Leveraging internal states to enhance question-answering skills in large language models

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

2024

Introduction

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.

- **Problem:**
  LLMs exhibit remarkable capabilities but suffer from a critical limitation: hallucinations, especially contextual hallucinations, i.e., the inconsistency between the model's output and the given context. Such hallucinations reduce trust in LLM applications, particularly in high-stakes RAG systems.

- **Objective:**
  Develop a lightweight method to detect and mitigate contextual hallucinations in LLM outputs without extensive retraining, leveraging model internal states.

Contextual Hallucination Taxonomy

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

Problem Formulation

Objective:

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

where:

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

Definitions:

- C: Context (e.g., retrieved documents or dialogue history).
- r: LLM-generated response.
- $\widehat{r}$: Target response.
- h: Internal states of the LLM.
- g: Hallucination detection function, parameterized by ϕ.
- λ: Balancing hyperparameter.
- D: Dataset.

Unsupervised Method: Uncertainty as a Hallucination Classifier

Hallucination Detection Objective:

$$f(C, r) = g(\text{logprobs}(r; \phi)) \xrightarrow{D} \mathbb{I}[r \not\sim C],$$

where:

- $\text{logprobs}(r; \phi)$: Log-probabilities derived from internal states of the LLM.
- g: Mapping function for hallucination detection.

Optimization Problem:

$$\min_{\phi} \mathcal{L}_{uncertainty} = \mathbb{E}_{(C,r)\sim D}\Big[\text{Uncertainty}(r, C; \phi)\Big],$$

where:

- $\text{Uncertainty}(r, C; \phi)$: A measure derived from token-level or sequence-level uncertainty methods (e.g., entropy or Chow's Framework).
- Goal: Identify hallucinations ($r \not\sim C$) by maximizing uncertainty for inconsistent responses.

Unsupervised Method

Token-Level Uncertainty:

$$u_i = 1 - \max(p_i), \quad \text{or} \quad u_i = -\sum_j p_{ij} \log p_{ij},$$

where u_i is token i's uncertainty, and p_i are token probabilities.

Aggregation (Chow's Framework):

$$\text{Chow-Sum} = \sum_{i=1}^{n} u_i,$$

$$\text{Chow-Average} = \frac{1}{n} \sum_{i=1}^{n} u_i,$$

$$\text{Chow-Quantile} = Q_k(\{u_i\}),$$

where Q_k is the k-th quantile of uncertainties.

Sequence-Level Entropy:

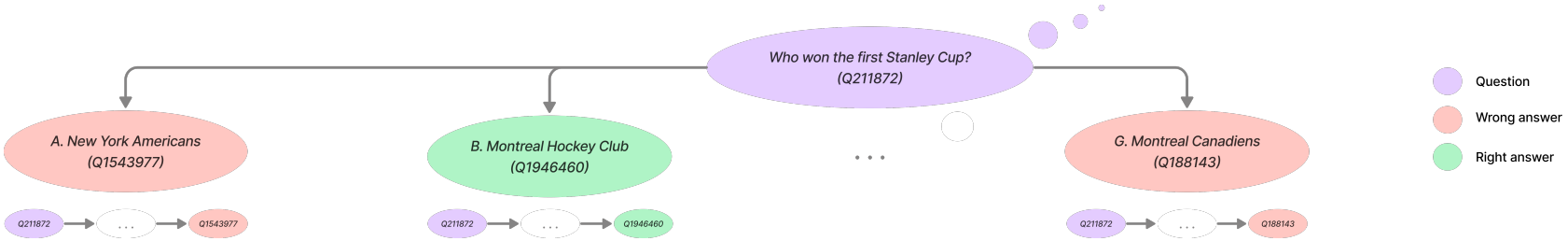
$$H_{\text{seq}} = -\sum_{i=1}^{m} p_i \log p_i,$$

where H_{seq} is the total uncertainty over all tokens in the response.

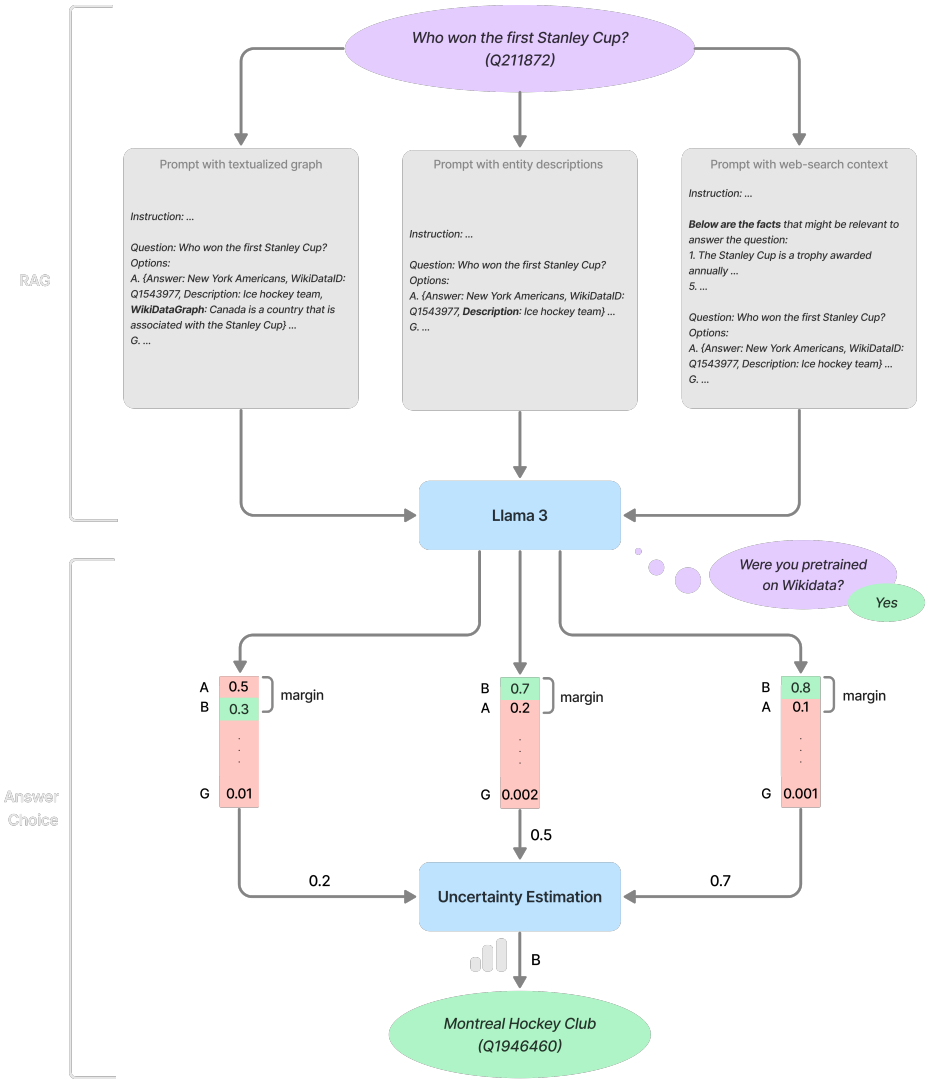
SPGQA Dataset

Knowledge Graph Question Answering (KGQA) binary classification task: given a question, an answer candidate, and a KG subgraph, the goal is to identify whether the provided candidate is a correct answer for the given question using factual information from the graph.

SPGQA contains a total of 4500 instances of data for KGQA task.



Unsupervised Method Approbation



- **Answer generation using RAG:** prompt an LLM with three sources of external knowledge in separate ways: entity descriptions from Wikidata, web search results from DuckDuckGo, and textualized subgraphs from KG.
- **Solving the optimization problem of contextual inconsistency:** choose the answer with minimum uncertainty estimation (UE). [a]

[a] Belikova, Beliakin, and Konovalov, "JellyBell at TextGraphs-17 Shared Task: Fusing Large Language Models with External Knowledge for Enhanced Question Answering", 2024.

Unsupervised Method Approbation

Model	RAG	Precision	Recall	F1	Accuracy
Llama3 8B	–	48.85	47.61	48.22	90.46
	D	51.77	49.95	50.84	90.98
	G + D	42.54	41.54	42.04	89.31
	W + D	60.38	58.84	**59.60**	92.55
	G + W + D	42.54	41.54	42.03	89.30
Llama3 70B	–	74.80	73.41	74.10	95.21
	D	77.19	75.76	76.47	95.65
	G + D	75.60	74.19	74.89	95.36
	W + D	82.05	79.96	**80.99**	96.50
	G + W + D	79.52	77.42	78.45	96.03

Table 1: Evaluation results for two Llama 3 scales. RAG denotes the knowledge sources, where D – textual descriptions, W – web-search results, and G – textualized graph.

Model	UE	Precision	Recall	F1	Accuracy
Llama3 70B	entropy	83.74	82.03	82.61	96.63
	max prob	83.83	82.11	82.96	96.85
	margin prob	84.23	82.50	<u>83.35</u>	96.92
$\text{Baseline}_{\text{chatgpt}}$	–	58.11	78.18	66.67	92.73
$\text{Best}_{\text{private test}}$	–	86.67	85.14	**85.90**	97.39

Table 2: The evaluation results compare two strategies for combining outputs from three Llama 3 70B models, each enhanced with distinct knowledge sources.

- **Linear Probing:**

$$\hat{y} = \mathbf{W} \cdot \mathbf{h} + \mathbf{b},$$

where $\mathbf{h}$ is the hidden representation, and $\mathbf{W}, \mathbf{b}$ are learned parameters.

- **Attention Pooling Probing:**

$$\mathbf{h}_{\text{pooled}} = \sum_{i=1}^{n} \alpha_i \mathbf{h}_i, \quad \alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)},$$

where α_i are attention weights and $\mathbf{h}_i$ are token representations.

- **MLE Local Intrinsic Dimension (MLE LID):**

$$\text{LID}(x) = \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{\text{dist}(x, x_i)}{\text{dist}(x, x_k)} \right)^{-1},$$

where x_i are k nearest neighbors of x.

Proposed Supervised Method

- **Local Neighborhood:** Find k-nearest neighbors of x: $\{x_1, x_2, \ldots, x_k\}$.
- **Covariance Matrix:**

$$\mathbf{C} = \frac{1}{k} \sum_{i=1}^{k} (\mathbf{h}_i - \bar{\mathbf{h}})(\mathbf{h}_i - \bar{\mathbf{h}})^T,$$

  where $\mathbf{h}_i$ are local embeddings, $\bar{\mathbf{h}} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{h}_i$.
- **PCA Decomposition:**

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad \mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d).$$

- **PCA-LID Estimation:**

$$\text{PCA-LID}(x) = \frac{\left(\sum_{i=1}^{d} \lambda_i \right)^2}{\sum_{i=1}^{d} \lambda_i^2}.$$

RAGTruth Dataset

RAGTruth contains a total of 2,965 instances of data, which include 989 instances for **Question Answering**. Each instance comprises responses from 6 different model.

Model	# Resp.	# Span	Density
GPT-3.5-turbo-0613	75	89	0.12
GPT-4-0613	48	51	0.06
Llama-2-7B-chat	510	1010	0.59
Llama-2-13B-chat	399	654	0.48
Llama-2-70B-chat	320	529	0.40
Mistral-7B-Instruct	378	594	0.59

- # Resp.: Hallucination counts in responses.
- # Span: Hallucination counts in spans.
- Density: Hallucination density is defined as the average number of hallucination spans per hundred words in the response.

Methods Evaluation

Method	**ROC-AUC**	**PR-AUC**
Linear Probing	0.839	0.628
Att. Pooling Linear Probing	**0.886**	**0.750**
MLE LID	0.806	0.613
PCA LID	<u>0.812</u>	<u>0.622</u>
Token-level Entropy	0.732	0.415
Chow-Quantile	0.780	0.604

Table 1: Evaluation Results for Mistral-7B-Instruct

Method	**ROC-AUC**	**PR-AUC**
Linear Probing	0.753	0.624
Att. Pooling Linear Probing	**0.783**	**0.685**
MLE LID	0.582	0.441
PCA LID	<u>0.572</u>	<u>0.465</u>
Token-level Entropy	0.588	0.462
Chow-Quantile	0.614	0.473

Table 2: Evaluation Results for Llama-2-7B-chat

Conclusion

Results:

- Implemented and tested supervised and unsupervised methods for hallucination detection.
- An unsupervised approach for effective leveraging of different knowledge sources in a KGQA task is proposed. The approach took 2nd place at the ACL workshop.
- A supervised approach for detecting hallucinations has been proposed. But the method has not been shown to be sufficiently effective.

Further development:

- Annotation of the new dataset.
- Interpretation of hidden states.
- Uncertainty estimation of internal states.
- Response sampling strategy development.

Thank you for your attention!

belikova.iua@phystech.edu