Unsupervised Hallucination Detection by Inspecting Reasoning Processes

Ponhvoan Srey **Xiaobao Wu*** **Anh Tuan Luu***
Nanyang Technological University
{ponhvoan002, xiaobao002, anhtuan.luu}@ntu.edu.sg

Abstract

Unsupervised hallucination detection aims to identify hallucinated content generated by large language models (LLMs) without relying on labeled data. While unsupervised methods have gained popularity by eliminating labor-intensive human annotations, they frequently rely on proxy signals unrelated to factual correctness. This misalignment biases detection probes toward superficial or non-truth-related aspects, limiting generalizability across datasets and scenarios. To overcome these limitations, we propose IRIS, an unsupervised hallucination detection framework, leveraging internal representations intrinsic to factual correctness. IRIS prompts the LLM to carefully verify the truthfulness of a given statement, and obtain its contextualized embedding as informative features for training. Meanwhile, the uncertainty of each response is considered a soft pseudolabel for truthfulness. Experimental results demonstrate that IRIS consistently outperforms existing unsupervised methods. Our approach is fully unsupervised, computationally low cost, and works well even with few training data, making it suitable for real-time detection. [1]

1 Introduction

In recent years, Large Language Models (LLMs) such as GPT-4 (Achiam et al., 2023) have demonstrated remarkable capabilities to generate coherent and relevant responses to user queries. Their success led to the broad adoption of LLMs in a wide variety of tasks, from coding to text summarization (Touvron et al., 2023; Yang et al., 2024; Driess et al., 2023). However, a critical concern arises because of their tendency to hallucinate, which refers to the phenomenon where LLMs generate texts that appear logical and coherent but contain fictitious information (Zhang et al., 2023b). As LLMs have been observed to confidently generate false information, it is challenging for users to identify hallucinations. This represents a significant drawback to the robustness of LLMs in real-world applications, resulting in growing interest in hallucination detection.

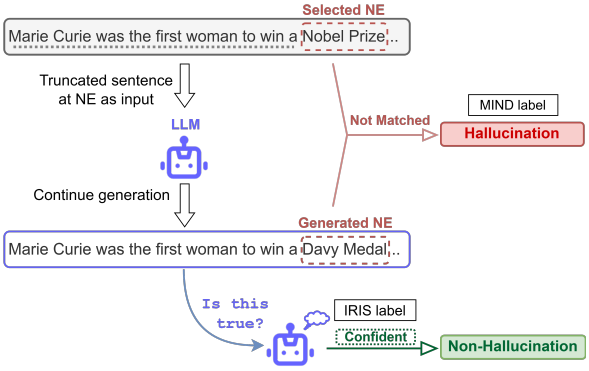

Figure 1: Comparison between MIND (Su et al., 2024) and our method. MIND incorrectly identifies a statement as hallucination. Our method extracts model internal knowledge by asking it to think carefully whether the statement is correct. Its confidence is obtained as a soft pseudolabel.

Current hallucination detection strategies are concerned with determining the truth or falsity of a generated statement (Su et al., 2024; Manakul et al., 2023; Min et al., 2023). A straightforward paradigm is to directly instruct a language model to determine the truth or falsehood of a given text (Li et al., 2024; Su et al., 2024). Others scrutinize the internal activations for latent signals that may correlate with factual accuracy (Su et al., 2024; Azaria and Mitchell, 2023; Burns et al., 2022). On the other hand, other methods regard the model uncertainty during text generation as a proxy for hallucination risk. Generally, state-of-the-arts methods aim to extract the inherent knowledge of LLMs to verify the veracity of statements, whether by directly querying the model, or by inspecting its

*Corresponding Authors.
[1]The code repository is available online at https://github.com/ponhvoan/iris.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22129–22141
November 4-9, 2025

internal states, or by measuring its uncertainty.

Unfortunately, these methods remain suboptimal for practical deployment. First, they are expensive and incur high computational overhead. Direct querying methods achieve satisfactory performance primarily with high-capacity commercial models, such as GPT-4 (Li et al., 2024; Azaria and Mitchell, 2023), rendering them impractical for open-source alternatives and cost-sensitive applications. Uncertainty-based techniques are hampered by ambiguous threshold determination and often require generating multiple samples per query to reach acceptable performance (Kuhn et al., 2023; Fadeeva et al., 2023; Chen et al., 2024). Similarly, internal activation approaches depend heavily on extensive manual annotations to train a probe. Second, methods that bypass the high computational demands risk introducing bias into the detection process. Burns et al. (2022) and Su et al. (2024) consider model internal activations under the unsupervised scenario. Such methods are limited in scope and do not inherently capture the notion of truth, biasing the probe towards non-truth-related signals. These limitations highlight critical gaps in current methodologies and motivate the need for more robust, scalable, and efficient frameworks for real-time hallucination detection. A key challenge is to obtain truth labels to train a classifier probe based on model internal states without resorting to human annotations.

In this work, we propose **Internal Reasoning for Inference of Statement veracity** (**IRIS**), an unsupervised hallucination detection method that inspects the uncertainty of the reasoning process. We assert that the model's confidence or uncertainty when verifying the truthfulness of a statement reveals the likelihood of hallucination. This uncertainty is regarded as a soft pseudolabel for statement accuracy. To illustrate, let us consider the statement "Marie Curie was the first woman to win a Davy Medal", which is a continuation of a truncated Wikipedia sentence (see Figure 1). As Marie Curie is a well-known public figure, LLMs would retain knowledge that she is also indeed the first woman to win a Davy Medal. When asked to assess the statement, LLMs would exhibit high confidence that it is correct, thus most likely a non-hallucination. IRIS regards this confidence score as a soft pseudolabel for truth. On the other hand, in MIND (Su et al., 2024), the label is derived by matching the original and generated NE, incorrectly resulting in "Hallucination". MIND biases the labels towards matching named entities, rather than the truth value of the statement.

Furthermore, the model internal states when thinking about the statement represent its latent knowledge, and contain information indicative of factual correctness. IRIS trains a lightweight probe on the contextualized embeddings of the model verification with the pseudolabels. Our advantage is that we only need one query to the model for each statement whereas uncertainty-based methods demand multiple samples.

IRIS alleviates the burden of human annotations and promotes real-time hallucination detection. To summarize, the contributions of this paper are as follows:

- We propose an unsupervised hallucination detection framework where the pseudolabels are related to the truthfulness of the statements and the features are model internal states.
- We demonstrate that the contextualized embeddings of the model's verification is more informative of the hallucination risk, compared to those of the statements themselves.
- We conduct extensive experiments and show that our method achieves an improvement of 3.2%, 7.0%, and 10.2% on unsupervised hallucination detection with the True-False, HaluEval2, and HELM datasets, respectively.

2 Methodology

In this section, we present the details of our proposed unsupervised hallucination method based on internal activations. We note that our method is designed to detect hallucinations in LLMs as the primary goal, but our proposal is general and is thus applicable to identifying falsehoods by leveraging the internal knowledge of LLMs.

2.1 Eliciting Internal Knowledge

Given a set $(s_1, \ldots, s_n)$ of statements, our goal is to determine whether each s_i is true. Here, we allow for the most general case where each statement could be either human-written or generated by an external language model that we no longer have access to. Previous research (Azaria and Mitchell, 2023; Ji et al., 2024) show that the LLMs contain internal knowledge, accessed via their internal states, on the truthfulness of statements. Additionally, we believe that the contextualized embeddings of the

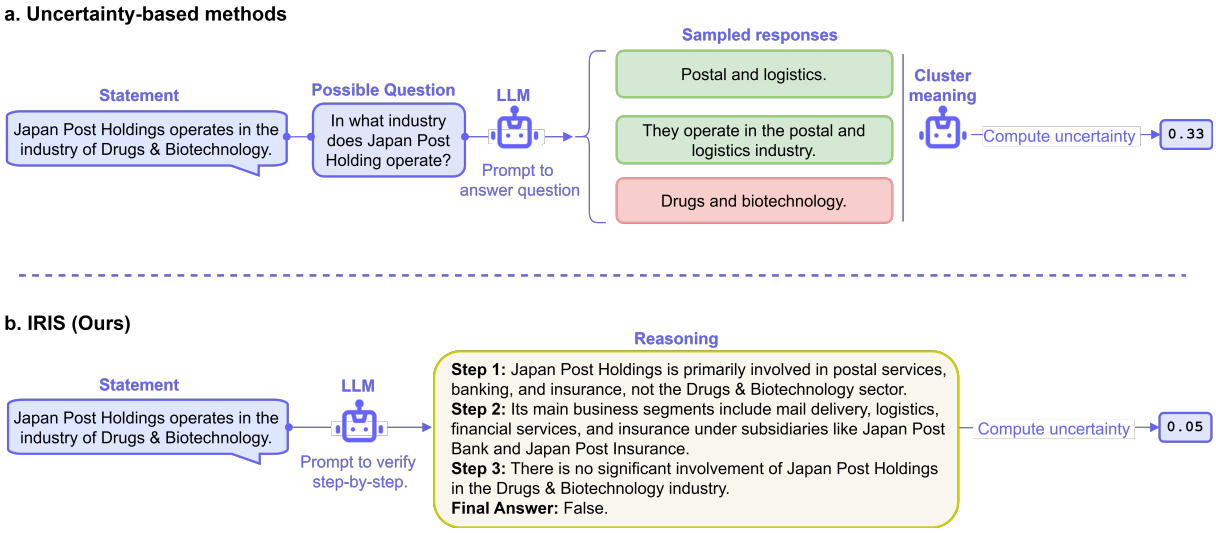

Figure 2: Illustrations of eliciting uncertainty. In uncertainty-based methods, given a statement, a question to which the statement could have been the answer is generated. Multiple responses from the LLM are then sampled, and the responses are aggregated based on meaning and the uncertainty is computed. Meanwhile, our approach only calls the LLM *once* using CoT prompting to facilitate a more thoughtful and informative response. The uncertainty is normalized so that 0 indicates hallucination.

response verifying the statements is a more reliable predictor of hallucination than those of the statements themselves. As such, we first prompt the model to carefully verify whether a statement is correct. Section 3 provides the results of using the embeddings of the original embeddings (SAPLMA) and of the verification response.

2.2 Uncertainty as Pseudolabels

Importantly, we require that our method does not rely on external supervision. To this end, we derive a pseudolabel for each statement. Ultimately, we wish to train a classifier probe on the internal states and these pseudolabels. To ensure that the classifier is endowed with the capacity to tell truth from falsehood, these pseudolabels should encompass some notion of truth.

In uncertainty-based hallucination detection, e.g. as formulated by Kuhn et al. (2023) and Duan et al. (2024), the underpinning assumption is that when an LLM is responding to a query, its level of uncertainty indicates its lack of knowledge and the likelihood of hallucinating. As the uncertainty in the response directly links to hallucination risk and captures some aspect of factuality, we want to leverage the same insight and make use of uncertainty as pseudolabels. However, these methods involve passing the query to the model and evaluating its output. This imposes three key demands. First, to measure an LLM's uncertainty regarding a statement, a question to which the statement might be the answer is generated either using the same LLM or a different one. However, this may propagate errors as the question may be vague and lead to answers inconsistent with the provided statement. Second, multiple calls to the LLM is required to sample responses. Lastly, an additional model is used to aggregate the responses according to semantic meanings to estimate uncertainty. Overall, this approach is compute heavy and less practical for real-time detection. In our case, we seek to bypass these requirements.

With this goal in mind, given a statement s_i, we prompt an LLM to evaluate its correctness. We argue that the uncertainty in its appraisal equates to the model's knowledge regarding the statement and indicates the truthfulness of the sentence. Further, we prompt the LLM to reason step-by-step, eliciting a more thoughtful and informative response as the model assesses the correctness of the statements. As illustrated in Figure 2, our approach involves calling the LLM only once, and the reasoning chain is examined to estimate the LLM's uncertainty regarding the provided statement.

To estimate the uncertainty, we explore two options: (a). entropy-based: calculate the entropy of the reasoning chain using token probabilities, and normalize the entropy such that a value of 1 indicates correctness; and (b). verbalized: the LLM expresses the confidence that the statement is cor-

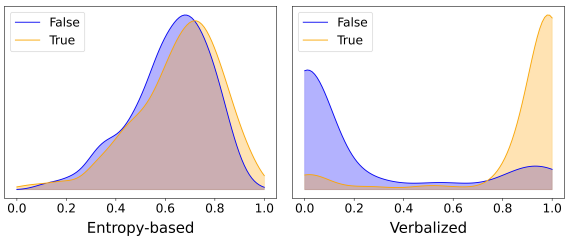

Figure 3: Distribution of confidence in determining the correctness of statements from the True-False dataset.

rect in numerical probabilities in the context of the preceding reasoning steps. The prompt templates are provided in Appendix A. Similar to the findings by Tian et al. (2023), we find that the verbalized confidence is better calibrated compared to token probabilities as shown in Figure 3.

2.3 Classifier Training

Suppose the evaluation of the statement s_i yields a response x_i, we obtain the contextualized embeddings $\phi(x_i)$ from the LLM. In line with SAPLAMA (Azaria and Mitchell, 2023) and MIND (Su et al., 2024), we utilize the last token's embeddings at the last layer. In Section 4.2, we study the efficacy of using embeddings at other layer depths.

For each $\phi(x_i)$, we generate a soft pseduolabel $\tilde{y}_i \in [0, 1]$, which is the uncertainty of the LLM's assessment of whether the statement s_i is true. We normalize $\tilde{y}_i$ so that $\tilde{y}_i = 1$ means s_i is correct, and vice versa. To address the issue of noisy labels, we apply soft bootstrapping (Reed et al., 2014), where new regression targets are generated for each mini-batch based on the classifier's current predictions. The updated target for the i-th statement is $t_i = \beta\tilde{y}_i + (1 - \beta)\hat{y}_i$, where q_i is the classifier's prediction, and $\beta \in (0, 1)$. Further, we employ the symmetric cross entropy loss (Wang et al., 2019) as the overall objective

$$l_i = l_{ce} + l_{rce} = H(\hat{y}_i, t_i) + \phi H(t_i, \hat{y}_i), \quad (1)$$

where $H(\hat{y}_i, t_i) = t_i \log \hat{y}_i + (1 - t_i) \log(1 - \hat{y}_i)$ is the standard cross entropy loss, $H(t_i, \hat{y}_i)$ is the reverse cross entropy loss, and ϕ is a hyperparameter balancing the losses. The cross entropy term aligns the predictions with the pseudolabels, but this causes the classifier to be sensitive to noisy labels. On the other hand, the reverse cross entropy term penalizes the classifier for being too confident on potentially incorrect targets. Their combination prevents overfitting to noisy labels, and allows for better generalization.

To reduce computational demands, the classifier is implemented as a small feedforward MLP with three fully-connected hidden layers of units $(256, 128, 64)$, each followed by ReLU activation. In the final layer, a sigmoid activation is applied. The Adam optimizer is used for training, with a learning rate of 10^{-2}, and weight decay of 10^{-5}. We present hyperparameter finetuning results in Appendix C. For all settings, the classifier is trained for 10 epochs with a patience of 5 epochs. In Appendix B, results on prompt sensitivity are reported.

3 Experiments

3.1 Datasets

We evaluate our method on three recent hallucination datasets: (i) **True-False** dataset (Azaria and Mitchell, 2023): a compilation of ∼6300 factual statements across 6 topics (Animals, Cities, Companies, Elements, Facts, and Inventions) and a set of LLM generated statements (Generated); (ii) a subset of **HaluEval2** with human annotations (Li et al., 2024), which consists of a total of ∼3800 ChatGPT responses to challenging queries related to Bio-Medical, Education, Finance, Open-Domain, and Science; (iii) **HELM** (Su et al., 2024): a collection of ∼3600 LLM-generated text continuation based on Wikipedia articles across six LLMs of varying sizes. Together, these datasets represent a balanced mix of LLM-generated and non-LLM statements, covering a wide range of topics. The topic or sub-dataset in each dataset is split 80-20 for training and testing.

3.2 Baselines

In the main experiments, we use **Llama-3.1-8B-Instruct** (Dubey et al., 2024) as the proxy model to extract knowledge from and to assess the correctness of statements. We evaluate our proposal against three different types of baselines.

The first is direct prompting, where the LLM is asked to determine whether a statement is factual under the zero-shot and few-shot (three demonstrations) settings. We utilize two prompting techniques: asking the model to directly determine correctness, and chain-of-thought (CoT) prompting. This results in **Zero-shot**, **Few-shot**, **CoT (zero-shot)**, and **CoT (few-shot)**. To compare to commercial LLMs, following the approach in HaluEval (Li et al., 2023a), results from querying **GPT-4o** are provided.

Secondly, we evaluate the datasets with two

Method	Animals	Cities	Companies	Elements	Facts	Generation	Invention	Average
Zero-shot‡	80.06	91.34	90.33	81.83	92.17	79.18	75.46	85.37
Few-shot‡	78.08	89.23	89.75	82.90	92.00	75.51	89.38	86.38
CoT (zero-shot)‡	**80.65**	92.04	91.67	86.88	92.82	83.27	80.48	87.54
CoT (few-shot)‡	75.69	92.39	91.00	86.34	91.19	78.78	83.11	86.65
EigenScore‡	57.45	63.04	53.57	51.15	51.28	60.23	53.67	56.06
SAR‡	64.68	58.04	62.74	53.15	58.28	70.76	58.56	59.86
CCS‡	67.33	51.71	57.08	70.97	79.67	57.14	67.61	63.16
MIND‡	51.49	53.08	53.75	44.62	54.47	36.73	61.36	52.36
IRIS‡(ours)	80.20	**93.84**	**94.17**	**89.25**	**93.50**	**87.76**	**90.91**	**90.38**
GPT-4o*	82.74	94.10	91.42	96.02	98.53	82.45	86.30	90.96
SAPLMA†	84.16	95.21	87.92	84.41	93.50	79.59	94.89	89.67
Ceiling†	81.19	95.21	94.17	91.40	95.21	87.76	90.34	91.25

(a) True-False

Method	Bio-Medical	Education	Finance	Open-Domain	Science	Average
Zero-shot‡	59.90	63.48	66.53	73.81	61.25	63.67
Few-shot‡	58.33	59.23	71.93	55.78	59.04	61.88
CoT (zero-shot)‡	56.28	61.29	67.58	63.61	56.83	60.77
CoT (few-shot)‡	59.18	62.11	58.05	63.95	55.52	58.87
EigenScore‡	41.80	46.58	45.15	58.05	35.01	43.03
SAR‡	**62.35**	63.01	63.94	57.56	**75.61**	65.98
CCS‡	53.61	60.54	61.94	54.24	53.00	56.91
MIND‡	40.36	45.58	75.66	30.51	45.50	50.74
IRIS‡(ours)	59.64	**63.95**	**78.84**	**93.22**	70.00	**70.57**
GPT-4o*	71.38	83.47	71.18	73.87	81.29	75.59
SAPLMA†	59.64	64.63	77.78	93.22	81.00	73.33
Ceiling†	72.89	67.35	77.25	93.22	81.50	76.74

(b) HaluEval2

Method	Falcon	GPT-J	LLB-7B	LLC-7B	LLC-13B	OPT-7B	Average
Zero-shot‡	56.81	64.51	57.17	56.34	61.34	64.49	60.21
Few-shot‡	58.73	66.43	60.88	48.09	44.12	64.66	57.71
CoT (zero-shot)‡	62.57	68.71	58.94	58.35	57.35	65.55	62.12
CoT (few-shot)‡	56.62	64.86	60.53	54.53	56.09	66.43	60.12
EigenScore‡	38.46	42.25	44.30	37.18	42.34	48.99	42.41
SAR‡	61.26	57.50	64.30	59.94	57.66	51.52	58.66
CCS‡	55.24	50.43	52.21	57.00	61.46	64.04	56.79
MIND‡	54.29	52.17	59.29	44.00	59.38	56.14	54.10
IRIS‡(ours)	**67.62**	**69.57**	**69.03**	**66.00**	**69.79**	**68.42**	**68.43**
GPT-4o*	70.25	76.57	70.97	57.34	62.39	77.56	69.63
SAPLMA†	73.33	85.22	56.64	76.00	68.75	81.58	73.61
Ceiling†	77.14	83.48	66.37	85.00	71.88	82.46	77.80

(c) HELM

Table 1: Accuracy results on the True-False, HaluEval2, and HELM datasets with three kinds of methods: Unsupervised (‡), Supervised (†), and Commercial (*). The best results of unsupervised methods are in **bold**.

recent uncertainty-based methods, **EigenScore** (Chen et al., 2024) and **SAR** (Duan et al., 2024). For each statement, the LLM is queried 10 times to determine its confidence of the statement correctness, and an uncertainty score is computed from these samples. To follow the unsupervised setting, sentences with uncertainty scores greater than the median for each dataset are considered incorrect.

Next, we compare with methods most closely related to our work — unsupervised hallucination detection based on LLM internal activations. In particular, we compare with Contrast-Consistent Search (**CCS**) (Burns et al., 2022) and **MIND** (Su et al., 2024). CCS converts every sentence into a pair of contrasting statements, and trains a probe such that the probabilities of each pair being true

(the statement itself and its negation) adds up to one. MIND truncates Wikipedia articles and prompts an LLM to continue the next sentence. The truth labels are automatically generated by matching the named entities of the response and of the succeeding sentence in the original article. MIND assumes access to the LLMs that were used to generate the statements and use their contextualized embeddings as the features for training the probe. To conform to the same setting as our method, we do not have such access, and instead make use of the proxy model. Note that MIND's probe is trained on its own automatically labeled dataset before testing on the validation sets.

Finally, we provide results of supervised baselines for comparison. **SAPLMA** (Azaria and Mitchell, 2023) utilizes the contextualized embeddings of the statements, and trains a probe using these and ground-truth labels. Lastly, the supervised **Ceiling** trains the probe on the embeddings of the model's statement verification and the ground-truth labels. All classifier-based methods employ the same MLP architecture.

Since the main task is binary classification, detection accuracy is used as the evaluation metric. For probe-based methods, the validation accuracy on a held-out set is computed.

4 Experimental Results and Analysis

4.1 Main Results

Table 1 presents the main results across all subtopics of the True-False, HaluEval2, and HELM datasets. Among all unsupervised methods using Llama-3.1-8B-Instruct, our method attains the highest average accuracy for all datasets, outperforming the best baseline by 3.2%, 7.0%, and 10.2% on the True-False, HaluEval2, and HELM datasets, respectively. Overall, our method is the most consistent, obtaining the best accuracy for 15 out of 18 sub-datasets. Although the commercial GPT-4o performs better, IRIS uses a much smaller 8B model, and is able to achieve comparable accuracy on the True-False and HELM datasets.

While Chain-of-Thought may help improve accuracy, it is not consistent. This shows that direct query as a method to detect hallucination is sensitive to the input prompt. Meanwhile, the unsupervised detection methods, CCS and MIND, mostly perform worse than direct query. In particular, MIND reached a training and validation accuracy of approximately 76% and 70% on their automatically annotated data. However, its accuracy drops significantly when tested on other datasets, indicating its inability to generalize well. Notably, the supervised ceiling exceeds the performance of SAPLMA. Recall that SAPLMA utilizes the embeddings of the statements themselves as the training features whereas the supervised ceiling uses those of the model verification process of the statements. These results reinforce our hypothesis that the latter is more informative than the former.

Train on	Test on					
	Animals	Cities	Bio-Med	Education	Falcon	GPT-J
Animals	80.20	91.78 (-2.06)	60.84 (+1.20)	67.35 (+3.40)	63.81 (-3.81)	68.70 (-0.87)
Cities	73.76 (-6.44)	93.84	57.83 (-1.81)	68.03 (+4.08)	57.14 (-10.48)	65.22 (-4.35)
Bio-Med	74.26 (-5.94)	86.99 (-6.85)	59.64	63.27 (-0.68)	62.86 (-4.76)	63.48 (-6.09)
Education	73.27 (-6.93)	89.04 (-4.80)	65.66 (+6.02)	63.95	62.86 (-4.76)	67.83 (-1.74)
Falcon	79.70 (-0.50)	90.75 (-3.09)	59.04 (-0.60)	59.18 (-4.77)	67.62	66.96 (-2.61)
GPT-J	74.75 (-5.45)	83.90 (-9.94)	60.84 (1.20)	59.86 (-4.09)	61.90 (-5.72)	69.57

Table 2: Accuracy of trained probes on out-of-distribution test set. The performance decrease and increase are highlighted in red and green, respectively.

4.2 Analysis

In this section, we analyze IRIS to evaluate its performance under various scenarios.

Out-of-Distribution (OOD) Setting. To test the performance of IRIS under the OOD setting, we consider two topic from each of the True-False, HaluEval2, and HELM datasets. Table 2 reports the OOD performance. On average, the accuracy decreases by 3.1%, but remains generally robust. With the True-False datasets used for training, the OOD test accuracy decreases slightly by 2.1%, compared to a drop of 3.7% and 3.6% when HaluEval2 and HELM are used. We can observe that when trained on easier datasets where the model has more knowledge, the test accuracy decreases less drastically or even improves slightly. For example, for "Animals", the average OOD accuracy only decreases marginally by 0.4%, and its test accuracy on "Bio-Medical" and "Education" improves by 1.2% and 3.4%.

Model Size. We test with different model sizes to understand the influence of model capacity and

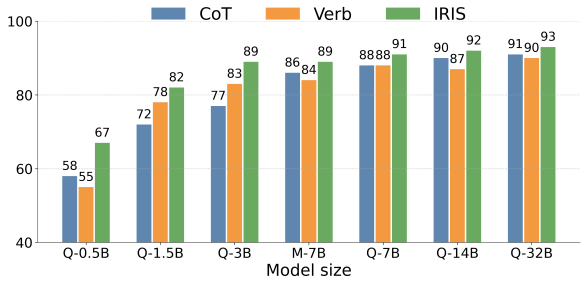

Figure 4: Average accuracy (%) on True-False dataset with models of different sizes. "Q" and "M" represent instruction-tuned Qwen-2.5 (Yang et al., 2024) and Mistral-v0.3 (MistralAI, 2024) models, respectively.

internal knowledge on the performance of IRIS. Figure 4 reports the accuracy of IRIS, along with direct prompting using CoT and verbalized confidence (Verb). For Verb, statements with a confidence score above 0.5 are considered factual. IRIS consistently performs the best, and the improvement is more significant in smaller models. With smaller models, the model knowledge regarding factual correctness decreases. The verbalized confidence is not as well-calibrated, and thus, using it to directly identify correct statements yields a significantly worse accuracy. Nonetheless, their hidden states contain richer contextual information compared to the response. Training a probe on the hidden states via the IRIS pipeline leads to greater gains, especially on the smallest 0.5B model where IRIS outperforms Verb by 12%.

Layer Depth. We investigate the contribution of each layer on the overall detection performance. As shown in Figure 5, the embeddings at different layers yield a range of accuracy. The average accuracy peaks with the middle layer, but for some dataset, such as "Companies", "Elements", and "Generated", the last layer gives better results. Our study does not provide conclusive evidence regarding which layer is ideal for hallucination discrimination. In previous works, Ji et al. (2024) found that the final layer is the best in recognizing falsehoods, whereas Azaria and Mitchell (2023) advocated for the intermediate layers. We believe that it ultimately depends on the dataset and LLM used. Therefore, we can employ an architecture to integrate the embeddings at all depths to learn better. However, our preliminary consideration of adding a small learnable module to fuse the embeddings did not improve performance. Further examination with more sophisticated architecture is required.

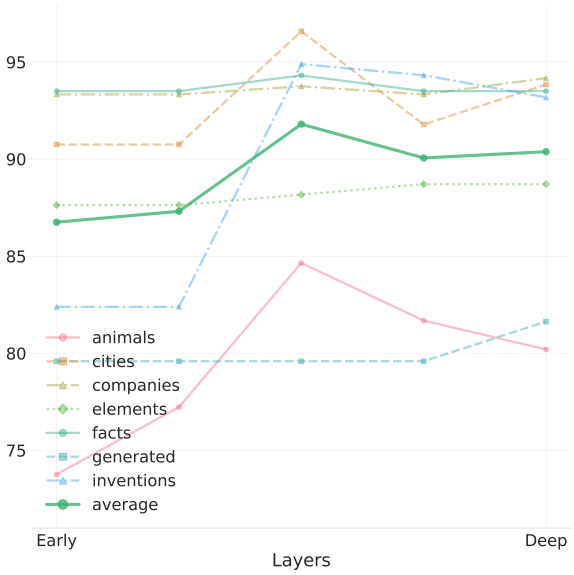

Figure 5: Accuracy (%) using embeddings at different layers of Llama-3.1-8B-Instruct.

# Data	32	64	128	256	All
Acc	84.63	87.86	88.49	89.83	90.38

Table 3: Average True-False accuracy (%) using different sizes of training data.

Training Data Size. IRIS relies on access to an unlabeled collection of statements as training data. In this subsection, we evaluate its performance with training data of different sizes. Table 3 reports the average True-False accuracy on the validation sets. The experiments indicate that our method improves with more data, but the gains plateau with more than 128 data points. Most importantly, our proposal works reasonably well even with as few as 32 statements. With this observation, given some demonstration statements from HELM, we prompt Llama-3.1-8B-Instruct to generate 32 statements of similar nature and topics, half of which it knows to be true, and the other half false. An example generated true statement is "The largest mammal by mass is the blue whale", and false statement is "Google's motto is 'Don't be evil'." Using the generated statements as training data, the average accuracy on the validation sets is 86.28%, comparable to when the actual True-False data is used. This implication reduces our reliance on having a large carefully crafted training dataset.

5 Related Work

In this section, we provide an overview of existing research on LLM hallucination detection. These methods can be broadly categorized into three groups: detection by direct query, by estimating consistency or uncertainty in the response, and by extracting the model's internal knowledge.

Direct Prompting. LLMs have been increasingly utilized to evaluate the quality of generated natural language (Fu et al., 2023; Wang et al., 2023; Liu et al., 2023). Coupled with the success of few-shot learning (Brown et al., 2020), prompt engineering, a simple strategy to elicit what LLMs know via better prompts, has garnered great attention (Sahoo et al., 2024). In hallucination detection, improved prompting is often incorporated to more accurately elicit LLMs world knowledge. Li et al. (2024) instruct GPT-4 (Achiam et al., 2023) directly to assess factual accuracy given a few demonstrations. Mündler et al. (2023) generate two responses based on the same context and prompt an LLM to evaluate whether there is a contradiction, which implies inaccuracy in one of the responses. Chain-of-Verification (CoVe) (Dhuliawala et al., 2024) decomposes a response into individual claims and constructs verification questions for each. For more complex tasks, Chain-of-Thought (CoT) prompting, which requests LLMs to respond to queries by reasoning step-by-step, results in more thoughtful responses (Kojima et al., 2022; Wei et al., 2022). Self-consistent CoT (Wang et al., 2022) samples multiple reasoning paths and chooses the most consistent response. Tree of Thoughts (ToT) (Yao et al., 2023) branches out the reasoning chain into intermediate thoughts and allows backtracking and lookahead. An obvious extension is to apply CoT prompting to LLMs to detect cases of hallucination.

Uncertainty Estimation. Another approach in hallucination detection is to determine the uncertainty of the generated sequence. The fundamental premise is that the higher the uncertainty in the response, the likelier the model has hallucinated (Fadeeva et al., 2023). Kuhn et al. (2023) propose semantic uncertainty to account for how syntactically different responses that share the same meaning should not increase the uncertainty. Claim-conditioned uncertainty further removes the impact of the uncertainty due to synonyms, different claim types, or order in the answer (Fadeeva et al., 2024). On top of sentence level uncertainty, Duan et al. (2024) examine uncertainty at the token level, and remove the influence of semantically insignificant tokens. Zhang et al. (2023a) weigh their uncertainty measure with attention values, and address over- and under-confidence issues by correcting the token probability with the inverse document frequency (IDF). Chen et al. (2024) devise EigenScore, an uncertainty metric that measures the spread of the response in the embedding space. Under the black-box setting, multiple responses are sampled and their consistency is estimated (Kuhn et al., 2023; Manakul et al., 2023; Lin et al., 2024). However, to estimate an uncertainty score, these approaches generally require multiple calls to the LLM, rendering them computationally heavy.

Internal Knowledge. Findings from recent research provide strong evidence that LLMs contain more knowledge about their own response (Pan et al., 2024; Wu et al., 2023, 2024b,a,c,d; Wu, 2025). Saunders et al. (2022) analyze the generation-discrimination gap, which is the discrepancy between the model's ability to produce high-quality outputs and its capacity to accurately evaluate those outputs. Li et al. (2023b) highlight that attention heads contain crucial information related to factuality, and modify attention activations to enforce more truthful generation. SAPLMA (Azaria and Mitchell, 2023) argues that the internal states are discriminative features to uncover statement truthfulness, and train a probe on the internal states with human annotations to label hallucinations. Likewise, Ji et al. (2024) describe similar findings. As human annotation is labor-intensive, some studies focus on the unsupervised setting. Burns et al. (2022) train a linear probe through unsupervised clustering by maximizing the distance between the embeddings of a statement and its negation. Meanwhile, MIND (Su et al., 2024) designs a training dataset that is automatically annotated by matching named entities in Wikipedia articles. However, as these approaches do not incorporate truth-related concept, the probe may be misled to discover directions biased towards named entities, for instance. This limits generalizability to other datasets.

6 Conclusion

In this work, we introduce IRIS, a novel unsupervised hallucination detection framework utilizing the internal states of LLMs. Specifically, an LLM is prompted to carefully verify the veracity of a given

statement, and the contextualized embeddings of its response are obtained. Then, the uncertainty of the answer is evaluated, and regarded as a soft label for the truthfulness of the statement. Finally, a classifier is trained on the embeddings and corresponding pseudolabels. IRIS demonstrates state-of-the-arts performance on recent hallucination benchmarks, improving over strong baselines by a large margin. Our method does not incur much computational overhead and does not require enormous training data, making it suitable for real-time detection.

Further investigation, using SAR scores as pseudolabels, shows the potential of incorporating more advanced uncertainty metrics into the IRIS pipeline (see Appendix E). Other uncertainty approaches that are not as computationally costly can be explored. A preliminary analysis reveals that performance depends on the layer depth of the embeddings. More complex probe architecture can be exploited to aggregate embeddings at different layers. In addition, evaluation of embeddings at different tokens may uncover more granular insights into how internal representations correlate with truthfulness.

Limitations

We believe our work has the following limitations:

Model Variety. This work uses instruction-tuned models, ranging from 500 million parameters to 32 billion parameters. A wider family of models and of different sizes could provide additional insights.

Data Coverage. The datasets used comprised single statements with clear-cut truth values for checking. In practice, we are more keen on checking entire passages, with a combination of factual statements that require verification, and non-factual claims that do not. A comprehensive system to effectively decompose the passages and filter statements before hallucination detection would make our pipeline more complete.

Interpretability of Internal Signals. Although our method benefits from accessing latent internal activations, the black-box nature of these signals poses challenges for interpretability. Understanding the exact relationship between these internal cues and the truthfulness of a statement remains an open problem, suggesting that future work should also focus on enhancing the transparency and explainability of the detection process.

Acknowledgments

We thank all the anonymous reviewers for their constructive feedback on our work. This research is supported by DSO grant DSOCL23216.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. 2023. Lm-polygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Preprint*, arXiv:2305.19187.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

MistralAI. 2024. Mistral-7b-instruct-v0.3.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. Are LLMs good zero-shot fallacy classifiers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14338–14364, Miami, Florida, USA. Association for Computational Linguistics.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiaobao Wu. 2025. Sailing ai by the stars: A survey of learning from rewards in post-training and test-time scaling of large language models. *arXiv preprint arXiv:2505.02686*.

Xiaobao Wu, Xinshuai Dong, Thong Nguyen, and Anh Tuan Luu. 2023. Effective neural topic modeling with embedding clustering regularization. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*.

Xiaobao Wu, Thong Thanh Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Anh Tuan Luu. 2024c. AKEW: Assessing knowledge editing in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15118–15133, Miami, Florida, USA. Association for Computational Linguistics.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. 2024d. Antileak-bench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. *arXiv preprint arXiv:2412.13670*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models, 2023. *arXiv preprint arXiv:2305.10601*.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Appendix

A Prompt Templates

Model Response

```
Determine whether the given statement is TRUE or FALSE.
Provide your reasoning and then exactly 'TRUE.' or 'FALSE.'
Do not output any extra text, code, or commentary.
The reasoning process and answer are enclosed within
<think> </think> and <answer> </answer> tags, respectively,
i.e., <think> reasoning process here </think> <answer>
answer here </answer>.

Here is the statement.
Statement: {statement} <think>
```

Verbalized Confidence

```
Provide the probability that the given statement is correct.
The probability must be between 0.00 and 1.00.
Statement: {statement}
```

Figure 6: Prompts to verify statement correctness and to obtain verbalized confidence.

B Prompt Sensitivity

To test sensitivity to different prompts, we consider 3 additional common prompting approaches on the True-False dataset. The first one is zero-shot CoT prompt (CoT-0). The other two are adversarial prompts, as suggested by Turpin et al. (2023): (i) Suggested Answer: we append "I think the answer is false" to the end of every query; and (ii) Always False: in the few-shot demonstration, the answer is always "False", regardless of the reason.

Prompt	Anim.	Cit.	Comp.	Elem.	Facts	Gen.	Invent.	Avg
CoT-0	82.18	92.12	94.58	90.86	95.93	83.67	94.32	91.16
Suggested	79.70	91.78	92.50	89.78	91.87	86.36	87.50	88.91
Always F	82.18	92.81	93.33	87.10	94.31	79.59	77.27	87.86
CoT	80.20	93.84	94.17	89.25	93.50	87.76	90.91	90.38

A slight drop in performance is observed for the adversarial prompts, but overall, IRIS is robust to these prompting techniques. With the adversarial prompts, the response is mostly correct, but the final answer is forced to comply with the demonstrations, resulting in a wrong answer. However, the hidden states contain useful information of correctness, and thus, the classifier can correctly identify using the hidden states. On the other hand, when these adversarial prompts are applied to get the verbalized confidence, the scores are all pushed to zero, rendering the pseudolabels ineffective for training the classifier probe.

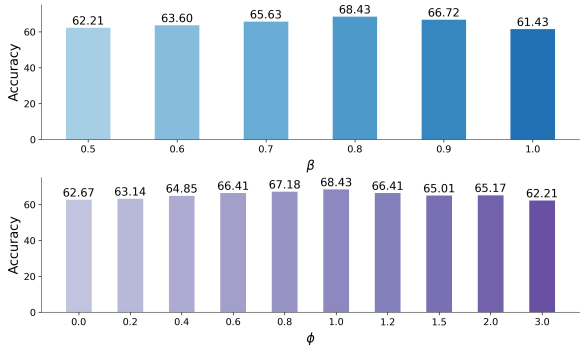

Figure 7: HELM average accuracy (%) with Llama-3.1-8B-Instruct for various values of β and ϕ.

C Ablation Study

Classifier Hyperparameters. In Figure 7, we provide results for various values of the hyperparameters β and ϕ. β denotes the importance given to the pseudolabel $\tilde{y}$ compared to the classifier's current state $\hat{y}$. β cannot be too low and requires careful tuning. Meanwhile, the accuracy is less sensitive to ϕ, except for when $\phi > 1$ where a considerable drop in performance is observed. Overall, the incorporation of soft bootstrapping ($\beta < 1$) and the symmetric cross entropy loss ($\phi > 0$) provides additional performance gains.

Loss Function. We further tested using L_1 and L_2 loss to train the classifier. The bootstrapping mechanism is maintained with $\beta = 0.8$. The accuracy with L_1 and L_2 loss are 62.51% and 58.16%, respectively. A significant drop in accuracy is observed, highlighting the benefit of using the cross entropy loss.

D Case Study

While CoT prompting generally improves performance for large models, smaller models do not always enjoy the same benefits. Here, we provide some examples from Llama-3.1-8B-Instruct when reasoning is helpful and when it is harmful.

Good case. For the statement `'Human uses for hyena include conservation, research.'`, the reasoning is provided as: `'Hyenas are primarily viewed as pests or scavengers and are not typically kept in captivity for conservation or research purposes. </think>\n\n<answer> FALSE. </answer>'` Without reasoning, the model incorrectly claims the statement is true.

Bad case. For `'The giant anteater uses walking for locomotion.'`, even though it is true, the reasoning is provided as: `'The giant anteater primarily uses its powerful front legs and long claws for walking, but it also uses its long, sharp claws for digging and its tail for balance. </think>\n\n<answer> FALSE. </answer>'`. Although the intermediate reasoning is correct, it deduced wrongly that the statement is false. However, its internal states reflect more accurately what the model knows, and the probe accordingly classifies the statement as correct.

E Future Extensions

Other Uncertainty Scores as Pseudolabels

IRIS offers a bridge between uncertainty-based methods and internal knowledge extraction. More advanced uncertainty metrics, such as SAR (Duan et al., 2024), can be flexibly incorporated to generate pseudolabels to further enhance IRIS. The performance is shown in Table 4. Compared to simply using SAR as a threshold (see Table 1), IRIS efficiently extracts the information embedded in the internal states to accurately assess the statement veracity.

Label	Anim.	Cit.	Co.	El.	Facts	Gen.	Inv.	Avg
SAR	70.79	73.29	78.25	59.68	85.37	79.59	77.84	73.88
Verb	80.20	93.84	94.17	89.25	93.50	87.76	90.91	90.38

(a) True-False

Label	Bio-Med	Edu	Fin.	Open	Sci.	Avg
SAR	67.47	68.03	77.25	93.22	76.50	74.38
Verb	59.64	63.95	78.84	93.22	70.00	70.57

(b) HaluEval2

Label	Falcon	GPT-J	LLB-7B	LLC-7B	LLC-13B	OPT-7B	Avg
SAR	65.71	61.74	68.14	63.00	67.71	70.18	66.10
Verb	67.62	69.57	69.03	66.00	69.79	68.42	68.43

(c) HELM

Table 4: SAR uncertainty as pseudolabels.

Specialized Dataset

Our preliminary investigation with mathematical reasoning datasets, i.e Arithmetic and GSM8K (Cobbe et al., 2021), shows that IRIS can potentially tackle specialized datasets. In particular, IRIS is applied to identify correct or incorrect answers to mathematical questions. IRIS achieves an accuracy of 87.2% and 73.4% on the Arithmetic and GSM8K datasets, respectively, compared to CoT prompting with 80.6% and 69.9%. However, on these datasets, it is more challenging to elicit verbalized confidence. When the prompting is done poorly, we observe that the model outputs a confidence of 0 for most queries, making the pseudolabels misguiding. In future work, experiments can be conducted on a wider range of reasoning datasets to understand the effectiveness of IRIS on these specialized domains.