SCOPE: Semantic Entropy Probes for LLM-as-a-Judge

Anna Deng[1,*] **Adithya Gnanasundar**[1,*] **Anthony Cui**[1]

Cole Blondin[1] **Kevin Zhu**[1] **Sean O'Brien**[1]

[1]Algoverse AI Research

1 Introduction

Since the advent of the Transformer architecture (Vaswani et al., 2017), the capabilities of Large Language Models (LLMs) have rapidly advanced, driving the emergence of LLM-as-a-Judge systems that enable automated evaluation of model outputs at scale without the need for costly manual annotations (Zheng et al., 2023). However, similarly to generative models, LLM-as-a-Judge are prone to hallucinations, outputting content that is not supported by the input or underlying facts (Maynez et al., 2020; Filippova, 2020; Ji et al., 2023). Thus, there is an increasing need for efficient and generalizable methods to detect and mitigate hallucinations in LLM-as-a-Judge systems to ensure that large-scale evaluations remain accurate and trustworthy.

A promising approach for detecting LLM hallucinations is semantic entropy (SE), which quantifies the variability in meaning across numerous model outputs generated from the same input prompt (Farquhar et al., 2024). If the outputs vary substantially in their conveyed meanings, the resulting high semantic entropy can serve as a strong indicator of hallucination or uncertainty. However, computing semantic entropy requires multiple forward passes for each input, which restricts its applicability in real-world, large-scale settings.

Recent work on Semantic Entropy Probes (SEPs) (Kossen et al., 2024) demonstrates that semantic entropy can be predicted from a model's hidden states, enabling the training of a simple linear probe to predict the semantic entropy for a given prompt. This probing approach eliminates the need for multiple generations per input, requiring only a single forward pass, and thus greatly reduces computational cost while preserving detection effectiveness.

In this work, we present ***S**emantic **E**ntr**o**py **P**robes for LLM **E**valuators* (SCOPE) a novel extension of the SEP framework to LLM-as-a-Judge tasks. Our approach applies semantic entropy probing to detect and mitigate hallucinations in LLM-as-a-Judge evaluations, focusing on rating–rationale tasks. We train a linear probe to predict semantic entropy labels from the LLM-as-a-Judge hidden states. Our method requires only a single forward pass per example, enabling scalable and efficient reliability assessment of LLM-as-a-Judge systems without repeated sampling. Beyond hallucination detection, SCOPE further supports an adaptive pipeline for mitigating uncertain judgments.

2 Methodology

As expressed by (Zheng et al., 2023; Gu et al., 2025), LLM-as-a-Judge is defined as an LLM tasked with evaluating model outputs. In our work, we formalize this definition specifically for rating-rationale tasks: let x be an input and y a candidate model output to be evaluated. An LLM-as-a-Judge J maps (x, y) to a numerical rating $r \in \mathbb{R}$, which is usually restricted to an integer in $[0, n]$ for $n \in \mathbb{N}$, and a rationale in natural language for that rating. We define an LLM-as-a-Judge hallucination as a case where J's rating or rationale is not supported by the input evidence or the ground truth for (x, y).

To train SCOPE, as outlined by (Kossen et al., 2024), we must obtain data points of $(h_p^l(x), H_{\text{SE}}(x))$ pairs, where x is an input query, $h_p^l(x) \in \mathbb{R}^d$ is the model hidden state at token position p and layer l, d is the hidden state dimension, and $H_{\text{SE}}(x) \in \mathbb{R}$ is the semantic entropy.

Measuring uncertainty of text generation models is a challenging task. Thus, (Farquhar et al., 2024) proposed semantic entropy, a method that aggregates token-level uncertainties across clusters of semantic equivalence. To calculate semantic entropy for a given query x, we sample model completions from the LLM, aggregate the generations into clusters $(C_1, \ldots, C_K)$ of equivalent seman-

[*]Equal contribution by author.

tic meaning, and then calculate semantic entropy (H_{SE}) by aggregating uncertainties within each cluster. (Farquhar et al., 2024) determined whether two generations have equivalent semantic meaning using natural language inference (NLI) models to predict entailment between the generations, with two generations being semantically equivalent if they entail each other. We add each generation to existing clusters if semantically equivalent and create new clusters otherwise, hence successfully separating the generations into clusters.

With these clusters, we can compute the uncertainty related to the distribution of clusters. Given an input context x, the joint probability of a generation s consisting of tokens $(t_1, \ldots, t_n)$ is given by the product of conditional token probabilities in the sequence, $p(s \mid x) = \prod_{i=1}^{n} p(t_i \mid t_{1:i-1}, x)$. The probability of the semantic cluster C is then the aggregate probability of all possible generations s belonging to that cluster, $p(C \mid x) = \sum_{s \in C} p(s \mid x)$. The uncertainty associated with the distribution over semantic clusters is hence the semantic entropy,

$$H[C \mid x] = \mathbb{E}_{p(C|x)}[-\log p(C \mid x)].$$

To estimate this value, (Farquhar et al., 2024) treated the K generated clusters $(C_1, \ldots, C_K)$ as Monte Carlo samples from the true distribution over semantic clusters $p(C|x)$, and thus the approximation for entropy becomes

$$H[C \mid x] \approx -\frac{1}{K} \sum_{k=1}^{K} \log p(C_k|x).$$

To train SCOPE with judging tasks, we access datasets with sample model responses, in particular, the UltraFeedback (Cui et al., 2024) and HH-RLHF (Bai et al., 2022) datasets. Each example in these datasets consists of a chain of user prompts and their corresponding model outputs. For each chain of model responses, we sample ten generations from `Gemma-2-9B-it` prompted to complete the judgment task on these responses.

Using DeBERTa (He et al., 2021) as our entailment model, we create clusters of semantic equivalence and calculate their semantic entropy. Following the methodology of (Kossen et al., 2024), we save the hidden states at the token before generating the model response (TBG) for each query.

With this data, we binarize semantic entropy scores indicating whether semantic entropy is high or low. We train linear logistic regression probes to predict these labels based on the TBG. In testing, SCOPE predicts the probability that an LLM-as-a-Judge generation for a given model response query x has high semantic entropy.

Sample LLM-as-a-Judge Input:

```
Human: How far can a man walk in a day?
Assistant: I'd say anything between 15-20
    miles.
```

Sample LLM-as-a-Judge Output:

```
Rating: 3
Rationale: The response provides a
    plausible but general range, lacking
    context for individual capabilities.
```

Figure 1: Example of an input from the HH-RLHF dataset and one of its resulting evaluations generated by the LLM-as-a-Judge. A total of ten generations are completed to estimate the uncertainty of the LLM-as-a-Judge in regards to this specific input.

We then apply these trained probes to flag unreliable LLM-as-a-Judge evaluations and extend them into an adaptive pipeline that not only detects hallucinations but also mitigates their impact. When a weaker judge produces an evaluation, SCOPE probes its hidden states to assess semantic entropy in a single forward pass. If the probe indicates low uncertainty, the judgment is accepted immediately, preserving efficiency. However, if SCOPE signals high uncertainty, the evaluation automatically escalates to a stronger model or, in high-stakes settings, to human review. This transforms SCOPE from a diagnostic tool into the backbone of a tiered judgment system, where judgments remain cheap, and only the most uncertain cases consume additional resources. Crucially, because the probe operates on existing hidden states, this escalation pipeline incurs almost no additional computational cost beyond the initial forward pass. Enabling detection and mitigation with negligible overhead, SCOPE offers a practical, scalable approach to ensure trustworthiness of LLM-as-a-Judge evaluations.

3 Conclusion

We introduce SCOPE, a novel extension of SEPs to LLM-as-a-Judge rating–rationale tasks. Utilizing a single forward pass using LLM-as-a-Judge hidden states, SCOPE enables a cheap binary classification of uncertainty for hallucination detection and supports an adaptive pipeline for hallucination mitigation.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *nature*, 630.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *Preprint*, arXiv:2010.05873.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Representations*. Under review.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *Computing Research Repository*, arXiv:2406.15927.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *Preprint*, arXiv:2005.00661.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762v1.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.