# Representation-based Broad Hallucination Detectors Fail to Generalize Out of Distribution

**Zuzanna Dubanowska[1], Maciej Żelaszczyk[1], Michał Brzozowski[1], Paolo Mandica[1], Michał Karpowicz[1]**
[1]Samsung AI Research Center, Warsaw, Poland
**Correspondence:** z.dubanowska@samsung.com

## Abstract

We critically assess the efficacy of the current SOTA in hallucination detection and find that its performance on the `RAGTruth` dataset is largely driven by a spurious correlation with data. Controlling for this effect, state-of-the-art performs no better than supervised linear probes, while requiring extensive hyperparameter tuning across datasets. Out-of-distribution generalization is currently out of reach, with all of the analyzed methods performing close to random. We propose a set of guidelines for hallucination detection and its evaluation.

## 1 Introduction

While LLMs (Grattafiori et al., 2024; Jiang et al., 2023; Bai et al., 2023; Biderman et al., 2023) have made significant progress in various scenarios, they still display undesirable behavior, known as *hallucinations*, which undermines their reliability in both critical and everyday tasks. Hallucinations have been extensively studied, with a wide range of detection methodologies proposed depending on the level of access to model internals. **Black-box methods** operate on model input-output behavior. These include frameworks verifying generated responses against context (Es et al., 2024; Saad-Falcon et al., 2024; Hu et al., 2024a), evaluating self-consistency across multiple response variants (Manakul et al., 2023) or multiple verification steps (Friel and Sanyal, 2023), querying the model to evaluate the truthfulness of it's own response (Kadavath et al., 2022) or training classifiers to flag inconsistencies (Ádám Kovács and Recski, 2025). **Grey-box methods** center on uncertainty quantification through metrics like token-level log probabilities (Hu et al., 2024b) or entropy of token distribution (Farquhar et al., 2024). Recent **white-box methodologies** (Du et al., 2024; Sriramanan et al., 2024) have focused on detecting hallucinations by leveraging internal representations, moving beyond text-based detection. A line of work learns to predict the hallucination by training linear probes on the hidden states of LLMs directly (Azaria and Mitchell, 2023; Li et al., 2023) or to approximate derived metrics like semantic entropy (Kossen et al., 2024) or semantic consistency (Chen et al., 2024). Among white-box methods, many recent approaches rely on analyzing the attention mechanisms of Transformer models, which has garnered significant research interest. Working under the assumption that *retrieval heads* (Wu et al., 2024) are an important mechanism for information propagation within LLMs, Gema et al. (2024) show that masking the retrieval heads leads to increased hallucination incidence and that contrasting the next-token predictions of the base model with those of the masked one can act as a mitigation mechanism. Sun et al. (2025) further explores this attention-based methodology, which we analyze in depth in Section 2. Treating all types of hallucinations as a single category is challenging, resulting in methods that perform well in some scenarios (e.g., hallucination detection in summarization tasks) but poorly in others. Retrieval-Augmented Generation (RAG) has gained traction as a method to potentially decrease hallucinations. Nevertheless, even with RAG, models can still fail to correctly attend to the context or overwrite it with parametric knowledge. Due to the limited availability of human-labeled RAG hallucination datasets, studying hallucinations related to contextual errors is difficult, with only a few attempts so far (Sun et al., 2025; Ravi et al., 2024; Ádám Kovács and Recski, 2025). We focus on the current SOTA approaches for detecting hallucinations using internal representations, with the goal of defining a clearer future trajectory for hallucination detection. Our contributions are as follows. **(A)** We verify hallucination detectors based on model internals via linear probes, a random forest classifier and sparse-autoencoder (SAE) probes. **(B)** Assess SOTA and find that most of its performance is due to spurious



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17563–17575
November 4-9, 2025 

correlation rather than genuine hallucinatory signal. **(C)** Analyze detection out-of-distribution and find that generalization remains a challenge. **(D)** Find that SAE features do not provide consistent benefits for hallucination detection.

## 2 State of the art

ReDeEP (Sun et al., 2025) employs internal model mechanisms to identify hallucinations, specifically those cases where external context serves as the ground truth but may be compromised by the model's parametric knowledge. It relies on copying heads (Elhage et al., 2021) and assumes that attention scores are an appropriate proxy for what information the model integrates from the context. Moreover, the parametric interventions of the model, measured by the Jensen-Shannon divergence between pre- and post-MLP representations, serve as an important hallucination indicator. They hypothesize that increased attention scores correlate with a decrease in the hallucination incidence, while elevated parametric interventions correspond to increased occurrence of hallucinations. ReDeEP comes in two flavors: token- and chunk-based, where the attention and parametric scores are measured on individual tokens or token chunks, respectively. ReDeEP achieves SOTA hallucination detection results on two publicly available RAG hallucinations datasets: RAGTruth and Dolly (AC).

Azaria and Mitchell (2023) introduce SAPLMA, another representation-based hallucination detection method. They investigate whether the LLM possesses some notion of truthfulness of generated statements in its hidden state. The authors postulate that if it does, this information can be used to detect and mitigate hallucinations. Based on experimental results, they hypothesize that the hidden states indeed encode signals of truth or falsehood, leading to the hypothesis that the model may "know" when it hallucinates. SAPLMA performs on-par with ReDeEP in hallucination detection on RAGTruth.

## 3 Spurious correlation explains SOTA performance

ReDeEP's evaluation focuses exclusively on the performance across all three subtasks in the RAGTruth dataset. This prompts us to investigate its performance on individual tasks. Table 1 reveals significant variability in performance across individual tasks, with overall performance generally lower than that on the entire dataset. To understand the disparity in performance between tasks, we examine the composition of the RAGTruth dataset (details in Appendix B).

| Model | Task Type | AUC | PCC | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| LLaMA-2 7B Chat | D2T | 0.3951 | -0.1246 | **0.7931** | **0.748** | **0.7699** |
| | QA | **0.6360** | **0.2122** | 0.4528 | 0.4615 | 0.4571 |
| | Summary | 0.5767 | 0.1121 | 0.4839 | 0.2941 | 0.3659 |
| | Overall | 0.7324 | 0.3978 | 0.7217 | 0.6770 | 0.6986 |
| LLaMA-2 13B Chat | D2T | **0.6679** | 0.1545 | **0.9357** | **0.9493** | **0.9424** |
| | QA | 0.6503 | **0.2259** | 0.3103 | 0.7500 | 0.4390 |
| | Summary | 0.5342 | 0.0424 | 0.2500 | 0.7879 | 0.3796 |
| | Overall | 0.8177 | 0.5484 | 0.5875 | 0.8599 | 0.6980 |
| LLaMA-3 8B Instruct | D2T | 0.5084 | -0.0072 | **0.8707** | 0.7652 | **0.8145** |
| | QA | **0.5974** | **0.1573** | 0.5514 | 0.7564 | 0.6378 |
| | Summary | 0.5593 | 0.1102 | 0.2475 | **0.8065** | 0.3788 |
| | Overall | 0.7534 | 0.4512 | 0.6465 | 0.8000 | 0.7151 |

Table 1: ReDeEP (token) performance on RAGTruth per task type. Best task results (excluding Overall) are bolded.

| Model | Task Type | Hallucination Rate |
|---|---|---|
| LLaMA-2 7B | D2T | 0.8596 |
| | QA | 0.5157 |
| | Summary | 0.4602 |
| LLaMA-3 8B Instruct | D2T | 0.8800 |
| | QA | 0.5200 |
| | Summary | 0.2200 |

Table 2: Hallucination rates per task type on LLaMA-2 7B and LLaMA-3 8B. See Appendix E, Table 10 for other models.

The fraction of model responses which human annotators labeled as hallucinations (hallucination rate) is notably higher for the D2T task (Table 2 and Table 10). This means that a random sample from the D2T task is far more likely to be hallucinatory than for other tasks, leading to a strong correlation between task type and the hallucination label, which may be exploited by ReDeEP.

We construct a *naïve classifier* which follows a heuristic: it predicts 1 (i.e., hallucinated sample) if the input's task type is D2T, and 0 for all other tasks. In Table 3, we compare the results of the naïve classifier with those reported in the ReDeEP paper for the RAGTruth test set. The AUC metrics are very similar, and the naïve classifier even outperforms ReDeEP in terms of the Pearson correlation between the true hallucination label and the model score. This suggests that the high scores on the overall RAGTruth benchmark are mostly due to the imbalance in the hallucination incidence between the tasks. Additionally, our naïve classifier can outperform SEP (Kossen et al., 2024) and SAPLMA (Azaria and Mitchell, 2023) - supervised methods.

The D2T task is highly specific as it relies on

| CLASSIFIER | AUC | PCC |
|---|---|---|
| naïve | 0.7119 | **0.4494** |
| SEP | 0.7143 | 0.3355 |
| SAPLMA | 0.7037 | 0.3188 |
| ReDeEP (token) | 0.7325 | 0.3979 |
| ReDeEP (chunk) | **0.7458** | 0.4203 |

Table 3: Model performance metrics on RAGTruth using LLaMA-2 7B.

| EVAL DATASET | HYPER-PARAMS | AUC | PCC | RECALL | F1 |
|---|---|---|---|---|---|
| RAGTruth | RAGTruth | 0.7541 | 0.4522 | 0.8008 | 0.7148 |
| | Dolly | 0.7230 | 0.4413 | 0.6390 | 0.6844 |
| Dolly | Dolly | 0.6223 | 0.2129 | 0.8684 | 0.5841 |
| | RAGTruth | 0.5005 | -0.0188 | 0.8684 | 0.5410 |

The EVAL DATASET column specifies the dataset on which the method is evaluated, whereas the HYPER-PARAMS column indicates the dataset used to fine-tune the hyperparameters of the method.

Table 4: Cross-dataset evaluation results of ReDeEP – LLaMA-3 8B.

prompts in the JSON format (Appendix B, Table 8). Therefore, detectors trained and tested on the entire RAGTruth dataset may in reality respond to the presence of JSON in the prompt. To verify that detecting the type of the task from model activations is possible, we collect model activations from the last token and the last layer of the Llama2-7B-Chat model and we use them to train a logistic regression to predict the JSON task. The trained linear probe achieves perfect prediction on the test set, with an AUC of 1.0.

The evaluation problems are not an issue with the RAGTruth dataset itself. In fact, the results in the dataset paper (Niu et al., 2024) are reported per task. This distinction also exists in methods proposed in (Song et al., 2024; Belyi et al., 2025; Ádám Kovács and Recski, 2025). However, aggregated metrics have also been reported in (Ravi et al., 2024; Sriramanan et al., 2024).

## 4 Hallucination detection with model internals

The fundamental question we should address first is this: *Is it possible to classify a response from an LLM as hallucinatory based on its internal states?* If this is possible, then it would be a natural benchmark against other methods based on model internals. To provide an answer, we extract the activations of the LLM in the prompt processing and answer generation phases and then use those activations as input to a classifier. Extraction is performed from the residual stream: pre-attention (resid_pre) and pre-MLP (resid_mid). We consider linear and non-linear probes (raw activations, details in Appendix D), and SAE probes (SAE features, details in Appendix D.1).

### 4.1 Experimental results

We assess ReDeEP's generalization capabilities in a cross-dataset scenario. As shown in Table 4, ReDeEP requires specific hyperparameter tuning to perform effectively on different datasets. Although this issue is only slightly noticeable when evaluating on RAGTruth with hyperparameters optimized for Dolly, it becomes glaringly apparent when evaluating on Dolly using hyperparameters optimized for RAGTruth, where the AUC is equivalent to that of a random classifier and the correlation with the hallucination label is close to zero. The high recall on Dolly can be attributed to the low classification threshold employed by ReDeEP (approximately 0.15).

The evaluation results of our probes on LLaMA models and RAGTruth dataset are presented in Tables 5 and 14. For completeness, we include results from additional model architectures in Appendix E (Tables 15 and 16). The performance on the SQuAD dataset is summarized in Tables 18 and 19, while the Dolly evaluation results are presented in Tables 20 and 21.

Given our experimental results, we find that performance across hallucination methods is highly fragmented: different classifiers perform best depending on the dataset, model, or task, with no consistent winner across settings. In particular, there is no clear advantage of SoTA detection methods over simple linear probes. In many cases, linear classifiers trained on model activations match or even outperform more complex methods like ReDeEP, SAPLMA or SAE-based classifiers. This suggests that current approaches may be overfitting to task-specific artifacts rather than capturing generalizable signals of hallucination.

To further evaluate robustness, we have assessed all methods in a cross-task setting on RAGTruth (cf. Table 7 and 17), as well as in the RAGTruth → SQuAD (cf. Table 6 and 22) and SQuAD → RAGTruth cross-dataset settings (cf. Table 24 and 25) specifically aimed at evaluating generalization capabilities. Across the board, performance dropped substantially, with no method demonstrating consistent transferability. Both SoTA and linear probes exhibit near-random performance when applied outside their training distribution. This supports our hypothesis that hallucination detectors are latching onto task- or dataset-specific cues. The failure

| Method | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| ReDeEP[a] | 0.6360 | 0.4528 | 0.4615 | 0.4571 | 0.3951 | 0.7931 | 0.7480 | **0.7699** | 0.5767 | 0.4839 | 0.2941 | 0.3659 | 0.7324 | 0.7217 | 0.677 | 0.6986 |
| Logistic Regression[b] | $0.6900_{\pm 0.02}$ | $\mathbf{0.6912}_{\pm 0.01}$ | $0.6900_{\pm 0.03}$ | $\mathbf{0.6901}_{\pm 0.01}$ | $\mathbf{0.6555}_{\pm 0.00}$ | $0.6555_{\pm 0.00}$ | $0.7611_{\pm 0.00}$ | $0.6777_{\pm 0.00}$ | $0.6376_{\pm 0.01}$ | $0.6380_{\pm 0.02}$ | $0.6376_{\pm 0.04}$ | $\mathbf{0.6378}_{\pm 0.01}$ | $\mathbf{0.7951}_{\pm 0.01}$ | $\mathbf{0.7951}_{\pm 0.01}$ | $0.7930_{\pm 0.01}$ | $\mathbf{0.7934}_{\pm 0.01}$ |
| Random Forest[c] | $0.6821_{\pm 0.01}$ | $0.6864_{\pm 0.01}$ | $0.6821_{\pm 0.01}$ | $0.6817_{\pm 0.01}$ | $0.5227_{\pm 0.00}$ | $0.5227_{\pm 0.00}$ | $0.9318_{\pm 0.00}$ | $0.5068_{\pm 0.00}$ | $\mathbf{0.6410}_{\pm 0.02}$ | $\mathbf{0.6598}_{\pm 0.02}$ | $\mathbf{0.6410}_{\pm 0.01}$ | $0.6376_{\pm 0.02}$ | $0.6994_{\pm 0.03}$ | $0.6994_{\pm 0.02}$ | $0.7191_{\pm 0.03}$ | $0.7050_{\pm 0.03}$ |
| SAE Classifier[d] | $\mathbf{0.7106}_{\pm 0.00}$ | $0.5325_{\pm 0.00}$ | $\mathbf{0.7885}_{\pm 0.00}$ | $0.6804_{\pm 0.00}$ | $0.6391_{\pm 0.00}$ | $\mathbf{0.8750}_{\pm 0.00}$ | $0.7967_{\pm 0.00}$ | $0.6170_{\pm 0.00}$ | $0.6182_{\pm 0.00}$ | $0.4821_{\pm 0.00}$ | $0.5294_{\pm 0.00}$ | $0.6150_{\pm 0.00}$ | $0.7105_{\pm 0.00}$ | $0.6655_{\pm 0.00}$ | $\mathbf{0.8540}_{\pm 0.00}$ | $0.7048_{\pm 0.00}$ |
| SAPLMA[e] | $0.5699_{\pm 0.02}$ | $0.3905_{\pm 0.02}$ | $0.500_{\pm 0.18}$ | $0.5178_{\pm 0.02}$ | $0.5476_{\pm 0.01}$ | $0.8200_{\pm 0.00}$ | $\mathbf{1.0000}_{\pm 0.00}$ | $0.4505_{\pm 0.00}$ | $0.5959_{\pm 0.03}$ | $0.3976_{\pm 0.03}$ | $0.5294_{\pm 0.03}$ | $0.5427_{\pm 0.03}$ | $0.7486_{\pm 0.00}$ | $0.6300_{\pm 0.03}$ | $0.7788_{\pm 0.05}$ | $0.6479_{\pm 0.03}$ |

[a] ReDeEP is a deterministic method. Results do not vary between runs.
[b] Best result - based on layer 15's activations.
[c] Best result - based on layer 5's activations.
[d] QA: layer 15, max activation. D2T: layer 12, last token of response, contrastive. Summary: layer 13, last token of response, contrastive. Overall: layer 13, max activation.
[e] Best result - based on layer 16's activations.

Table 5: RAGTruth evaluation results - LLaMA-2 7B.

| Method | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| ReDeEP | - | - | - | - | - | - | - | - | - | - | - | - | 0.5163 | 0.4878 | 0.7092 | **0.578** |
| Logistic Regression | **0.5086** | **0.5000** | 0.5100 | 0.3100 | 0.5072 | 0.5000 | 0.5100 | 0.2300 | 0.5424 | **0.5100** | 0.5300 | 0.2800 | 0.5139 | **0.5100** | 0.5100 | 0.3300 |
| Random Forest | 0.5077 | **0.5000** | 0.5100 | **0.4800** | 0.5178 | 0.5100 | 0.5200 | 0.4300 | 0.5049 | 0.5000 | 0.5000 | 0.3400 | 0.5033 | 0.5000 | 0.5000 | 0.4500 |
| SAE classifier | 0.4889 | 0.4915 | **0.8112** | 0.4291 | **0.5273** | **0.5148** | 0.8531 | **0.4705** | 0.4856 | 0.4901 | **0.8671** | 0.3974 | 0.4845 | 0.4841 | 0.5315 | 0.4833 |
| SAPLMA | 0.5000 | 0.0833 | 0.0051 | 0.4256 | 0.4388 | 0.2603 | **1.000** | 0.2084 | **0.5583** | 0.3671 | 0.5686 | **0.5113** | **0.5520** | 0.2610 | **1.000** | 0.2124 |

Results presented for layer / hyperparameter combinations performing best on RAGTruth on the Overall task (Table 5), except for the SAE classifier for which the activation on the last token of the response has been used instead of the max activation.

Table 6: Evaluation results | train: RAGTruth | eval: SQuAD - LLaMA-2 7B.

to generalize even across similarly structured QA tasks underscores the limitations of using internal activations as a reliable signal for hallucination detection.

### 4.2 Are we really measuring the right thing?

The problem of measuring general hallucinations based on model internals remains open. Spurious correlation is part of a larger issue. There has been some doubt about hallucination detectors. (Levinstein and Herrmann, 2024) demonstrate how SAPLMA (Azaria and Mitchell, 2023) does not predict truth but rather another spurious phenomenon, such as *Sentence is true and contains no negation.* Even minor changes to the test dataset, like negating sentences, make SAPLMA's accuracy random. Contrast-Consistent Search (CCS) (Burns et al., 2024) finds a direction in the activation space that satisfies logical consistency properties, such as having opposite truth values for a statement and its negation. However, these axioms are insufficient, and the CCS probe identifies sentences with negations (Levinstein and Herrmann, 2024; Farquhar et al., 2023). Another line uses uncertainty quantification metrics, such as perplexity, length-normalized entropy (Sun et al., 2025), semantic entropy (Kossen et al., 2024), and P(true) (Kadavath et al., 2022). However, uncertainty metrics correlate with sequence length, skewing evaluations (Santilli et al., 2024). The overall goal in hallucination detection is reminiscent of the controversial status of polygraphs (Lykken, 1998). There is no consensus on whether hallucination behavior in a general context can be detected using current MI methods. While simpler behaviors like refusal (Arditi et al., 2024) or sentiment (Tigges et al., 2023) have been explained by a single linear direction, the hallucination phenomenon is less clearly defined.

### 4.3 Evaluation guidelines

We propose a number of best practices for future work in mechanistic interpretability of hallucinations. First, consider a rigorous mathematical definition of a hallucination. In the absence of one, it is challenging to design a detector. Then consider what follows. **(I)** Check against naïve classifiers based on training set features (see Section 3) and **(II)** against simple linear probes. **(III)** A detector method should be trained or tuned on a specific dataset and evaluated on a different one. This also applies to unsupervised methods, like ReDeEP, which depend on dataset-specific hyperparameters for optimal performance (see Appendix I in ReDeEP (Sun et al., 2025)). An example of that would be training / tuning on SQuAD and evaluating on RAGTruth. **(IV)** Verify if the suspected *truth circuit* satisfies logical requirements like negation and de Morgan rules. **(V)** Attempt to highlight the incorrect part of the answer. This can be done using external BERT models, as seen in (Ádám Kovács and Recski, 2025). The RAGTruth dataset has span-level labels that allow testing this. To our knowledge, there is no purely activation-based detection method evaluated in this manner.

Also worth considering is that LLMs are alike imagination engines - hallucination enables the exploration of ideas and options. In this light, detecting hallucinations could be related to notions of orthogonality between the generated answer and

| Method | Eval task | Question Answering | | | | Data-to-text writing | | | | Summarization | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| Logistic Regression | QA | 0.5720 | 0.5700 | 0.5700 | 0.5700 | 0.5596 | 0.5500 | 0.5600 | 0.5500 | 0.5280 | 0.5300 | 0.5300 | 0.5200 |
| | D2T | 0.5140 | 0.5600 | 0.5100 | 0.4000 | 0.5564 | 0.6400 | 0.5600 | 0.5600 | 0.5113 | 0.5200 | 0.5100 | 0.4400 |
| | Summ. | 0.5332 | 0.5400 | 0.5200 | 0.4800 | 0.4464 | 0.4700 | 0.4500 | 0.3900 | 0.6006 | 0.6000 | 0.6000 | 0.6000 |
| Random Forest | QA | 0.5886 | 0.5900 | 0.5800 | 0.5900 | 0.5402 | 0.5200 | 0.5400 | 0.5000 | 0.5006 | 0.5000 | 0.5000 | 0.5000 |
| | D2T | 0.4910 | 0.4800 | 0.4900 | 0.4300 | 0.5000 | 0.4000 | 0.5000 | 0.4400 | 0.5080 | 0.5200 | 0.5100 | 0.4100 |
| | Summ. | 0.5115 | 0.5100 | 0.5100 | 0.4900 | 0.5055 | 0.5000 | 0.5100 | 0.3700 | 0.5177 | 0.5300 | 0.5200 | 0.4900 |
| SAE Classifier | QA | 0.7055 | 0.5256 | 0.7885 | 0.6742 | 0.5000 | 0.3467 | 1.000 | 0.2574 | 0.5051 | 0.3490 | 1.000 | 0.2688 |
| | D2T | 0.5000 | 0.8200 | 1.000 | 0.4505 | 0.5000 | 0.8200 | 1.000 | 0.4505 | 0.5000 | 0.8200 | 1.000 | 0.4505 |
| | Summ. | 0.5000 | 0.3400 | 1.000 | 0.2537 | 0.5000 | 0.3400 | 1.000 | 0.2537 | 0.5642 | 0.4231 | 0.4314 | 0.5638 |
| SAPLMA | QA | 0.5836 | 0.4058 | 0.5385 | 0.5498 | 0.4584 | 0.3467 | 1.000 | 0.2574 | 0.4790 | 0.4667 | 0.1346 | 0.4907 |
| | D2T | 0.5929 | 0.8200 | 1.000 | 0.4505 | 0.5330 | 0.8200 | 1.000 | 0.4505 | 0.6534 | 0.8704 | 0.7642 | 0.5953 |
| | Summ. | 0.5670 | 0.3651 | 0.9020 | 0.4144 | 0.5592 | 0.3400 | 1.000 | 0.2537 | 0.5661 | 0.3421 | 0.5098 | 0.4880 |

Results presented for layer / hyperparameter combinations performing best on the Overall task (Table 5).

Table 7: RAGTruth cross-task evaluation results - LLaMA-2 7B.

the context provided to the model.

We find the hurdles in hallucination detection to be similar to the challenges in evaluating adversarial defenses (Carlini et al., 2019). (Carlini, 2020) suggest that instead of trying to find a method for all cases, it is better to restrict the scope. For hallucination detection, we should shift the attention to finding specific, clearly defined sub-hallucinations. **(1)** A method to verify if a given entity is known to the model (Ferrando et al., 2024). **(2)** A mechanism by which LLMs switch between contextual and parametric knowledge (Zhao et al., 2024; Minder et al., 2025).

## 5 Conclusion

We have assessed general hallucination detectors and found that SOTA performance on the RAGTruth dataset may be overstated due to evaluation problems. On top of that, even in more rigorous settings, SOTA is often outperformed in-distribution by linear probes. Furthermore, each of the considered methods performs no better than random out-of-distribution. This underscores the challenge in general hallucination detection, where current methods based on model internals are unlikely to generalize to unseen data. In light of this challenge, we propose a set of guidelines for future hallucination detection methods, in particular: restricting the scope of the hallucination and focusing on specific hallucination spans rather than general labels - potentially promising angles of attack.

## Limitations

This work is limited in that out-of-distribution generalization is currently not achieved by any method we are aware of but the introduction of new detection methods might change this state of affairs, so it is not a definite statement on the impossibility of such methods. Additionally, while we have striven to provide a comprehensive assessment of cross-dataset performance, it may be possible to show that there are approaches which work significantly better than chance on specific dataset combinations, or for narrowed down specifications of what a hallucination is, e.g. only untruthful answers, like is done in the Truthful QA dataset (Lin et al., 2022).

## References


Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *Preprint*, arXiv:2406.11717.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2025. Luna: A lightweight evaluation model to catch language model hallucinations with high accuracy and low cost. In *Proceedings of the 31st International Conference on Computational Linguistics:*

*Industry Track*, pages 398–409, Abu Dhabi, UAE. Association for Computational Linguistics.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *Preprint*, arXiv:2404.14082.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2024. Discovering latent knowledge in language models without supervision. *Preprint*, arXiv:2212.03827.

Nicholas Carlini. 2020. Are adversarial example defenses improving? an empirical study of the evolution of image recognition defenses.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On evaluating adversarial robustness. *Preprint*, arXiv:1902.06705.

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. In *The Twelfth International Conference on Learning Representations*.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. 2025. Not all language model features are one-dimensionally linear. *Preprint*, arXiv:2405.14860.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. 2023. Challenges with unsupervised llm knowledge discovery. *Preprint*, arXiv:2312.10029.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do I know this entity? knowledge awareness and hallucinations in language models. *Preprint*, arXiv:2411.14257.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *Preprint*, arXiv:2310.18344.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *Preprint*, arXiv:2406.04093.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2024. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. *Preprint*, arXiv:2410.18860.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024a. Refchecker: Reference-based fine-grained hallucination checker and benchmark for large language models. *Preprint*, arXiv:2405.14486.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 2024b. Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in*

*Natural Language Processing*, pages 1950–1959, Miami, Florida, USA. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

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

B.A. Levinstein and D.A. Herrmann. 2024. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

David T Lykken. 1998. *A Tremor in the Blood: Uses and Abuses of the Lie Detector, 2d ed*. Perseus, New York.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Julian Minder, Kevin Du, Niklas Stoehr, Giovanni Monea, Chris Wendler, Robert West, and Ryan Cotterell. 2025. Controllable context sensitivity and the knob behind it. *Preprint*, arXiv:2411.07404.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *Preprint*, arXiv:2407.08488.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Preprint*, arXiv:2311.09476.

Andrea Santilli, Miao Xiong, Michael Kirchhof, Pau Rodriguez, Federico Danieli, Xavier Suau, Luca Zappella, Sinead Williamson, and Adam Golinski. 2024. On the protocol for evaluating uncertainty in generative question-answering tasks. In *Neurips Safe Generative AI Workshop 2024*.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall,

Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *Preprint*, arXiv:2310.15154.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *Preprint*, arXiv:2404.15574.

Yu Zhao, Alessio Devoto, Giwon Hong, Xiaotang Du, Aryo Pradipta Gema, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. 2024. Steering knowledge selection behaviours in llms via sae-based representation engineering. *Preprint*, arXiv:2410.15999.

Ádám Kovács and Gábor Recski. 2025. Lettucedetect: A hallucination detection framework for rag applications. *Preprint*, arXiv:2502.17125.

## A Hallucination types

Hallucinations are a byproduct of all stages of the capability acquisition process in LLMs. They originate from artifacts in pre-training data, the training and alignment procedures, the decoding strategies (Huang et al., 2025). In-context learning (Brown et al., 2020) was introduced as a strategy to mitigate hallucinations in LLMs by grounding generation in externally provided information at inference time. Yet, hallucinations occur even in the presence of adequate context. Recent analyses (Huang et al., 2025) suggest these failures stem not only from missing or irrelevant information, but from how the models internally manage and integrate contextual input with their parametric knowledge. These hallucinations may reflect breakdowns in *contextual awareness* - where the relevant context is available but poorly attended to, lost in long inputs or overriden by conflicting parametric knowledge - and *contextual alignment* - where information is misattributed or incorrectly decoded. Our work examines whether such failures can be detected from the model's internal activations.

## B RAGTruth

Given the ambiguity of the *hallucination* concept and the potential multitude of hallucination types, the characteristics of specific datasets used for training or hyperparameter tuning of detection methods become crucial. To the best of our knowledge, the `RAGTruth` dataset (Niu et al., 2024) is the only publicly-available manually-annotated RAG-based hallucination dataset. It comprises 2 965 prompts and 17 790 responses to those prompts from 6 LLMs: `GPT-3.5-Turbo-0613`, `GPT-4-0613`, `LLaMA-2-7B-Chat`, `LLaMA-2-13B-Chat`, `LLaMA-2-70B-Chat` and `Mistral-7B-Instruct`. The data is broken down into 3 tasks:

- Question Answering (`QA`): answer questions related to daily life, pre-selected from the `MS MARCO` dataset (Bajaj et al., 2016).
- Data-to-text Writing (`D2T`): provide an objective description of a randomly sampled business from the `Yelp Open Dataset` where part of the prompt contains JSON-formatted data.
- Summarization (`Summary`): summarize a piece of news from either the `CNN/Daily Mail` dataset (Hermann et al., 2015; Nallapati et al., 2016) or an undisclosed news platform.

An example of a prompt and response for the `LlaMA-2-7B-Chat` model with the hallucinated part of the response marked is visible in Table 8. The responses from the models are generated for non-zero temperature and so, are not consistently reproducible. The evaluation of hallucination detection and mitigation methods on `RAGTruth` focuses on the specific rollout present in the dataset and implicitly ignores the possibility that the model may not generate hallucinatory responses for a specific prompt for the majority of sampled rollouts not present in the dataset.

In general, the distinct characteristics of the tasks (`QA`, `D2T` and `Summary`) provide a hint that each of them may contain data artifacts not present in the other ones. Additionally, hallucination incidence varies significantly between the tasks, being markedly higher for `D2T`. Those two factors combined suggest that the artifacts may correlate with the hallucination label. Consequently, the evaluation of hallucination detection methods should take this into account. At the very least, it should provide performance metrics per task rather than aggregated for the whole dataset.

## C SQuAD

Stanford Question Answering Dataset (`SQuAD`) (Rajpurkar et al., 2018) is a reading comprehension benchmark comprising of context passages based

| PROMPT | ANSWER |
|---|---|
| Instruction:<br>Write an objective overview (...) based only on the provided structured data in the JSON format.<br>(...)<br>Structured data:<br>{<br>'name': 'Apna Indian Kitchen',<br>'address': '718 State St',<br>'city': 'Santa Barbara',<br>'state': 'CA',<br>(...),<br>'attributes': {<br>(...),<br>'OutdoorSeating': True,<br>'WiFi': 'no',<br>(...)<br>},<br>'business_stars': 4.5,<br>'review_info': [(. . .)]<br>}<br>Overview: | Apna Indian Kitchen is a highly rated restaurant located in Santa Barbara, California, serving a variety of gluten-free, gastropub, vegan, and Indian cuisine. According to customer reviews, the restaurant offers tasty and authentic Indian food with great service. Customers have praised the menu options, including mixed veg spicy and butter chicken medium spicy dishes, roti, and naan. The restaurant has a cozy atmosphere with outdoor seating and free WiFi. While the restrooms could use some attention, the serving sizes are perfect, and the staff is friendly. Customers have given the restaurant a 4.5-star rating, with many leaving positive reviews highlighting the quality of the food and service. |

Table 8: RAGTruth D2T example.

on a set of Wikipedia articles and questions created by crowdworkers (see Appendix C, Table 9 for an example). SQuAD's consistent format - answering questions directly from context - helps minimize the chance of spurious task-related correlations. It resembles the QA task in RAGTruth, where the model answers questions based on short context excerpts but, contrary to RAGTruth, SQuAD only contains human-generated reference answers and not model generated answers reviewed by humans. In our evaluation procedure, we prompted each model to answer questions based on context. We subsequently evaluated the generated answers as hallucinatory or non-hallucinatory by comparing them to reference answers using LLaMA-3 70B (Grattafiori et al., 2024) as a judge.

## D Activation probes

Linear probes have gained traction in the MI community as, under the *linear representation* hypothesis, they are sufficient to detect features in a model's representations (Bereska and Gavves, 2024). We train a separate Logistic Regression classifier for each layer's pre-attention and pre-MLP activations, using the last token's features to predict hallucinations across the model.

Given the high class imbalance (the hallucination rate in responses is approximately 10-20% depending on model and dataset) and a line of work in MI hinting at existence of non-linear features in LLMs (Engels et al., 2025), we decided to train a non-linear probe, a Random Forest classifier, on model activations alongside the linear probe.

### D.1 SAE probes

SAEs have garnered considerable attention in the MI community thanks to their ability to find at least some human-interpretable features (Templeton et al., 2024; Gao et al., 2024). It has been shown that SAE features may also store some meta information on the activity of the base model. For instance, they can, to an extent, regulate the strength of attending to the context vs. relying on parametric knowledge in LLMs (Zhao et al., 2024). Similarly, SAE activations carry information about the uncertainty of the LLM about an entity it is asked about (Ferrando et al., 2024). We extract SAE activations using the SAEs provided by (Zhao et al., 2024) for layers 12, 13, 14 and 15 of the LLaMA-2 7B model. All activations are rescaled per feature to the $[0; 1]$ range. Two extraction methods are used:

| CONTEXT | QUESTION | ANSWER |
|---|---|---|
| In many societies, beer is the most popular alcoholic drink. Various social traditions and activities are associated with beer drinking, such as playing cards, darts, or other pub games; attending beer festivals; engaging in zythology (the study of beer); visiting a series of pubs in one evening; visiting breweries; beer-oriented tourism; or rating beer. Drinking games, such as beer pong, are also popular. A relatively new profession is that of the beer sommelier, who informs restaurant patrons about beers and food pairings. | What is a popular drinking game where beer is often considered? | Beer pong |

Table 9: SQuAD dataset example.

activations at the last token of the response and the maximum activations over the prompt and the response. Additionally, the activations are provided to the classifiers in two flavors. In the first approach, they are directly treated as input. In the second one, a contrastive representation is calculated:

$$\hat{\mathbf{a}} = \mathbf{a}_H - \mathbf{a}_C \tag{1}$$

where $\hat{\mathbf{a}} \in \mathbb{R}^d$, $\mathbf{a}_H$ are the SAE activations for hallucinatory samples and $\mathbf{a}_C$ are the corresponding activations for non-hallucinatory samples. The dimensionality of SAE dictionary elements is $d$. We choose the top $k$ elements from $\hat{\mathbf{a}}$ with the highest magnitude on the train set. In our experiments, we set $k = 4\,096$ to match the dimensionality of the raw representation from the LLM in order to make the results more comparable with classification based on raw activations.

## E Additional experimental results

We present additional experimental results in Tables 10-25.

| Model | Task Type | Hallucination Rate |
|---|---|---|
| GPT-3.5 Turbo (0613) | D2T | 0.2633 |
| | QA | 0.0758 |
| | Summary | 0.0573 |
| GPT-4 (0613) | D2T | 0.2807 |
| | QA | 0.0425 |
| | Summary | 0.0785 |
| LLaMA-2 13B Chat | D2T | 0.9516 |
| | QA | 0.4034 |
| | Summary | 0.3128 |
| LLaMA-2 70B Chat | D2T | 0.8354 |
| | QA | 0.3236 |
| | Summary | 0.2248 |
| LLaMA-2 7B Chat | D2T | 0.8596 |
| | QA | 0.5157 |
| | Summary | 0.4602 |
| LLaMA-3 8B Instruct | D2T | 0.8800 |
| | QA | 0.5200 |
| | Summary | 0.2200 |
| Mistral 7B Instruct | D2T | 0.9274 |
| | QA | 0.3822 |
| | Summary | 0.6543 |

Table 10: Hallucination rates on RAGTruth per model and task type.

| Model | AUC | PCC |
|---|---|---|
| GPT-4 (0613) | 0.7757 | 0.3403 |
| GPT-3.5 Turbo (0613) | 0.7623 | 0.3372 |
| Mistral 7B Instruct | 0.7267 | 0.4777 |
| LLaMA-2 7B Chat | 0.7119 | 0.4494 |
| LLaMA-2 13B Chat | 0.8086 | 0.6526 |
| LLaMA-2 70B Chat | 0.7594 | 0.5342 |
| LLaMA-3 8B Instruct | 0.7084 | 0.4437 |

Table 11: The AUC and PCC scores on naïve classifier across models on RAGTruth.

| Classifier | AUC | PCC |
|---|---|---|
| naïve | 0.8086 | **0.6526** |
| SEP | 0.8089 | 0.5276 |
| SAPLMA | 0.8029 | 0.3956 |
| ReDeEP (token) | 0.8181 | 0.5478 |
| ReDeEP (chunk) | **0.8244** | 0.5566 |

Table 12: Model performance metrics on RAGTruth using LLaMA-2 13B.

| Classifier | AUC | PCC |
|---|---|---|
| naïve | 0.7281 | **0.4824** |
| SEP | 0.7004 | 0.3713 |
| SAPLMA | 0.7092 | 0.4054 |
| ReDeEP (token) | **0.7522** | 0.4493 |
| ReDeEP (chunk) | 0.7285 | 0.3964 |

Table 13: Model performance metrics on RAGTruth using LLaMA-3 8B.

| | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| ReDeEP | 0.5974 | 0.5514 | **0.7564** | 0.6378 | **0.5084** | **0.8707** | **0.7652** | **0.8145** | 0.5593 | 0.2475 | **0.8065** | 0.3788 | **0.7534** | 0.6465 | **0.8000** | **0.7151** |
| Logistic Regression[a] | **0.8750** | **0.8928** | **0.8750** | **0.8685** | 0.5000 | 0.4347 | 0.5000 | 0.4651 | 0.5166 | 0.5197 | 0.5166 | 0.5165 | 0.7001 | 0.7045 | 0.7000 | 0.6984 |
| Random Forest[b] | 0.8257 | 0.8257 | 0.8257 | 0.8257 | 0.5000 | 0.4347 | 0.5000 | 0.4651 | **0.6410** | **0.6598** | 0.6410 | **0.6376** | 0.8221 | **0.8221** | 0.8221 | 0.8221 |

[a] Best result - based on layer 14's activations.
[b] Best result - based on layer 8's activations.

Table 14: RAGTruth evaluation results - LLaMA-3 8B.

| | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| SAPLMA | **0.7161** | 0.5405 | 0.5000 | 0.6768 | 0.5277 | 0.2778 | 0.2632 | 0.5157 | **0.8103** | 0.6000 | 0.3750 | 0.7187 | **0.7303** | 0.4231 | 0.3837 | 0.6346 |
| Logistic Regression[a] | 0.6863 | **0.7544** | **0.6920** | **0.7011** | **0.5546** | 0.5571 | **0.5571** | **0.5571** | 0.6428 | **0.7829** | **0.6448** | **0.7136** | 0.6483 | 0.6789 | **0.6401** | **0.6513** |
| Random Forest [b] | 0.5000 | 0.5000 | 0.5000 | 0.4100 | 0.5111 | **0.8642** | 0.5125 | 0.5441 | 0.5714 | 0.7428 | 0.5717 | 0.6108 | 0.5097 | **0.8861** | 0.5092 | 0.4543 |

[a] Best result - based on layer 18's activations.
[b] Best result - based on layer 22's activations.

Table 15: RAGTruth evaluation results - Phi3.5 Mini.

| | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| SAPLMA | **0.8905** | 0.6897 | 0.6452 | **0.7917** | 0.5823 | **0.8933** | **1.0000** | 0.4718 | **0.6957** | 0.6356 | **0.8721** | 0.5864 | **0.8746** | 0.7710 | **0.9124** | **0.7904** |
| Logistic Regression[a] | 0.6451 | 0.6400 | 0.6300 | 0.6400 | 0.6120 | 0.6120 | 0.6940 | **0.7010** | 0.6291 | **0.6292** | 0.6291 | **0.6292** | 0.7818 | 0.8060 | 0.7817 | 0.7862 |
| Random Forest[b] | 0.6224 | **0.6200** | **0.6200** | 0.6100 | **0.5439** | 0.5439 | 0.8917 | 0.5207 | 0.6371 | 0.6371 | 0.6498 | **0.6245** | 0.7429 | **0.8064** | 0.7429 | 0.7451 |

[a] Best result - based on layer 17's activations.
[b] Best result - based on layer 24's activations.

Table 16: RAGTruth evaluation results - Mistral 7B.

| | | Question Answering | | | | Data-to-text writing | | | | Summarization | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Eval task | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| Logistic Regression | QA | 0.6477 | 0.6500 | 0.6500 | 0.6500 | 0.5117 | 0.5300 | 0.5300 | 0.4500 | 0.5104 | 0.5400 | 0.5100 | 0.2300 |
| | D2T | 0.5479 | 0.7000 | 0.5500 | 0.5600 | 0.5000 | 0.4300 | 0.5000 | 0.4700 | 0.5878 | 0.5600 | 0.5900 | 0.5000 |
| | Summ. | 0.5104 | 0.5400 | 0.5100 | 0.2300 | 0.5113 | 0.5600 | 0.5100 | 0.1300 | 0.4666 | 0.3200 | 0.4700 | 0.3600 |
| Random Forest | QA | 0.7386 | 0.7400 | 0.7400 | 0.7400 | 0.6338 | 0.5600 | 0.6300 | 0.5500 | 0.6146 | 0.5800 | 0.6100 | 0.5800 |
| | D2T | 0.4903 | 0.4900 | 0.4900 | 0.4500 | 0.5000 | 0.4300 | 0.5000 | 0.4700 | 0.5392 | 0.5300 | 0.5400 | 0.5100 |
| | Summ. | 0.5331 | 0.5600 | 0.5300 | 0.4700 | 0.4861 | 0.4800 | 0.4900 | 0.1800 | 0.5000 | 0.3300 | 0.5000 | 0.3900 |

Results presented for layer / hyperparameter combinations performing best on the Overall task (Table 14).

Table 17: RAGTruth cross-task evaluation results - LLaMA-3 8B.

| Method | AUC | Precision | Recall | F1 |
|---|---|---|---|---|
| ReDeEP | 0.5163 | 0.4946 | **0.9716** | 0.6555 |
| Logistic Regression | **0.6862** | 0.6872 | 0.6859 | **0.6862** |
| Random Forest | 0.6581 | 0.6656 | 0.6581 | 0.6541 |
| SAE Classifier | 0.6684 | **0.7609** | 0.4895 | 0.6578 |

Table 18: SQuAD evaluation results – LLaMA-2 7B.

| Method | AUC | Precision | Recall | F1 |
|---|---|---|---|---|
| ReDeEP | 0.5851 | 0.5330 | 0.8821 | 0.6645 |
| Logistic Regression | 0.7000 | 0.7045 | 0.7000 | 0.6984 |
| Random Forest | 0.6687 | 0.6687 | 0.6687 | 0.6687 |

Table 19: SQuAD evaluation results – LLaMA-3 8B.

| Method | AUC | Precision | Recall | F1 |
|---|---|---|---|---|
| ReDeEP | 0.5741 | 0.5714 | 0.8889 | 0.6957 |
| Logistic Regression | 0.8055 | 0.7946 | 0.8055 | 0.7963 |
| Random Forest | 0.8055 | 0.7946 | 0.8055 | 0.7963 |

The results presented are based on a train/test split of the Dolly dataset, which is different that in (Sun et al., 2025) where the whole dataset is used as the test set.

Table 20: Dolly evaluation results – LLaMA-2 7B.

| Method | AUC | Precision | Recall | F1 |
|---|---|---|---|---|
| ReDeEP | 0.6852 | 0.5000 | 1.0000 | 0.6667 |
| Logistic Regression | 0.6900 | 0.7386 | 0.6944 | 0.7000 |
| Random Forest | 0.7500 | 0.8750 | 0.7500 | 0.7619 |

Table 21: Dolly evaluation results – LLaMA-3 8B.

| | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| ReDeEP | - | - | - | - | - | - | - | - | - | - | - | - | 0.5851 | 0.5385 | 0.7642 | 0.6318 |
| Logistic Regression | 0.5116 | 0.5000 | 0.3200 | 0.3900 | 0.5026 | 0.5000 | 0.4300 | 0.4600 | 0.4959 | 0.5000 | 0.0 | 0.0 | 0.5038 | 0.5000 | 0.2600 | 0.3400 |
| Random Forest | 0.4969 | 0.5000 | 0.1500 | 0.2300 | 0.5246 | 0.5000 | 0.2500 | 0.3300 | 0.5009 | 0.5000 | 0.1900 | 0.2700 | 0.4978 | 0.5000 | 0.2500 | 0.3300 |

Table 22: Evaluation results | train: RAGTruth | eval: SQuAD - LLaMA-3 8B.

| | Question Answering | | | | Data-to-text writing | | | | Summarization | | | | Overall | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 | AUC | Precision | Recall | F1 |
| Logistic Regression | 0.5011 | 0.5000 | 0.4600 | 0.4900 | 0.5126 | 0.5000 | 0.4200 | 0.4600 | 0.4994 | 0.5000 | 0.5100 | 0.3900 | 0.5097 | 0.5100 | 0.5097 | 0.5098 |
| Random Forest | 0.4820 | 0.5000 | 0.2200 | 0.4100 | 0.5294 | 0.5000 | 0.4200 | 0.4400 | 0.5022 | 0.5000 | 0.2600 | 0.3500 | 0.5120 | 0.5091 | 0.5120 | 0.5063 |

Table 23: Evaluation results | train: RAGTruth | eval: SQuAD - Phi3.5 Mini.

| Method | Eval task | AUC | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Logistic Regression | QA | 0.5061 | 0.5100 | 0.5100 | 0.5000 |
| | D2T | 0.5410 | 0.5600 | 0.5100 | 0.1600 |
| | Summ. | 0.4916 | 0.4900 | 0.4900 | 0.4800 |
| | Overall | 0.4475 | 0.4300 | 0.4500 | 0.3700 |
| Random Forest | QA | 0.5138 | 0.5200 | 0.5100 | 0.4900 |
| | D2T | 0.5027 | 0.5300 | 0.5000 | 0.1400 |
| | Summ. | 0.4990 | 0.4800 | 0.5000 | 0.3600 |
| | Overall | 0.4838 | 0.5200 | 0.5300 | 0.4800 |
| SAE Classifier | QA | 0.5000 | 0.3467 | 1.0000 | 0.2574 |
| | D2T | 0.5000 | 0.8200 | 1.0000 | 0.4505 |
| | Summ. | 0.5000 | 0.3400 | 1.0000 | 0.2537 |
| | Overall | 0.5000 | 0.3467 | 1.0000 | 0.2574 |

Results presented for layer / hyperparameter combinations performing best on SQuAD. For SAEs, only activations for the last token of the response were considered.

Table 24: Evaluation results | train: SQuAD | eval: RAGTruth - LLaMA-2 7B.

| Method | Eval task | AUC | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Logistic Regression | QA | 0.5053 | 0.5100 | 0.5100 | 0.4400 |
| | D2T | 0.5000 | 0.4400 | 0.5000 | 0.4700 |
| | Summ. | 0.5589 | 0.6200 | 0.5600 | 0.3000 |
| | Overall | 0.5373 | 0.6300 | 0.5400 | 0.4500 |
| Random Forest | QA | 0.4887 | 0.4400 | 0.4900 | 0.3700 |
| | D2T | 0.4924 | 0.4400 | 0.4900 | 0.4600 |
| | Summ. | 0.5128 | 0.6100 | 0.5100 | 0.2100 |
| | Overall | 0.5254 | 0.5700 | 0.5300 | 0.4500 |

Table 25: Evaluation results | train: SQuAD | eval: RAGTruth - LLaMA-3 8B.