Análise de Falhas na Detecção de Alucinação em Textos Jurídico-Policiais em Português

Ricardo Rodrigues Barcelar[1], Thiago Meirelles Ventura[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Cuiabá – MT – Brazil

ricardo.barcelar@sou.ufmt.br, thiago@ic.ufmt.br

Abstract. *This paper documents the failures of hallucination detection mechanisms when applied to Brazilian police legal documents generated by large language models. Three evaluation pipelines were tested using the Lynx model as auditor, with Self-Consistency voting (k=3) on AWS SageMaker. Results show high recall but low precision across all configurations, with false positives driven by two patterns: referential rigidity and normative opacity. The findings indicate that current factuality verification tools fail to distinguish institutional inferences from fabricated content in this specialized domain.*

Resumo. *Este artigo documenta as falhas dos mecanismos de detecção de alucinação quando aplicados a textos jurídico-policiais brasileiros gerados por modelos de linguagem. Três pipelines de avaliação foram testadas, utilizando o modelo Lynx como auditor com votação por Self-Consistency (k=3) em infraestrutura AWS SageMaker. Os resultados revelam alto recall, porém precisão baixa em todas as configurações, com falsos positivos causados por dois padrões: rigidez referencial e opacidade normativa. Os achados indicam que as ferramentas atuais de verificação de factualidade não distinguem adequadamente inferências institucionais de conteúdo fabricado neste domínio especializado.*

1. Introdução

A ascensão dos Grandes Modelos de Linguagem (LLMs) tem ampliado a automação de tarefas cognitivas em contextos organizacionais, cujas implicações se estendem às atividades no setor público (Chen et al., 2024). No domínio da segurança pública, a capacidade desses modelos de converter narrativas de Boletins de Ocorrência em documentos formais, como Portarias de Instauração de Inquérito Policial, representa um avanço na celeridade investigativa. Entretanto, a natureza probabilística das arquiteturas de Transformers introduz vulnerabilidades, sendo a mais perniciosa o fenômeno das alucinações, como a geração de conteúdos gramaticalmente fluentes, mas factualmente infiéis ao contexto de origem.

No escopo desta pesquisa, define-se alucinação sob o prisma da fidelidade contextual. Trata-se da inclusão de fatos, entidades ou dados novos no texto produzido que não possuam lastro na narrativa original fornecida como premissa. Em cenários jurídico-policiais, o impacto dessas distorções ultrapassa a mera falha técnica. Qualquer fabricação de nomes de suspeitos, números de protocolo ou detalhes de evidências pode incorrer em danos irreparáveis a direitos individuais, além de fundamentar falsas acusações que comprometem a integridade dos procedimentos policiais.

A incerteza sobre a real eficiência dos mecanismos atuais de detecção de alucinação constitui o problema central deste estudo. Embora existam *frameworks*

propostos para auditoria estatística de modelos de linguagem, a maioria das abordagens é avaliada em *datasets* baseados em textos enciclopédicos em língua inglesa, como o WikiBio (Lebret et al., 2016) e suas derivações para detecção de alucinação (Manakul et al., 2023), onde a distinção entre verdade e mentira é facilitada pela distância semântica entre afirmações.

O domínio jurídico-policial brasileiro ainda impõe desafios que os estudos tradicionais negligenciam. Os documentos possuem uma rigidez estrutural que exige a inserção de novos conteúdos funcionais, como enquadramentos penais, termos institucionais (ex.: Instituto Médico Legal, Ministério Público, etc) e determinações procedimentais, que não constam no contexto inicial que ensejou tal documento (ex.: Boletim de Ocorrência), mas que são componentes obrigatórios do texto oficial.

O objetivo deste trabalho é documentar as falhas na detecção de alucinações em textos jurídico-policiais em português, expondo os limites do estado da arte em cenários de alta especialização. Diferente da literatura que celebra sucessos incrementais, este artigo adota a perspectiva de análise de resultados negativos, investigando os modos de falha que impedem a automação segura desses documentos.

2. Trabalhos Relacionados

A ocorrência de alucinações em modelos de linguagem tem sido documentada como um obstáculo à confiabilidade de sistemas baseados em Processamento de Linguagem Natural, particularmente em domínios que exigem rigor factual. Dahl et al., (2024) argumentam que, embora LLMs consigam gerar argumentos legais plausíveis, sua confiabilidade permanece limitada devido à geração de elementos inexistentes nos autos, caracterizando alucinação em cenários legais. Conforme discutido por Bender et al. (2021), modelos de linguagem operam por predição estatística sem compromisso intrínseco com a veracidade. Estudos como os de Ji et al. (2022) reforçam que a ausência de mecanismos explícitos de grounding contribui para inconsistências factuais, especialmente em tarefas que demandam precisão terminológica e aderência a evidências documentais.

A literatura recente tem evoluído com benchmarks e *frameworks* que ampliam o escopo de avaliação da factualidade. Iniciativas como HalluLens (Bang et al., 2025) e FaithJudge (Tamber et al., 2025) propõem protocolos sistemáticos de mensuração, enquanto AgentHallu (Liu et al., 2026) explora cenários de raciocínio multi-etapas. Contudo, tais propostas ainda apresentam heterogeneidade metodológica e não se consolidaram como baselines amplamente adotadas. Em contraste, abordagens como Lynx (Ravi et al., 2024), SelfCheckGPT (Manakul et al., 2023) e G-Eval (Liu et al., 2023) representam paradigmas recorrentes em estudos empíricos.

Abordagens baseadas em modelos generalistas, como SelfCheckGPT e G-Eval, dependem de modelos externos da família GPT, introduzindo variabilidade e reduzindo o controle experimental, enquanto soluções como LLM Guard (Goyal et al., 2024) operam como *pipelines* heurísticas de difícil reprodutibilidade.

Neste estudo, adotou-se o modelo Lynx como auditor principal. Sua arquitetura supervisionada modela explicitamente relações de inferência entre premissas e evidências, com suporte a explicações estruturadas, o que favorece a análise de

consistência factual em textos complexos.

3. Metodologia

Esta seção detalha o desenho experimental adotado para investigar as falhas na detecção de alucinações em textos jurídico-policiais escritos em português. O estudo afasta-se da busca por métricas de acurácia otimistas para concentrar-se na análise sistemática dos padrões de erro, investigando como a rigidez referencial e a opacidade normativa dos mecanismos de verificação comprometem a confiabilidade da detecção de factualidade nesse domínio especializado.

3.1. Caracterização do Corpus e Domínio Técnico

Esta pesquisa utiliza dados reais de campo da Polícia Judiciária Civil de Mato Grosso (PJC-MT). O corpus é composto por 667 narrativas de Boletins de Ocorrência, devidamente anonimizados, versando predominantemente sobre crimes de violência contra a mulher, idoso e crianças, registrados entre os anos de 2024 e 2025.

A escolha deste domínio visa explorar estratégias de detecção de alucinações em um cenário onde a precisão de entidades (nomes, datas, menções a documentos, organizações, etc) é crítica, assim como o uso de jargões jurídico-policiais brasileiros é onipresente (Costa, 2014).

3.2. Infraestrutura em Nuvem

A execução dos experimentos foi operacionalizada na infraestrutura Amazon Web Services (AWS), utilizando o Amazon SageMaker Studio como ambiente unificado de experimentação. O ambiente foi provisionado em uma instância ml.g6.xlarge, equipada com uma GPU NVIDIA L4 (24 GB de VRAM), 4 vCPUs e 16 GiB de RAM, além de um disco NVMe local de 233 GB utilizado para armazenamento dos pesos dos modelos e dados experimentais.

A arquitetura implantada na instância compreendeu três componentes de serviço: (i) o modelo Lynx (4B parâmetros, quantização FP8), hospedado via vLLM com configuração otimizada para inferência em lote, incluindo controle de utilização de memória GPU (88%), limite de 24 sequências simultâneas e janela de contexto de 6.144 tokens, ajustada ao tamanho dos documentos jurídico-policiais; (ii) um API Gateway intermediário, responsável pelo agrupamento automático de requisições em lote, com janela temporal de 50ms e lotes de até 8 requisições, permitindo que o protocolo de *Self-Consistency* com k=3 execuções independentes por sentença fosse processado de forma paralela; e (iii) um serviço de reconhecimento de entidades nomeadas baseado no modelo LeNER-Br (BERT-base), executado em CPU, utilizado exclusivamente pela *pipeline* P3.

A adoção da AWS não teve papel apenas instrumental, mas viabilizou a execução do protocolo experimental nas condições requeridas pelo estudo. Em ambiente local, a combinação entre Lynx, Self-Consistency com k=3, avaliação por sentença e processamento de 667 documentos produziu tempos de execução na escala de dias e timeouts recorrentes, considerando vários ciclos experimentais. O uso do SageMaker Studio com GPU NVIDIA L4, vLLM e batching permitiu paralelizar milhares de

chamadas ao auditor e reduzir a execução para escala de horas, preservando rastreabilidade dos notebooks, parâmetros e artefatos. Como contrapartida, a solução introduz custos de infraestrutura, dependência de serviços gerenciados e necessidade de controle sobre dados sensíveis, mitigada neste estudo pela anonimização do corpus.

3.3. Preparação dos Dados

Para viabilizar a avaliação de verificações de factualidade e da ocorrência de alucinações em modelos de Inteligência Artificial (IA), foi construído um dataset sintético composto por 667 pares de narrativa de um boletim de ocorrência e uma portaria produzida por IA. A portaria consiste em um ato administrativo de natureza normativa ou ordinatória, emitida por autoridade policial com a finalidade de organizar, instruir ou disciplinar serviços internos e procedimentos investigativos, sempre fundamentado em legislação prévia. Trata-se, portanto, de um documento com características estruturais e formais específicas.

A construção do dataset foi realizada por meio da inferência de um modelo previamente ajustado para a geração de portarias, derivado de Qwen3-4B-Instruct-2507 (Yang et al., 2025). Esse modelo foi empregado em uma *pipeline* na qual a narrativa do boletim de ocorrência foi utilizada como entrada, resultando na geração automatizada das portarias correspondentes.

3.4. Protocolo de Avaliação

O protocolo de avaliação proposto fundamenta-se na verificação da fidelidade do texto gerado por modelos de linguagem em relação ao contexto de referência que subsidiou sua produção. Para mensurar a ocorrência de alucinações de forma sistemática, foram conduzidas múltiplas rodadas experimentais com variação de parâmetros, organizadas em três *pipelines* distintas.

A primeira *pipeline* foi composta por duas etapas principais: (i) Realizou-se a segmentação de cada par contexto/portaria em sentenças menores por meio de expressões regulares, abordagem que apresentou desempenho superior na segmentação estrutural de documentos jurídico-policiais em comparação com bibliotecas consolidadas como NLTK e spaCy; e (ii) Cada sentença foi submetida a um processo de auditoria conduzido pelo modelo Lynx. Para mitigar a variabilidade das respostas e capturar a incerteza epistêmica, aplicou-se a técnica de *Self-Consistency*, com k=3 execuções independentes por sentença. O escore de não-conformidade foi definido como a média dos votos de alucinação produzidos nessas execuções.

A segunda *pipeline* manteve o procedimento de auditoria com *Self-Consistency*, porém suprimiu a etapa de segmentação. Nesse caso, o texto gerado foi avaliado integralmente pelo modelo Lynx, preservando-se o mesmo critério de cálculo do escore de não-conformidade. Essa configuração permitiu analisar o impacto da granularidade da avaliação na detecção de alucinações.

A terceira *pipeline* incorporou mecanismos adicionais de verificação simbólica e semântica, organizados em três etapas. (i) Manteve-se a segmentação em sentenças por expressões regulares. (ii) Aplicou-se reconhecimento de entidades nomeadas por meio de um modelo de NER, seguido da verificação de correspondência entre as entidades

extraídas e aquelas presentes no contexto. (iii) Por fim, apenas nos casos em que as etapas anteriores não produziram evidência conclusiva, a sentença foi submetida à auditoria pelo modelo Lynx, novamente com aplicação de *Self-Consistency* com k=3. O escore de não-conformidade foi calculado de forma análoga às demais *pipelines*.

A validação dos resultados utilizou amostragem estratificada com anotação humana parcial, em razão da inviabilidade de anotação integral do corpus. Em consonância com Card et al. (2020), que destacam a necessidade de atenção ao tamanho amostral, à variância das estimativas e ao poder estatístico em avaliações humanas de PLN baseadas em amostras limitadas, os escores de não-conformidade foram estratificados nos intervalos 0, 0.01-0.20, 0.21-0.50, 0.51-0.70, 0.71-0.90 e 0.91-1.00, com seleção aleatória de até 25 sentenças por intervalo para validação manual. Cada instância foi classificada por um anotador humano como "ok", quando o julgamento do pipeline foi considerado correto, ou "falha", quando incorreto. Adotou-se critério binário, no qual $s > 0$ indica alucinação, uma vez que o domínio jurídico-policial penaliza falsos negativos. Assim, $s = 0$ indica conformidade com o contexto.

A partir das amostras anotadas, foram computadas métricas de desempenho de classificação (*Precision, Recall, F1-score e Accuracy*) para cada *pipeline*. Os intervalos de confiança de 95% foram estimados por meio de *bootstrap* não-paramétrico com 2.000 reamostragens, permitindo avaliar a estabilidade das estimativas frente ao tamanho amostral. Complementarmente, calculou-se a taxa de acerto por faixa de escore, de modo a identificar as regiões do espaço de escores nas quais cada *pipeline* apresenta maior ou menor confiabilidade.

Os três *pipelines* foram aplicados sobre o mesmo dataset, garantindo comparabilidade direta entre os resultados. O código-fonte das *pipelines*, os *scripts* de análise estatística e os dados de amostragem estratificada foram organizados em notebooks reprodutíveis, assegurando a replicabilidade integral do protocolo experimental e estão disponíveis em https://github.com/ricardobarcelar/analise-falhas-deteccao-alucinacao

4. Resultados e Discussão

A avaliação comparativa das três *pipelines* foi conduzida sobre amostras estratificadas, com 72, 57 e 82 instâncias anotadas para as *pipelines* P1, P2 e P3, respectivamente. Os resultados são apresentados na Tabela 1.

Tabela 1. Desempenho das *pipelines*

Pipeline	***n***	***Precision***	***Recall***	***F1-score***	***Accuracy***
P1 (Lynx+SC, c/ segmentação)	72	0,319 [0,188-0,463]	0,882 [0,700-1,000]	0,468 [0,305-0,615]	0,527 [0,416-0,652]
P2 (Lynx+SC, s/ segmentação)	57	0,469 [0,285-0,645]	0,714 [0,500-0,889]	0,566 [0,378-0,709]	0,596 [0,473-0,719]
P3 (NER+Lynx, c/ segmentação)	82	0,210 [0,109-0,320]	0,750 [0,529-0,944]	0,328 [0,184-0,461]	0,402 [0,304-0,512]

Os intervalos de confiança de 95% indicaram maior incerteza nas estimativas de precisão e F1-score, em razão do tamanho reduzido das amostras anotadas, sem alterar, contudo, o padrão geral de baixo desempenho em precisão nas três pipelines.

A *pipeline* P1 apresentou elevado *recall* (0,8824), indicando alta sensibilidade na detecção de alucinações. No entanto, a precisão foi baixa (0,3191), com número expressivo de falsos positivos, em especial entre as configurações baseadas exclusivamente no Lynx. A *pipeline* P2 apresentou melhor equilíbrio entre as métricas. Houve aumento da precisão (0,4688) e da acurácia (0,5965), além de maior F1-score (0,5660), indicando uma redução dos falsos positivos. A *pipeline* P3, por sua vez, apresentou o pior desempenho global. Apesar de *recall* relativamente alto (0,7500), a precisão foi a menor entre os experimentos (0,2105), com elevado número de falsos positivos, superando expressivamente o de falsos negativos.

A análise por faixas de escore, apresentada na Tabela 2, reforça esse padrão. A faixa s=0 apresentou altas taxas de acerto em todas as *pipelines*, indicando consistência na identificação de textos fiéis. Em contraste, as faixas intermediárias concentraram os maiores erros, com queda acentuada de desempenho, especialmente na *pipeline* P3.

Tabela 2. Faixa de escore obtidas pelas *pipelines*

Faixa	**P1**		**P2**		**P3**	
	n	***Acerto***	***n***	***Acerto***	***n***	***Acerto***
$s = 0$	25	92,0%	25	76,0%	25	84,0%
0,21 – 0,50	25	32,0%	25	36,0%	25	16,0%
0,51 – 0,70	17	29,4%	6	83,3%	25	24,0%
0,91 – 1,00	5	40,0%	1	100,0%	7	28,6%

Nota: Nenhuma das pipelines produziu instâncias anotadas no intervalo de 0,01–0,20 conforme previsto no protocolo de estratificação.

A estruturação do experimento em três *pipelines* permitiu avaliar a detecção de factualidade sob diferentes circunstâncias: o fracionamento em sentenças individuais (P1), a análise do texto integral (P2) e a integração de mecanismos determinísticos via NER (P3). Os resultados sugerem que as dificuldades observadas não decorrem primariamente da arquitetura das estratégias de auditoria, mas da limitação dos mecanismos de aferição em processar as particularidades do domínio jurídico-policial.

O contraste entre P1 e P2 indica que a segmentação em sentenças isoladas amplia a sensibilidade ao custo de aumentar falsos positivos: ao perder o contexto global do documento, o auditor tende a classificar fragmentos procedimentais como afirmações factuais não fundamentadas. A avaliação do texto integral, por sua vez, preserva relações intra-documentais que permitem ao modelo reconhecer com maior acurácia a função pragmática de cada trecho. O desempenho inferior da P3 revela que verificações baseadas em correspondência de entidades são particularmente frágeis neste domínio, onde variações de grafia e inferências contextuais são frequentes.

A análise dos erros revelou dois padrões recorrentes que explicam a predominância de falsos positivos. O primeiro, caracterizado como rigidez referencial, manifesta-se quando o auditor exige correspondência literal entre o texto gerado e o contexto, penalizando inferências institucionais legítimas. Um exemplo ocorre quando a IA inclui "Junte-se cópia da decisão proferida em audiência de custódia, referente à situação jurídica do conduzido". Embora coerente com a prática policial, a ausência de menção explícita à audiência de custódia no contexto original leva o detector a rotulá-lo como alucinação.

O segundo padrão, caracterizado como opacidade normativa, ocorre quando o auditor trata construções obrigatórias do domínio como afirmações factuais verificáveis. Frases como "DETERMINAR ao Escrivão de Polícia responsável pelo feito que adote as seguintes providências: I - Juntar aos autos o relatório final mencionado", que constituem atos performativos inerentes à formalização do documento, são indevidamente classificadas como alucinação.

Em todas as *pipelines*, o modelo errou majoritariamente ao apontar alucinações onde elas não existiam, confirmando que a terminologia e as estruturas do domínio jurídico-policial interferem sistematicamente na capacidade do auditor de discriminar entre informação inventada e prática institucional legítima.

5. Considerações Finais

Este estudo documentou falhas na detecção de alucinações em textos jurídico-policiais em português brasileiro, a partir de uma avaliação executada em infraestrutura AWS. O uso do Amazon SageMaker Studio viabilizou o processamento das pipelines na escala de horas. Os resultados indicaram que mecanismos tradicionais de verificação de factualidade apresentam alta sensibilidade, mas baixa precisão quando aplicados a documentos marcados por convenções normativas, vocabulário técnico e inferências institucionais próprias do sistema de justiça.

A análise dos erros revelou padrões recorrentes de falso positivo. Na rigidez referencial, o auditor exige correspondência literal com o contexto e classifica como alucinação inferências legítimas do domínio jurídico-policial. Na opacidade normativa, determinações, encaminhamentos e atos performativos próprios desse tipo documental, embora juridicamente esperados, são tratados pelo auditor como afirmações factuais sem lastro explícito no boletim de ocorrência. Esses padrões explicam a predominância de falsos positivos observada nas pipelines avaliadas.

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. A natureza estocástica dos LLMs implica variabilidade nos julgamentos, ainda que mitigada pelo protocolo de Self-Consistency. Além disso, o tamanho das amostras de validação humana restringe a generalização das estimativas, e o uso de portarias sintéticas pode introduzir vieses de estilo, estrutura e distribuição de erros distintos daqueles observados em documentos produzidos por humanos.

Por fim, a opção por avaliar apenas o Lynx decorreu de uma decisão de escopo experimental. O estudo buscou analisar padrões de falha em um auditor aberto, executável em infraestrutura controlada na AWS e compatível com avaliação reprodutível em lote, sem pretensão de comparar exaustivamente frameworks de

detecção de alucinação. Trabalhos futuros devem avaliar outros modelos, como G-Eval, SelfCheckGPT e benchmarks recentes, além de desenvolver auditores especializados por meio de Supervised Fine-Tuning (SFT) e Direct Preference Optimization (DPO), apoiados por datasets anotados do domínio jurídico-policial.

Declaração sobre uso de Inteligência Artificial

Ferramentas de Inteligência Artificial generativa foram utilizadas como apoio em etapas específicas deste trabalho. O Claude.ai (Anthropic) foi empregado como copiloto na implementação dos *scripts* Python e *notebooks* de avaliação. O NotebookLM (Google) auxiliou na organização e revisão da literatura. O Gemini (Google) foi utilizado como apoio à redação e revisão textual.

Em todos os casos onde a Inteligência Artificial interveio na escrita dos conteúdos estes foram novamente revisados, corrigidos e validados pelos autores, que assumem integral responsabilidade pelo conteúdo final. Nenhuma ferramenta de IA generativa é listada como autora do trabalho.

Referências

BANG, Yejin; JI, Ziwei; SCHELTEN, Alan; HARTSHORN, Anthony; FOWLER, Tara; ZHANG, Cheng; CANCEDDA, Nicola; FUNG, Pascale. HalluLens: LLM hallucination benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna: ACL, 2025. p. 24128–24156. DOI: 10.18653/v1/2025.acl-long.1176

BENDER, E. M.; GEBRU, T.; MCMILLAN-MAJOR, A.; SHMITCHELL, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, p. 610–623, 2021.

CARD, D.; HENDERSON, P.; KHANDELWAL, U.; MOORE, R.; SMITH, N. A. With Little Power Comes Great Responsibility: Efficient and Effective Evaluation in NLP. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 9263–9274, 2020.

CHEN, Wilbur Xinyuan; SRINIVASAN, Suraj; ZAKERINIA, Saleh. Displacement or complementarity? The labor market impact of generative AI. Harvard Business School Working Paper, n. 25-039, 2024.

COSTA, Maria Izabel Plath da. Terminologia jurídico-policial: seleção e validação de termos em textos-base e mapa-domínio. Tradterm, São Paulo, Brasil, v. 24, p. 301–324, 2015. DOI: 10.11606/issn.2317-9511.tradterm.2014.96574.

DAHL, M.; MAGESH, V.; SUZGUN, M.; HO, D. E. Large legal fictions: profiling legal hallucinations in large language models. Journal of Legal Analysis, v. 16, n. 1, p. 64–93, 2024. DOI: 10.1093/jla/laae003.

GOYAL, Shubh; HIRA, Medha; MISHRA, Shubham; GOYAL, Sukriti; GOEL, Arnav;

DADU, Niharika; DB, Kirushikesh; MEHTA, Sameep; MADAAN, Nishtha. LLMGuard: guarding against unsafe LLM behavior. Proceedings of the AAAI Conference on Artificial Intelligence, v. 38, n. 21, p. 23790–23792, 2024. DOI: 10.1609/aaai.v38i21.30566

JI, Ziwei; LEE, Nayeon; FRIESKE, Rita; et al. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, v. 55, n. 12, 2022. DOI: 10.1145/3571730.

LEBRET, Rémi; GRANGIER, David; AULI, Michael. Neural text generation from structured data with application to the biography domain. In: SU, Jian; DUH, Kevin; CARRERAS, Xavier (ed.). Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016. p. 1203–1213. DOI: 10.18653/v1/D16-1128.

LIU, Yang; ITER, Dan; XU, Yichong; WANG, Shuohang; XU, Ruochen; and ZHU, Chenguang. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. p. 2511-2522. DOI: 10.18653/v1/2023.emnlp-main.153.

LIU, Xuannan; YANG, Xiao; ZEKUN, Li; LI, Peipei; HE, Ran. AgentHallu: Benchmarking Automated Hallucination Attribution of LLM-based Agents. arXiv preprint arXiv:2601.06818, 2026.

MANAKUL, Potsawee; LIUSIE, Adian; GALES, Mark. SelfCheckGPT: zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. p. 9004–9017. DOI: 10.18653/v1/2023.emnlp-main.557

RAVI, S. S. et al. Lynx: An Open Source Hallucination Evaluation Model. arXiv preprint arXiv:2407.08488, 2024.

TAMBER, Manveer Singh; BAO, Forrest Sheng; XU, Chenyu; LUO, Ge; KAZI, Suleman; BAE, Minseok; LI, Miaoran; MENDELEVITCH, Ofer; QU, Renyi; LIN, Jimmy. Benchmarking LLM faithfulness in RAG with evolving leaderboards. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track. Suzhou: ACL, 2025. p. 799–811. DOI: 10.18653/v1/2025.emnlp-industry.54.

YANG, An et al. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025.