LLMs Não São Garantido - Avaliando LLMs: Métricas, Benchmarks, Técnicas Automáticas e Considerações Éticas

Livy Real [1,2], **Daniela Vianna**[2],
André Luiz da Costa Carvalho[1], **Altigran Soares da Silva**[1]

[1]Instituto de Computação, Universidade Federal do Amazonas (UFAM)
Manaus - AM - Brazil

[2]Jusbrasil
Salvador - BA - Brazil

{livyreal,andre,alti}@icomp.ufam.edu.br, daniela.vianna}@jusbrasil.com.br

Abstract.

The rapid advancement of Large Language Models (LLMs) demands robust and comprehensive evaluation methodologies to assess their capabilities, reliability, and safety. This advanced 2-hour tutorial will explore the multifaceted landscape of LLM evaluation, moving beyond traditional NLP metrics to include modern benchmarks, human-in-the-loop approaches, and cutting-edge techniques such as "LLM-as-a-Judge". We will address the challenges of evaluating complex emergent behaviors, factual accuracy, reasoning, and ethical considerations such as bias and toxicity. Participants will gain practical insights into selecting appropriate evaluation strategies and understanding the limitations of current methods, empowering them to critically assess LLM performance across a range of real-world scenarios.

Resumo. *Os rápidos avanços nos Large Language Models (LLMs) exigem metodologias de avaliação robustas e abrangentes para verificar suas capacidades, confiabilidade e segurança. Este tutorial avançado de 2 horas aprofundar-se-á na multifacetada paisagem da avaliação de LLMs, indo além das métricas tradicionais de PLN, para cobrir benchmarks modernos, abordagens com intervenção humana e técnicas de ponta como "LLM-as-a-Judge". Discutiremos os desafios de avaliar comportamentos emergentes complexos, precisão factual, raciocínio e considerações éticas como viés e toxicidade. Os participantes obterão insights práticos sobre a seleção de estratégias de avaliação apropriadas e a compreensão das limitações dos métodos atuais, capacitando-os a avaliar criticamente o desempenho de LLMs em vários cenários do mundo real.*

1. Introdução

A avaliação de Modelos de Linguagem de Grande Escala (Large Language Models – LLMs) constitui uma etapa central no ciclo de pesquisa e desenvolvimento desses sistemas, transcendendo a verificação pontual de desempenho. Trata-se de um processo essencial para fins de comparação entre arquiteturas, mitigação de riscos, validação de segurança em aplicações práticas e progresso científico da

área [Meva and Kukadiya 2025]. A natureza generativa desses modelos, aliada à complexidade da linguagem natural e às capacidades emergentes, impõe desafios avaliativos que extrapolam as metodologias tradicionais de Processamento de Linguagem Natural (PLN) [Minaee et al. 2025].

Entre tais capacidades destacam-se o aprendizado em contexto (in-context learning) e o seguimento de instruções (instruction following), que tornam os modelos capazes de generalizar tarefas a partir de poucos exemplos ou mesmo sem exemplos explícitos após o ajuste por instruções. Tais propriedades dificultam a aplicação de métricas clássicas, as quais não são suficientes para captar atributos como criatividade, factualidade, coerência semântica e utilidade prática. O raciocínio passo a passo, por exemplo, demanda critérios avaliativos que ultrapassam a precisão da resposta final, incorporando aspectos como validade lógica, consistência argumentativa e adequação ao contexto.

Além disso, o uso crescente de LLMs em domínios de alta criticidade, como aplicações médicas, jurídicas e educacionais, torna imperativa a adoção de frameworks de avaliação multifacetados, capazes de abarcar não apenas o desempenho técnico, mas também segurança, eficiência e impacto social. A multiplicidade de respostas válidas que esses modelos podem gerar para uma mesma entrada evidencia a necessidade de métricas mais matizadas, subjetivas e semanticamente alinhadas à percepção humana. A simples correspondência de strings deixa de ser suficiente, sendo substituída por abordagens que consideram a qualidade linguística, contextual e interpretativa da saída [Minaee et al. 2025, Meva and Kukadiya 2025].

É importante destacar que a avaliação não ocorre apenas de forma post hoc, mas desempenha papel ativo durante o treinamento dos modelos. O aprendizado por reforço com feedback humano (Reinforcement Learning from Human Feedback – RLHF) depende da qualidade e do alinhamento das métricas utilizadas como função de recompensa. Métricas mal calibradas ou desalinhadas com valores humanos podem levar à otimização de aspectos secundários ou irrelevantes, prejudicando a utilidade, segurança ou confiabilidade do modelo.

Este tutorial propõe uma jornada crítica e estruturada pelo campo da avaliação de LLMs, abordando os dilemas técnicos, metodológicos e éticos que o acompanham.

2. Esboço do Tutorial

- **Introdução à Avaliação de LLMs (20 minutos)**: A avaliação de LLMs é essencial para garantir segurança, utilidade e alinhamento em aplicações reais, indo além da precisão técnica. Modelos generativos demandam métricas sofisticadas que captem nuances como criatividade, coerência e factualidade. A avaliação divide-se em intrínseca vs. extrínseca e humana vs. automática, e seu avanço requer frameworks híbridos que combinem eficiência computacional com profundidade interpretativa humana, especialmente em contextos críticos como saúde e direito. [Meva and Kukadiya 2025, Minaee et al. 2025, Gao et al. 2025, Dierk et al. 2025]
  - Por que a avaliação de LLMs é crítica e desafiadora?
  - As propriedades únicas dos LLMs que complicam a avaliação (capacidade gerativa, comportamentos emergentes, escala, não-determinismo).

  - Breve visão geral dos tipos de avaliação: Intrínseca vs. Extrínseca, Humana vs. Automática.
- **Métricas Tradicionais de PLN e Suas Limitações para LLMs (20 minutos)**: As métricas tradicionais de avaliação de LLMs apresentam limitações significativas por não refletirem com precisão a percepção humana de qualidade textual. Elas penalizam paráfrases e sinônimos, ignoram o raciocínio lógico, não detectam alucinações factuais e falham em avaliar aspectos emergentes como segurança, viés e seguimento de instruções. Embora úteis para comparações superficiais, essas métricas não capturam a complexidade e profundidade das capacidades dos LLMs, exigindo abordagens mais robustas e complementares para uma avaliação fiel do desempenho real dos modelos. [Peyrard 2019, Mathur et al. 2020]
  - Revisão das Métricas Clássicas: Perplexidade, BLEU, ROUGE, METEOR, Acurácia, Precisão, Recall, F1-Score.
  - Quando e Por Que São Insuficientes: Discutir as limitações para modelos gerativos (por exemplo, falta de compreensão semântica, múltiplas saídas válidas) e para avaliar comportamentos matizados (por exemplo, erros factuais, coerência).
  - Exemplos Práticos: Ilustrar cenários onde as métricas tradicionais podem ser enganosas.
- **Benchmarks e Datasets Modernos para LLMs (20 minutos)**: Benchmarks são conjuntos de dados usados para avaliar LLMs em tarefas como compreensão, QA, sumarização, raciocínio e segurança. Embora muitos tragam placares numéricos, é fundamental considerar limitações metodológicas e incluir análises qualitativas para garantir avaliações confiáveis e alinhadas a critérios éticos. [Meva and Kukadiya 2025, AISERA 2025]
  - Benchmarks de Propósito Geral: MMLU, HELM, BIG-bench (Hard), GLUE/SuperGLUE (revisitados no contexto de LLMs).
  - Benchmarks para Tarefas Específicas:
    * Question Answering (QA): Natural Questions, TriviaQA.
    * Sumarização: XSum, CNN/DailyMail.
    * Raciocínio e Código: GSM8K, HumanEval, MBPP.
  - Benchmarks de Segurança e IA Responsável: TruthfulQA, BBQ, ToxiGen/RealToxicityPrompts.
  - Desafios com Benchmarks: Contaminação de dados, saturação de conjuntos de teste, a ”falácia do leaderboard”.
- **Além dos Benchmarks: Metodologias de Avaliação Avançadas (30 minutos)**: A avaliação humana é considerada o “padrão ouro” para medir a qualidade de saídas de LLMs em tarefas de geração de linguagem natural, por capturar nuances como fluência, coerência e criatividade que métricas automáticas não detectam. Metodologias comuns incluem escalas Likert, ranking, testes A/B e anotação de erros. Apesar de sua importância, enfrenta desafios como custo elevado, subjetividade, viés cultural, dificuldades de escalabilidade e necessidade de treinamento rigoroso dos avaliadores. Esses limites motivam o avanço de métodos automáticos e híbridos para complementar a avaliação humana. [Gao et al. 2025, Dierk et al. 2025]
  - Avaliação Humana como Padrão Ouro: Metodologias, concordância entre anotadores, desafios de escalabilidade.

  - Abordagens ”LLM-as-a-Judge”:
    - Conceito e motivação (escalabilidade, custo-benefício).
    - Metodologias: Comparações pareadas, avaliação de resposta única, prompt engineering para judges.
    - Benefícios e limitações: Alucinações do judge, viés em modelos judge, reprodutibilidade.
    - Estudos de caso e pesquisa recente.
- **Considerações Éticas e Direções Futuras na Avaliação de LLMs (30 minutos)**: A avaliação de LLMs deve ser contínua, ética e integrada ao ciclo de desenvolvimento. Transparência e interpretabilidade são fundamentais para superar o caráter de “caixa-preta” desses sistemas. O uso dual levanta preocupações quanto a abusos, exigindo estratégias eficazes de mitigação. A regulamentação e a padronização são urgentes para garantir comparabilidade e promover a responsabilização industrial. Avanços futuros devem priorizar meta-avaliação, monitoramento em tempo real, adaptação por domínio, alinhamento com valores humanos e análise de impactos sociais de longo prazo, adotando uma abordagem multidisciplinar para enfrentar a complexidade dos modelos. [Jiao et al. 2025]
  - Viés e Justiça (Fairness): Identificar e mitigar vieses nas saídas de LLMs, métricas de fairness.
  - Precisão Factual e Alucinações: Técnicas avançadas de detecção, verificabilidade, citabilidade.
  - Conteúdo Tóxico e Nocivo: Detecção e prevenção.
  - Interpretabilidade e Explicabilidade: Avaliação de modelos black-box.
  - O Cenário em Evolução: Lacunas de pesquisa, problemas abertos e o papel dos princípios de IA responsável na avaliação.
  - Avaliação Adversarial e de Robustez: Teste de estresse, injeções de prompt, jailbreaking.

3. Biografia dos Autores

Livy Real (Apresentadora) pesquisadora na área de Processamento de Linguagem Natural e Inteligência Artificial, atuando na interface entre indústria e academia, e entre linguística e computação. Atualmente, é pesquisadora na Jusbrasil e Colaboradora de Pesquisa na Universidade Federal do Amazonas (UFAM), com foco em avaliação de LLMs. Atuou como pesquisadora de pós-doutoramento na IBM Research Brazil de 2014 a 2017. Possui Doutorado em Linguística pela UFPR (2014) com período sanduíche no LaBRI (França), Mestrado e Bacharelado em Linguística e Letras, respectivamente, pela UFPR. Sua expertise inclui semântica formal, semântica computacional e o desenvolvimento de recursos linguísticos para o português.

Daniela Vianna cientista de dados na Jusbrasil, onde utiliza técnicas de aprendizado de máquina para promover inovação em tecnologia jurídica. É doutora em Ciência da Computação pela Rutgers University, nos Estados Unidos, e possui mestrado e bacharelado pela Universidade Federal Fluminense (UFF), no Brasil. Seus principais interesses incluem Recuperação de Informação, Processamento de Linguagem Natural e Modelos de Linguagem de Grande Escala.

André Luiz da Costa Carvalho Professor na Universidade Federal do Amazonas (UFAM), com Doutorado em Informática pela mesma instituição. Coordena o laboratório de Inteligência Artificial Aplicada a Teste de Software (IATS). Sua pesquisa concentra-se na aplicação de técnicas de aprendizado de máquina, em particular deep learning, modelos transformadores e LLMs, para processamento de texto e recuperação de informação. Liderou diversos projetos de pesquisa com financiamento público e privado (por exemplo, Motorola, Samsung, Nokia) e foi reconhecido com o Prêmio José Mauro Castilho de Melhor Artigo no XXII Simpósio Brasileiro de Banco de Dados.

Altigran Soares da Silva (Apresentador) Professor Titular no Instituto de Computação da Universidade Federal do Amazonas (IComp/UFAM) e Bolsista de Produtividade em Pesquisa do CNPq (Nível 1B). Possui Doutorado em Ciência da Computação pela UFMG (2002). Seus interesses de pesquisa abrangem Gestão de Dados, Recuperação de Informação, Mineração de Dados, Aprendizado de Máquina e Modelos de Linguagem, com ênfase em web, mídias sociais, direito, mercado financeiro e saúde. Coordenou inúmeros projetos e publicou extensivamente. Desempenhou papéis de liderança significativos em organizações de computação brasileiras (CNPq, CAPES, SBC) e foi cofundador de empreendimentos de tecnologia de sucesso adquiridos pela Google (Akwan), Linx Sistemas (Neemu) e JusBrasil (Teewa).

References

AISERA (2025). Llm evaluation: Key metrics, best practices and frameworks. `https://aisera.com/blog/llm-evaluation/`. Accessed July 2025.

Dierk, C., Healey, J., and Dogan, D. (2025). Evaluating llms in experiential context. In *Workshop on Human-centered Evaluation and Auditing of Language Models (HEAL), CHI 2025*.

Gao, M., Hu, X., Ruan, J., Pu, X., and Wan, X. (2025). Llm-based nlg evaluation: Current status and challenges.

Jiao, J., Afroogh, S., Xu, Y., and Phillips, C. (2025). Navigating llm ethics: Advancements, challenges, and future directions.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Meva, D. D. and Kukadiya, H. (2025). Performance evaluation of large language models: A comprehensive review. *International Research Journal of Computer Science*, 12:109–114.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2025). Large language models: A survey.

Peyrard, M. (2019). Studying summarization evaluation metrics in the appropriate scoring range. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.