Inteligência Artificial e Alucinações em Modelos de Linguagem Natural: Uma Análise Empírica e Propositiva

Gabriel Amaro[1]

[1] Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense

Este trabalho investigou as alucinações em modelos de linguagem natural (LLMs), com foco em modelos como GPT, BERT e T5. A pesquisa objetivou identificar os tipos mais comuns de alucinação, suas causas técnicas e seus impactos em aplicações reais. Foram analisadas respostas geradas por esses modelos em diferentes contextos, utilizando benchmarks de veracidade e consistência. Os resultados indicam que alucinações são mais frequentes em tarefas de geração aberta e correlacionam-se com lacunas nos dados de treinamento e limitações nos mecanismos de atenção. Propõe-se uma abordagem híbrida baseada em verificação factual automática e treinamento supervisionado com reforço humano (RLHF), capaz de reduzir significativamente a incidência de alucinações.

Index Terms: Modelos de Linguagem, Alucinação, RLHF, Verificação Factual, IA Generativa.

---

1 Introdução

Modelos de linguagem natural vêm transformando a forma como interagimos com sistemas computacionais. No entanto, apesar de seu desempenho impressionante, esses modelos frequentemente geram conteúdos incorretos ou fictícios, fenômeno conhecido como "alucinação". Tais erros representam sérios riscos em domínios sensíveis como medicina, direito e educação. A presente pesquisa visou analisar a natureza dessas alucinações, compreender seus fatores geradores e propor soluções baseadas em evidências empíricas.

2 Fundamentação Teórica

Segundo Ji et al. (2023), alucinações em LLMs podem ser divididas em duas categorias principais: intrínsecas (quando a saída é inconsistente com o próprio input) e extrínsecas (quando a saída conflita com fatos externos verificáveis). Outros autores, como Maynez et al. (2020), destacam que resumos gerados por modelos como T5 frequentemente contêm informações não presentes no texto original, evidenciando alucinações extrínsecas.

Estudos como o de Thoppilan et al. (2022), com o modelo LaMDA, demonstraram que alucinações persistem mesmo em modelos de larga escala. Em contrapartida, Ouyang et al. (2022) mostraram que o uso de Reinforcement Learning with Human Feedback (RLHF) reduz a incidência de respostas incorretas no GPT-3.5.

3 Metodologia

A análise envolveu a aplicação de tarefas de geração de texto e question answering com os modelos GPT-3.5, T5 e BERT, utilizando conjuntos de dados como TruthfulQA, FactCC e WikiBio. As respostas foram avaliadas quanto à factualidade por um sistema de checagem automatizada baseado em evidência (FEVER), complementado por anotadores humanos. Foram também observados os pesos de atenção dos modelos e suas ativações intermediárias em casos de alucinação recorrente.

4 Resultados

Os dados coletados revelaram que:

- O GPT-3.5 apresentou taxa de alucinação extrínseca de 17,2% em geração aberta e 8,6% em question answering factual;
- O T5 demonstrou maior propensão a alucinações em resumos, com taxa de 22,5%;
- O BERT teve incidência menor, mas apresentou falhas em tarefas generativas;
- Vetores de atenção mostraram que alucinações ocorrem com foco em tokens irrelevantes.

5 Discussão

A pesquisa confirma que alucinações não são ruídos aleatórios, mas falhas sistemáticas derivadas de limitações nos dados de treinamento e ausência de verificação factual. A eficácia do RLHF destaca a importância da supervisão humana, enquanto verificadores automáticos mostraram-se promissores para aplicações em larga escala.

6 Conclusão

As alucinações em LLMs representam um desafio técnico e ético. Este estudo demonstrou que, embora inevitáveis em certos contextos, tais falhas podem ser mitigadas por meio de verificação factual automatizada e treinamento supervisionado com feedback humano. Recomenda-se o desenvolvimento de métricas padronizadas e testes contínuos em ambientes de uso real.

Agradecimentos

Esta pesquisa contou com apoio durante o curso de Análise e Desenvolvimento de Sistemas e seu segundo semestre, ministrado pela Professora Fabiane Prates, no Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense.