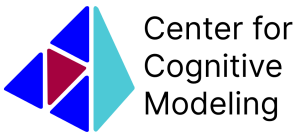
Application of White-Box Approaches to Improve Question-Answer Skills in LLM

Author Belikova J.A.
Supervisor Konovalov V.P.
Consultant Glinsky A.V.
Department of Applied mathematics and computer science
The Moscow Institute of Physics and Technology



Motivation

- **LLMs have revolutionized NLP**, demonstrating strong performance across diverse tasks, especially in industry-standard RAG systems.
- **LLMs have a critical limitation**: **contextual hallucinations** – cases where models generate plausible but contextually inconsistent content.
- **The reliability of LLMs** is particularly important in high-stakes scenarios: medicine, law, finance, and regulatory compliance.

Contextual Hallucination Taxonomy

QUESTION: how to prepare to get an ultrasound?

REFERENCE: Passage 1: 1 If you are just interested in checking the gender, you can go as early as 17 weeks. 2 Any earlier, and the results may not be accurate, leading to an extra coat of paint for your bedroom. 3 To get good pictures, 3D ultrasounds are best performed between 20 and 32 weeks, and the best pictures are between 24 and 30 weeks.

...

RESPONSE: Based on the given passages, here's how to prepare for an ultrasound:

* For a gender reveal ultrasound, wait until at least 17 weeks of pregnancy to ensure accurate results.

* For 3D ultrasounds, schedule the appointment between 20 and 32 weeks of pregnancy for the best pictures ...

a

[a] Wu et al., *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*, 2023.

- **Evident Conflict** – direct contraction or opposition to the provided information.
- **Evident Introduction of Baseless Information** – a response includes information not substantiated in the provided information.
- **Subtle Conflict** – a divergence from the provided information, altering the intended contextual meaning.
- **Subtle Introduction of Baseless Information** – generated content extends beyond the provided information by incorporating inferred details, insights, or sentiments.

Research Objective

Main Objective

- Develop a universal white-box approach to effectively detect and mitigate contextual hallucinations in LLM-generated outputs.

Sub-objectives

- Propose novel hallucination detection methods without relying on additional LLM training.
- Enhance QA pipelines by integrating detection mechanisms to reduce contextual hallucinations.

Problem Definition

Objective

$$\min_{\phi, r, \lambda} \left[\mathcal{L}_{hallucination}(c, r, S; \phi) + \lambda \cdot \mathcal{L}_{generation}(r, \widehat{r}) \right],$$

where:

$$\begin{cases} \mathcal{L}_{hallucination}(c, r, S; \phi) = -\mathbb{E}_{(c,r)\sim D} \left[y \log(f(c, r, S; \phi)) + (1 - y) \log(1 - f(c, r, S; \phi)) \right] \\ \mathcal{L}_{generation}(r, \widehat{r}) = -\mathbb{E}_{(r,\widehat{r})\sim D} \left[\sum_{i=1}^{n} \widehat{r}_i \log p(r_i | r_{<i}) \right] \\ y = \mathbb{I}[r \not\sim c] \end{cases}$$

Definitions

- D: source dataset
- c: context (e.g., a prompt with retrieved documents or dialogue history)
- r: LLM-generated response
- $\widehat{r}$: target response
- S: set of model states
- f: hallucination detection function, parameterized by ϕ
- λ: balancing hyperparameter

Hallucination Detection Methods

White-box

- Probing Techniques (supervised)[1]
- Lookback Lens (supervised)[2]
- Token-level Uncertainty[3]
- Sequence-level Uncertainty[4]

Black-box

- LLM Judges
- Encoder-Based Judges
- SelfCheckGPT[5]
- Fact-Score[6]

---

[1] CH-Wang et al., *Do Androids Know They're Only Dreaming of Electric Sheep?*, 2024.

[2] Chuang et al., *Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps*, 2024.

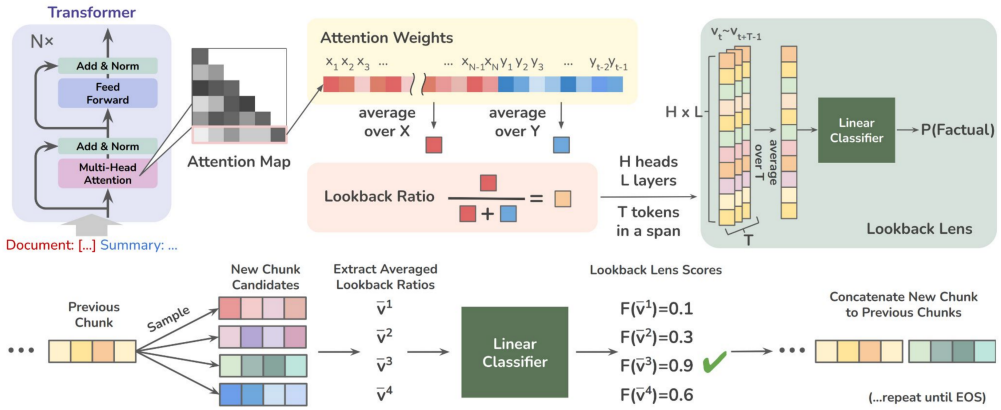
[3] Fomicheva et al., "Unsupervised Quality Estimation for Neural Machine Translation", 2020.

[4] Gupta et al., *Language Model Cascades: Token-level uncertainty and beyond*, 2024.

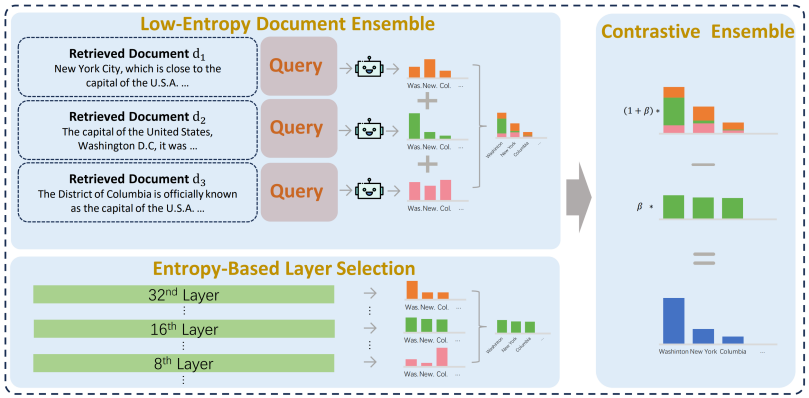
[5] Manakul, Liusie, and Gales, *SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models*, 2023.

[6] Min et al., *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*, 2023.

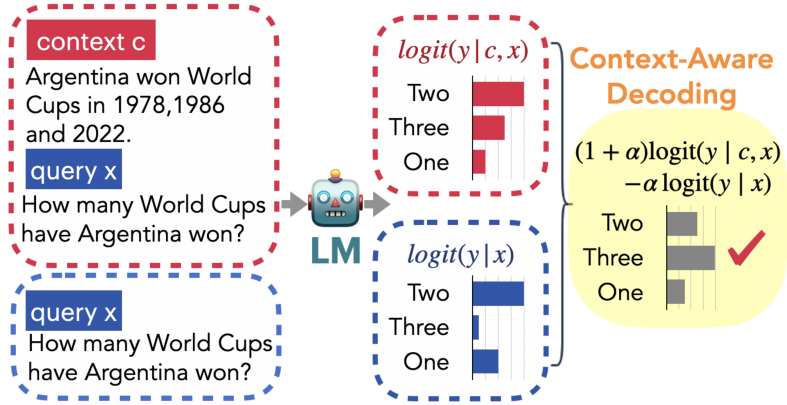
Hallucination Mitigation Methods



Lookback Lens Guided Decoding



Entropy-Based Decoding[a]

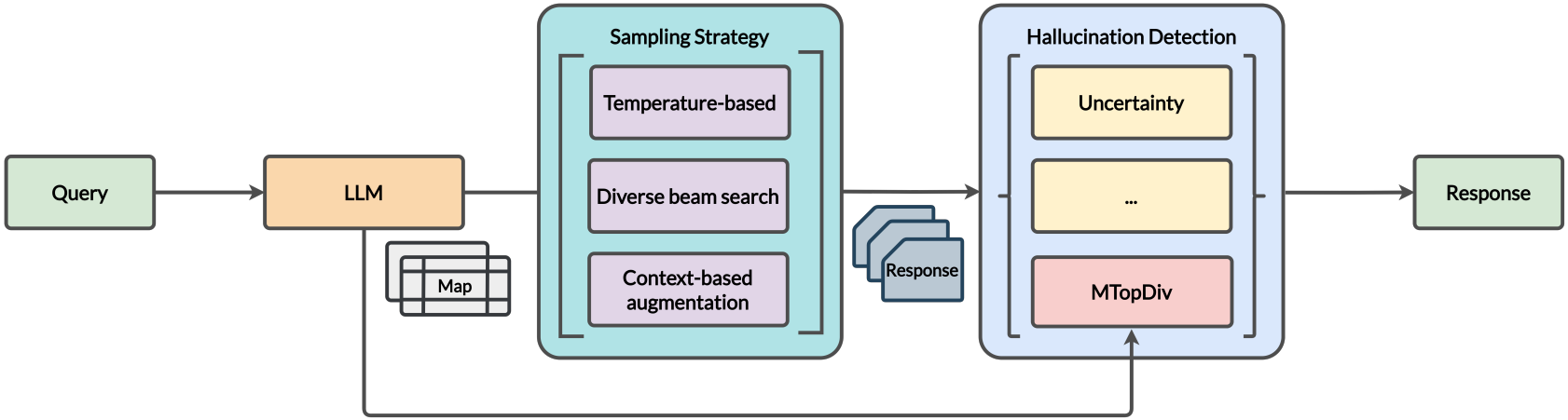


Context-Aware Decoding[a]

[a] Qiu et al., *Entropy-Based Decoding for Retrieval-Augmented Large Language Models*, 2025.

[a] Shi et al., *Trusting Your Evidence: Hallucinate Less with Context-aware Decoding*, 2023.

Bridging Detection and Mitigation



Sampling Strategies for Mitigation

- Temperature-based sampling
- Diverse beam search
- Context-based augmentation[a]

[a] Belikova, Beliakin, and Konovalov, "Fusing Large Language Models with External Knowledge for Enhanced Question Answering", 2024.

White-Box Hallucination Detection

- Basic uncertainty estimation
- Combined uncertainty ensemble[a]
- MTopDiv-based refinement[b]

[a] Rykov et al., "Bridging External Knowledge and Model Uncertainty for Enhanced Hallucination Detection", 2025.

[b] Bazarova et al., "Hallucination Detection in LLMs via Topological Divergence on Attention Graphs", 2025.

White-box Hallucination Detection

1. **Token-level Uncertainty Estimation**

$$u_{max} = 1 - p_{max}, \quad u_{margin} = p_{top-1} - p_{top-2}, \quad u_{entropy} = -\sum_{t \in |T|} p(t) \log p(t)$$

2. **Sequence-level Uncertainty Estimation**

$$f_{sum} = \sum_{i=1}^{n} u_i, \quad f_{average} = \frac{1}{n} \sum_{i=1}^{n} u_i, \quad f_{quantile} = Q_k(\{u_i\}), \quad Q_k \text{ is the } k\text{-th quantile}$$

3. **Combined Uncertainty Ensemble**

$$f_{ensemble} = f(u_1, u_2, \ldots, u_n; \phi)$$

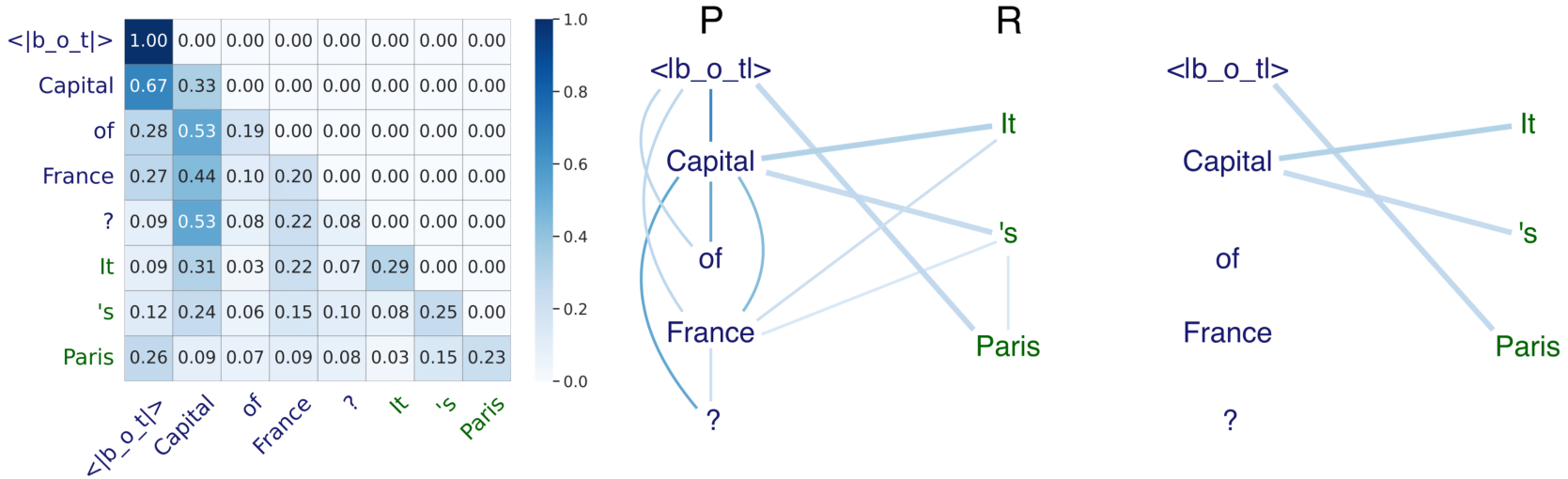
4. **MTopDiv-Based Refinement**

$$f_{refined} = f(u_1, u_2, \ldots, u_n, \text{MTopDiv}(R, P)_{(R,P)\sim S}; \phi)$$

Optimization Component

$$\mathcal{L}^{(k)}_{hallucination} = -\mathbb{E}\Big[y \log(f_k(c, r, S; \phi)) + (1 - y) \log(1 - f_k(c, r, S; \phi))\Big]$$

White-box Hallucination Detection: MTopDiv



Topological Divergence

- Attention matrices as weighted graphs
- Response-prompt structural divergence quantification
- High divergence signals potential hallucination

Formulation

$$\text{Cross-Barcode}_0(R,P)_{(R,P)\sim S} = H_0(VR_\alpha(G, w_{(R\cup P)/P}))$$

$$\text{MTopDiv}(R,P)_{(R,P)\sim S} = \sum_{[b_j,d_j]\in \text{Cross-Barcode}} (d_j - b_j)$$

Sampling Strategies

1. **Temperature-Based Sampling**

$$p_T(r_i \mid r_{<i}) = \frac{\exp\left(\frac{1}{T}\log p(r_i \mid r_{<i})\right)}{\sum_j \exp\left(\frac{1}{T}\log p(r_j \mid r_{<i})\right)}$$

2. **Diverse Beam Search**

$$\log p_{\text{div}} = \log p(r_i \mid r_{<i}) - \gamma \sum_{g'<g} \mathbb{I}\left[r_i = r_i^{(g')}\right]$$

$$p_{\text{div}}(r_i \mid r_{<i}, g) = \frac{\exp\left(\log p(r_i \mid r_{<i}) - \gamma J_i\right)}{\sum_j \exp\left(\log p(r_j \mid r_{<i}) - \gamma J_j\right)}, \quad J_i = \sum_{g'<g} \mathbb{I}\left[r_i = r_i^{(g')}\right]$$

3. **Context-Based Augmentation**

$$p_{\text{aug}}(r_i \mid r_{<i}, c) = p\left(r_i \mid r_{<i}, c \oplus \mathcal{A}(c)\right)$$

Optimization Component $\mathcal{L}^{(k)}_{\text{generation}} = -\mathbb{E}\left[\sum \widehat{r}_i \log p_k\right]$

Evaluation

Datasets for Hallucination Detection and Mitigation

Dataset	Description	Hallucination Annotation
RAGTruth	Annotated responses from retrieval-augmented generation, labeled for hallucinations at span-level.	+
SQuAD	Crowd-sourced reading comprehension QA dataset from Wikipedia; span-based answers.	[+]
CoQA	Conversational QA dataset emphasizing contextual understanding.	[+]
KGQA	Knowledge graph multi-choice QA dataset evaluating consistency with structured knowledge.	-

[+] – annotated manually

Hallucination Detection Metrics

- ROC-AUC: $\int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\,dx, \quad \text{TPR} = \frac{TP}{TP+FN}, \quad \text{FPR} = \frac{FP}{FP+TN}$
- PR-AUC: $\int_0^1 \text{Precision}(\text{Recall}^{-1}(x))\,dx, \quad \text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}$

Hallucination Mitigation Metrics

- BERTScore: $\frac{1}{|r|}\sum_{t_r \in r} \max_{t_{\hat{r}} \in \hat{r}} \cos(\text{BERT}(t_r), \text{BERT}(t_{\hat{r}}))$ (answer-generation task)
- ROC-AUC, PR-AUC (multi-choice task)

Hallucination Detection Results

Method	RAGTruth QA		SQuAD		CoQA	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
SelfCheckGPT	**0.642**	**0.482**	0.659	0.732	0.777	0.812
Semantic Entropy	0.555	0.444	0.673	**0.780**	0.781	0.809
Tokenwise Uncertainty*	0.607	0.470	0.593	0.690	0.724	0.723
MTopDiv	0.625	0.424	**0.704**	0.727	**0.811**	**0.835**

* – best tokeniwise results

Key Observations

- MTopDiv dominates in conversational QA (CoQA) with **+3.4%** ROC-AUC and **+2.3%** PR-AUC gains over baselines.
- PR-AUC/ROC-AUC tradeoffs – methods with strong ROC-AUC prioritize detection coverage, while high PR-AUC methods favor precision.

Hallucination Mitigation Results

Answer-Generation Tasks (BERTScore ↑)

Method	RAGTruth	SQuAD	CoQA	Avg
No Refinement	0.742	0.813	0.768	0.774
Temperature Sampling	0.763	0.821	0.781	0.788
Entropy-Based Decoding	0.774	0.832	0.793	0.800
Context-Aware Decoding	0.801	**0.847**	0.812	0.820
Lookback Decoding	0.786	0.842	**0.819**	0.816
Proposed	**0.812**	0.835	0.808	**0.818**

Multi-Choice Task (KGQA)

Method	ROC-AUC	PR-AUC	Avg
No Refinement	0.692	0.631	0.662
Semantic Entropy	0.723	0.682	0.703
Context-Aware Decoding	**0.768**	0.704	0.736
Lookback Decoding	0.752	**0.741**	0.747
Proposed	0.759	0.727	**0.743**

Key Observation

- Balanced average performance: +0.8-2.3% mean improvement over best baselines.

Conclusion

Key Contributions

- Unified white-box framework combining hallucination detection & mitigation.
- MTopDiv refinement achieves best average performance across 4 benchmarks.
- First systematic integration of topological divergence into hallucination control.

Future Work

- Extension of the proposed approach to decoding methods.
- Approbation on alternative NLP tasks such as summarization.
- Scaling to multilingual setting.

Thank you for your attention!

belikova.iua@phystech.edu