagem (Gao et al., 2024), é apenas a avaliação humana que garante o alinhamento de LLMs ou de qualquer outro tipo de avaliação (meta-avaliação) a valores humanos. Por valores humanos, entende-se não apenas ética e responsabilidade social, mas justamente a qualidade da máquina em agir exatamente como um humano agiria (e não apenas 'parecido').

4.2 Avaliações tradicionais

Em PLN, as tarefas podem ser divididas, de maneira geral, em dois grandes grupos: classificação e geração. Tradicionalmente, cada um desses tipos de tarefas é avaliado com uma abordagem específica, utilizando métricas distintas.

Nas tarefas de classificação, o sistema atribui rótulos (classes) a entradas de texto, como na análise de sentimentos ou na categorização de um documento. A avaliação costuma ser feita por métricas como acurácia, precisão, revocação e F1-score, que medem o quanto as previsões do modelo coincidem com as respostas esperadas. Tais métricas têm mais relação com classes, distribuição de instâncias nas classes do que com linguagem propriamente dita.

Já nas tarefas de geração, o modelo produz novas sequências de texto, como em tradução automática, sumarização ou resposta a perguntas. Nesse caso, a avaliação é mais complexa, pois envolve tanto a fluência quanto a adequação do texto, recorrendo a métricas automáticas como BLEU, ROUGE e METEOR, além de julgamentos humanos de qualidade. A maioria das métricas automáticas de geração tende a avaliar a corretude sintática e/ou semântica do texto gerado em relação a um texto de referência.

Neste capítulo, não revisitaremos em detalhe o funcionamento de cada uma dessas métricas, já discutidas com maior profundidade na Seção Métricas: Medindo a performance. Para ver mais sobre tarefas de geração de linguagem, você também pode consultar o Capítulo Geração de linguagem natural.

Cabe destacar, no entanto, que as métricas tradicionais de PLN, embora adequadas para modelos anteriores e tarefas específicas, mostram-se insuficientes diante das capacidades emergentes dos LLMs. Em estágios iniciais da tecnologia, a avaliação costumava concentrar-se em camadas superficiais do texto, como correção lexical ou gramaticalidade. Com os LLMs, que já atingem alta precisão nesses aspectos, torna-se necessário avaliar dimensões mais complexas da linguagem, como consistência factual, completude e utilidade em contextos específicos. Assim, o próprio objeto de avaliação se transforma, exigindo métricas novas e mais abrangentes, capazes de capturar a natureza multifacetada da geração de linguagem em larga escala.

4.2.1 Descompasso com a percepção humana

Um dos principais desafios das métricas tradicionais é o seu descompasso com a percepção humana de qualidade. Métricas baseadas em sobreposição de palavras, como BLEU e ROUGE, penalizam severamente paráfrases e sinônimos, mesmo que semanticamente equivalentes à referência (Mathur et al., 2020; Peyrard, 2019). Por exemplo, um texto gerado por um LLM que utiliza um vocabulário diferente, mas que transmite o mesmo significado que o texto de referência, pode receber uma pontuação artificialmente baixa nessas métricas, apesar de ser considerado de alta qualidade por um avaliador humano (Mathur et al., 2020; Minaee et al., 2025). O BLEU, por exemplo, ignora a sequência de palavras, resultando na mesma pontuação para frases com as mesmas palavras, mas com significados diferentes devido à ordem (Mathur et al., 2020). Ainda que esses sejam problemas que a área de geração de linguagem sempre tenha enfrentado (Gatt; Krahmer, 2018), eles se acentuam com o advento das LLMs.

As métricas tradicionais também não foram desenhadas para capturar aspectos complexos da linguagem, como fluência ou criatividade (Gao et al., 2024). Elas são sensíveis a pequenas variações lexicais que não afetam a qualidade percebida pelo ser humano (Mathur et al., 2020). A otimização exclusiva para essas métricas pode levar a modelos que "enganam" a métrica em vez de realmente melhorar a qualidade da

geração de texto (Gao et al., 2024). Isso demonstra que as métricas tradicionais, ao focar na estrutura de superfície (*surface structure* (Chomsky, 1965)), falham em capturar a equivalência semântica e podem levar a conclusões enganosas sobre o desempenho real do sistema (Mathur et al., 2020; Minaee et al., 2025).

4.2.2 Dificuldade em avaliar aspectos complexos

Além do descompasso com a percepção humana, as métricas tradicionais demonstram uma dificuldade intrínseca em avaliar aspectos mais complexos e emergentes dos LLMs.

Factualidade e Alucinações

LLMs são propensas a "alucinações", elas geram informações plausíveis, textos estatisticamente possíveis, mas factualmente incorretos, irrelevantes ou sem sentido (Meva; Kukadiya, 2025). Por exemplo, um LLM pode afirmar que "Thomas Edison inventou a internet" ou fornecer detalhes clínicos incorretos sobre um paciente específico. Tais erros não são detectados por métricas tradicionais, pois o texto pode ser gramaticalmente correto e fluente (Gao et al., 2024). A factualidade, especialmente em cenários de raciocínio passo a passo, exige que a informação possa ser fundamentada em fontes confiáveis ou que utilize corretamente números e restrições fornecidos na consulta. Verificar o conhecimento factual e o conhecimento de senso comum parametrizados, isto é, o conhecimento real e factual que a LLM adquiriu através de seu treinamento, permanece um desafio em aberto.

Raciocínio e Senso Comum

Tanto o processamento da informação de forma que pareça um raciocínio passo-a-passo quanto a capacidade de armazenar o chamado 'senso comum' são das mais relevantes e impressionantes capacidades emergentes que as LLMs apresentam. Por serem 'novidade', as métricas clássicas são incapazes de medir a capacidade de inferência, raciocínio complexo ou compreensão profunda dos LLMs (Meva; Kukadiya, 2025). A avaliação do raciocínio passo a passo é crucial para entender e melhorar o desempenho dos LLMs em problemas complexos como lógica, matemática e ciência. As métricas tradicionais não foram desenhadas para medir esses aspectos cruciais do processo de raciocínio. LLMs podem gerar saídas que soam plausíveis, mas carecem de coerência lógica, especialmente quando se baseiam em correlações estatísticas em vez de raciocínio causal (Xiong et al., 2025).

Segurança e Viés (*Bias*)

As métricas tradicionais também não foram projetadas para detectar conteúdo tóxico, preconceituoso ou prejudicial, a menos que sejam explicitamente treinadas para isso (Meva; Kukadiya, 2025). A avaliação da segurança e do viés requer abordagens específicas que analisem a imparcialidade, a inofensividade e a recusa de respostas inadequadas (Jiao et al., 2024). A otimização para métricas de sobreposição lexical não garante que o modelo não produzirá conteúdo problemático.

Seguimento de instruções

A capacidade de um LLM de seguir instruções complexas ou com múltiplas restrições é um aspecto crucial para sua utilidade (Minaee et al., 2025). Por esta ser uma nova capacidade da tecnologia, métricas tradicionais não a avaliam. É especialmente difícil avaliar o quão bem o modelo adere a essas instruções quando estas envolvem nuances ou múltiplos passos (Gao et al., 2024). LLMs podem falhar em compreender contextos longos ou instruções em conversas multiturno, quando as instruções são inconsistentes ou contêm informações incorretas. Métricas recentes que avaliam sistemas de diálogo podem avaliar a adequação de uma resposta ao contexto do turno anterior ou ainda a progressão do diálogo (Yeh et al., 2021), características que se assemelham ao *instruction following*, porém, tais métricas são também mensuradas através de modelos treinados para isso, assemelhando-se ao novo paradigma de avaliação que vamos discutir ao longo do capítulo, o *LLM as a Judge*.

4.2.3 Exemplos práticos

Para ilustrar as limitações das métricas tradicionais, vamos visitar o trabalho de (Li et al., 2025) que compara diferentes tipos de avaliação em uma tarefa específica do domínio do direito.

Os autores propõem o CaseGen, um conjunto de dados para avaliação de geração de documentos de casos jurídicos na área penal chinesa. O *benchmark* tem 500 casos reais, anotados por especialistas do

direito. Os autores avaliam 9 diferentes LLMs para gerar parte desses casos. Para avaliar como essas LLMs foram, especialistas em direito avaliaram detidamente 50 casos. Esses 50 casos anotados foram comparados com diferentes métricas, como ROUGE-L, BertScore e o paradigma LLM-as-a-Judge (uma LLM avaliando outra LLM). Esse modelo dá notas de 0 a 10 para cada documento gerado considerando critérios como precisão dos fatos, lógica e correção legal.

Veja na Tabela abaixo a comparação entre a anotação humana e três métricas de avaliação usando diferentes medidas de calcular a correlação entre anotação. As correlações de Kendall, Pearson e Spearman são métricas de correlação entre avaliações automáticas e avaliações humanas frequentemente usadas quando precisamos medir a proximidade das avaliações (Shaqiri et al., 2023).

Tabela 4.1: Comparação entre correlações das métricas automáticas e avaliações humanas.

Metrics	LLM Score	Rouge-L	BERTScore
Kendall	**0.667**	0.333	0.166
Pearson	**0.726**	0.264	0.239
Spearman	**0.750**	0.375	0.250

O interessante para a gente é que as notas automáticas se aproximam bastante das avaliações de juristas humanos (correlação de até 0,75 quando compara-se a anotação humana à anotação do LLM juíz), muito melhor do que as métricas tradicionais, que ficaram abaixo de 0,4.

Esse exemplo demonstra que, embora as métricas tradicionais sejam fáceis de computar e úteis para certas comparações de superfície, elas não são necessariamente alinhadas às anotações humanas, exigindo a complementação com abordagens mais robustas e, muitas vezes, híbridas. É ainda relevante notar que a avaliação com LLM chega a 0.75 de correlação, mas não chega a 0.8 nem a 1. Por isso, a avaliação humana ainda é **tecnicamente** a mais confiável.

4.2.4 Avaliação humana: O padrão ouro?

Como vimos, a avaliação humana é amplamente reconhecida como o "padrão ouro" para a avaliação de LLMs, especialmente para tarefas de Geração de Linguagem Natural (NLG), onde a subjetividade e a nuance são inerentes (Dierk et al., 2025; Gao et al., 2024).

4.2.5 Importância e metodologias comuns

A avaliação humana é crucial porque é capaz de capturar aspectos de qualidade que as métricas automáticas não conseguem. Isso inclui a fluência, coerência, relevância semântica profunda, criatividade, factualidade e o seguimento de instruções complexas (Gao et al., 2024). A capacidade dos avaliadores humanos de compreender o contexto, interpretar significados implícitos e julgar a adequação de uma resposta para um determinado propósito a torna indispensável (Gao et al., 2024).

Metodologias comuns de avaliação humana incluem:

- **Escalas Likert** – Utilizadas para avaliar a qualidade em diversas dimensões, como fluência, relevância, coerência e adequação. Os avaliadores atribuem uma pontuação em uma escala graduada (e.g., de 1 a 5 ou 1 a 100) para cada critério. Esta abordagem permite quantificar julgamentos subjetivos e pode ser aprimorada com a coleta de explicações em linguagem natural dos avaliadores para capturar nuances e calibrar as pontuações.
- ***Ranking*** – Envolve a comparação direta entre diferentes saídas de modelos para uma mesma entrada. Os avaliadores classificam as respostas em ordem de preferência, indicando qual é superior ou se há um empate (Vongthongsri, 2025). Essa metodologia pode ser mais estável do que a pontuação absoluta, pois foca na comparação relativa (Gao et al., 2024), porém ela nem sempre é reutilizável. Se, em um primeiro momento, analistas comparam dois modelos e, finalmente, temos um terceiro modelo em jogo, a primeira avaliação dificilmente será útil para decidirmos algo sobre este novo modelo.
- **Testes A/B** – Uma forma de ranking em que os avaliadores escolhem a preferência entre duas opções (A ou B) (Río; Vaahtio, 2024). É útil para comparar o desempenho de dois modelos ou variações de um mesmo modelo, como diferentes *prompts* , por exemplo. Geralmente testes A/B são feitos com grupos diferentes de usuários que não sabem que estão participando de um teste. Quando uma avaliação é feita através de teste A/B é importante considerar tanto a relevância estatística das

amostras quanto como coletar o *feedback* dos grupos antes de implementar o teste. Rodar um teste sem ter uma hipótese, sem saber coletar e interpretar o *feedback* pode ser inócuo.

- **Análise de erros** – Consiste em classificar os tipos de erros presentes nas saídas dos LLMs (e.g., erros gramaticais, factuais, de coerência, alucinações, omissões). Esta metodologia fornece avaliação detalhada que pode ser usada para depurar e melhorar os modelos. A coleta de anotações de erros em respostas de LLMs é desafiadora devido à natureza subjetiva de muitas tarefas de PLN e à dificuldade em obter anotações objetivas para erros complexos. Veja a Subseção Análise de Erro deste livro para saber mais sobre análise de erro.

4.2.6 Desafios da avaliação humana

Apesar de ser o "padrão ouro", a avaliação humana apresenta uma série de desafios que limitam sua escalabilidade e consistência.

Custo e Tempo

O processo de avaliação humana é notoriamente caro e demorado (Dierk et al., 2025). A obtenção de anotações de qualidade pode levar dias ou semanas, o que introduz latência no ciclo de *feedback* de desenvolvimento do modelo. Isso torna impraticável realizar uma avaliação humana completa para cada ajuste ou iteração do modelo (Gao et al., 2024).

Subjetividade e Variabilidade

Diferentes avaliadores podem ter opiniões distintas sobre a qualidade de uma mesma saída, levando a variabilidade nos julgamentos (Dierk et al., 2025; Jiao et al., 2024; Liu et al., 2024). Para algumas tarefas, como a avaliação da "criatividade" de uma história, até mesmo humanos podem discordar sobre o que é "melhor" (Gao et al., 2024). A subjetividade pode ser mitigada com diretrizes claras e perguntas mais objetivas.

Viés do avaliador

As experiências, expectativas e origens culturais dos avaliadores podem influenciar seus julgamentos, introduzindo vieses (Jiao et al., 2024). Por exemplo, um avaliador de uma cultura específica pode ter uma percepção diferente do que é "polido" em comparação com outro. É crucial ter um grupo de avaliadores diversificado para mitigar esses vieses. Além disso, a forma como a tarefa é apresentada aos avaliadores pode influenciar os resultados (e.g., "Modelo A foi ajustado para ser mais conciso. Qual é melhor?" pode induzir a um viés em favor da concisão) (Gao et al., 2024). Um viés conhecido é o viés da concordância ("acquiescence bias") (Cronbach, 1942), que caracteriza a tendência do avaliador em concordar com afirmações ou responder de forma afirmativa, por exemplo, concordar com as etiquetas e as saídas dos sistemas automáticos quando exposto a eles.

Escalabilidade

É difícil aplicar a avaliação humana em larga escala e de forma contínua, especialmente com o rápido avanço e o grande volume de saídas dos LLMs (Dierk et al., 2025). A avaliação humana é frequentemente reservada para comparações finais ou verificações periódicas, enquanto métricas automáticas são frequentemente usadas durante o desenvolvimento dos sistemas (Gao et al., 2024).

Treinamento e Calibração de avaliadores

Para garantir consistência e alta qualidade nas anotações, os avaliadores precisam de diretrizes claras e treinamento rigoroso (Gao et al., 2024). A calibração é essencial para garantir que diferentes avaliadores apliquem os critérios de forma consistente (Liu et al., 2024). A falta de controle de qualidade pode levar a anotações de baixa qualidade, exigindo a inclusão de perguntas de teste para filtrar trabalhos de baixa qualidade (Gao et al., 2024). Veja as Seções Procedimentos e estratégias de anotação e revisão e Concordância entre-anotadores deste livro para mais sobre estratégias de anotação e concordância entre anotadores.

Em suma, embora a avaliação humana seja insubstituível para capturar a qualidade genuína dos LLMs em tarefas complexas, seus desafios inerentes impulsionam a pesquisa em métodos de avaliação automática

mais sofisticados e em abordagens híbridas que combinam o melhor dos dois mundos. Além disso, a maioria das avaliações por humano tem sido feita com *crowdworkers* (Hedderich; Oulasvirta, 2025), trabalhadores terceirizados, geralmente leigos e (mal) pagos por quantidade de anotação feita. Essa prática impacta diretamente a qualidade da avaliação, além de levantar questões éticas preocupantes (Gonzalez-Cabello et al., 2025; Jiang; Wagner, 2024) tanto socialmente quanto tecnicamente.

4.3 *Benchmarks* e *datasets* de avaliação

Benchmarks, no contexto da avaliação de LLMs, são conjuntos de dados e tarefas padronizadas que servem como referência para comparar o desempenho de diferentes modelos. Eles são cruciais para o avanço da pesquisa, permitindo uma avaliação sistemática e reprodutível das capacidades dos LLMs (Aisera, 2024; Minaee et al., 2025). Atualmente, temos muitos *benchmarks* disponíveis, especialmente para o inglês. Visitaremos apenas alguns para familiarizar o leitor com esses recursos.

4.3.1 *Benchmarks* genéricos (General-Purpose)

GLUE e SuperGLUE

GLUE (General Language Understanding Evaluation) e SuperGLUE (Super General Language Understanding Evaluation) são *benchmarks* históricos que impulsionaram a pesquisa em PLN. Enquanto GLUE focava em tarefas de compreensão de linguagem mais básicas, como inferência em linguagem natural (NLI), SuperGLUE foi desenvolvido para introduzir desafios mais complexos, exigindo raciocínio linguístico avançado, compreensão contextual, recuperação de conhecimento e aprendizado multi-tarefa (Wang et al., 2018). SuperGLUE inclui tarefas como resolução de correferência e question answering, refletindo problemas mais próximos do raciocínio humano (Wang et al., 2019b). No entanto, o GLUE original enfrentou o problema de saturação, onde o desempenho dos modelos se aproximava do nível humano, limitando sua capacidade de diferenciar modelos de ponta (Wang et al., 2019b). SuperGLUE buscou superar essa limitação, oferecendo um desafio mais rigoroso e com maior "espaço" para melhorias dos modelos.

HELM

HELM (Holistic Evaluation of Language Models) representa uma abordagem mais abrangente para a avaliação de LLMs. Diferente de *benchmarks* que se concentram apenas na acurácia, HELM avalia múltiplos aspectos do desempenho do modelo, incluindo acurácia, calibração, robustez, justiça (fairness) e eficiência (Aisera, 2024). Essa perspectiva multifacetada é essencial para compreender o comportamento dos LLMs em cenários diversos e complexos, especialmente em aplicações de alto risco (Meva; Kukadiya, 2025).

MMLU

MMLU (Massive Multitask Language Understanding) é um *benchmark* projetado para avaliar o conhecimento de LLMs em uma vasta gama de domínios. Ele abrange 57 áreas diversas, incluindo matemática, humanidades, ciências sociais e STEM (Ciência, Tecnologia, Engenharia e Matemática) (Aisera, 2024; Meva; Kukadiya, 2025).

4.3.2 *Benchmarks* para tarefas específicas

Question Answering (QA)

Aqui trataremos de *benchmarks* para avaliação de QA em larga escala. O Capítulo Perguntas e Respostas trata com bastante atenção a tarefa.

Para tarefas de Question Answering, *benchmarks* como SQuAD (Stanford Question Answering Dataset), *Natural Questions* e TriviaQA são amplamente utilizados (Houamegni; Gedikli, 2025).

O SQuAD é um *benchmark* fundamental, composto por mais de 100.000 pares de perguntas e respostas, onde as respostas são trechos de passagens fornecidas. SQuAD 2.0 introduziu perguntas sem resposta, exigindo que os modelos identificassem quando uma pergunta não podia ser respondida com o contexto dado (Rajpurkar et al., 2016).

Já o *Natural Questions* contém consultas geradas por usuários reais do Google Search, espelhando cenários de busca do mundo real onde as respostas podem não estar explicitamente em um único parágrafo (Kwiatkowski et al., 2019).

Por fim, o TriviaQA é um *dataset* mais desafiador que SQuAD, com 950.000 pares de perguntas e respostas de documentos da Wikipedia e da web, frequentemente exigindo raciocínio multi-hop e recuperação de informações entre documentos (Joshi et al., 2017).

Para atender à necessidade de avaliação de modelos em múltiplos idiomas, foram criadas extensões específicas para os principais *benchmarks* de Q&A. Essas versões mantêm o formato do *dataset* original, mas utilizam textos e perguntas em outras línguas.

- Para o SQuAD: Existem *datasets* como o XQuAD, que é uma tradução humana profissional do SQuAD v1.1 para 10 idiomas, incluindo árabe, grego, tailandês e turco. Outro exemplo é o MLQA (Multilingual Question Answering), que oferece um *benchmark* de extração de respostas em 7 idiomas, como espanhol, hindi e chinês.
- Para o *Natural Questions*: A principal extensão é o MKQA (Multilingual Knowledge Questions and Answers). Ele utiliza as perguntas do Natural Questions e as apresenta em 26 idiomas. As respostas devem ser encontradas nas versões da Wikipedia correspondentes a cada um desses idiomas.
- Para o TriviaQA: Foi desenvolvido o Mling-TriviaQA, que adapta o formato desafiador do original para 10 idiomas diferentes. As perguntas e os documentos de evidência foram coletados ou traduzidos para cada uma das línguas incluídas no *benchmark*.

Sumarização

Para a sumarização de texto, *datasets* como CNN/DailyMail[6] e XSum (Narayan et al., 2018) são referências relevantes. Aqui exploraremos apenas esses *datasets*, mas você pode consultar o Capítulo Sumarização Automática para se aprofundar sobre a tarefa.

O CNN/DailyMail é um *corpus* amplamente utilizado para sumarização de notícias, contendo aproximadamente 300 mil notícias escritas em inglês, acompanhadas dos destaques de cada uma.

O XSum (Extreme Summarization Dataset) é focado em sumarização abstrativa, com resumos mais concisos e frequentemente reescritos, em vez de apenas extraídos do texto original.

Raciocínio

A avaliação das capacidades de raciocínio[7] de uma LLM também pode ser o foco de um *benchmark*.

GSM8K (Grade School Math 8K)

Um *dataset* de 8.500 problemas matemáticos de nível fundamental, linguisticamente diversos, que exigem raciocínio multi-passos e operações aritméticas básicas (II, 2025; Said, 2025).

Big-Bench Hard (BBH)

Um subconjunto de tarefas desafiadoras do BIG-bench, projetado para testar as aptidões do modelo além dos *benchmarks* convencionais de PLN, incluindo raciocínio de alto nível (Meva; Kukadiya, 2025; Said, 2025).

4.3.3 *Benchmarks* para segurança e confiabilidade

A segurança e a confiabilidade emergem como dimensões centrais na avaliação contemporânea de LLMs, especialmente em aplicações de alto impacto social, como os domínios jurídico ou médico. Diferentemente dos *benchmarks* voltados apenas à performance ou ao 'raciocínio', esses conjuntos de avaliação buscam medir a robustez ética e factual dos modelos, isto é, sua capacidade de evitar a produção de conteúdo falso, enviesado ou potencialmente nocivo.

[6] https://huggingface.co/datasets/abisee/cnn_dailymail

[7] Usamos o termo raciocínio de forma operacional, em linha com a literatura recente sobre LLMs. Não se deve entender que tais modelos realizem processos cognitivos análogos ao raciocínio humano, o termo aqui refere-se à capacidade de manipular representações linguísticas e numéricas por meio de correlações estatísticas complexas aprendidas durante o treinamento dos LLMs.

Esses *benchmarks* abordam desafios que extrapolam a precisão técnica, como a tendência dos modelos a reproduzirem vieses presentes nos dados de treinamento, a gerarem informações incorretas com aparência de verdade ou a responderem de forma inadequada a *prompts* sensíveis. Assim, a avaliação de segurança e confiabilidade opera em um duplo eixo: veracidade factual e responsabilidade social.

Nos últimos anos, a literatura tem consolidado uma série de *benchmarks* voltados especificamente para essas dimensões, que hoje se tornaram indispensáveis para aferir se um LLM é apto a operar de forma segura e ética. Entre os principais, destacam-se o TruthfulQA, voltado à verificação da veracidade das respostas; o BBQ (*Bias Benchmark for Question Answering*), que mede o grau de viés social reproduzido por modelos; e o ToxiGen e o RealToxicityPrompts, que avaliam a propensão de um sistema a gerar conteúdo tóxico, ofensivo ou discriminatório.

TruthfulQA

Avalia a veracidade das respostas dos modelos, medindo sua resistência à geração de informações falsas ou equívocos comuns (Aisera, 2024; Lin et al., 2022).

BBQ (*Bias Benchmark for Question Answering*)

Mede o viés social em modelos de Question Answering, ajudando a identificar e mitigar preconceitos (Jan et al., 2025; Sánchez, 2024).

ToxiGen/RealToxicityPrompts

Avaliam a propensão dos modelos a gerar texto tóxico ou prejudicial, sendo fundamentais para garantir a segurança do conteúdo gerado (Gehman et al., 2020; Jan et al., 2025).

4.3.4 A importância de ir além do placar

Embora os *benchmarks* e os placares (leaderboards) forneçam uma maneira padronizada de comparar modelos, é crucial ir além da simples pontuação. A saturação de *benchmarks*, nos quais os modelos atingem ou superam o desempenho humano, pode mascarar limitações reais e a falta de generalização da aplicação dos modelos em cenários do mundo real (Meva; Kukadiya, 2025; Wang et al., 2019b). Desafios como a sensibilidade a pequenas mudanças nos *prompts* (*prompt* sensitivity) e a estocasticidade na geração de texto introduzem variabilidade nos resultados, o que torna a reprodutibilidade dos resultados de placares e *benchmarks* para modelos baseados em LLMs um problema mais complexo do que na avaliação de sistemas clássicos (Meva; Kukadiya, 2025). Além disso, a contaminação de dados (*benchmark leakage*), em que dados de avaliação são inadvertidamente incluídos no treinamento, pode inflar artificialmente as métricas de desempenho (Wang et al., 2019b). Portanto, uma análise qualitativa aprofundada e a consideração de múltiplos *benchmarks* são essenciais para uma avaliação holística e precisa.

4.4 Avaliação baseada em modelos (*Model-based evaluation*)

Nessa seção, trataremos do novo paradigma LLM-as-a-Judge, das vantagens, de possíveis abordagens para utilizar esse paradigma, bem como de suas desvantagens e pontos de atenção a serem considerados.

A ideia de utilizar um LLM para avaliar as saídas de outro LLM, conhecida como "LLM-as-a-Judge", emergiu como uma alternativa promissora à avaliação humana, especialmente para textos abertos e complexos (Li et al., 2024; Vongthongsri, 2025). Essa abordagem envolve instruir um LLM (geralmente mais potente ou especificamente treinado para a tarefa de avaliação) a pontuar ou criticar a saída de outro modelo com base em critérios definidos (Li et al., 2024).

4.4.1 Vantagens

Escalabilidade e custo potencialmente menor

A avaliação humana é um processo notoriamente caro e demorado, o que a torna impraticável para avaliações contínuas ou em larga escala. LLMs como avaliadores oferecem uma solução mais escalável e potencialmente mais econômica para essa limitação (Li et al., 2024; Vongthongsri, 2025).

Capacidade de fornecer *feedback* em linguagem natural

Diferente das métricas automáticas tradicionais que fornecem apenas pontuações numéricas, os LLMs avaliadores podem gerar *feedback* detalhado em linguagem natural, explicando o porquê de uma determinada pontuação ou crítica. Isso pode ser útil para depurar e melhorar os modelos, se utilizado dentro de um fluxo híbrido, no qual esses *feedbacks* sejam realmente consumidos, porém o simples fato de uma LLM gerar um feedback, uma explicação sobre sua resposta, não garante que a resposta esteja correta.

Potencial para avaliar aspectos mais abstratos

LLMs avaliadores têm o potencial de avaliar aspectos mais abstratos da qualidade do texto, como coerência, relevância, tom e criatividade, que são difíceis de capturar com métricas clássicas (Vongthongsri, 2025).

Avaliação de interações dinâmicas e contextuais

Uma vantagem crucial é a capacidade de avaliar sistemas de IA da mesma forma que são utilizados. Por exemplo, em vez de depender de *datasets* estáticos, que são inadequados para medir a qualidade de um diálogo, um LLM avaliador pode analisar uma conversa em tempo real. Ele pode avaliar a interação passo a passo ou o diálogo completo, proporcionando uma avaliação muito mais fiel à experiência real do usuário em um *chatbot* ou assistente virtual[8].

4.4.2 Abordagens

Avaliação baseada em referência vs. sem referência

Algumas abordagens de LLM-as-a-Judge podem operar com ou sem textos de referência. Métricas baseadas em *embeddings* de LLMs, por exemplo, frequentemente não exigem referências, calculando a similaridade semântica diretamente entre a saída do modelo e a consulta do usuário ou o contexto de uso. No entanto, a avaliação sem referência pode ser mais suscetível a falhas e vieses.

Prompting para avaliação

A criação de *prompts* eficazes é fundamental. Esses *prompts* instruem o LLM avaliador sobre os critérios de avaliação, o formato da saída (e.g., escalas Likert, comparações binárias, análise de erros) e o texto a ser avaliado (Li et al., 2024). A qualidade do *prompt* pode influenciar drasticamente a avaliação.

Fine-tuning de LLMs para tarefas de avaliação

Além do *prompting*, LLMs podem ser especializados (fine-tuned) especificamente para tarefas de avaliação, treinando-os para prever pontuações humanas ou para identificar tipos específicos de erros.

LLM-as-a-Jury (LLM como um júri)

Para aumentar a robustez e mitigar os vieses de um único avaliador, esta abordagem utiliza um painel de múltiplos LLMs para julgar uma mesma saída. A tarefa é enviada a vários "LLMs juízes", que podem ser modelos diferentes ou instâncias do mesmo modelo com configurações distintas. As avaliações individuais são então agregadas (por exemplo, através de uma votação majoritária ou pela média das pontuações) para chegar a um veredito de consenso. Essa técnica visa replicar a dinâmica de um painel de especialistas humanos, resultando em uma avaliação final mais estável e confiável.

Avaliação por decomposição de tarefas

Esta abordagem quebra a avaliação de uma tarefa complexa em subtarefas mais simples e verificáveis. Por exemplo, para avaliar a qualidade de um resumo, em vez de julgá-lo diretamente, o processo poderis ser: 1) Gerar um conjunto de perguntas e respostas com base no documento original (fonte). 2) Pedir a um LLM que responda às mesmas perguntas usando apenas o resumo gerado. 3) A avaliação final, comparar as

[8] Atualmente, também têm sido propostos *frameworks* dinâmicos para a análise de agentes baseados em LLMs, como o clembench (Chalamalasetti et al., 2023), cujo diferencial é ser uma avaliação interativa diretamente motivada pela Teoria do Diálogo, para avaliar como fenômenos do diálogo humano são ou não reproduzidos pelos LLMs.

respostas obtidas a partir do resumo com as respostas obtidas a partir do documento original, medindo assim a fidelidade e a retenção de informação do resumo.

4.4.3 Desafios e Limitações

Apesar das vantagens, a avaliação baseada em modelos apresenta desafios significativos.

Viés do modelo avaliador

O LLM avaliador pode herdar ou amplificar vieses presentes em seus dados de treinamento, ou até mesmo exibir um "viés narcisista", favorecendo respostas geradas por si mesmo (Panickssery et al., 2024) ou por modelos da mesma família (Vongthongsri, 2025). Estudos mostram que avaliadores generativos tendem a atribuir pontuações mais altas ao conteúdo gerado pelo mesmo modelo subjacente.

Sensibilidade ao *prompt*

Pequenas alterações na formulação do *prompt* de avaliação podem levar a diferenças substanciais nos resultados, tornando a avaliação inconsistente e não robusta (Zhao et al., 2025). Para lidar com essa questão, um método chamado PromptEval (Polo et al., 2024) foi proposto. A abordagem reconhece que avaliações baseadas em um número limitado de *prompts* podem não capturar as reais habilidades dos LLMs, levando a classificações inconsistentes. Em vez de depender de um único "*prompt*", o PromptEval estima a distribuição de desempenho do modelo através de um grande conjunto de variações de "*prompts*". Isso permite a criação de métricas de avaliação mais robustas, como a performance mediana ou em quantis específicos (por exemplo, o desempenho nos 5% piores *prompts* ou nos 5% melhores), fornecendo um panorama mais completo e confiável da capacidade do modelo, que é menos suscetível a alterações em um "*prompt*" individual.

Concordância com humanos

Embora alguns estudos sugiram que LLMs podem atingir alta concordância com julgamentos humanos (até 85% para GPT-4, por exemplo) (Bencke et al., 2024; Kim et al., 2025; Vongthongsri, 2025), outros indicam que a acurácia da avaliação por LLMs não é garantida em todos os cenários (Pombal et al., 2025; Wang et al., 2024). A correlação pode variar dependendo da tarefa e do modelo avaliador. No entanto, a discussão é mais complexa, pois replicar o julgamento humano pode significar replicar também seus vieses. Um estudo sobre o tema ilustra essa questão, mostrando que, embora juízes humanos demonstrem menos viés de gênero que os LLMs, ambos são igualmente vulneráveis ao viés de autoridade (favorecendo respostas com referências falsas), ao "viés de beleza" (preferindo textos mais bem formatados) e à dificuldade em detectar desinformação (Chen et al., 2024a). Isso demonstra que o verdadeiro desafio não é apenas alcançar a concordância, mas desenvolver métodos de avaliação que superem as falhas inerentes tanto a humanos quanto a máquinas.

Custo computacional

Embora potencialmente mais barato que a avaliação humana em larga escala, o uso de LLMs potentes como juízes ainda pode incorrer em custos computacionais significativos, especialmente para avaliações contínuas (Meva; Kukadiya, 2025), avaliações em que uma única instância é avaliada usando diferentes critérios ou no uso de LLM-as-a-Jury, o uso de diferentes modelos para avaliar o mesmo sistema.

Janela de contexto finita

Os modelos de linguagem possuem um limite na quantidade de texto (contexto) que conseguem processar de uma só vez. Essa restrição se torna um grande obstáculo ao avaliar tarefas que envolvem textos longos, como a sumarização de documentos extensos ou a análise de um histórico de diálogo completo. A necessidade de truncar ou dividir o texto para que ele caiba na janela de contexto pode levar à perda de informações e prejudicar a capacidade do LLM de realizar uma avaliação holística e precisa.

Não determinismo

As pontuações de LLMs juízes podem ser não determinísticas, ou seja, a mesma saída pode receber pontuações diferentes em avaliações repetidas, a menos que sejam aplicadas técnicas para garantir (algum) determinismo (Vongthongsri, 2025).

Viés de posição e verbosidade

LLMs avaliadores frequentemente exibem um "viés de posição", preferindo a primeira saída, ou eventualmente a segunda saída, em comparações pareadas, e um "viés de verbosidade", preferindo textos mais longos, mesmo que não sejam necessariamente de maior qualidade (Chen et al., 2024a).

Viés de leniência (*Leniency bias*)

Este fenômeno é caracterizado pela propensão dos modelos em avaliar respostas como "corretas" em situações de incerteza, ou quando seus critérios de avaliação não estão perfeitamente alinhados com as instruções fornecidas. Um estudo de 2024 quantifica essa tendência ao estimar a probabilidade de um juiz marcar uma resposta como correta quando, na verdade, ele não está aplicando os critérios de avaliação corretamente (Thakur et al., 2025). Os resultados demonstram que a maioria dos modelos juízes exibe um viés de leniência consideravelmente alto, indicando que, na dúvida, eles tendem a julgar positivamente. Essa propensão é mais pronunciada em modelos menores.

Outros vieses

Vários outros tipos de vieses relacionados foram identificados mais recentemente (Kamruzzaman et al., 2024; Park et al., 2024) como *Length Bias* (viés de extensão), *Concreteness Bias* (a tendência de atribuir maior credibilidade a respostas que contêm detalhes específicos, como citações, números e terminologias complexas), *Empty Reference Bias* (o modelo prefere conteúdo alucinado que apenas parece estar associado à instrução original), o Content Continuation Bias (o modelo favorece respostas que continuam o texto de entrada em vez de seguir a instrução principal) e o *Nested Instruction Bias* (tendência a responder a perguntas aninhadas dentro do texto, ignorando a instrução principal).

4.4.4 Exemplos de *frameworks* de *LLM as a Judge*

G-Eval

Um método que gera uma série de passos de avaliação a partir de critérios definidos, utilizando um LLM para preencher um formulário e determinar a pontuação final (Vongthongsri, 2025).

Prometheus

Uma suíte de LLMs avaliadores de código aberto, otimizados para avaliação multilíngue, que podem fornecer avaliação direta e comparação pareada (Kim et al., 2023).

AutoEval

Abordagens que visam automatizar a avaliação, como a investigação do uso de LLMs como juízes para questões abertas legais, mostrando alta correlação com pontuações humanas em exames padronizados (Samuylova, 2025).

Avaliação jurídica baseada em QA

Este método avalia resumos jurídicos usando um LLM em um processo de Pergunta e Resposta (QA). O LLM primeiro cria perguntas com base em um resumo de referência e depois usa o resumo candidato para respondê-las. Finalmente, o modelo compara as duas respostas para gerar uma pontuação, focando na qualidade da argumentação legal em vez de apenas na sobreposição de palavras (Xu; Ashley, 2023).

DeepEval

É um *framework* de código aberto para avaliação de LLMs[9], frequentemente descrito como o "Pytest para LLMs", por aplicar uma abordagem de testes unitários à avaliação de saídas de modelos de linguagem. Ele oferece uma suíte com mais de 14 métricas, como G-Eval, RAGAS, relevância da resposta e detecção de alucinações. O DeepEval se integra a outras ferramentas do ecossistema, como LangChain e LlamaIndex, e é complementado pela plataforma em nuvem Confident AI[10], que adiciona funcionalidades de gerenciamento de *datasets* e detecção de regressão nos testes.

4.5 Aspectos emergentes e avançados da avaliação

Como já comentado, com as LLMs, surgem também novos desafios para entender como eles realmente funcionam e se comportam. Já não basta medir apenas a precisão ou a fluência das respostas, avaliar se os modelos são robustos, justos, confiáveis e eficientes também torna-se necessário. Nesta seção, exploramos essas novas dimensões da avaliação, trazendo, brevemente, exemplos de possíveis formas de detecção e mensuração desses desafios.

4.5.1 Robustez e Ataques adversariais

A robustez de um LLM refere-se à sua capacidade de manter o desempenho e a segurança diante de entradas ruidosas, com erros ortográficos, ou maliciosamente construídas (ataques adversariais). A avaliação da robustez é crucial para a implantação segura e responsável de LLMs em domínios de alto risco, onde erros podem ter consequências graves. Ataques adversariais podem incluir:

- Perturbações em nível de caractere: Alterações sutis nos *prompts* que podem induzir o modelo a comportamentos indesejados (Zhao et al., 2025).
- *Jailbreak prompts*: *Prompts* deliberadamente elaborados para contornar os mecanismos de segurança dos LLMs e induzi-los a gerar conteúdo prejudicial. Novos *frameworks* estão sendo desenvolvidos para avaliar autonomamente a robustez dos LLMs, utilizando *prompts* adversariais refinados e diretrizes de conhecimento restritas ao domínio, muitas vezes na forma de grafos de conhecimento.

4.5.2 Avaliação de justiça, viés e toxicidade

A avaliação de justiça (fairness), viés (bias) e toxicidade é um campo em crescimento, impulsionado pela necessidade de garantir que os LLMs não perpetuem ou amplifiquem preconceitos presentes em seus dados de treinamento (Japkowicz; Boukouvalas, 2024; Jiao et al., 2024).

A identificação e a quantificação de vieses em LLMs têm-se tornado uma frente essencial da pesquisa contemporânea. Diversas ferramentas e técnicas vêm sendo desenvolvidas para identificar e mensurar diferentes tipos de vieses, como de gênero, raça ou culturais, presentes nas saídas desses modelos (Jan et al., 2025; Japkowicz; Boukouvalas, 2024; Kamruzzaman et al., 2024). *Benchmarks* específicos, como o BBQ (*Bias Benchmark for Question Answering*) e o ToxiGen, têm sido utilizados para avaliar tanto a reprodução de estereótipos sociais quanto a propensão dos modelos a gerar discurso de ódio (Gehman et al., 2020; Jan et al., 2025).

A própria definição de viés é um desafio. Ela varia conforme o contexto cultural e exige cuidado para distinguir entre padrões legítimos do mundo real e estereótipos prejudiciais (Meva; Kukadiya, 2025). Para enfrentar essa complexidade, novas abordagens vêm sendo propostas, como o *framework de auto-coerência*, no qual os próprios modelos analisam seus padrões de saída em busca de inconsistências e vieses implícitos (Japkowicz; Boukouvalas, 2024). Ainda assim, a mitigação efetiva desses problemas requer monitoramento contínuo e o uso de técnicas de debiasing, voltadas à correção progressiva das distorções detectadas (Sánchez, 2024).

Uma estratégia frequentemente explorada é o uso de *datasets* sintéticos para tentar corrigir vieses complexos. No entanto, essa solução pode introduzir novos riscos metodológicos: em vez de eliminar o viés original, o modelo avaliador pode aprender a reproduzir os mesmos padrões ou artefatos (artifacts) gerados pelo modelo que criou o *dataset* sintético. O resultado é a introdução de um novo tipo de viés, que compromete a validade da avaliação e revela uma limitação importante das atuais estratégias de mitigação baseadas em dados sintéticos (Chen et al., 2024a).

[9] https://github.com/confident-ai/deepeval
[10] https://www.confident-ai.com/

4.5.3 Factualidade, alucinações e citabilidade

A factualidade e a prevenção de "alucinações" são desafios críticos para os LLMs (Meva; Kukadiya, 2025). Métodos para verificar a veracidade da informação gerada incluem:

Detecção de alucinações

Técnicas como FactSelfCheck permitem a detecção de alucinações em nível de fato, analisando a consistência factual entre múltiplas respostas do LLM (Sawczyn et al., 2025).

Citabilidade e Atribuição de fontes

A capacidade de um LLM de citar corretamente suas fontes é fundamental para a confiabilidade. Métodos como LLM-Cite exploram a capacidade do LLM de gerar URLs de citação potenciais para uma determinada afirmação, que são então verificadas em tempo real (Joshi et al., 2025). No entanto, a atribuição de fontes ainda enfrenta desafios, especialmente em domínios específicos, como o jurídico (Zhang et al., 2025b).

Avaliação de fundamentação (*Groundedness*) via NLI

Essa abordagem avalia se uma afirmação gerada por um LLM é factualmente suportada por uma fonte específica, usando a Inferência de Linguagem Natural (NLI) para classificar a relação como corroboração (acarretamento) ou contradição. O trabalho de (Trautmann et al., 2024) mostra que, embora métodos híbridos complexos (como combinar o uso de LLM-as-a-Judge com métodos de NLI) sejam os mais precisos, são também muito caros e lentos. O trabalho destaca que uma abordagem mais simples e eficiente, como perguntar diretamente a um LLM potente (e.g. GPT-4) se a fonte suporta a afirmação, já oferece uma performance robusta, servindo como um forte *baseline*.

4.5.4 Eficiência computacional e custos

A eficiência computacional é uma dimensão cada vez mais importante na avaliação de LLMs, dada a sua crescente escala e os custos associados à inferência e implantação (Galileo AI, 2024; Meva; Kukadiya, 2025).

Métricas

As métricas de eficiência incluem:

1. Latência: O tempo entre a submissão de um *prompt* e o recebimento da resposta completa (Google Cloud, 2025).
2. Throughput: O volume de requisições ou *tokens* que o sistema pode processar por unidade de tempo (Google Cloud, 2025).
3. Consumo de Energia: A energia gasta durante a inferência, um fator crucial para a sustentabilidade (Meva; Kukadiya, 2025);
4. Custo de Inferência: O custo monetário associado à execução do modelo (Meva; Kukadiya, 2025).

A otimização dessas métricas é vital para a viabilidade de aplicações práticas, especialmente aquelas que exigem respostas em tempo real ou operam em larga escala (Galileo AI, 2024; Google Cloud, 2025). Técnicas como test-time compute (TTC) e otimizações de arquitetura visam melhorar a eficiência sem comprometer a acurácia (Meva; Kukadiya, 2025).

4.6 Considerações éticas e o Futuro da avaliação de LLMs

À medida que os LLMs passam a ocupar papéis centrais em produtos, pesquisas e até decisões governamentais, a avaliação deixa de ser apenas uma questão técnica e se torna também uma questão ética e social. Avaliar modelos implica também em discutir responsabilidade, transparência e possíveis impactos negativos de seu uso. Nesta seção, exploramos como a avaliação deve incorporar princípios de responsabilidade, interpretabilidade e regulação, preparando o caminho para práticas mais seguras e sustentáveis no futuro da inteligência artificial.

Aqui trataremos especificamente da questão de uso de LLMs para a avaliação, porém você pode ver a Seção Uso Responsável e Boas Práticas e o Capítulo Questões éticas em IA e PLN deste livro para mais discussões sobre o uso responsável de IA.

4.6.1 Responsabilidade na avaliação

A implantação generalizada de LLMs levanta questões críticas sobre a responsabilidade quando esses modelos falham ou causam danos. Robustos *frameworks* de avaliação são essenciais para verificar se um LLM atende a padrões de confiabilidade, ética e desempenho em aplicações reais. A avaliação deve, portanto, ser uma parte integral do ciclo de vida do desenvolvimento do LLM, e não apenas uma análise *post-hoc* (Meva; Kukadiya, 2025).

4.6.2 Transparência e Interpretabilidade

A "caixa preta" dos LLMs, em que as decisões e as saídas são difíceis de interpretar, é um desafio significativo. A necessidade de entender por que um modelo toma certas decisões ou gera certas saídas é fundamental para a confiança e a responsabilidade social (Sánchez, 2024). A transparência contextual, por exemplo, pode envolver a inclusão de disclaimers sobre a natureza probabilística das respostas ou a indicação das fontes utilizadas.

4.6.3 Dualidade do uso

A capacidade dos LLMs de gerar conteúdo complexo e convincente também levanta preocupações sobre seu uso malicioso. A "dualidade do uso" refere-se ao potencial de modelos, originalmente projetados para fins benéficos, serem manipulados para gerar conteúdo prejudicial, como discurso de ódio, golpes ou até mesmo código malicioso para ataques cibernéticos. A avaliação deve, portanto, incluir a análise da suscetibilidade dos modelos a tais manipulações e o desenvolvimento de estratégias de mitigação de seu uso malicioso.

4.6.4 O papel da regulação e de padrões na avaliação

A rápida evolução dos LLMs exige o desenvolvimento de padrões e regulamentações para guiar sua avaliação e implantação responsáveis. Há uma necessidade crescente de protocolos de avaliação mais padronizados e transparentes para permitir comparações significativas e rastrear o progresso ao longo do tempo (Meva; Kukadiya, 2025; National Institute of Standards and Technology, 2024). A privacidade dos dados, a reprodutibilidade dos resultados e a capacidade de realizar avaliações sem utilizar recursos de terceiros (*on-premise*) são considerações importantes em contextos industriais ou governamentais (National Institute of Standards and Technology, 2024).

4.6.5 Direções futuras

O futuro da avaliação de LLMs é definido tanto por oportunidades de avanço quanto por desafios fundamentais que permanecem não resolvidos. As direções prioritárias incluem:

Estabelecimento de meta-avaliação

(Evaluating the Evaluators) Talvez a direção mais crítica seja a investigação sobre a própria validade das métricas e dos *frameworks* de avaliação. A questão de "quando uma avaliação é boa o suficiente?" ainda carece de uma resposta definitiva. A pesquisa futura deve se concentrar no desenvolvimento de métodos rigorosos para validar os próprios sistemas de avaliação, garantindo que suas conclusões sejam robustas, replicáveis e verdadeiramente indicativas do desempenho do modelo.

Avaliação contínua e em tempo real

Existe uma lacuna significativa entre a avaliação em ambientes de pesquisa e as necessidades da indústria (Heller, 2024). A implementação de *frameworks* que permitam a avaliação contínua e em tempo real é uma necessidade urgente para monitorar o desempenho de modelos em produção, detectar degradações (*drifts*) e garantir a segurança. Superar os obstáculos técnicos e de custo para viabilizar essa monitorização é um dos principais focos para a aplicação prática e responsável de LLMs (Meva; Kukadiya, 2025).

Avaliação adaptativa e personalizada

Métricas e abordagens que se adaptam a diferentes domínios, casos de uso e preferências do usuário (Meva; Kukadiya, 2025; PrototypeJam, 2025).

Desenvolvimento de métricas alinhadas com valores humanos

O objetivo de alinhar LLMs com "valores humanos" é mais complexo do que simplesmente buscar concordância com avaliadores humanos. A pesquisa deve abordar a distinção filosófica entre mimetizar as preferências e os vieses de um grupo de anotadores e alinhar-se a princípios éticos mais amplos e universais. Isso envolve não apenas o desenvolvimento de métricas mais sofisticadas, mas também uma reflexão sobre quais valores devem ser codificados e como (Meva; Kukadiya, 2025).

Avaliação de impacto social a longo prazo

Uma avaliação ampla, especialmente tratando-se do uso de IA em ambientes governamentais, deve considerar o impacto social e ético dos LLMs em um horizonte de tempo mais longo (Meva; Kukadiya, 2025; National Institute of Standards and Technology, 2024).

Abordagem multidisciplinar

A complexidade dos LLMs exige uma abordagem que transcende a ciência da computação. A avaliação deve ser inerentemente multidisciplinar, integrando insights da linguística, psicologia, ética e ciências sociais para criar uma compreensão holística do desempenho técnico dos modelos e do seu impacto social a longo prazo. Isso inclui o desenvolvimento de avaliações adaptativas, que se ajustem a diferentes contextos culturais, domínios de aplicação e casos de uso específicos (Jiao et al., 2024; Meva; Kukadiya, 2025).

4.7 Considerações finais

Neste capítulo, abordamos a avaliação de Grandes Modelos de Linguagem (Large Language Models), um campo dinâmico e multifacetado, que exige ir além das abordagens convencionais da linguística computacional e do processamento de linguagem natural. O avanço desses modelos trouxe novos desafios metodológicos, éticos e epistemológicos, forçando a revisão de pressupostos sobre o que significa avaliar o desempenho em sistemas que produzem linguagem de forma aberta e contextual.

As métricas automáticas tradicionais continuam a oferecer um ponto de partida valioso, sobretudo pela rapidez e pela comparabilidade que proporcionam. No entanto, sua limitação em capturar aspectos semânticos e, especialmente, pragmáticos torna indispensável a incorporação de camadas adicionais de análise. A avaliação humana, ainda que cara e limitada em escala, permanece o padrão-ouro por sua capacidade de julgar nuances, contextos e valores universais e éticos.

Nesse novo cenário, o uso de modelos como avaliadores, o paradigma LLM-as-a-Judge, representa uma via promissora. Ao unir a eficiência da automação com a profundidade interpretativa dos julgamentos humanos, essas abordagens híbridas apontam para um futuro de avaliações mais contínuas, contextualizadas e adaptativas. No entanto, desenvolver um LLM-juiz justo e eficiente demanda tempo, meta-avaliação e supervisão humana constante. Além disso, o LLM-juiz não é garantia de alinhamento da avaliação a valores humanos, ele é apenas um método 'mais alinhado' do que métodos tradicionais.

O desafio atual é consolidar *frameworks* eficientes, robustos e rápidos capazes de equilibrar essas três dimensões: automação, julgamento humano e meta-avaliação por modelos, de forma ética, transparente e alinhada a valores humanos. Somente assim será possível garantir que a avaliação de LLMs cumpra não apenas sua função técnica, mas também seu papel social: orientar o desenvolvimento de sistemas de linguagem que sejam realmente úteis, justos e responsáveis. O assunto parece muito complexo e é: a avaliação de sistemas de inteligência artificial precisa ser factível, confiável, tecnicamente viável e socialmente responsável. Da mesma forma que o estado da arte da arquitetura de modelos evolui e continua evoluindo, o estado da arte de como avaliá-los está em constante evolução.

Referências

AISERA. **LLM Evaluation: Key Metrics and Frameworks**. https://aisera.com/blog/llm-evaluation/, 2024.

BENCKE, L. et al. **Can we trust LLMs as relevance judges?** Anais do XXXIX Simpósio Brasileiro de Bancos de Dados. **Anais**...Porto Alegre, RS, Brasil: SBC, 2024. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/30724>

BENDER, E. M. et al. **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** . Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. **Anais**...: FAccT '21.New York, NY, USA: Association for Computing Machinery, 2021. Disponível em: <https://doi.org/10.1145/3442188.3445922>

BOWMAN, S. R.; DAHL, G. **What Will it Take to Fix Benchmarking in Natural Language Understanding?** (K. Toutanova et al., Eds.)Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. **Anais**...Online: Association for Computational Linguistics, jun. 2021. Disponível em: <https://aclanthology.org/2021.naacl-main.385/>

CHALAMALASETTI, K. et al. **clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents**. (H. Bouamor, J. Pino, K. Bali, Eds.)Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. **Anais**...Singapore: Association for Computational Linguistics, dez. 2023. Disponível em: <https://aclanthology.org/2023.emnlp-main.689>

CHEN, G. H. et al. Humans or LLMs as the Judge? A Study on Judgement Bias. **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, 2024.

CHOMSKY, N. **Aspects of the Theory of Syntax**. Cambridge, MA: MIT Press, 1965.

CRONBACH, L. J. Studies of acquiescence as a factor in the true–false test. **Journal of Educational Psychology**, v. 33, p. 401–415, 1942.

DIERK, C.; HEALEY, J.; DOGAN, M. D. **Evaluating LLMs in Experiential Context: Insights from a Survey of Recent CHI Publications**. Human-centered Evaluation and Auditing of Language Models Workshop (HEAL), CHI '25. **Anais**...Yokohama, Japan: ACM, 2025. Disponível em: <https://heal-workshop.github.io/chi2025_papers/43_Evaluating_LLMs_in_Experien.pdf>

GALILEO AI. **Generative AI and LLM Insights**. https://galileo.ai/blog/, 2024.

GAO, M. et al. **LLM-based NLG Evaluation: Current Status and Challenges**. ArXiv, 2024. Disponível em: <https://arxiv.org/abs/2402.01383>

GATT, A.; KRAHMER, E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, v. 61, n. 1, p. 65–170, 2018.

GEHMAN, S. et al. **RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models**. (T. Cohn, Y. He, Y. Liu, Eds.)Findings of the Association for Computational Linguistics: EMNLP 2020. **Anais**...Online: Association for Computational Linguistics, nov. 2020. Disponível em: <https://aclanthology.org/2020.findings-emnlp.301/>

GONZALEZ-CABELLO, M. et al. Fairness in crowdwork: Making the human AI supply chain more humane. **Business Horizons**, v. 68, n. 5, p. 645–657, 2025.

GOOGLE CLOUD. **Best Practices with Large Language Models**. https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompt-best-practices?hl=en, 2025.

HEDDERICH, M. A.; OULASVIRTA, A. Explaining crowdworker behaviour through computational rationality. **Behaviour & Information Technology**, v. 44, n. 3, p. 552–573, 2025.

HELLER, J. **Legal AI benchmarking: CoCounsel – from code to courtroom: The meticulous testing of CoCounsel's professional-grade AI**., 23 out. 2024. Disponível em: <https://www.thomsonreuters.com/en-us/posts/innovation/legal-ai-benchmarking-cocounsel/>. Acesso em: 20 ago. 2025

HOUAMEGNI, L. R. P.; GEDIKLI, F. **Evaluating the Effectiveness of Large Language Models in Automated News Article Summarization**., 2025. Disponível em: <https://arxiv.org/abs/2502.17136>

II, S. M. W. **GSM8K Benchmark**. Klu; https://klu.ai/glossary/GSM8K-eval, 2025.

JAN, E. et al. **Multitask-Bench: Unveiling and Mitigating Safety Gaps in LLMs Fine-tuning**. (O. Rambow et al., Eds.)Proceedings of the 31st International Conference on Computational Linguistics. **Anais**...Abu Dhabi, UAE: Association for Computational Linguistics, jan. 2025. Disponível em: <https://aclanthology.org/2025.coling-main.606/>

JAPKOWICZ, N.; BOUKOUVALAS, Z. **Machine Learning Evaluation: Towards Reliable and Responsible AI**. [s.l.] Cambridge University Press, 2024.

JIANG, L.; WAGNER, C. **How Low is Low? Crowdworker Perceptions of Microtask Payments in Work versus Leisure Situations**. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. **Anais**...: CHI '24.New York, NY, USA: Association for Computing Machinery, 2024. Disponível em: <https://doi.org/10.1145/3613904.3642601>

JIAO, J. et al. Navigating LLM Ethics: Advancements, Challenges, and Future Directions. **ArXiv**, v. abs/2406.18841, 2024.

JOSHI, M. et al. **TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**. (R. Barzilay, M.-Y. Kan, Eds.)Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers. **Anais**...Association for Computational Linguistics, 2017. Disponível em: <https://doi.org/10.18653/v1/P17-1147>

JOSHI, N.; TALY, A.; MUPPALLA, D. **LLM-Cite: Cheap Fact Verification with Attribution via URL Generation**., 2025. Disponível em: <https://openreview.net/forum?id=qb2QRoE4W3>

KALOULI, A.-L. et al. Curing the SICK and Other NLI Maladies. **Computational Linguistics**, v. 49, n. 1, p. 199–243, mar. 2023.

KAMRUZZAMAN, M.; SHOVON, MD.; KIM, G. **Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.)Findings of the Association for Computational Linguistics: ACL 2024. **Anais**...Bangkok, Thailand: Association for Computational Linguistics, ago. 2024. Disponível em: <https://aclanthology.org/2024.findings-acl.530/>

KIM, D. K. et al. **Analyzing Offensive Language Dataset Insights from Training Dynamics and Human Agreement Level**. (O. Rambow et al., Eds.)Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025. **Anais**...Association for Computational Linguistics, 2025. Disponível em: <https://aclanthology.org/2025.coling-main.653/>

KIM, S. et al. Prometheus: Inducing Fine-grained Evaluation Capability in Language Models. **arXiv preprint arXiv:2310.08491**, 2023.

KWIATKOWSKI, T. et al. Natural Questions: a Benchmark for Question Answering Research. **Transactions of the Association of Computational Linguistics**, 2019.

LI, D. et al. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. **CoRR**, v. abs/2411.16594, 2024.

LI, H. et al. **CaseGen: A benchmark for multi-stage legal case documents generation**., 2025. Disponível em: <https://arxiv.org/abs/2502.17943>

LIN, S.; HILTON, J.; EVANS, O. **TruthfulQA: Measuring How Models Mimic Human Falsehoods**. (S. Muresan, P. Nakov, A. Villavicencio, Eds.)Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Dublin, Ireland: Association for Computational Linguistics, 2022. Disponível em: <https://aclanthology.org/2022.acl-long.229/>

LIU, Y. et al. **HD-Eval: Aligning Large Language Model Evaluators Through Hierarchical Criteria Decomposition**. (L.-W. Ku, A. Martins, V. Srikumar, Eds.)Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok,

Thailand, August 11-16, 2024. **Anais**...Association for Computational Linguistics, 2024. Disponível em: <https://doi.org/10.18653/v1/2024.acl-long.413>

MATHUR, N.; BALDWIN, T.; COHN, T. **Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics**. Proceedings of the Conference on Empirical Methods in Natural Language Processing. **Anais**...Association for Computational Linguistics, 2020. Disponível em: <https://aclanthology.org/2020.acl-main.448/>

MEVA, DR. D.; KUKADIYA, H. Performance Evaluation of Large Language Models: A Comprehensive Review. **International Research Journal of Computer Science**, v. 12, p. 109–114, mar. 2025.

MINAEE, S. et al. **Large Language Models: A Survey**., 2025. Disponível em: <https://arxiv.org/abs/2402.06196>

NARAYAN, S.; COHEN, S. B.; LAPATA, M. **Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization**. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. **Anais**...Brussels, Belgium: 2018.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **A Plan for Global Engagement on AI Standards**. [s.l.] U.S. Department of Commerce, National Institute of Standards; Technology, jul. 2024. Disponível em: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf>.

PANICKSSERY, A.; BOWMAN, S. R.; FENG, S. LLM Evaluators Recognize and Favor Their Own Generations. **Proceedings of the 38th International Conference on Neural Information Processing Systems**, 2024.

PARK, K. et al. **OffsetBias: Leveraging Debiased Data for Tuning Evaluators**., 2024. Disponível em: <https://arxiv.org/abs/2407.06551>

PEYRARD, M. **Studying Summarization Evaluation Metrics in the Appropriate Scoring Range**. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). **Anais**...Online: Association for Computational Linguistics, 2019. Disponível em: <https://aclanthology.org/P19-1502/>

POLO, F. M. et al. Efficient Multi-Prompt Evaluation of LLMs. **Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)**, 2024.

POMBAL, J. et al. M-Prometheus: A Suite of Open Multilingual LLM Judges. **CoRR**, v. abs/2504.04953, 2025.

PROTOTYPEJAM. **Lake Merritt: AI Evaluation Workbench**. https://prototypejam.github.io/lake_merritt/, 2025.

RAJPURKAR, P. et al. **SQuAD: 100,000+ Questions for Machine Comprehension of Text**. (J. Su, K. Duh, X. Carreras, Eds.)Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. **Anais**...Austin, Texas: Association for Computational Linguistics, nov. 2016. Disponível em: <https://aclanthology.org/D16-1264>

RIBEIRO, M. T. et al. **Beyond Accuracy: Behavioral Testing of NLP Models with CheckList**. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. **Anais**...Online: Association for Computational Linguistics, jul. 2020. Disponível em: <https://aclanthology.org/2020.acl-main.442>

RÍO, B. G. DEL; VAAHTIO, T. **Improving LLM systems with A/B testing**. https://www.flow-ai.com/blog/improving-llm-systems-with-a-b-testing, 2024.

SAID, H. **40 Large Language Model Benchmarks and The Future of LLMs**. Arize AI; https://arize.com/blog/llm-benchmarks-mmlu-codexglue-gsm8k, 2025.

SAMUYLOVA, E. **LLM-as-a-Judge: A Complete Guide to Using LLMs for Evaluations**. https://www.evidentlyai.com/llm-guide/llm-as-a-judge, 2025.

SÁNCHEZ, L. C. **Ethical Considerations and Best Practices in LLM Development**. https://neptune.ai/blog/llm-ethical-considerations, 2024.

SAWCZYN, A. et al. FactSelfCheck: Fact-Level Black-Box Hallucination Detection for LLMs. **arXiv**, 2025.

SCHNABEL, T. et al. **Evaluation methods for unsupervised word embeddings**. (L. Màrquez, C. Callison-Burch, J. Su, Eds.)Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. **Anais**...Lisbon, Portugal: Association for Computational Linguistics, set. 2015. Disponível em: <https://aclanthology.org/D15-1036/>

SHAQIRI, M. et al. Differences between the correlation coefficients Pearson, Kendall and Spearman. **5th International Conference of Natural Science and Mathematics**, nov. 2023.

THAKUR, A. S. et al. JUDGING THE JUDGES: EVALUATING ALIGNMENT AND VULNERABILITIES IN LLMs-AS-JUDGES. **arXiv preprint arXiv:2406.12624v5**, 2025.

TRAUTMANN, D. et al. **Measuring the Groundedness of Legal Question-Answering Systems**. Proceedings of the Natural Legal Language Processing Workshop 2024. **Anais**...Singapore: Association for Computational Linguistics, 2024. Disponível em: <https://aclanthology.org/2024.nllp-1.14>

VONGTHONGSRI, K. **G-Eval Simply Explained: LLM-as-a-Judge for LLM Evaluation**. Confident AI; https://www.confident-ai.com/blog/g-eval-the-definitive-guide, 2025.

WANG, A. et al. **GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding**. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. **Anais**...Brussels, Belgium: Association for Computational Linguistics, nov. 2018. Disponível em: <https://aclanthology.org/W18-5446/>

WANG, A. et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. Em: **Proceedings of the 33rd International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2019.

WANG, R. et al. Can LLMs Replace Human Evaluators? An Empirical Study of LLM-as-a-Judge in Software Engineering. **Proc. ACM Softw. Eng.**, v. 2, n. ISSTA, jun. 2025.

WANG, Y. et al. **PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization**. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. **Anais**...OpenReview.net, 2024. Disponível em: <https://openreview.net/forum?id=5Nn2BLV7SB>

XIONG, K. et al. **Com^2 : A Causal-Guided Benchmark for Exploring Complex Commonsense Reasoning in Large Language Models**. (W. Che et al., Eds.)Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). **Anais**...Vienna, Austria: Association for Computational Linguistics, jul. 2025. Disponível em: <https://aclanthology.org/2025.acl-long.785/>

XU, H.; ASHLEY, K. **A question-answering approach to evaluating legal summaries**., 2023. Disponível em: <https://arxiv.org/abs/2309.15016>

YEH, Y.-T.; ESKÉNAZI, M.; MEHRI, S. A Comprehensive Assessment of Dialog Evaluation Metrics. **ArXiv**, v. abs/2106.03706, 2021.

ZHANG, K. et al. **CitaLaw: Enhancing LLM with Citations in Legal Domain**. (W. Che et al., Eds.)Findings of the Association for Computational Linguistics: ACL 2025. **Anais**...Vienna, Austria: Association for Computational Linguistics, jul. a2025. Disponível em: <https://aclanthology.org/2025.find

ings-acl.583/>

ZHANG, S. et al. **Instruction Tuning for Large Language Models: A Survey**., b2025. Disponível em: <https://arxiv.org/abs/2308.10792>

ZHAO, Y. et al. **One Token to Fool LLM-as-a-Judge**., 2025. Disponível em: <https://arxiv.org/abs/2507.08794>