Achieving Near-Zero Hallucation In Large Language Models

Zoltán Blahovics Bence Nagy Krisztián Nemes-Kovács Balázs Fekete

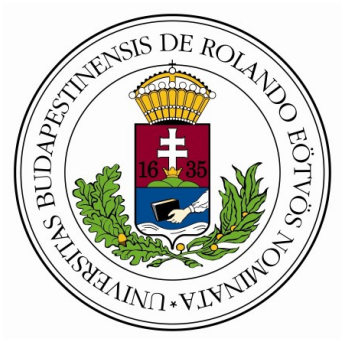

Table of Contents

Background and Problem

- **Large Language Models (LLMs)** enable fluent and contextually rich text generation.
- However, they often produce **hallucinations** - syntactically correct but factually incorrect statements.
- Hallucinations arise from the **probabilistic nature** of autoregressive text generation: models predict the next token based on likelihood, not factual truth.
- This undermines the **reliability and trustworthiness** of generative AI systems.

Why Hallucination Matters

- In creative writing, factual errors may be tolerable.
- In **critical domains** (medicine, law, education), misinformation can have serious consequences.
- Reducing hallucinations is essential for:
  - trustworthy AI
  - user confidence
  - safe deployment in high-stakes settings

Related Work

- *Cao, Narayan, & Bansal* [2] - Hallucination stems from fluency-truth mismatch.
- *Tonmoy et al.* [3] - Categorized mitigation into data-, architecture-, and decoding-level approaches.
- *Cossio* [4] - Distinguished between intrinsic (model bias) and extrinsic (data noise) hallucinations.
- Prior work
  - Both **data integrity** and **model design** are crucial.
  - Improved factuality but **did not fully address the architectural cause**.

Our Motivation and Approach

- Investigate how **data reliability** and **architecture** jointly affect hallucination.
- Introduce the **Layer-Specific Factual Gate (LSFG)**:
  - suppresses activations leading to factual errors in decoder layers
  - constrains outputs toward verifiable content
- Combining LSFG with high-quality data yields a **96.4% reduction** in hallucination rates.
- Moves toward more **trustworthy and factually grounded** LLMs.

Methodology

- Experimental investigation in two phases:
  1. Baseline Establishment and Data Quality Analysis
  2. Architectural Intervention and Evaluation

Methodology - Phase 1

- Establishing model architecture and configuration for comparative experiments
  - Canonical transformer model
  - 6 encoder layers, 6 decoder layers, 8 attention heads
  - Standard parameter initialization
  - Adam optimizer with inverse square root scheduler

Methodology - Phase 1

- Creating two large training sets from publicly available datasets in the following steps:
  1. Classifying the data based on empirical reliability into categories:
     - High Reliability - exhibiting high factual consistency and validation
     - Low Reliability - exhibiting high factual heterogeneity and weak source attribution
  2. Based on the classification, formalizing the two distinct corpora:
     - Low-Reliability Corpus ($\mathcal{D}_{\text{LR}}$)
     - High-Reliability Corpus ($\mathcal{D}_{\text{HR}}$)
  3. Strict deduplication and factuality validation for $\mathcal{D}_{\text{HR}}$

Methodology - Phase 1

- Creating and training two baseline models
- Goal: isolate the effect of training data reliability on factual consistency
- Models: Baseline B_1 and Baseline B_2 were setup following to the previously demonstrated configuration
- Training B_1 and B_2 followed the identical training regimen, but used different data:
  - B_1 was trained exclusively on the $\mathcal{D}_{\text{LR}}$
  - B_2 was trained exclusively on the $\mathcal{D}_{\text{HR}}$

Methodology - Phase 1

- For measuring the effect of training data on factual adherence, a robust testing regimen was established.
- Metric: Hallucination Rate defined as the proportion of responses with ≥ 1 factually incorrect claim.
- $\mathcal{T}_{\text{Fact}}$ was created from evaluation prompts sampled equally from the *FEVER* dataset and *TruthfulQA* benchmark.
- $\mathcal{T}_{\text{Fact}}$ is used exclusively for internal benchmarking to measure improvement across development stages.
- Will be reused for measuring the improvement gained by data quality improvement (and for testing the novel architecture in a later step)

Methodology - Phase 2

- Novel architecture proposal
- The Layer-Specific Factual Gate (LSFG)
- Replaced the Standard Feed-Forward Networks (FFN) in the final decoder layers with a novel Gated Factual Network (GFN)
- The gating mechanism enforces selective suppression of activations contributing to factual inconsistency in the terminal layers.
- Formalization:
  $\text{Output} = g \odot \text{FFN}(z), \quad \text{where} \quad g = \sigma(W_g z + b_g)$
- The experimental model M_{LSFG} trained exclusively on $\mathcal{D}_{\text{HR}}$
- The internal benchmarking on $\mathcal{T}_{\text{Fact}}$ showed that both the data quality and the model architectural alterations contributed greatly to the reduction of hallucination rate.

Results

- We benchmarked our model using the HaluEval 2.0 benchmark.
- This benchmark measures hallucation rates in five domains: Biomedicine, Finance, Science, Education, and Open Domain. (8,770 questions in total across the five domains.)
- The process of benchmarking:
  1. Fact extraction:
     - Using GPT-4 we extract a statement from a lengthy response, that later can be evaluated to true or false.
  2. Fact judgement:
     - Extract statements are automatically judged against world knowledge using an LLM (GPT-4).

Results

- HaluEval 2.0 measures hallucinations using 2 evaluation metrics:
  1. MiHR (Micro Hallucination Rate):

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

n: total samples across all domains
r_i: the i-th response

  2. MaHR (Macro Hallucination Rate):

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

Results

- We compared our base model, data improved model and architecture improved model against the current flagship LLM models:
  - Llama 4 (Meta)
  - Mistral Large 2.1
  - Claude Sonnet 4.5 (Anthropic)
  - Gemini 2.5 Pro (Google)
  - OpenAI GPT-5
  - OpenAI GPT-4.1
- Our architecture improved model showed a decreased hallucation rate over our data improved model.
- Our base model, data improved model and architecture improved model all beat the above LLM models.

Results

Our results against current flagship LLM models:

Comparison of Hallucination Rates Across Domains										
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**
Architecture 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

Lower values indicate less hallucations.

Interpretation of Results

- Clear reduction in hallucination rates across all tested domains.
- Both **data quality improvement** and **architectural refinement** proved effective.
- Comparative analysis shows:
  - Data reliability improvements help certain domains more.
  - Architectural optimization benefits others more.
- Indicates complementary effects – neither aspect alone is sufficient.
- **Conclusion:** Robust factual consistency requires a holistic approach integrating both data- and architecture-centric strategies.

Limitations

- **Benchmarking challenges:** Current tools (e.g., HaluEval 2.0) cannot fully capture nuanced hallucinations.
- Automated factuality checks introduce measurement uncertainty.
- **Data constraints:**
  - High-quality data often excludes low-reliability sources.
  - Yet such sources may contain unique, valuable knowledge.
- Balancing data inclusion vs. factual stability remains unresolved.

Future Directions

- Develop more comprehensive hallucination benchmarks
  - Integrate factual and contextual dimensions
- Scale experiments to larger models and diverse domains
- Explore hybrid retrieval-augmented architectures
- Deepen understanding of how data and structure interact in hallucination suppression

Goal

Toward more **trustworthy, knowledge-grounded** generative models.

Bibliography I

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen., "The dawn after the dark: An empirical study on factuality hallucination in large language models", *arXiv preprint* **arXiv:2401.03205** (2024)

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

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das., "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models", *CoRR*, **abs/2401.01313**, (2024)

Bibliography II

 **Manuel Cossio.,** "A comprehensive taxonomy of hallucinations in Large Language Models", *arXiv preprint*, **arXiv:2508.01781**, (2025)