
Appendix of “Robust Hallucination Detection in LLMs via Adaptive Token Selection”

Anonymous Author(s)

Affiliation

Address

email

1 A Implementation Details

2 A.1 Setup

3 For the main results, our hallucination detector $f_\theta(\cdot)$ utilises a two-layer MLP with a hidden dimension
4 of 256. The first linear layer is followed by a BatchNorm and ReLU activations and the second
5 linear layer is accompanied by a sigmoid output. As shown in Figure 3b in our paper, our method
6 is relatively insensitive to variations in the number of layers and the hidden dimension size. At
7 the training stage, we use Adam optimiser. For the input representations of our detector, we utilise
8 representations extracted from a single layer and the layer index is determined by the results of the
9 validation set.

10 We implement our method using PyTorch 2.6.0 [3] and transformers 4.51.3 [4] and conduct all
11 experiments on NVIDIA A100 GPUs. Based on experimental records, the estimated total training and
12 inference time for each dataset under the main experimental settings is as follows: 0.40 GPU-hours
13 for LLaMA-3.1-8B, 0.45 GPU-hours for Mistral-Nemo-Instruct (12B) and 0.65 GPU-hours for
14 LLaMA-3.13-Instruct-70B with 8-bit quantisation. It is important to note that all these results are
15 based on the assumption that the semantic uncertainty score $P_{\text{uncertainty}}^c$ is readily available. Otherwise,
16 computing this score involves generating multiple responses and evaluating entailment across them,
17 which introduces nontrivial latency, approximately 7.6 seconds per question for LLaMA-3.1-8B, 8.6
18 seconds for Mistral-Nemo-Instruct (12B), and 46.1 seconds for LLaMA-3.13-Instruct-70B.

19 Our code is available at <https://anonymous.4open.science/r/HaMI-NeurIPS/>

20 A.2 Prompts for generation and evaluation

21 All prompts utilised in our experiments refer to those used by Farquhar *et al.* [2]. For the QA tasks,
22 we prompt selected LLaMA and Mistral models to generate answers without context in a zero-shot
23 manner. Figure 1 presents the used input prompt. To obtain response labels, we prompt GPT-4.1 [1]
24 with OpenAI API to evaluate the quality of the response. The prompts are illustrated in Figure 2. It
25 is observed that GPT-4.1 can make mistakes for the positive samples as checked with the gold answers.
26 Therefore, we ask GPT-4.1 to rejudge samples labelled as positive and discard samples if the result is
27 inconsistent with the first result.

28 B Study on k for the top k highest hallucination score selection

29 As presented in Section 4.1 in our paper, we select token instances with the top k highest hallucination
30 scores as the salient tokens in each sequence. k is dynamically determined by the generated token
31 length l as follows:

$$k = \lfloor r_k \times l \rfloor + 1, \quad (1)$$

Prompts for QA Tasks

Answer the following question in a single but complete sentence.\n

Question: Which politician won the Nobel Peace Prize in 2009?\n

Answer:

Figure 1: Prompts for QA tasks without context.

Prompts for GPT-4.1 Labelling

We are assessing the quality of answers to the following question:
{Qusetion} \n

The expected answer is: {Expected_Answer}.\n

- or: The following are expected answers to this question:
{Expected Answers}.\n

The proposed answer is: {predicted_answer}\n

Within the context of the question and your own knowledge, is the
proposed answer correct or the same as the expected answer?
Respond only with yes or no.\nResponse:

Figure 2: GPT-4.1 evaluation prompts

32 where r_k is set to 0.1 by default. In this section, we investigate the sensitivity of our method to the
 33 choice of r_k in the adaptive token selection process. As shown in Figure 3, HaMI exhibits stable
 34 performance across a set of r_k , ranging from 0.00 to 0.20, on all four datasets with LLaMA-3.1-8B
 35 and Mistral-Nemo-Instruct (12B). Notably, smaller r_k values already yield strong results, which
 36 implies that only a few high-salience tokens are sufficient for effective hallucination discrimination.

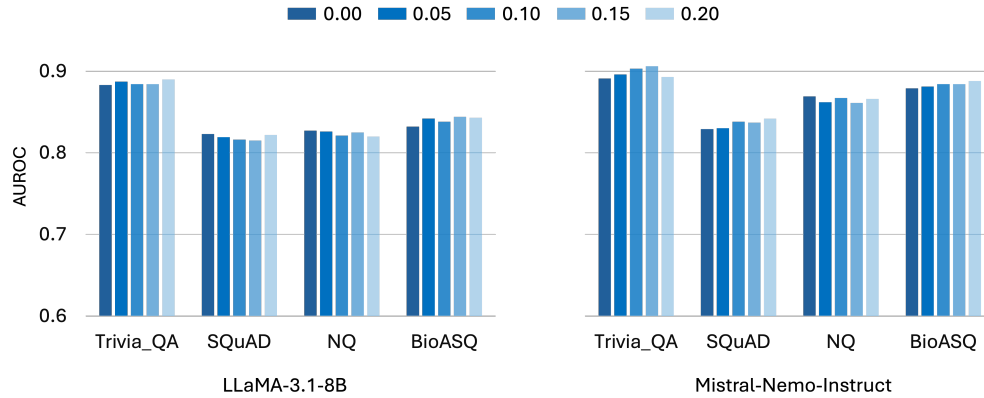


Figure 3: Study on different choice of r_k across four datasets with two LLMs.

37 C Limitations and Broader Impacts

38 **Broader impacts.** With the growing deployment of LLMs, ensuring the truthfulness of their outputs
39 has become a critical challenge, especially in high-stakes domains such as finance and healthcare. To
40 enable effective hallucination detection, in this work, we propose a practical and robust approach
41 that leverages internal representations to identify tokens with a high likelihood of being hallucinated
42 and enables the following detection. We hope that our proposed method, HaMI, can contribute to
43 safer and more reliable real-world deployment of LLMs. Although our research focuses on QA tasks,
44 HaMI can be extended to other tasks as well. For example, in the summarisation task, truthfulness
45 can be assessed sentence by sentence. Verifying the correctness of each generated sentence can be
46 reformulated as an entailment query against the source context, allowing hallucination detection to be
47 performed in a QA-like manner. Furthermore, the similarity between a generated sentence and its
48 corresponding context could serve as an uncertainty metric for internal representation enhancement.

49 **Limitations.** Apart from the aforementioned capabilities of HaMI, there remain some concerns.
50 Unlike some black-box uncertainty-based approaches, our method requires access to internal repre-
51 sentations, limiting its deployments to open-source LLMs. Moreover, we also explore integrating
52 uncertainty metrics into the original representations to enhance their discriminative capability on
53 correct and incorrect generations. While effective, the integration strategy is relatively simple and
54 we acknowledge that more sophisticated fusion methods could be explored for better detection
55 performance in future research.

56 References

- 57 [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altschmidt,
58 S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- 59 [2] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models
60 using semantic entropy. *Nature*, 630(8017):625–630, 2024. 1
- 61 [3] A. Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint*
62 *arXiv:1912.01703*, 2019. 1
- 63 [4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf,
64 M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L.
65 Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural
66 language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*
67 *Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for
68 Computational Linguistics. 1