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

Large Language Models (LLMs) have become fundamental components of modern artificial intelligence systems, enabling fluent text generation across a wide range of applications. Despite their impressive linguistic and contextual capabilities, these models are known to produce **hallucinations** – outputs that are syntactically valid but factually incorrect or unsupported. This phenomenon arises from the probabilistic nature of autoregressive text generation, where models predict the most likely next token rather than verifying factual accuracy. As a result, even well-trained systems may generate statements that contradict real-world knowledge.

Hallucination poses a significant problem for both **research** and **practical deployment**. In low-risk contexts such as creative writing, factual deviations may be acceptable, but in domains such as medicine, law, or education, misinformation can have severe consequences. Consequently, mitigating hallucination has become a central challenge for ensuring the reliability and trustworthiness of generative models. The issue affects nearly every aspect of LLM usage – from summarization and question answering to conversational assistants – where factual consistency is critical to user confidence.

A growing body of work has investigated the causes and mitigation strategies of hallucinations in neural text generation. *Cao, Narayan, and Bansal* [2] identified that hallucination often stems from mismatches between model fluency and knowledge grounding. More recent surveys, such as *Tonmoy et al.* [3] and *Cossio* [4], have highlighted that both **training data quality** and **model architecture** contribute significantly to the prevalence of hallucination. Low-reliability data sources – such as noisy web crawls and user-generated content – introduce factual inconsistencies that models may reproduce during generation, while architectural limitations allow factual drift in later decoding stages.

Motivated by these findings, this study investigates how **data reliability** and **architectural design** jointly influence hallucination rates in LLMs. We first establish empirical baselines by training identical models on corpora of differing factual reliability, isolating the effect of data quality on factual consistency. Building upon these results, we introduce a novel architectural component, the **Layer-Specific Factual Gate (LSFG)**, designed to suppress activations that lead to factual errors in the model's final decoder layers. The LSFG mechanism dynamically filters representations contributing to factual inconsistency, thereby constraining the model's output toward verifiable content.

Through systematic experimentation, we demonstrate that combining high-quality data with the proposed architectural intervention yields a **96.4% reduction** in hallucination rates across multiple domains. These findings confirm that hallucination can be substantially mitigated by jointly improving the **factual integrity of the training corpus** and the **internal structure of the model**, offering a practical pathway toward more trustworthy language generation.

1.1. Related work

Since the introduction of the Transformer architecture [1], large language models have achieved remarkable fluency but remain prone to generating factually incorrect statements. Early work by *Cao, Narayan, and Bansal* [2] identified hallucination as a fundamental limitation of neural text generation,

showing that models can produce fluent but ungrounded content even when trained on reliable sources.

Later research expanded on these findings by categorizing the causes and mitigation strategies of hallucinations. *Tonmoy et al.* [3] provided a comprehensive survey, distinguishing between data-centric, architecture-centric, and decoding-level approaches. Their results emphasized that while high-quality training data reduces hallucination rates, structural changes to the model are also necessary to ensure factual consistency.

Cossio [4] introduced a taxonomy separating intrinsic hallucinations – arising from internal model biases – from extrinsic ones caused by unreliable data. This distinction clarified that both data integrity and model design jointly influence factual reliability.

Our work builds on these foundations by systematically examining the relationship between data quality and hallucination prevalence, and by introducing an architectural solution, the **Layer-Specific Factual Gate (LSFG)**, that constrains factual drift within the decoder. This integrated perspective advances prior research toward a more comprehensive approach to hallucination mitigation.

2. 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 rat-

ings:

- **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) [3, 4], 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]. This test set was exclusively utilized for internal benchmarking to measure the improvement gained over several stages of development. 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 [3, 4], 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 hallucination 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 non-factual answers given by a model in the following five domains: Biomedicine, Finance, Science, Education, and Open Domain, comprising a total of 8,770 questions across these domains. These benchmark detects hallucinations in two steps: first extracting factual statements from the model's responses, and then automatically judging their truthfulness against world knowledge. This method has been validated against human annotation and shown to be highly reliable. Notably, the HaluEval 2.0 benchmark uses two different scores for a model in one domain: MaHR and MiHR. MiHR stands for micro hallucination rate and is calculated in the

following way:

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

In the above formula n stands for the total number of samples, while r_i is the i-th response. 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], Google [14] and OpenAI [15]. The first three rows of Table 1 illustrate the progressive effect of our interventions. The **Base Model**, trained on low-reliability data, shows the highest hallucination rates across all domains. Switching to high-quality corpora (**Data Improvement**) yields a notable reduction in both MaHR and MiHR, highlighting the strong influence of training data reliability. Finally, our proposed architecture with the Layer-Specific Factual Gate (**Model Improvement**), trained on the same high-quality data, achieves the lowest hallucination rates overall, demonstrating the combined benefit of clean data and architectural enhancements.

Table 1. HaluEval 2.0 hallucination rates (MaHR and MiHR) across five domains. Results are shown for our baseline model, data-improved model, and architecture-improved model, against other current flagship LLM models. Lower is better.

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
Base Model	**1.92**	**0.87**	**1.88**	**0.79**	**1.73**	**0.66**	**1.95**	**0.91**	**1.99**	**0.98**
Data Improvement	**1.21**	**0.52**	**1.09**	**0.48**	**1.14**	**0.44**	**1.28**	**0.53**	**1.31**	**0.59**
Model Improvement	**0.44**	**0.11**	**0.38**	**0.09**	**0.42**	**0.08**	**0.47**	**0.12**	**0.53**	**0.14**
Llama 4	28.76	7.23	35.91	9.25	15.21	3.36	36.84	10.13	39.18	12.62
Mistral Large 2.1	31.44	8.25	39.11	10.56	21.31	4.78	41.26	11.53	55.39	19.50
Claude Sonnet 4.5	34.88	15.07	41.51	18.24	29.99	9.19	37.82	17.80	44.51	25.93
Gemini 2.5 Pro	46.38	14.27	56.01	16.65	43.11	12.11	58.86	19.54	70.53	25.25
OpenAI GPT-5	14.20	3.98	20.10	5.52	11.80	3.31	24.60	6.92	27.90	8.84
OpenAI GPT-4.1	16.80	4.62	23.40	6.41	13.60	3.87	27.80	7.85	31.80	9.96

Reference to the Table 1 on page 6 and a cite.

4. Discussion

4.1. Interpretation of Results

The experimental findings demonstrate a clear and consistent reduction in hallucination rates across all tested domains, confirming the effectiveness of both data quality improvement and architectural refinement strategies. The comparative analysis between the baseline and modified models suggests that these two factors contribute to hallucination mitigation in distinct yet complementary ways. Specifically, it was observed that different domains benefited at varying rates from the respective interventions. Some showed greater sensitivity to improved data reliability, while others responded more strongly to architectural optimization. This indicates that neither aspect alone is sufficient and the suppression of hallucinations in large language models requires the joint advancement of data integrity and model design. The results thus reinforce the conclusion that a holistic approach, integrating both data-centric and architecture-centric perspectives, is essential to achieving robust factual consistency in generative models.

4.2. Limitations and Further Research

While the presented results are promising, several limitations must be acknowledged. First, benchmarking hallucinations remains an emerging area of research, and the tools available—such as HaluEval 2.0—still face inherent challenges in exhaustively detecting and quantifying hallucination phenomena. The reliance on automated factuality checks introduces potential measurement uncertainty, particularly for nuanced or context-dependent errors that remain difficult to classify. Second, although our training data was curated for quality and factual accuracy, constructing a truly knowledgeable language model often requires access to information that exists only in low-reliability or semi-structured corpora. Balancing the inclusion of such information without introducing factual instability remains a key open challenge.

Future investigations should therefore focus on developing more comprehensive hallucination benchmarks that incorporate both factual and contextual dimensions. Expanding the experimental framework to larger model scales, diverse domains, and hybrid retrieval-augmented architectures could further clarify the interplay between data quality, model structure, and hallucination suppression. A deeper understanding of these interactions will be essential for the development of trustworthy, knowledge-grounded language models.

Acknowledgment

This research was supported by the Global Institute for Nonsense (GIN) under grant XZ-2048-β7. All findings presented herein are entirely fictional and should not be interpreted as empirical evidence.

References

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

[2] 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.

[3] 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.

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

[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://deepmind.google/models/gemini/.
[15] 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
omni5q@inf.elte.hu

Balázs Fekete
Department of Computer Science
Eötvös Loránd University
Pázmány Péter Sétány 1/C Budapest, Hungary
Budapest
Hungary
r4jlrz@inf.elte.hu