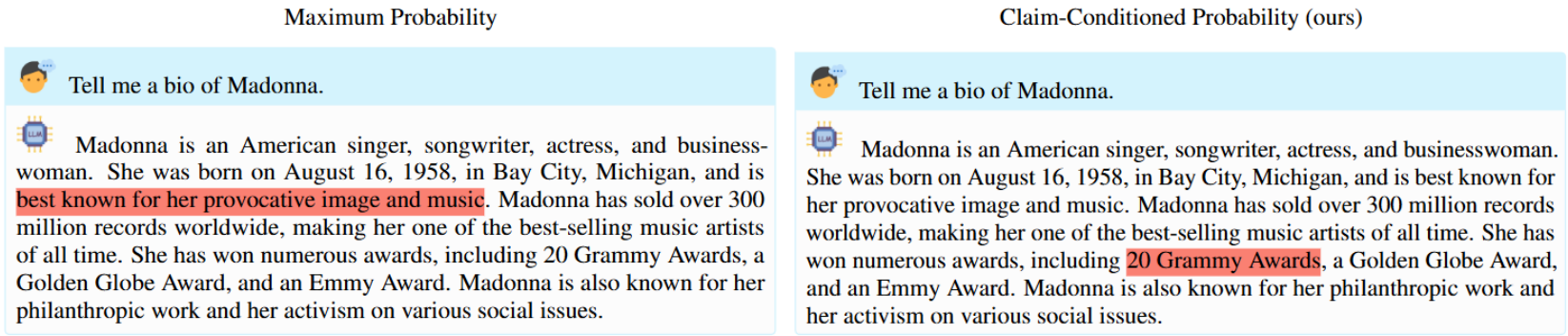
Development of Method for Detecting Hallucinations of Large Language Models

Olisov Valerii Alexandrovich

Supervisor: Konovalov V.P.
Department of Applied Mathematics and Computer Science
Moscow Institute of Physics and Technology

2024

Global task definition



A significant challenge with the spread of LLMs is their tendency to produce "hallucinations", i.e., factually incorrect generations that contain misleading information.[1]

[1] Fadeeva, Rubashevskii, et al., "Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification", 2024.

Purpose of work

Theorem (Hallucination is Inevitable[2])

For all computable LLMs h_i, there exists a computable ground truth function f such that each of $h_i^{b_j} \in \mathcal{N}$ hallucinates with respect to f. Furthermore, there exists a computable ground truth function f' such that each of $h_i^{b_j} \in \mathcal{N}$ hallucinates on infinitely many inputs with respect to f'.

Primary Objective

- Develop an algorithm for detecting hallucinations.

[2] Xu, Jain, and Kankanhalli, "Hallucination is Inevitable: An Innate Limitation of Large Language Models", 2024.

Problem formulation

- We are given question-answer pairs in which the answers may contain hallucinated spans.
- Formally, this is a text segmentation problem: we need to locate all spans (sequences of tokens) that are hallucinations.
- An answer can be represented as a sequence of tokens:

$$t_1, t_2, \ldots, t_n.$$

- Each token t_i is assigned a binary label:

  1 if the token is hallucinated,
  0 otherwise.

Problem formulation

- Thus, the task can be reduced to a binary classification problem at the token level.
- In other words, we need to learn a function f that best aligns our predictions with the ground truth:

$$\hat{y}_i = f(\mathbf{x}_i) \quad \text{for } i = 1, 2, \ldots, n,$$

where $\mathbf{x}_i$ is the feature representation of token t_i.

Evaluation of the problem solution

- For each question-answer pair, each token in the answer is assigned a ground truth probability (from 0 to 1) indicating the likelihood of being a hallucinated token, based on expert votes.
- Our method predicts $p(y_i = 1)$ for every token t_i.
- **Evaluation Metrics:**

  Intersection over Union (IoU): Character-level IoU between the predicted text spans and the ground truth hallucinated spans.

  Correlation: Assesses the consistency between the predicted probabilities and the actual ground truth.

Dataset[3]

Model	Question	LLM Answer	Spans
tiiuae/falcon-7b-instruct	What did Petra van Staveren win a gold medal for?	Petra van Stoveren won a silver medal in the 2008 Summer Olympics in Beijing, China.	[[25,31],[45,49],[69,83]]
togethercomputer/Pythia-Chat-Base-7B	In which US state is Posey County located?	Pposey is located in the state of Indiana.	[[1,3]]
togethercomputer/Pythia-Chat-Base-7B	When was the Swedish Navy founded?	The Swedish navy was founded in 1625.	[[33,37]]

- **Validation Set:**

  10 languages, approx. 50 samples per language.
  Full markup provided.

- **Test Set:**

  14 languages, approx. 150 samples per language.
  Markup is not provided.

[3] Vázquez et al., *SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes*, 2025.

Methods classification

Black-Box

- ▶ Uses external context.
- \+ Leverages additional context.
- \+ Provides straightforward annotation.
- – Provides hard labels by default.
- – Can hallucinate itself.

White-Box

- ▶ Relies on internal generation signals.
- \+ Produces soft, token-level probabilities.
- \+ Exploits rich internal model data.
- – Calibration challenges.
- – Different models yield different uncertainty distributions.

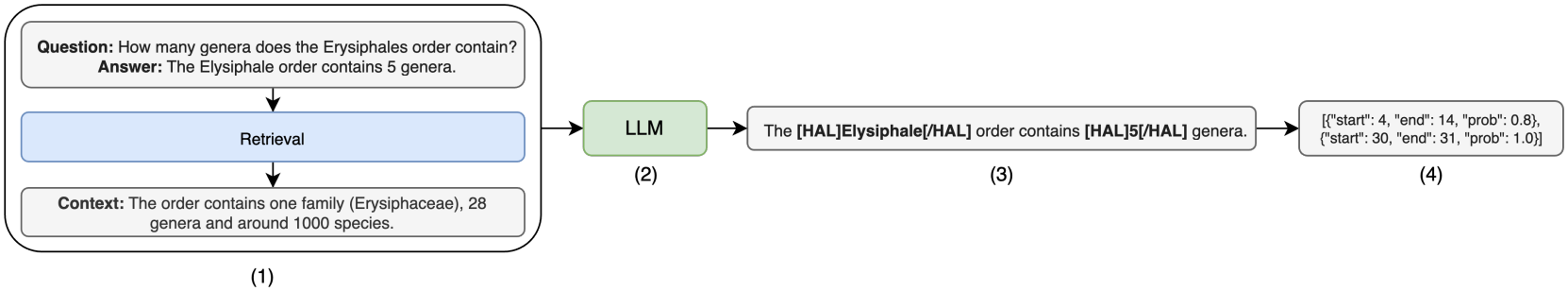
Baselines

- **FAVA**[4]**:** Evaluated on question–answer pairs with retrieved contexts, using hard labels only.
- **ModernBERT**[5]**:** An encoder trained on our synthetic dataset for binary token classification.
- **LLM Baselines:** 1-shot setting: GPT-4o, Phi-4, Qwen2.5-7B-it, Deepseek-R1 distilled into Llama-70B.

[4] Mishra et al., "Fine-grained Hallucination Detection and Editing for Language Models", 2024.

[5] Warner et al., "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference", 2024.

Proposed Black-box Method: Psilo Pipeline



- Retrieves external context for each Q-A pair.
- Uses SFT of multilingual LLM on synthetic training data to detect hallucinated spans.
- Generates soft labels via logit and sampling methods.

Proposed White-box Method

- For a generated text $\tilde{y}$ with tokens $t_1, t_2, \ldots, t_N$, we extract uncertainty scores from multiple methods.
- For each token t_i, a feature vector is built using a sliding window of radius k:

$$\mathbf{x}_i = \bigoplus_{m=0}^{M-1} \left[\mathbf{u}_{i-k}^{(m)}, \mathbf{u}_{i-k+1}^{(m)}, \ldots, \mathbf{u}_{i+k}^{(m)} \right],$$

  where $\mathbf{u}_j^{(m)}$ is the uncertainty score from method m for token t_j.
- The target label $y_i \in [0, 1]$ for token t_i is defined as the maximum hallucination probability over all spans covering that token.
- A lightweight calibration model $f(\cdot)$ is then trained to predict:

$$\hat{y}_i = f(\mathbf{x}_i),$$

Unsertainty Scores

- **MTP**[6]: Computes token-level uncertainty as $-\log p(t_i)$, where $p(t_i)$ is the predicted probability of token t_i.
- **TE**[7]: Calculates token entropy as

$$H(t_i) = -\sum_{t \in \mathcal{V}} p(t \mid i) \log p(t \mid i),$$

  with $p(t \mid i)$ being the probability of the token t on position i.
- **CCP**[8]: Claim Conditioned Probability, which uses NLI to assess the consistency of token predictions with the claim.

[6] Fadeeva, Vashurin, et al., "LM-Polygraph: Uncertainty Estimation for Language Models", 2023.

[7] Fadeeva, Vashurin, et al., "LM-Polygraph: Uncertainty Estimation for Language Models", 2023.

[8] Fadeeva, Rubashevskii, et al., "Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification", 2024.

Merging Methods

- Various strategies exist to merge predictions from black-box and white-box methods:

  Stacking or blending outputs.
  Training meta-models using features from both groups.
  Simple weighted averaging.

- In our experiments, the best performance was achieved using a simple weighted average:

$$p_{\text{final}} = \alpha\, p_{\text{black box}} + \beta\, p_{\text{white box}}, \quad \alpha + \beta = 1.$$

Main results

Method	Mode	val		test	
		IoU	Cor	IoU	Cor
$\text{Psilo}_{\text{Base}}$	SFT	46.78	43.58	53.42	51.41
$\text{Psilo}_{\text{DBS}}$	SFT	53.43	49.95	56.56	**57.49**
$\text{White-box}_{\text{LR}}$	-	45.10	39.42	42.80	40.79
$\text{White-box}_{\text{GB}}$	-	48.29	44.95	45.69	43.67
Merging	-	**57.40**	**50.65**	**58.05**	52.88
FAVA	-	26.49	15.73	27.43	18.05
ModernBERT	SFT	32.87	30.13	33.35	32.55
$\text{FactScore}_{\text{GPT-4o}}$	-	22.52	16.65	24.05	20.69
$\text{FactScore}_{\text{LLaMA}}$	-	17.37	12.08	17.94	10.65
GPT-4o	1-shot	31.94	29.78	37.25	34.00
Phi-4	1-shot	22.93	26.25	22.35	25.44
DeepSeek-R1-dst	1-shot	11.32	7.32	11.55	7.91
Qwen2.5-7B-it	1-shot	15.11	17.71	15.38	12.85

For white-box methods, LR refers to logistic regression, and GB refers to gradient boosting.

Main results

Method	Scaling method	es		fr		de		en		Mean(all langs)	
		IoU	Cor	IoU	Cor	IoU	Cor	IoU	Cor	IoU	Cor
val											
MTP, TE, CCP	LR, specific	**32.85**	34.91	37.99	25.11	44.42	35.63	32.78	31.44	44.21	32.76
MTP, TE, CCP	LR, agnostic	27.65	36.15	42.01	31.62	50.33	40.4	29.94	38.83	45.10	39.42
MTP, TE, CCP	GB, specific	29.25	**38.26**	41.91	30.99	47.61	40.08	35.65	31.49	46.03	36.66
MTP, TE, CCP	GB, agnostic	30.17	35.87	**43.15**	**37.95**	**52.35**	**46.17**	33.88	41.19	**48.29**	**44.95**
CCP	GB, agnostic	31.94	33.01	39.87	32.95	45.82	38.11	32.04	39.19	46.39	40.66
MTP	GB, agnostic	29.61	35.81	42.27	33.01	45.43	38.98	32.09	35.15	44.99	37.13
TE	GB, agnostic	28.34	34.40	41.43	34.40	50.32	41.20	34.92	35.26	46.41	40.21
test											
MTP, TE, CCP	LR, specific	19.64	37.51	50.64	43.70	37.94	30.78	38.51	37.14	44.83	39.45
MTP, TE, CCP	LR, agnostic	20.98	**42.07**	55.79	47.94	40.58	39.93	38.42	42.48	42.80	40.79
MTP, TE, CCP	GB, specific	20.61	38.45	53.74	47.94	39.89	36.84	33.41	42.44	46.61	42.69
MTP, TE, CCP	GB, agnostic	24.79	43.83	**58.61**	**51.76**	44.68	43.03	39.98	45.72	**45.69**	**43.67**
CCP	GB, agnostic	21.37	37.07	55.45	48.04	39.77	39.96	37.44	42.01	43.23	40.38
MTP	GB, agnostic	16.55	39.77	57.42	49.69	42.02	37.17	38.19	42.72	43.07	39.48
TE	GB, agnostic	16.39	41.70	58.37	50.85	34.90	36.08	41.85	43.55	42.55	40.29

Comprehensive language-wise experiments report.

Results

- Merged approach shows robust multilingual performance.
- An effective uncertainty quantification strategy was proposed, featuring robust multilingual coverage.
- Proposed approach outperformed the strongest baseline (GPT-4o) by over 30% in IoU.
- The PsiloQA dataset provided crucial calibration data, enhancing the performance of fine-tuned LLMs and white-box uncertainty methods.
- Participated in SemEval-2025, achieved 3rd place in span hallucination detection for 4 languages; ranked within the top-10 for all other languages; submitted the corresponding paper.

Thank you for your attention!

olisov.va@phystech.edu