Achieving Near-Zero Hallucation In Large Language Models

Abstract. This paper presents a research methodology focused on the mitigation of hallucinations in modern large language models. The initial phase involved the development and training of a models following the framework established in Google's "Attention Is All You Need" [1] paper. The degree of hallucination present in the model outputs was then systematically measured with respect to the training data. These measurements were obtained after multiple training iterations conducted on models of identical size but trained on datasets differing in quality and factual reliability, thereby yielding models with varying levels of knowledge representation. The results indicated that both data quality and model architecture contributed significantly to the prevalence of hallucinations. Consequently, architectural modifications were introduced, after which a model was trained on high-quality data. This revised configuration achieved a reduction of hallucinations in 96.4% of test cases.
This research paper is for a Research Methodology course at ELTE University. The data in it is made up, and should not be taken seriously or referenced.

1. Introduction

Your introduction

1.1. Related work

Works related to your paper

2. Experimental Design and Methodology

The experimental investigation was conducted in two sequential phases: first, establishing a robust performance baseline and quantifying the relationship between training data reliability and hallucination prevalence [2–4]; and second, implementing and evaluating the novel architectural intervention.

2.1. Phase 1: Baseline Establishment and Data Quality Analysis

2.1.1. Model Initialization and Data Corpus Formalization

All comparative experiments utilized the **Canonical Transformer Model** as the foundational architecture, strictly instantiated according to the specifications of [1]. A standard configuration was employed (e.g., **6 encoder layers, 6 decoder layers, and 8 attention heads**) [1]. All initializations used a standard parameter initialization and an Adam optimizer with the inverse square-root learning rate schedule [1].

The experimental corpora were constructed from large-scale, publicly available datasets, segregated into two categories based on empirical reliability ratings:

- **Low-Reliability Corpus** ($\mathcal{D}_{\text{LR}}$): Comprised of tokens sampled from *Reddit (2020 snapshot)*, *Common Crawl WET files*, and *WikiAnswers community data*. Prior work ([2–4]) has demonstrated high factual heterogeneity and weak source attribution in these domains.
- **High-Reliability Corpus** ($\mathcal{D}_{\text{HR}}$): Comprised of tokens from *English Wikipedia (2023-06 dump)*, the *Stanford Question Answering Dataset (SQuAD v2)* [5], and the *Natural Questions Open dataset* [6]. These sources underwent strict deduplication and factuality validation, exhibiting high human-rated factual consistency in preliminary audits.

2.1.2. Differential Baseline Training and Factual Integrity Quantification

Two baseline models were trained to empirically isolate the causal effect of training data reliability on factual consistency. Both followed an **identical training regimen** (e.g., same number of training steps, batch size, and dropout rate).

- **Baseline** B_1: Trained exclusively on the Low-Reliability Corpus ($\mathcal{D}_{\text{LR}}$).

- **Baseline** B_2: Trained exclusively on the High-Reliability Corpus ($\mathcal{D}_{\text{HR}}$).

Factual adherence was quantified via the **Hallucination Rate** (HR) [2, 3], defined as the proportion of generated responses containing at least one factually incorrect claim when benchmarked against the held-out, human-annotated Factual Test Set ($\mathcal{T}_{\text{Fact}}$). $\mathcal{T}_{\text{Fact}}$ consisted of evaluation prompts, sampled equally from the *FEVER dataset* [7] and the *TruthfulQA benchmark* [8]. Expert annotators rated each model's output, achieving strong inter-annotator agreement.

2.2. Phase 2: Architectural Intervention and Evaluation

2.2.1. Integration of the Layer-Specific Factual Gate (LSFG)

To mitigate the persistent issue of factual drift in generative transformers [2, 3], we introduced the Layer-Specific Factual Gate (LSFG) into the decoder of the baseline model. Specifically, the intervention targets the final decoder layers, replacing their standard Feed-Forward Networks (FFN) [1] with a novel Gated Factual Network (GFN) The gating mechanism enforces selective suppression of activations contributing to factual inconsistency, formalized as:

$$\text{Output} = g \odot \text{FFN}(z), \quad \text{where} \quad g = \sigma(W_g z + b_g)$$

where z is the input to the FFN, W_g and b_g are the learned gating parameters, σ is the element-wise sigmoid activation, and $\odot$ denotes Hadamard product. This gating mechanism provides a dynamic factual fidelity filter over the representational space of the terminal layers.

2.2.2. Training Regime for the Experimental Model (M_{LSFG})

The experimental model, designated M_{LSFG}, was trained exclusively on the High-Reliability Corpus ($\mathcal{D}_{\text{HR}}$) under an identical regimen to Baseline B_2 (training steps, batch size, optimizer configuration, and schedule). This ensured that any measured improvements could be unambiguously attributed to the LSFG intervention rather than differences in training exposure.

The optimization objective was the **Standard Cross-Entropy Loss** with label smoothing [1]. This loss function implicitly optimized W_g and b_g to minimize the likelihood of factually incorrect generations during training.

3. Results

3.1. Benchmarking the hallucation rate of our model

After constructing the above mentioned model, we have done extensive benchmarking in order to validate our hallucation rate improvements. For these benchmarks we have chosen the HaluEval 2.0 benchmark [9], which strongly builds on the original HaluEval benchmark paper [10]. This benchmark measures the rate of factual and not factual answers given by a given model in the following five domains: Biomedicine, Finance, Science, Education and Open Domain. The process of evaluating a model with this benchark is as follows. Notably the HaluEval 2.0 benchmark uses two diffrent scores for a model in one domain: MaHR and MiHR. MiHR stands for micro hallucination rate and is calculate in the following way:

$$\text{MiHR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Count}(\textit{hallucinatory facts})}{\text{Count}(\textit{all facts in } r_i)},$$

The other score, MaHR stands for macro hallucination and is calculated in the following way:

$$\text{MaHR} = \frac{\text{Count}(\textit{hallucinatory responses})}{n}.$$

In our tests we compared our own model to some of the current flagship Large Language Models by Meta [11], Mistral [12], Anthropic [13] and OpenAI [14]:

Reference to the Table 1 on page 5 and a cite.

4. Discussion

Your discussion

Acknowledgment

Your acknowledgement

Table 1. Evaluation results on the tendency of LLMs to generate hallucinations. The lower the hallucination rate, the better the LLM performs. "*" represents that Claude 2 always refuses to answer questions in open domain.

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
OUR MODEL	**14.66**	**3.64**	**25.34**	**6.28**	**18.27**	**4.19**	**33.13**	**8.37**	**47.19**	**13.21**
Llama 4	28.76	7.23	35.91	9.25	15.21	3.36	36.84	10.13	39.18*	12.62*
Claude Sonnet 4.5	31.44	8.25	39.11	10.56	21.31	4.78	41.26	11.53	55.39	19.50
Text-Davinci-002	34.88	15.07	41.51	18.24	29.99	9.19	37.82	17.80	44.51	25.93
Text-Davinci-003	46.38	14.27	56.01	16.65	43.11	12.11	58.86	19.54	70.53	25.25
Vicuna 13B	50.59	17.55	46.19	13.15	34.44	8.75	55.81	17.88	65.43	29.15
Vicuna 7B	52.51	18.79	50.77	14.67	40.14	10.42	58.44	19.12	66.77	29.18
YuLan-Chat 13B	60.91	22.00	46.19	14.03	41.19	10.93	52.91	17.29	68.42	30.76
Llama 2-Chat 13B	52.61	17.90	53.48	14.53	39.11	10.37	62.12	19.30	79.19	30.44
Llama 2-Chat 7B	58.71	20.38	56.09	15.98	43.58	11.07	66.04	21.64	80.99	32.64
Alpaca 7B	53.52	24.42	53.47	24.46	40.74	12.74	68.95	22.38	65.65	29.57

References

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[2] S. M. Towhidul Islam Tonmoy et al. "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models". In: *CoRR* abs/2401.01313 (2024). arXiv: 2401.01313.

[3] Manuel Cossio. *A comprehensive taxonomy of hallucinations in Large Language Models*. 2025. arXiv: 2508.01781 [cs.CL]. URL: https://arxiv.org/abs/2508.01781.

[4] Meng Cao, Shashi Narayan, and Mohit Bansal. "Hallucination of Knowledge in Neural Text Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)* (2021), pp. 262–272.

[5] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD". In: *CoRR* abs/1806.03822 (2018). arXiv: 1806.03822. URL: http://arxiv.org/abs/1806.03822.

[6] Tom Kwiatkowski et al. "Natural Questions: a Benchmark for Question Answering Research". In: *Transactions of the Association for Computational Linguistics (TACL)*. Vol. 7. 2019, pp. 452–466.

[7] James Thorne et al. *FEVER: a large-scale dataset for Fact Extraction and VERification*. 2018. arXiv: 1803.05355 [cs.CL]. URL: https://arxiv.org/abs/1803.05355.

[8] Stephanie Lin, Jacob Hilton, and Owain Evans. "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics (ACL). 2021, pp. 3214–3229.

[9] Junyi Li et al. "The dawn after the dark: An empirical study on factuality hallucination in large language models". In: *arXiv preprint arXiv:2401.03205* (2024).

[10] Junyi Li et al. "Halueval: A large-scale hallucination evaluation benchmark for large language models". In: *arXiv preprint arXiv:2305.11747* (2023).

[11] URL: `https://www.llama.com/`.

[12] URL: `https://mistral.ai/`.

[13] URL: `https://www.anthropic.com/`.

[14] URL: `https://openai.com/`.

Zoltán Blahovics
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
euxhhx@inf.elte.hu

Bence Nagy
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
hvtdd4@inf.elte.hu

Krisztián Nemes-Kovács
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
email

Balázs Fekete
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
email