

A Robust Evaluation of Probe Robustness: Lessons for Reliable OOD Uncertainty Quantification

Joe Stacey¹, Hadas Orgad², Kentaro Inui^{3,4,5},
Benjamin Heinzerling^{5,4}, Nafise Sadat Moosavi¹

¹University of Sheffield, ²Kempner Institute at Harvard University,

³MBZUAI, ⁴Tohoku University, ⁵RIKEN,

{j.stacey, n.s.moosavi}@sheffield.ac.uk, hadasorgad@fas.harvard.edu,
kentaro.inui@mbzuai.ac.ae, benjamin.heinzerling@riken.jp

Abstract

Recent work has shown that the hidden states of large language models contain signals useful for uncertainty estimation, motivating a growing interest in efficient probe-based approaches. Yet it remains unclear how robust existing methods are, with prior work reporting conflicting conclusions under substantially different evaluation settings. We address this by introducing ProbeDrift, a systematic evaluation framework for supervised uncertainty probes covering a wide range of OOD settings across models, tasks, and distributional shifts. Using ProbeDrift, we train over 2,000 probes to disentangle the effect of key design choices, showing poor robustness of current methods beyond near-OOD settings. We find that robustness is driven by design decisions that have a largely invisible effect in-distribution, including the choice of feature type, aggregation strategy, and training signal. We argue that robust uncertainty estimation requires robust evaluation. To support this, we release ProbeDrift as a lightweight Python library that contains the train and test splits underpinning our extensive evaluation.¹ We also show how insights from our evaluation can directly lead to more robust methods through a simple Hybrid Back-Off (HBO) strategy.

1 Introduction

Supervised Uncertainty Quantification (UQ) methods train probes to predict the reliability of model outputs from internal features such as hidden states or attention (Azaria and Mitchell, 2023; Orgad et al., 2025; He et al., 2024; Obeso et al., 2025). These UQ methods achieve strong in-distribution performance, providing evidence that model internals contain signals related to correctness and uncertainty. However, there is mixed evidence about the robustness of these probes, with some work

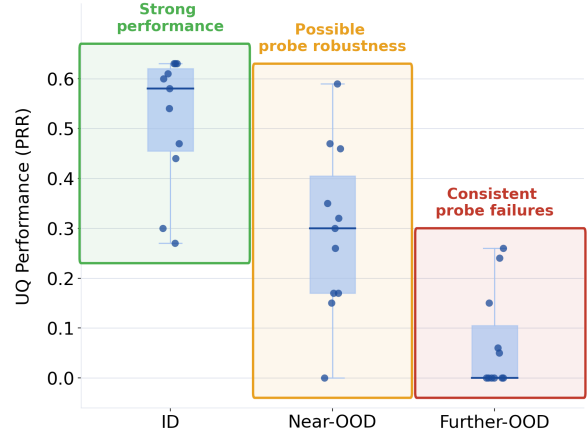


Figure 1: Uncertainty Quantification methods quickly lose robustness outside of their training distribution, and these failures are often not visible in existing OOD evaluation. We show that robustness is possible in a near-OOD setting with the right probe configuration. Results are from Table 1 for Llama-3.1-8B, with negative PRR scores floored at 0 for visualisation.

reporting strong out-of-distribution (OOD) performance (Shelmanov et al., 2025; Obeso et al., 2025; Vazhentsev et al., 2025b) while other work finds poor robustness (Levinstein and Herrmann, 2024; Orgad et al., 2025). Since probing methods differ in their features, architecture, and aggregation strategies, existing OOD findings often do not generalise beyond the specific methods being tested. Different experimental setups, baselines, and definitions of distributional shifts make comparisons between prior works even more challenging. It is therefore unclear whether strong robustness claims reflect genuine improvements or differences in evaluation.

In this work we introduce ProbeDrift, a systematic evaluation framework for measuring the robustness of supervised UQ probes across progressively stronger distributional shifts. ProbeDrift enables consistent comparisons across probe designs, tasks, models, and OOD settings, addressing the fragmented evaluation used in prior work. Using this

¹<https://github.com/joestacey/ProbeDrift>

framework, we train and evaluate over 2,000 probes while varying key design dimensions including representation layer, feature type, training signal, and token aggregation strategy.

Our evaluation reveals that probes can generalise under limited distributional shifts, but this generalisation quickly degrades for larger shifts. We find that probes trained on middle-layer hidden states generalise better than those from the final layer, and aggregating information across response tokens is better than relying on single-token features. We also find that the choice of correctness function used during training can substantially affect robustness, particularly for long-form generations. These design choices often have a minimal impact in-distribution, and remain hidden without systematic OOD evaluation.

Ultimately, we argue that robust uncertainty quantification requires robust evaluation. Insufficient evaluation can mask generalisation failures and obscure probe design choices which become critical under larger distributional shifts. In support of this, we provide ProbeDrift as a lightweight Python library to make our extensive evaluation easy to reproduce. We further show how insights from comprehensive evaluation can directly guide the design of more robust methods through a simple Hybrid Back-Off (HBO) strategy which substantially improves robustness.

To summarise our contributions:

1. We introduce ProbeDrift, a systematic evaluation framework for assessing the robustness of supervised UQ probes across progressively stronger distributional shifts. Using ProbeDrift, we comprehensively evaluate existing supervised UQ methods, releasing this as a lightweight Python library to support future work.
2. We disentangle key design factors, including the choice of architecture, feature, correctness function and token aggregation strategy, finding choices with little effect in-distribution can substantially impact robustness.
3. We show that robustness conclusions for supervised UQ probes depend strongly on the evaluation setting, highlighting the need for systematic robustness evaluation across progressively stronger distributional shifts.
4. We introduce Hybrid Back-Off (HBO) as an illustrative example to demonstrate how insights from ProbeDrift can directly guide the

development of more robust methods.²

2 Background

There is mixed evidence about the robustness of uncertainty quantification probes. Several recent methods report strong generalisation across tasks and datasets (Shelmanov et al., 2025; Obeso et al., 2025; Vazhentsev et al., 2025b), while other work observes OOD failures (CH-Wang et al., 2024; Levinstein and Herrmann, 2024; Orgad et al., 2025). For instance, Levinstein and Herrmann (2024) find that probes do not generalise to texts containing negations, while Orgad et al. (2025) find that probes rarely outperform unsupervised baselines when testing performance OOD. CH-Wang et al. (2024) report similar findings, showing that an ensemble of probes with linear classification heads do not always generalise well. However, these findings are specific to the probe configurations studied, for example the transformer probe architecture used in Shelmanov et al. (2025), the density-based features used in Vazhentsev et al. (2025b), or the single-token aggregation used in Levinstein and Herrmann (2024). It therefore remains difficult to draw general conclusions about the robustness of supervised UQ probes.

UQ probing methods rely on a variety of features encompassing hidden states (Orgad et al., 2025; Huang et al., 2025; He et al., 2024; Kossen et al., 2024; Azaria and Mitchell, 2023) and attention-based features (Shelmanov et al., 2025; Ni et al., 2026; Chuang et al., 2024). Probe architectures vary from linear probes (Chuang et al., 2024) and multi-layer perceptrons (Azaria and Mitchell, 2023) to transformers (Shelmanov et al., 2025), while token aggregation strategies include last-token aggregation (Azaria and Mitchell, 2023; Kossen et al., 2024) and mean pooling over response tokens (Shelmanov et al., 2025; Vazhentsev et al., 2025b). ProbeDrift enables these design dimensions to be evaluated systematically under consistent OOD settings.

OOD evaluation varies widely between different works. This can involve training on a range of datasets with another dataset kept as a hold-out for evaluation (Vazhentsev et al., 2025b; Azaria and Mitchell, 2023), testing on synthetic data generated for different domains (Shelmanov et al., 2025), or training with a single dataset and evaluating on the same or different tasks (Orgad et al., 2025). As

²<https://github.com/joestacey/RobustUQProbes>

each unique OOD setting measures different levels of distributional shift, these inconsistencies make it challenging to compare conflicting claims about robustness. ProbeDrift addresses this by providing a systematic robustness evaluation framework spanning progressively stronger distributional shifts, enabling more reliable and reproducible comparisons across methods and evaluation settings.

Complementary concurrent work analyses layer-wise input-output behaviour in gated MLP neurons, including effects on model internal confidence (Gerstner and Schuetze, 2026). In contrast, we focus on the OOD robustness of supervised UQ probes.

3 ProbeDrift: Evaluating Probe Robustness Under Distribution Shift

In this section we introduce the ProbeDrift evaluation framework for measuring the robustness of UQ probes across progressively stronger distributional shifts. Using ProbeDrift, we evaluate existing methods across a range of datasets and OOD settings, identifying where robustness holds and where it breaks down.

3.1 Experimental Setup

ProbeDrift evaluates UQ methods across five settings of increasing difficulty. For each evaluation dataset, we test probes after training on: 1) the corresponding *in-distribution* training data, 2) *leave-one-out* evaluation across a range of datasets from both the same task and a different task, 3) a *single same-task* dataset, 4) a *range of different-task* datasets, and 5) a *single different-task* dataset. Training data size is held constant across experiments, while the training data becomes progressively more OOD relative to the test data.

Datasets for Training and Evaluation. For evaluation, we consider four different question answering (QA) datasets, and two summarisation datasets. The QA datasets include both long-form and short-form answers. For QA we use SciQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), and COQA (Reddy et al., 2019), which contain shorter-form answers, alongside PubmedQA (Jin et al., 2019), which contains substantially longer generations. For summarisation we use XSum (Narayan et al., 2018) and CNN/DailyMail (Hermann et al., 2015; Nallapati et al., 2016), which also involve longer-form generations. We additionally use SamSum (Gliwa et al., 2019), TruthfulQA (Lin et al., 2022),

MMLU (Hendrycks et al., 2021) and MedQUAD (Abacha and Demner-Fushman, 2019) as additional training datasets for our leave-one-out experimentation. The leave-one-out evaluation in ProbeDrift mirrors the out-of-distribution evaluation performed by Vazhentsev et al. (2025b), which we extend to cover a wide variety of OOD settings. See Appendix Q for further details.

Baselines. We evaluate a representative set of probing-based UQ methods spanning hidden-state probes (SAPLMA (Azaria and Mitchell, 2023)), density-based probes (SATMD, SATRMD (Vazhentsev et al., 2025b)), attention-based probes (Uhead (Shelmanov et al., 2025), Lookback-Lens (Chuang et al., 2024)), and hybrid methods combining supervised probes with the unsupervised Maximum Sequence Probability (MSP) baseline (Vazhentsev et al., 2025b; Fomicheva et al., 2020). We also include MSP as a simple but competitive unsupervised baseline. We report SAPLMA probes using hidden states from the middle and top layers and Uhead configurations with one or two layers. Details are provided in Appendix A.

Metrics. Evaluating uncertainty quantification for generative tasks requires a correctness function that measures the quality of model outputs relative to reference answers. We primarily use AlignScore (Zha et al., 2023), estimating factual consistency using the entailment predictions of a RoBERTa model (Liu et al., 2019) fine-tuned on NLI and claim verification. AlignScore provides a continuous measure of output correctness, which we use as the target signal when evaluating uncertainty estimates. This score correlates closely with human judgements of correctness for short-form generations (Santilli et al., 2025), but without the cost of generating LLM-as-a-Judge responses. For evaluation, we further experiment with LLM-as-a-Judge, ROUGE-L, and exact match as alternative correctness functions. In Section 4, we investigate how the choice of correctness function used during training affects robustness, including training with LLM-as-a-Judge supervision. We create feature vectors by averaging feature representations from each output token, with further aggregation strategies tested in Section 4.2.

Performance is evaluated using the Prediction-Rejection Ratio (PRR) (Vashurin et al., 2025; Malinin et al., 2017), which measures how effectively an uncertainty estimate identifies unreliable outputs. PRR computes the area under the curve

Method	Llama-3.1-8B					Gemma-2-9B				
	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask
		near OOD	→ more OOD	→	most OOD		near OOD	→ more OOD	→	most OOD
Short-form generation datasets										
SAPLMA (middle)	0.63	0.46	0.26	0.35	0.32	0.58	0.35	0.11	0.30	0.21
SAPLMA (top)	0.54	0.26	-0.12	-0.14	0.18	0.56	0.34	-0.03	0.06	0.08
Uhead (1 layer)	0.63	0.17	0.24	0.16	0.08	0.66	0.34	0.18	-0.11	0.16
Uhead (2 layers)	0.63	0.15	0.15	-0.27	-0.36	0.65	0.33	0.20	-0.31	0.12
SATMD	0.30	0.35	-0.15	0.21	0.15	0.41	0.27	-0.06	0.04	0.22
SATRMD	0.47	0.17	-0.11	0.21	0.13	0.53	0.33	0.21	0.27	0.34
Lookback-lens	0.27	0.00	-0.03	-0.13	0.19	0.34	-0.05	0.09	-0.04	0.16
SATMD-MSP	0.44	0.32	0.06	0.24	0.23	0.62	0.17	0.15	-0.18	0.41
SATRMD-MSP	0.60	0.30	0.05	0.20	0.21	0.64	0.40	0.41	0.24	0.52
HUQ-SATMD	0.58	0.59	-0.14	0.22	0.25	0.64	0.59	0.00	0.10	0.14
HUQ-SATRMD	0.61	0.47	-0.08	0.24	0.18	0.65	0.53	0.17	0.16	0.22
MSP	0.57	0.57	0.57	0.57	0.57	0.64	0.64	0.64	0.64	0.64
Long-form generation datasets										
SAPLMA (middle)	0.32	0.03	0.07	0.12	0.04	0.32	0.01	0.10	-0.05	0.04
SAPLMA (top)	0.20	0.00	0.04	0.05	-0.00	0.22	-0.05	0.08	-0.03	0.03
Uhead (1 layer)	0.31	-0.04	0.05	-0.08	-0.01	0.28	-0.06	-0.07	0.05	0.02
Uhead (2 layers)	0.27	-0.05	0.02	-0.03	0.02	0.23	-0.09	-0.13	-0.13	0.02
SATMD	0.14	-0.02	-0.07	0.04	0.11	0.07	0.02	0.04	-0.07	0.08
SATRMD	0.16	0.03	-0.08	0.05	0.07	0.20	-0.15	0.01	-0.13	0.06
Lookback-lens	0.14	0.01	0.10	-0.04	-0.01	0.22	-0.04	0.10	-0.08	-0.04
SATMD-MSP	0.18	0.03	-0.03	0.08	0.12	0.18	-0.08	0.02	-0.08	0.09
SATRMD-MSP	0.17	0.05	-0.01	0.07	0.10	0.21	-0.15	0.02	-0.08	0.03
HUQ-SATMD	0.02	-0.02	-0.05	0.04	0.09	-0.01	-0.10	0.06	-0.04	0.08
HUQ-SATRMD	0.16	0.01	-0.08	0.03	0.04	0.20	-0.10	0.05	-0.04	-0.10
MSP	-0.01	-0.01	-0.01	-0.01	-0.01	-0.10	-0.10	-0.10	-0.10	-0.10

Table 1: PRR results (higher is better) across in-domain (ID) and out-of-domain (OOD) evaluation settings. Abbreviations: LOO = leave-one-out; 1D-SameTask = one dataset, same task; DiffTask = multiple datasets, different task; 1D-DiffTask = one dataset, different task. We show supervised methods, followed by hybrid methods and the MSP. Each result is an average across three different test sets. Blue highlights indicate that a corresponding evaluation was reported in the method’s paper. An analysis of multi-seed variance is also provided in Appendix D.

(AUC) for Prediction-Rejection curves, where uncertain examples are progressively rejected to measure the average correctness of the remaining examples. This score is normalised by an oracle (the best possible AUC given the generations), adjusting for a baseline that randomly predicts uncertainty:

$$PRR = \frac{AUC_{unc} - AUC_{rnd}}{AUC_{oracle} - AUC_{rnd}} \quad (1)$$

A PRR score of 1 means that the uncertainty estimate is attaining the performance of the oracle, while a score of 0 shows chance-level performance. Scores below 0 are worse than chance-level. PRR allows for the use of a continuous correctness function such as AlignScore.

3.2 Results

UQ probes achieve strong in-distribution performance but largely fail to generalise out-of-distribution (see Table 1). In near-OOD settings for datasets with shorter generations, the probe design matters considerably, and the best methods reach a PRR of 0.46–0.59 for Llama (SAPLMA middle and HUQ-SATMD respectively). However, as OOD-

ness increases, no method exceeds a PRR of 0.35 for Llama. We find largely consistent trends across each dataset tested (averaged for Table 1), although there are some exceptions such as poor robustness on SciQ for Uhead and on TriviaQA for SATRMD in a leave-one-out setting for Llama. Hybrid methods, which combine supervised probes with unsupervised estimates, show strong near-OOD robustness but this advantage also deteriorates under larger shifts. Notably, no probing method exceeds the unsupervised MSP baseline for any OOD setting with the exception of HUQ-SATMD in the leave-one-out setting for Llama.

When evaluating with different correctness functions, probe performance consistently degrades beyond near-OOD settings. For LLM-as-a-Judge evaluation, despite very similar findings for SAPLMA when using the last layer hidden state (see Appendix E), we see less OOD degradation in the case of SAPLMA when using the middle layer hidden state (see Table 2). For long-form tasks, when training with AlignScore, all methods achieve close to chance-level performance across all OOD conditions. In Section 4 we show that

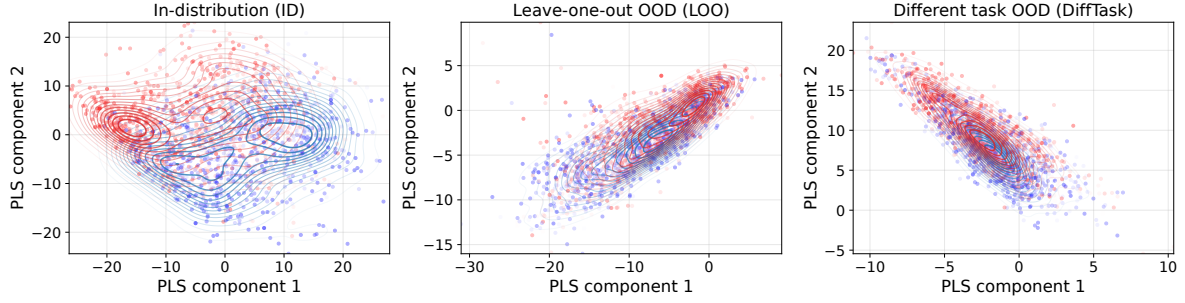


Figure 2: Visualisation of Llama 3.1-8B hidden states (layer 16) evaluated on TriviaQA, either for in-distribution, leave-one-out or different task OOD settings. Hidden states are down-projected to 2d subspaces using PLS models fitted to predict AlignScore from the training examples. As OOD-ness increases, hidden states of more correct generations (blue) and less correct generations (red) become less separable. Kernel density contour lines show the density of generations above and below the median AlignScore.

Metric	ID		LOO		ID-SameTask		DiffTask		ID-DiffTask	
	PRR	%	PRR	%	PRR	%	PRR	%	PRR	%
Short-form generation datasets										
AlignScore	0.63	100%	0.46	73%	0.26	41%	0.35	56%	0.32	51%
ROUGE-L	0.70	100%	0.49	70%	0.30	43%	0.33	47%	0.35	50%
Exact match	0.70	100%	0.48	69%	0.30	43%	0.33	47%	0.31	44%
LLM-as-a-judge	0.68	100%	0.55	81%	0.31	46%	0.48	71%	0.45	66%
Long-form generation datasets										
AlignScore	0.32	100%	0.03	9%	0.07	22%	0.12	38%	0.04	12%
ROUGE-L	0.30	100%	0.00	0%	0.04	13%	0.11	37%	0.01	3%
LLM-as-a-judge	0.33	100%	0.14	42%	0.14	42%	0.14	42%	0.08	24%

Table 2: PRR results (higher is better) for SAPLMA (middle) when evaluated across four correctness functions, for Llama-3.1-8B. Each value is shown alongside its percentage of the ID score for that metric. Exact match is excluded for long-form generations as the metric is uninformative in this setting.

the generalisability of probes for long-form tasks is highly dependent on the correctness function used during training, where the use of LLM-as-a-Judge for both training and evaluation can lead to substantially better reported robustness compared to using AlignScore.

We find that the strongest OOD performance is achieved by a simple probing approach (SAPLMA), and that implementation choices, such as layer selection, prove more consequential than adopting more recently proposed methods. Methods that appear robust under near-OOD evaluation also do not always maintain this advantage under broader distribution shifts (see SATMD). Finally, in Table 1 we show the evaluation settings closest to those used in the source papers, finding that prior evaluation tends to favour near-OOD conditions where robustness is more achievable.

3.3 Understanding Robustness Failures

To illustrate why probe performance deteriorates under distribution shift, particularly for long-form tasks, we visualise the hidden states using two-

component Partial Least Squares (PLS) regression models that are fitted to predict the rank-normalised AlignScore. The PLS regression is used as a dimensionality reduction method, representing a strongly-regularised linear regression probe that demonstrates how uncertainty information in the hidden states becomes less accessible to probes trained on OOD data.

Figure 2 shows that correct and incorrect generations are partially separable in the in-distribution setting, indicating that correctness signals are accessible in the representation space. This structure weakens under leave-one-out training and largely collapses under different-task training. Since linear probes can still achieve good performance when trained and evaluated on each distribution separately (Appendix K), the uncertainty information itself does not disappear under distribution shift. Instead, the relationship between hidden-state representations and correctness appears to change across distributions, reducing the ability of probes trained on one distribution to generalise to another. This effect is especially pronounced for datasets with

longer generations (Appendix K), consistent with the weaker robustness observed in Table 1.

While this section demonstrates the poor robustness for long-form generations using AlignScore, we show in Section 4 that robustness is heavily dependent on the correctness metric.

3.4 Limitations of Existing Robustness Evaluation

Prior work evaluates probe robustness under substantially different OOD settings, making robustness claims difficult to compare directly. Some works focus on relatively limited “near-OOD” settings, while others evaluate under larger distributional shifts, either within the same task or across tasks. Evaluation protocols also vary in their use of synthetic data, supervised and unsupervised baselines, and the range of distributional shifts considered.

Across the surveyed literature (with full details of the surveyed evaluation setups provided in Appendix L), no two works evaluate on the same combination of OOD datasets. Many papers evaluate robustness under only a single type of distributional shift (54%) and omit strong supervised or unsupervised baselines (61% and 46% respectively). We also find that positive robustness claims are often reported under synthetic evaluation settings, suggesting that robustness evaluation should extend beyond this setting alone.

These inconsistencies motivate the need for systematic robustness evaluation across progressively stronger distributional shifts.

4 What makes a probe robust?

The results in Section 3 show that probe performance varies substantially under distribution shift. To better understand what drives these robustness differences, we systematically analyse key probe design factors, including the probe architecture, feature representations, correctness functions, and token aggregation strategies.

Guided by the findings from ProbeDrift, we perform a more detailed analysis on a representative subset of datasets, using SciQ and TriviaQA as datasets with shorter generations, and PubmedQA as a dataset with longer generations. We refer to this reduced evaluation setup as ProbeDrift-Light, focusing on three representative settings: in-distribution (ID), leave-one-out OOD evaluation (LOO), and OOD evaluation after training only on

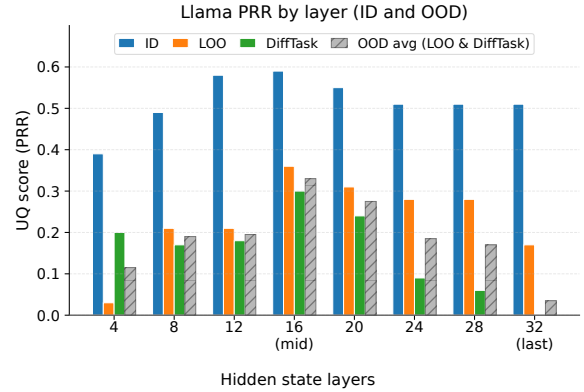


Figure 3: PRR by layer for Llama showing ID performance and two OOD settings, leave-one-out evaluation (LOO), and evaluation when training on a different task (DiffTask). OOD avg provides an average of LOO and DiffTask. Negative results floored at 0 for visualisation.

Features used	ID	LOO	DiffTask	OOD avg
Hidden state (mid. layer)	0.58	0.29	0.25	0.27
Hidden state (last layer)	0.52	0.17	-0.05	0.06
Attention (last token)	0.54	0.28	0.13	0.21
Attention (last two tokens)	0.55	0.25	0.15	0.20
Lookback lens	0.19	0.11	-0.04	0.04
Token probabilities	0.43	0.27	-0.30	-0.01

Table 3: Feature ablations, average from Llama and Gemma (PRR). OOD avg is the mean of LOO (leave-one-out) and DiffTask (different task).

datasets from a different task (DiffTask). We show in Appendix C that the ProbeDriftLight evaluation is representative of the full ProbeDrift.

We use the SAPLMA architecture in this section as it provides a simple and effective probe based on model hidden states. Results in this section are an average across both Llama-3.1-8B (Grattafiori et al., 2024) and Gemma-2-9B (Team et al., 2024) for each dataset.³

4.1 Choice of features

To isolate the effect of different uncertainty signals, we fix the probe architecture and vary only the features used to train the probe. We investigate different features including attention to the previous token, attention to the last two tokens (Shelmanov et al., 2025) and the ratio of attention to the context (Chuang et al., 2024). We also consider the model hidden states (Azaria and Mitchell, 2023) and token probability distributions (Shelmanov et al., 2025).

Across most features, in-distribution and leave-one-out performance is broadly similar (Table 3). Differences emerge under larger distributional

³We use the same models as Vazhentsev et al. (2025b), whose methods are our primary baseline in Section 5.

shifts. Hidden states from the middle layer outperform attention-based features in the DiffTask setting, while the lookback ratio and token probabilities are less effective as stand-alone signals.

Layer choice also has a strong effect on robustness (see Figure 3). While most layers perform comparably in-distribution, OOD performance degrades substantially at early and final layers, while middle-layer representations remain consistently more robust. This gap widens as distributional shift increases. Notably, using final-layer representations remains a common baseline in prior work (Vazhentsev et al., 2025b; Shelmanov et al., 2025; Vazhentsev et al., 2025a), despite its weaker robustness. In Appendix I we further verify these findings across a range of different LLMs and evaluation metrics.

4.2 Choice of Token Aggregation

Token representations can be aggregated differently when creating the probe features. While recent work explores new variations (McKenzie et al., 2025; Huang et al., 2025; Kantamneni et al., 2025), the most common approach is to use the representation from the last context token (Bürger et al., 2024; Azaria and Mitchell, 2023; Servedio et al., 2025; Kossen et al., 2024), the final response token (Kossen et al., 2024), or an average representation from the response tokens (Vazhentsev et al., 2025b; Shelmanov et al., 2025; Chuang et al., 2024). We compare the in-distribution performance and robustness of each approach, including a baseline where a response token is randomly sampled.

The best robustness is achieved when taking an average representation from the response tokens (see Table 4). While similar in-distribution performance is achieved when using the last response token, we find this is less robust when tested further OOD. Interestingly, we find that randomly sampling a token from the response achieves similar robustness to using the final response token. In Appendix J we show that these findings are consistent across a range of LLMs and evaluation metrics.

4.3 Training Signals from Different Correctness Functions

We experiment with changing the correctness function used during training from AlignScore to LLM-as-a-Judge, investigating the impact that this has on robustness. For datasets with shorter generations, training with LLM-as-a-Judge leads to modest robustness improvements when evaluating with

Method	ID	LOO	DiffTask	OOD avg
Response tokens only				
Avg. of response tokens	0.58	0.29	0.25	0.27
Last response token	0.57	0.30	0.09	0.20
Random response token	0.42	0.24	0.20	0.22
With context tokens				
Avg. of all tokens	0.33	0.09	0.06	0.07
Avg. of context tokens	0.23	0.10	0.04	0.07
Last context token	0.46	0.25	0.07	0.16

Table 4: Different strategies to aggregate representations across context and response tokens. Best results in bold.

Training Signal	ID	LOO	DiffTask
Short-form			
AlignScore	0.69	0.52	0.35
LLM-as-a-Judge	0.66	0.56	0.36
Long-form			
AlignScore	0.39	0.04	0.19
LLM-as-a-Judge	-0.08	-0.11	0.13

Table 5: PRR results (higher is better) for SAPLMA (middle) evaluated with AlignScore, comparing two training signals for Llama 3.1-8B, either training with AlignScore or LLM-as-a-Judge.

AlignScore (Table 5). Larger improvements are observed when both training and evaluation use LLM-as-a-Judge (see Appendix F).

For datasets with longer generations, this dependence becomes substantially stronger. Probes trained with LLM-as-a-Judge achieve near-chance performance when evaluated using AlignScore. When evaluation also uses LLM-as-a-Judge, we observe more robust probe behaviour for long-form tasks (see Appendix F). This suggests that robustness for longer generations is particularly sensitive to the correctness signal used during training and evaluation.

4.4 Probe architectures

We next examine how probe architectures impact robustness. Using middle-layer hidden states as features, we compare the different probe architectures introduced in Section 3. We also experiment with simple modifications to the SAPLMA architecture, either adding or removing layers.

Across architectures we observe differences in in-distribution performance, but with little variation in robustness under distribution shift. The exception is linear probe architectures, which we find can be substantially less robust, although this depends on their training procedure. Overall, we find that architectural complexity and the choice of training hyper-parameters do not lead to substantially better probe robustness. Additional details are provided

Method	Llama-3.1-8B					Gemma-2-9B				
	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask
Short-form generation datasets										
SATMD-MSP	0.44	0.32	0.06	0.24	0.23	0.62	0.17	0.15	-0.18	0.41
SATRMD-MSP	0.60	0.30	0.05	0.20	0.21	0.64	0.40	0.41	0.24	0.52
HUQ-SATMD	0.58	0.59	-0.14	0.22	0.25	0.64	0.59	0.00	0.10	0.14
HUQ-SATRMD	0.61	0.47	-0.08	0.24	0.18	0.65	0.53	0.17	0.16	0.22
MSP	0.57	0.57	0.57	0.57	0.57	0.64	0.64	0.64	0.64	0.64
SAPLMA (mid)	0.63	0.46	0.26	0.35	0.32	0.58	0.35	0.11	0.30	0.21
HBO	0.66	0.61	0.57	0.57	0.57	0.67	0.61	0.64	0.64	0.64
Long-form generation datasets										
SATMD-MSP	0.18	0.03	-0.03	0.08	0.12	0.18	-0.08	0.02	-0.08	0.09
SATRMD-MSP	0.17	0.05	-0.01	0.07	0.10	0.21	-0.15	0.02	-0.08	0.03
HUQ-SATMD	0.02	-0.02	-0.05	0.04	0.09	-0.01	-0.10	0.06	-0.04	0.08
HUQ-SATRMD	0.16	0.01	-0.08	0.03	0.04	0.20	-0.10	0.05	-0.04	-0.10
MSP	-0.01	-0.01	-0.01	-0.01	-0.01	-0.10	-0.10	-0.10	-0.10	-0.10
SAPLMA (mid)	0.32	0.03	0.07	0.12	0.04	0.32	0.01	0.10	-0.05	0.04
HBO	0.20	0.02	-0.01	0.03	-0.01	0.03	-0.05	-0.10	-0.08	-0.10

Table 6: PRR results for HBO compared to existing hybrid methods, MSP and SAPLMA using middle layer hidden states. Results are tested in-distribution, and across the four OOD settings in ProbeDrift.

in Appendix M.

5 Using ProbeDrift to Analyse and Improve Hybrid Robustness

Hybrid methods have previously been proposed to combine the strong in-distribution performance of supervised probes with the broader robustness of unsupervised uncertainty estimates (Vazhentsev et al., 2025b). ProbeDrift allows us to evaluate these approaches under progressively stronger distributional shifts, revealing the weakness of existing hybrid methods beyond a near-OOD setting.

Motivated by this observation, we investigate whether robustness can be improved by adapting the contribution of these signals based on how far examples lie from the training distribution. We focus on methods introduced by Vazhentsev et al. (2025b), which combine supervised SATRMD probes with the unsupervised MSP baseline. SATRMD-MSP incorporates MSP as an additional feature, while HUQ-SATRMD relies on MSP for in-distribution predictions, backing off to a combination of MSP and the supervised SATRMD method for OOD examples.

5.1 A Hybrid Back-Off Strategy

Building on our finding that supervised probe robustness degrades as distributional shift increases, we introduce a Hybrid Back-Off (HBO) strategy that progressively shifts from supervised to unsupervised uncertainty estimates. Following Vazhentsev et al. (2025b), we estimate how OOD each

test example is using Mahalanobis distance. HBO combines the robust SAPLMA probe with MSP, relying increasingly on MSP as examples move further from the training distribution. Further details are provided in Appendix B.

5.2 Results

We find that HBO outperforms each of the hybrid baselines in almost every OOD setting for short-form generations, with larger improvements under stronger distributional shifts (see Table 6). Despite its simplicity, HBO is substantially more robust than existing approaches, illustrating how insights from comprehensive robustness evaluation can directly lead to improved robustness. As HBO was introduced based on the strong performance of unsupervised methods for short-form generations, this method does not generalise well to long-form generations, which we leave for future work.

6 Evaluation Recommendations

Based on our extensive evaluation of supervised UQ probes, we identify several evaluation practices for improving the reliability of robustness claims in future work. These recommendations are informed by our experimental findings and by the gaps in existing evaluation practices highlighted in Section 3.4 and Appendix L.

We suggest future work should:

1. Use robust baselines, including an unsupervised method using token probabilities and an MLP trained on middle-layer hidden states.

2. Show robustness for both near and far distributional shifts (e.g. both leave-one-out evaluation and training on another task).
3. Evaluate with OOD test sets used directly by previous work, rather than creating bespoke evaluation for each paper.
4. Measure robustness on both short-form and long-form generations, highlighting the room for improvement on long-form tasks.
5. Include more negative robustness results, showing the points of failure to help spur improvements in subsequent research.

7 Discussion and Conclusion

Conflicting evidence about the robustness of UQ probes has made it difficult to assess whether recent methods genuinely improve OOD robustness or simply benefit from easier evaluation settings. In this work, we introduce ProbeDrift, a systematic evaluation framework for analysing robustness across progressively stronger distributional shifts. Our results show that robustness conclusions can change substantially depending on both the evaluation regime and seemingly minor implementation choices. Methods that appear robust under limited “near-OOD” evaluation often degrade under broader shifts, while several commonly used evaluation settings fail to expose these weaknesses. We further find that the strongest OOD performance is achieved by a simple probing approach (SAPLMA), with implementation choices such as layer selection proving more consequential than adopting more recently proposed methods.

More broadly, our findings suggest that progress on robust uncertainty estimation is currently constrained less by probe architecture and more by the lack of systematic robustness evaluation. We find that feature selection and token aggregation decisions are important for a probe’s robustness, and we also show the importance of the correctness function used in both training and evaluation, which is especially the case for long-form generations.

ProbeDrift provides a reusable framework for evaluating robustness across models, tasks, and distributional shifts, enabling more reliable comparisons between methods and helping to identify design choices that genuinely improve robustness. Beyond evaluation, we show that these insights can directly guide method development by introducing HBO, a method that adaptively shifts towards unsupervised uncertainty estimates under larger

distributional shifts. We hope ProbeDrift will support future work on both developing and evaluating more robust uncertainty estimation methods.

Limitations

In the paper we highlight how architectural advances are unlikely to lead to improvements in robustness. This is based on our experimentation testing robustness across a range of different architectures, all of which show similar levels of robustness (except for linear probes which can be less robust). While we cannot rule out the existence of a probe architecture that might improve robustness, we believe this is a less promising research direction than investigating how to create better probe inputs and training signals.

This paper involved extensive experimentation, training over 2,000 different probes (comprising more than 10k GPU hours), and therefore it will be difficult for others to fully replicate all of our experiments. We provide ProbeDriftLight to reduce the experimentation required for future OOD evaluation, which involves evaluation on three different datasets for two different OOD settings. Even with ProbeDriftLight, this additional OOD evaluation is still likely to increase the experimental burden for future work.

Due to the large computational requirement for this work, we do not perform experiments with larger LLMs. Instead, the model sizes we use are comparable with existing work (Vazhentsev et al., 2025b; Shelmanov et al., 2025; Orgad et al., 2025; Chuang et al., 2024).

Throughout this work we train with a fixed number of training examples (1,800), keeping this consistent so different experiments can be compared on a like-for-like basis. In Appendix P we show similar findings when increasing the size of the training data (by two or four times), but we do not verify each of our findings when increasing the training data size as this would require substantially more experiments.

The OOD splits in ProbeDrift have been created to generate varying degrees of out-of-distribution, testing robustness across different OOD test sets. However, these degrees of OOD conflate several factors that are difficult to separate. For example, test sets vary in diversity, response length depends on the task and dataset, and a change in task usually brings a change in domain and task difficulty. To mitigate the effect of any single dataset’s artifacts,

ProbeDrift averages results across multiple datasets for each OOD setting.

In this paper we focus on supervised probing approaches due to both their efficiency and strong performance. We compare the performance of these probing approaches to MSP, which we found to be the most competitive out of a range of unsupervised methods (see Appendix Table 23 and Table 24). We do not explore other uncertainty quantification methods, such as deep ensembles (Lakshminarayanan et al., 2017) and MC Dropout (Gal and Ghahramani, 2016).

Finally, HBO involves combining predictions from MSP and SAPLMA using their rank-normalised uncertainty scores, which assumes the whole test set is available at inference time rather than scoring examples one at a time. As PRR evaluation already requires ranking of the test set by uncertainty, the rankings required by HBO are always available.

Acknowledgements

We are grateful to Lihu Chen, Favour Aghaebe, Artem Shelmanov and Artem Vazhentsev for their valuable feedback.

This work was supported by an ISPF International Collaboration Award funded by the Royal Society, as well as by the Japan Science and Technology Agency under Grant No. JST BOOST JP-MJBY24F9. We also acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing. AI tools (Claude and ChatGPT) were used to provide feedback on the writing in this draft and to help refactor the project code.

References

- Asma Ben Abacha and Dina Demner-Fushman. 2019. [A question-entailment approach to question answering](#). *BMC Bioinform.*, 20(1):511:1–511:23.
- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. [The falcon series of open language models](#). *Preprint*, arXiv:2311.16867.
- 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.
- Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katya Mirylenka, Michael Glass, and Xiao Shou. 2025. [SIMBA UQ: Similarity-based aggregation for uncertainty quantification in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15880–15894, Suzhou, China. Association for Computational Linguistics.
- Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. 2024. [Truth is universal: Robust detection of lies in llms](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. [Do androids know they’re only dreaming of electric sheep?](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4401–4420, Bangkok, Thailand. Association for Computational Linguistics.
- Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. [Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. [Unsupervised quality estimation for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:539–555.
- Yarin Gal and Zoubin Ghahramani. 2016. [Dropout as a bayesian approximation: Representing model uncertainty in deep learning](#). In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.
- Sebastian Gerstner and Hinrich Schuetze. 2026. [Weakening neurons: A newly discovered read-write functionality in transformers with outsize influence](#).
- Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. [SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization](#). In *Proceedings of the 2nd Workshop on*

- New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- 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.
- Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024. [LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10218–10230, Bangkok, Thailand. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Karl Moritz Hermann, Tomas Kocisky, 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 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.
- Zhiqi Huang, Vivek Datla, Chenyang Zhu, Alf Samuel, Daben Liu, Anoop Kumar, and Ritesh Soni. 2025. [Confidence-based response abstinence: Improving LLM trustworthiness via activation-based uncertainty estimation](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainLP 2025)*, pages 184–193, Suzhou, China. Association for Computational Linguistics.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Subhash Kantamneni, Joshua Engels, Senthoooran Rajamanoharan, Max Tegmark, and Neel Nanda. 2025. [Are sparse autoencoders useful? a case study in sparse probing](#). In *Forty-second International Conference on Machine Learning*.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. [Large language models must be taught to know what they don't know](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. [Simple and scalable predictive uncertainty estimation using deep ensembles](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Benjamin A. Levinstein and Daniel A. Herrmann. 2024. [Still no lie detector for language models: probing empirical and conceptual roadblocks](#). *Philosophical Studies*, 182(7):1539–1565.
- 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.
- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. [Generating with confidence: Uncertainty quantification for black-box large language models](#). *Trans. Mach. Learn. Res.*, 2024.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Andrey Malinin, Anton Ragni, Kate Knill, and Mark Gales. 2017. [Incorporating uncertainty into deep learning for spoken language assessment](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–50, Vancouver, Canada. Association for Computational Linguistics.
- Alex McKenzie, Urja Pawar, Phil Blandfort, William Bankes, David Krueger, Ekdeep Singh Lubana, and Dmitrii Krashennnikov. 2025. [Detecting high-stakes interactions with activation probes](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

- 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.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. [Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Jingwei Ni, Ekaterina Fadeeva, Tianyi Wu, Mubashara Akhtar, Jiaheng Zhang, Elliott Ash, Markus Leippold, Timothy Baldwin, See-Kiong Ng, Artem Shelmanov, and Mrinmaya Sachan. 2026. [ReProbe: Efficient test-time scaling of multi-step reasoning by probing internal states of large language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11667–11689, San Diego, California, United States. Association for Computational Linguistics.
- Oscar Obeso, Andy Ardit, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. 2025. [Real-time detection of hallucinated entities in long-form generation](#). *Preprint*, arXiv:2509.03531.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Taffjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. [2 olmo 2 furious](#). *Preprint*, arXiv:2501.00656.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szepke, Hadas Kotek, and Yonatan Belinkov. 2025. [LLMs know more than they show: On the intrinsic representation of LLM hallucinations](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [Coqa: A conversational question answering challenge](#). *Trans. Assoc. Comput. Linguistics*, 7:249–266.
- Jie Ren, Stanislav Fort, Jeremiah Z. Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. 2021. [A simple fix to mahalanobis distance for improving near-ood detection](#). *CoRR*, abs/2106.09022.
- Andrea Santilli, Adam Golinski, Michael Kirchhof, Federico Danieli, Arno Blaas, Miao Xiong, Luca Zappella, and Sinead Williamson. 2025. [Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 743–759, Vienna, Austria. Association for Computational Linguistics.
- Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2025. [Are the hidden states hiding something? testing the limits of factuality-encoding capabilities in LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6089–6104, Vienna, Austria. Association for Computational Linguistics.
- Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. 2025. [A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35700–35719, Suzhou, China. Association for Computational Linguistics.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- 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](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Chaodong Tong, Qi Zhang, Jiayang Gao, Lei Jiang, Yanbing Liu, and Nannan Sun. 2025. [Halunet: Multi-granular uncertainty modeling for efficient hallucination detection in llm question answering](#). *Preprint*, arXiv:2512.24562.
- Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun,

- Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. [Benchmarking uncertainty quantification methods for large language models with LM-polygraph](#). *Transactions of the Association for Computational Linguistics*, 13:220–248.
- Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2025a. [Unconditional truthfulness: Learning unconditional uncertainty of large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35673–35694, Suzhou, China. Association for Computational Linguistics.
- Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025b. [Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. [Long-form factuality in large language models](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. [Reasoning models know when they’re right: Probing hidden states for self-verification](#). In *Second Conference on Language Modeling*.

A Probing Baselines

In this section we introduce the different supervised UQ methods that we include in Table 1.

SAPLMA (Azaria and Mitchell, 2023) trains a multi-layer perceptron (MLP) using the model hidden states, and we include variations of this method using either the middle or the final layer hidden states. Similar MLP probing architectures are common in existing work (Su et al., 2024; Zhang et al., 2025).

We also include baselines that train probes using attention features. Recent work by Shelmanov et al. (2025) introduces Uhead, a transformer-based probe using attention maps and token probabilities as features. We experiment with two Uhead architectures, reflecting the different hyper-parameter choices proposed in Shelmanov et al. (2025), either using one or two transformer layers for the probe. While Uhead was originally introduced to verify claim spans, we instead apply this approach to evaluate the confidence of each instance. We do this to allow for direct comparability to other methods, performing evaluation at an instance-level rather than at a claim level. Implementing Uhead at an instance level also avoids the need for claim-level training data.

Lookback-lens (Chuang et al., 2024) uses attention, training probes with the proportion of attention paid to the context tokens. We implement lookback-lens probes using sklearn’s LinearRegression architecture rather than the paper’s sklearn LogisticRegression architecture, as the LinearRegression sklearn implementation allows for training with soft target scores. Finally, the SATMD and SATRMD (Vazhentsev et al., 2025b) baselines use density-based features, training probes using the Mahalanobis distance of hidden states from a subset of training examples, using an average distance per layer. SATRMD is a modification of SATMD using the Relative Mahalanobis distance (Ren et al., 2021), which normalises the Mahalanobis distance with a separate reference set.

B Hybrid Back-Off (HBO)

HBO quantifies how far each test example lies from the training distribution using the Mahalanobis distance (MD), using this to adaptively weight the contributions of supervised and unsupervised uncertainty estimates. We describe below how an OOD ranking is computed which then determines this weighting.

Let r denote the rank of a test instance MD (taken from a combination of the training MDs and the added test instance MD). We define R as the normalised rank $r/(N+1)$ where N is the size of the training data. We then use R to decide how much to weight the rank-normalised supervised and unsupervised UQ estimates UQ_{sv} and UQ_{usv} . We describe these weights as W_{sv} and W_{usv} :

$$W_{usv} = \begin{cases} R + 0.5, & \text{if } R \leq 0.5, \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

$$W_{sv} = 1 - W_{usv} \quad (3)$$

$$UQ_{hyb} = W_{sv} * UQ_{sv} + W_{usv} * UQ_{usv} \quad (4)$$

This approach provides an even weighting between the rank-normalised supervised and unsupervised estimates for examples closest to the training distribution, while relying more on the unsupervised estimate for OOD examples. Finally, based on the results from Section 3, we use SAPLMA as a robust probe for UQ_{sv} , while MSP is used for UQ_{usv} . We introduce HBO for short-form generations, based on the strong performance of the MSP baseline in this setting.

HBO was developed to be intentionally simple, and was designed using the SciQ dataset as a validation set where we tested two configurations: 1) the HBO configuration used in the paper, 2) a simplified configuration of HBO where $UQ_{usv} = R$ and $UQ_{sv} = 1 - R$. No additional hyper-parameter tuning was performed, and there was no further development of this method. We show the full evaluation in Tables 30, 31 and 32, showing the performance of both HBO and the simplified configuration on SciQ, all datasets excluding SciQ, and the full ProbeDrift evaluation. While the simplified version of HBO performs well, it does not match the performance of the main configuration on short-form generation tasks.

C Comparing the reduced and full ProbeDrift evaluation

To validate that ProbeDriftLight is representative of the full ProbeDrift evaluation, we compare results across both frameworks for each supervised method presented in Table 1. We report the Mean Absolute Difference (MAD) in PRR scores between the two evaluations across all methods and

OOD settings (see Table 7). As shown in Table 7, differences are small across both short-form and long-form datasets, with a maximum MAD of 0.093 (long-form generations for the LOO setting), suggesting that ProbeDriftLight provides a reliable approximation of the full ProbeDrift evaluation at reduced computational cost.

	ID	LOO	DiffTask
Short-form	0.065	0.060	0.049
Long-form	0.088	0.093	0.075

Table 7: Mean Absolute Difference between the full ProbeDrift evaluation, and the reduced ProbeDrift evaluation comparing each of the results in Table 1.

D Random Seed Variance Analysis

The main results in this paper are reported using a single random seed. In this section we show that while variance increases under larger distributional shifts, the overall pattern of degradation remains consistent across seeds.

We report standard deviations both for a single dataset (see Table 8 and Table 9) and when averaging across the three datasets used in ProbeDriftLight (see Table 10). Variance is low in both cases, and decreases considerably when averaging across multiple datasets. For example, the standard deviation for SAPLMA (middle) drops from 0.10 in the DiffTask setting for a single dataset to 0.05 when averaged across three, supporting the reliability of the single-seed results reported in the main paper.

	ID	LOO	DiffTask
1 seed	0.63	0.57	0.35
5 seed avg.	0.62	0.51	0.29
Std dev	0.03	0.07	0.10

Table 8: 1 seed vs. 5 seeds using a single dataset (SciQ) for SAPLMA (middle) for Llama.

	ID	LOO	DiffTask
1 seed	0.61	0.24	-0.33
5 seed avg.	0.57	0.08	-0.02
Std dev	0.03	0.12	0.18

Table 9: 1 seed vs. 5 seeds using a single dataset (SciQ) for SAPLMA (top) for Llama.

	ID	LOO	DiffTask
1 seed	0.59	0.36	0.30
5 seed avg.	0.58	0.31	0.22
Std dev	0.01	0.03	0.05

Table 10: 1 seed vs. 5 seeds averaged across 3 datasets for SAPLMA (middle) for Llama.

E Evaluation with Different Correctness Functions

To understand the impact of changing the correctness function used for evaluation, we provide results showing the performance of SAPLMA (middle) across four different correctness functions: AlignScore, ROUGE-L, exact match (also referred to as accuracy) and LLM-as-a-Judge (see Table 2). In Table 13 we also show the performance of SAPLMA (top) for these same correctness functions.

We find that AlignScore, ROUGE-L and exact match are highly correlated. For SAPLMA (top), LLM-as-a-Judge evaluation is similar to the AlignScore evaluation, however for SAPLMA (middle) we see better robustness reported using LLM-as-a-Judge (although performance still degrades further OOD in line with each of the other evaluation metrics).

To better understand these differences for SAPLMA (middle), we compare LLM-as-a-Judge evaluation to AlignScore evaluation for each specific dataset (see Table 12). We find that the differences are particularly noticeable on specific datasets such as SciQ and XSum where we find substantially better robustness using LLM-as-a-Judge.

Finally, in Table 15 we provide the full results from Table 1 when using exact match, ROUGE-L and AlignScore for evaluation, demonstrating how closely correlated these three correctness functions are. We only provide results for short-form generations, as the exact-match correctness function is only meaningful in this setting.

F Training with Different Correctness Functions

Table 17 provides results comparing AlignScore and LLM-as-a-Judge as training signals for SAPLMA (middle). These results include evaluation across multiple different correctness functions.

For short-form generations, training with LLM-as-a-Judge improves robustness, with the largest

Dataset	Exact match	ROUGE-L	AlignScore	LLM-as-a-Judge
SciQ	0.44 [0.30, 0.58]	0.86 [0.72, 0.96]	0.84 [0.68, 0.96]	0.80 [0.68, 0.90]
TriviaQA	0.76 [0.64, 0.88]	0.92 [0.82, 1.00]	0.84 [0.68, 0.96]	0.94 [0.86, 1.00]
XSum	0.00 [0.00, 0.00]	0.48 [0.24, 0.72]	0.32 [0.04, 0.56]	0.52 [0.30, 0.74]

Table 11: Somers’ D alignment with human preferences for each correctness function, by dataset, with 95% bootstrap confidence intervals.

Eval data	Llama 3.1-8B				
	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask
PRR evaluation using AlignScore					
SciQ	0.63	0.57	0.27	0.35	0.35
TriviaQA	0.74	0.48	0.46	0.36	0.33
CoQA	0.52	0.32	0.03	0.33	0.27
XSum	0.32	0.07	0.13	0.12	0.09
PubmedQA	0.39	0.04	0.06	0.19	0.08
CNN/DailyMail	0.24	-0.03	0.02	0.05	-0.04
PRR evaluation using LLM-as-a-judge					
SciQ	0.69	0.73	0.24	0.63	0.61
TriviaQA	0.79	0.49	0.52	0.38	0.38
CoQA	0.56	0.41	0.18	0.44	0.37
XSum	0.38	0.16	0.22	0.29	0.07
PubmedQA	0.36	0.32	0.17	0.06	0.37
CNN/DailyMail	0.25	-0.07	0.03	0.08	-0.19

Table 12: PRR results (higher is better) for SAPLMA (middle layer) on Llama 3.1-8B, evaluated per dataset across in-domain (ID) and out-of-domain (OOD) settings, using AlignScore and LLM-as-a-judge as correctness functions. Abbreviations: LOO = leave-one-out OOD; 1D-SameTask = one dataset, same task (OOD); DiffTask = multiple datasets, different task (OOD); 1D-DiffTask = one dataset, different task (OOD).

improvements observed after evaluating with LLM-as-a-Judge itself. However, we also see improvements over training with AlignScore when AlignScore is used for evaluation. This suggests that LLM-as-a-Judge provides a more generalisable training signal.

G Human Evaluation of Correctness Functions

We perform a small human evaluation to check the appropriateness of the different correctness functions used in this work. This involves an author of the paper assessing pairs of test examples, measuring whether human preferences align with preferences from each correctness function. We measure the alignment to human preferences using Somers’ D, where we treat the human preferences as the reference variable. As a result, we do not consider examples where human judgement has no preference (81% of pairs for SciQ), leaving a sample of 50 pairs per dataset. The human evaluator only had access to the prompt, reference answer, and the LLM output when giving their preference, and did not see any of the correctness function scores.

As a result of this small sample size, this evaluation can be seen as a sanity check on our choice of correctness functions.

Table 11 suggests that ROUGE-L, AlignScore and LLM-as-a-Judge are all appropriate correctness functions for short-form generations, however determining correctness for long-form generations is substantially more challenging. Table 18 shows the pairs used for each dataset, including the number of pairs where human annotation expressed no preference.

H Evaluation with AlignScore for Long-Form Generations

In Table 29, we show each of the different design choices using the PubmedQA long-form dataset in ProbeDriftLight after training with AlignScore. We demonstrate consistent OOD failures across each of these different design choices, suggesting the limitations of using AlignScore in this setting.

I Layer Choice Analysis

Table 26 verifies that the findings from Section 4.1 hold consistently across a range of LLMs and evaluation metrics. Across all six models tested, in-distribution performance is broadly similar across layers, while OOD performance degrades substantially at early and final layers. Middle-layer representations (including the middle layer itself and for nearby layers) are consistently more robust across AlignScore, ROUGE-L, and exact match accuracy, with the exception of Qwen-2.5-3B-IT which performs better for later layers.

J Aggregation Choice Analysis

Table 28 verifies that the findings from Section 4.2 hold across different evaluation metrics. Across AlignScore, ROUGE-L, exact match, and LLM-as-a-Judge, averaging over response tokens achieves the best PRR for most OOD settings. Context-based aggregation strategies consistently underperform response-token strategies across all metrics.

Metric	ID		LOO		1D-SameTask		DiffTask		1D-DiffTask	
	PRR	%	PRR	%	PRR	%	PRR	%	PRR	%
					near OOD	→	most OOD			
Short-form generation datasets										
AlignScore	0.54	100%	0.26	48%	-0.12	<0%	-0.14	<0%	0.18	33%
ROUGE-L	0.60	100%	0.28	47%	-0.12	<0%	-0.04	<0%	0.13	22%
Exact match	0.59	100%	0.26	44%	-0.11	<0%	0.01	2%	0.10	17%
LLM-as-a-judge	0.60	100%	0.31	52%	-0.13	<0%	-0.24	<0%	0.25	42%
Long-form generation datasets										
AlignScore	0.20	100%	0.00	0%	0.04	20%	0.05	25%	-0.00	0%
ROUGE-L	0.20	100%	-0.07	<0%	0.03	15%	-0.03	<0%	0.00	0%
LLM-as-a-judge	0.11	100%	0.04	36%	-0.06	<0%	0.05	45%	0.00	0%

Table 13: PRR results (higher is better) for SAPLMA (top) across four correctness functions used for evaluation (AlignScore is used for training), for Llama 3.1-8B. Each value is shown alongside its percentage of the ID score for that metric. Exact match is excluded for long-form generations as the metric is uninformative in this setting.

Architecture	ID	LOO	DiffTask
Middle-layer hidden states			
SAPLMA	0.58	0.29	0.25
Uhead (1 layer)	0.52	0.26	0.21
Uhead (2 layers)	0.50	0.23	0.13
Linear regression (GD)	0.52	0.24	0.26
Linear regression (CF)	0.39	0.12	0.05
Linear regression (CF) + PCA	0.50	0.22	0.16
Linear regression (CF, Ridge)	0.44	0.17	0.11
SAPLMA variations			
SAPLMA (2 layer)	0.57	0.29	0.24
SAPLMA (3 layers)	0.57	0.26	0.24
SAPLMA (5 layers)	0.60	0.30	0.21

Table 14: Architecture comparison on ProbeDriftLight using middle-layer hidden states (single-column version). OOD: LOO = leave-one-out; DiffTask = multiple datasets, different task. The best result (measuring PRR) in each column is in bold. Results are an average across both Llama and Gemma models. CF = closed-form; GD = gradient-descent.

Additionally, we compare using an average from each of the response tokens to results using only the last response token, testing this across a variety of different LLMs (see Table 27). These were the two strategies with the highest in-distribution performance in Table 4. This analysis shows how using the average aggregation is more robust for all of the models further OOD (in the DiffTask setting).

K Hidden State Visualisations

This section shows the Partial Least Squares regression visualisations for both SciQ and PubmedQA. As expected, for SciQ the graphs show how less correct generations are less easily distinguishable from more correct generations in further OOD settings (see Figure 4).

For PubmedQA (see Figure 5) there are two distinct groups of examples, which upon closer inspection

represent examples where the model provides a yes/no answer without further elaboration (the smaller group, which have low AlignScores), and where the model provides an elaboration alongside its answer (the larger group). We find that separating these two types of answers is insufficient for good performance. Within the larger group of examples, for the in-distribution graph we can see that the correct and less correct examples are somewhat distinguishable, although not to the same extent as for SciQ and TriviaQA. This is explained by the lower PRR observed on PubmedQA (0.39 for SAPLMA, see Table 25), compared to SciQ and TriviaQA (0.63 and 0.74 respectively). Similar to the other datasets, these examples become less distinguishable in further OOD settings.

Table 19 shows the Spearman’s rank correlation coefficients for both the training and test data for the PLS regression models.

L Full Analysis of Past Experimentation

Table 33 provides the full list of papers analysed for their OOD experimentation and claims. We include probing papers from 2024 onwards, by which point SAPLMA (Azaria and Mitchell, 2023) was an established approach. For each paper, we investigate the different distributional shifts tested, considering evaluation in a leave-one-out setting (which we describe as near-OOD), evaluation after training on a single same-task dataset, and evaluation after training only on different-task datasets. We also highlight work that evaluates OOD performance on at least 3 different test sets, and papers that do not rely entirely on synthetic evaluation. Finally, we consider which papers include a strong unsupervised baseline (using the model soft probabilities, such as MSP or Perplexity), or a strong

	ID		LOO		1D-SameTask		DiffTask		1D-DiffTask	
			near OOD		→	more OOD	→	most OOD		
AlignScore — Short-form generation datasets										
SAPLMA (middle)	0.63	100%	0.46	73%	0.26	41%	0.35	56%	0.32	51%
SAPLMA (top)	0.54	100%	0.26	48%	-0.12	<0%	-0.14	<0%	0.18	33%
Uhead (1 layer)	0.63	100%	0.17	27%	0.24	38%	0.16	25%	0.08	13%
Uhead (2 layers)	0.63	100%	0.15	24%	0.15	24%	-0.27	<0%	-0.36	<0%
SATMD	0.30	100%	0.35	117%	-0.15	<0%	0.21	70%	0.15	50%
SATRMD	0.47	100%	0.17	36%	-0.11	<0%	0.21	45%	0.13	28%
Lookback-lens	0.27	100%	0.00	0%	-0.03	<0%	-0.13	<0%	0.19	70%
SATMD-MSP	0.44	100%	0.32	73%	0.06	14%	0.24	55%	0.23	52%
SATRMD-MSP	0.60	100%	0.30	50%	0.05	8%	0.20	33%	0.21	35%
HUQ-SATMD	0.58	100%	0.59	102%	-0.14	<0%	0.22	38%	0.25	43%
HUQ-SATRMD	0.61	100%	0.47	77%	-0.08	<0%	0.24	39%	0.18	30%
Avg. supervised	0.50	100%	0.22	45%	0.03	7%	0.06	11%	0.10	20%
Avg. hybrid	0.56	100%	0.42	75%	-0.03	<0%	0.23	40%	0.22	39%
MSP	0.57		0.57		0.57		0.57		0.57	
ROUGE-L — Short-form generation datasets										
SAPLMA (middle)	0.70	100%	0.49	70%	0.30	43%	0.33	47%	0.35	50%
SAPLMA (top)	0.60	100%	0.28	47%	-0.12	<0%	-0.04	<0%	0.13	22%
Uhead (1 layer)	0.73	100%	0.33	45%	0.23	32%	0.18	25%	0.01	1%
Uhead (2 layers)	0.74	100%	0.30	41%	0.12	16%	-0.18	<0%	-0.12	<0%
SATMD	0.32	100%	0.26	81%	-0.26	<0%	0.27	84%	0.24	75%
SATRMD	0.55	100%	0.31	56%	-0.06	<0%	0.26	47%	0.16	29%
Lookback-lens	0.33	100%	0.13	39%	-0.09	<0%	-0.03	<0%	0.32	97%
SATMD-MSP	0.63	100%	0.04	6%	-0.15	<0%	0.19	30%	0.16	25%
SATRMD-MSP	0.71	100%	0.29	41%	0.00	0%	0.18	25%	0.10	14%
HUQ-SATMD	0.63	100%	0.60	95%	-0.18	<0%	0.26	41%	0.24	38%
HUQ-SATRMD	0.70	100%	0.59	84%	-0.02	<0%	0.27	39%	0.17	24%
Avg. supervised	0.57	100%	0.30	53%	0.02	3%	0.11	20%	0.16	27%
Avg. hybrid	0.67	100%	0.38	57%	-0.09	<0%	0.23	34%	0.17	25%
MSP	0.63		0.63		0.63		0.63		0.63	
Accuracy — Short-form generation datasets										
SAPLMA (middle)	0.70	100%	0.48	69%	0.30	43%	0.33	47%	0.31	44%
SAPLMA (top)	0.59	100%	0.26	44%	-0.11	<0%	0.01	2%	0.10	17%
Uhead (1 layer)	0.74	100%	0.32	43%	0.23	31%	0.15	20%	0.03	4%
Uhead (2 layers)	0.75	100%	0.31	41%	0.10	13%	-0.21	<0%	-0.06	<0%
SATMD	0.27	100%	0.26	96%	-0.17	<0%	0.25	93%	0.18	67%
SATRMD	0.56	100%	0.34	61%	-0.09	<0%	0.22	39%	0.09	16%
Lookback-lens	0.29	100%	0.10	34%	-0.03	<0%	-0.03	<0%	0.40	138%
SATMD-MSP	0.63	100%	0.10	16%	-0.09	<0%	0.15	24%	0.10	16%
SATRMD-MSP	0.71	100%	0.32	45%	-0.02	<0%	0.11	15%	0.03	4%
HUQ-SATMD	0.69	100%	0.66	96%	-0.13	<0%	0.23	33%	0.17	25%
HUQ-SATRMD	0.75	100%	0.64	85%	-0.04	<0%	0.21	28%	0.08	11%
Avg. supervised	0.56	100%	0.30	53%	0.03	6%	0.10	18%	0.15	27%
Avg. hybrid	0.69	100%	0.43	62%	-0.07	<0%	0.17	25%	0.10	14%
MSP	0.69		0.69		0.69		0.69		0.69	

Table 15: PRR results (higher is better) for Llama 3.1-8B across three correctness functions (AlignScore, ROUGE-L, and exact match accuracy), for short-form generation datasets only. Each value is shown alongside its percentage of the ID score for that method. Average rows show the mean PRR across supervised and hybrid methods respectively, with percentages computed relative to the averaged ID score.

supervised baseline (training a probe with a middle layer hidden state, which for Table 33 we define as any hidden state that is not from the first or last layer).

M Probe Architectures

We experiment with a range of different probe architectures when using the middle layer hidden state from the model (which the best performing

SAPLMA baseline used). Similar to the other results in Section 4, we consider an average from Gemma-2-9B and Llama 3.1-8B across SciQ, TriviaQA and PubmedQA. We experiment with the SAPLMA architecture, a four layer MLP with hidden layers of size 256, 128 and 64, with a single output logit on the final layer. We also experiment with the two Uhead architectures used by Shelmanov et al. (2025) after their hyper-parameter tun-

Dataset	Scoring judgements			Binary judgements	
	0	(0, 1)	1	0	1
COQA	9.6%	8.5%	82.0%	14.1%	85.9%
PubmedQA	5.2%	17.4%	77.4%	10.0%	90.0%
SciQ	4.2%	4.5%	91.3%	7.0%	93.0%
TriviaQA	35.2%	3.9%	61.0%	37.5%	62.5%

Table 16: Score distributions for scoring LLM-as-a-Judge judgements vs binary judgements. We show results for our QA datasets, as binary judgements are less suitable for summarisation tasks.

Eval Metric	ID		LOO		DiffTask	
	PRR	%	PRR	%	PRR	%
<i>Short-form — Trained with AlignScore</i>						
AlignScore	0.69	100%	0.52	77%	0.35	52%
ROUGE-L	0.78	100%	0.57	73%	0.32	40%
Accuracy	0.77	100%	0.57	75%	0.32	41%
LLM-judge	0.74	100%	0.61	82%	0.51	68%
<i>Short-form — Trained with LLM-as-a-judge</i>						
AlignScore	0.66	100%	0.56	86%	0.36	55%
ROUGE-L	0.62	100%	0.56	90%	0.39	62%
Accuracy	0.57	100%	0.53	92%	0.35	62%
LLM-judge	0.86	100%	0.72	84%	0.55	64%
<i>Long-form — Trained with AlignScore</i>						
AlignScore	0.39	100%	0.04	10%	0.19	49%
ROUGE-L	0.40	100%	-0.17	<0%	0.09	22%
LLM-judge	0.36	100%	0.32	89%	0.06	17%
<i>Long-form — Trained with LLM-as-a-judge</i>						
AlignScore	-0.08	-	-0.11	-	0.13	-
ROUGE-L	-0.35	-	-0.41	-	-0.02	-
LLM-judge	0.69	100%	0.31	45%	0.46	67%

Table 17: PRR results (higher is better) for SAPLMA (middle) trained with different training signals, evaluated with four correctness functions, for Llama 3.1-8B with ProbeDriftLight. Accuracy is excluded as an evaluation metric for long-form generations (as this is not applicable in these cases). Each value is shown alongside its percentage of the ID score for that metric.

ing. The single layer Uhead architecture involves a single transformer layer with 16 attention heads, while the 2 layer Uhead has 4 attention heads. Both architectures have 768 dimensional hidden states. We additionally experiment with a single linear layer (Chuang et al., 2024), and a linear layer after applying PCA (Vazhentsev et al., 2025b). Based on the success of SAPLMA, we also experiment with removing either one or two linear layers (removing those with the largest dimensionality), or introducing an additional layer with a 512 dimensional hidden state (resulting in 5 layers with 512, 256, 128 and 64 dimensional hidden states respectively).

We find that all the probes tested (with the exception of linear probes) achieve very similar levels

of robustness (see Table 14). When evaluating in a leave-one-out OOD setting, all other architectures are within 0.07 percentage points. We conclude that the choice of architecture makes less difference to the probe’s robustness (although interestingly we see larger differences for in-distribution performance). For linear probes, their robustness depends on the training procedure, with probes trained with gradient descent achieving similar robustness to other probing architectures, while closed-form methods (including when trained with PCA or ridge regression regularisation) have worse robustness.

Our results also suggest that additional hyper-parameter tuning for the probes is unlikely to further improve robustness. We find similar performance across each type of probe with their different parameters (for example comparing the two Uhead configurations). Moreover, as hyper-parameter tuning is also likely to be conducted on an in-distribution validation set, we argue that this is unlikely to lead to substantial improvements in robustness.

N LLM-as-a-Judge Setup

For our LLM-as-a-Judge we allow for either a correct judgement (a score of 1), an incorrect judgement (a score of 0), or a partially correct judgement with a score between 0 and 1. We do this to avoid class imbalances which cause higher variance with the PRR metric, which considers the extent to which more correct instances are more confident than less correct instances. We show the difference in distribution between both approaches in Table 16, where the binary-judgement LLM-as-a-Judge setup mirrors the judge in Orgad et al. (2025).

We use GPT-5 (gpt-5-2025-08-07) (Singh et al., 2026) for the LLM-as-a-Judge. The soft scoring approach used in the paper uses the prompts in Figure 6. Human preferences are used to assess the

Dataset	Pairs seen (incl. ties)	# Ties	Sample size	Tie rate
SciQ	261	211	50	80.8%
TriviaQA	109	59	50	54.1%
XSum	91	41	50	45.1%

Table 18: Number of pairs sampled for the human evaluation, before and after excluding pairs where human annotation expressed no preference.

	Train ρ	Test ρ
SciQ		
ID	0.7421	0.5163
LOO	0.5432	0.4633
DiffTask	0.5528	0.2518
TriviaQA		
ID	0.5988	0.5385
LOO	0.5733	0.2601
DiffTask	0.5528	0.2738
PubmedQA		
ID	0.3754	0.3526
LOO	0.6221	-0.0674
DiffTask	0.5528	-0.0903

Table 19: Spearman’s rank correlation coefficient for PLS regression models for both the training and test data.

Training set size	ID	LOO	DiffTask
SAPLMA (middle)			
Normal (1x)	0.59	0.36	0.30
x2	0.59	0.37	0.17
x4	0.62	0.40	0.32
SAPLMA (top)			
Normal (1x)	0.51	0.17	-0.10
x2	0.52	0.06	0.02
x4	0.56	0.22	0.09

Table 20: PRR results (higher is better) for SAPLMA (middle) and SAPLMA (top), comparing training set size (Normal, 2x, and 4x the standard size) for Llama-3.1-8B with ProbeDriftLight.

LLM-as-a-Judge, comparing the LLM-as-a-Judge to other correctness functions (see Appendix G):

O The performance of unsupervised baselines

Table 23 and Table 24 show the performance of a range of unsupervised methods for Llama-3.1 and Gemma-2 respectively. We find that MSP (Maximum Sequence Probability) performs best for short-form generations on average across both models, and we therefore consider this as our main unsupervised baseline. For long-form generations all methods perform poorly (with close to random chance performance).

The unsupervised baselines include MSP, Per-

plexity, Lexical Similarity (Fomicheva et al., 2020), the DegMat, Eccentricity and EigValLaplacian methods proposed by Lin et al. (2024), Semantic Entropy (Kuhn et al., 2023) and SAR (Duan et al., 2024).

P Scaling up Training Data Size

Throughout this work, we fix the number of training examples to allow a like-for-like comparison between experiments, training probes with 1,800 training instances. In Table 20 we also experiment with training with more data, using either two or four times the amount of training data.

When increasing the size of the training data, there are small improvements in in-distribution performance. There are also some improvements in out-of-distribution performance, although these are less consistent. We find similar behaviour when increasing the size of the training data, with probes on the middle layer hidden states being consistently more robust than probes trained on the last layer hidden states.

Q Training and Evaluation Datasets used OOD

Table 22 provides a comprehensive list of each of the datasets used for training for each OOD setting and dataset used for evaluation. In each case, we fix the number of training instances to 1,800. We evaluate on at most 2,000 test instances (following Vazhentsev et al. (2025b)).

R Additional Baseline for Long-Form Answers

We additionally consider the probing method proposed by Obeso et al. (2025) specifically for long-form generations. This approach involves additional annotation for each training instance to create hallucination labels for different entities. Therefore, we use the probes provided by Obeso et al. (2025) which are trained on a sample of 10,000 instances from LongFact (Wei et al., 2024), LongFact++ and TriviaQA. The training process involves using a

Dataset	Avg. words
Short-form generation datasets	
SciQ	1.5
TriviaQA	2.3
COQA	2.5
Avg. Short form	2.1
Long-form generation datasets	
XSum	20.9
PubmedQA	37.1
CNN	54.4
Avg. Long form	37.5

Table 21: Average dataset answer lengths.

LoRA adapter to change the behaviour of the original baseline model in addition to training a linear probe. For evaluation, we consider each test instance as a single entity.

As we present in Table 25, despite the much larger training sample, the additional entity-level annotation, and the use of the LoRA adapter, we do not find more robust behaviour from these probes. Similar to the other probes considered, performance on the long-form datasets is similar to random chance performance.

S Answer Length by Dataset

Table 21 shows the answer lengths (number of words) for each dataset, showing which datasets are categorised as long-form and short-form based on the length of their answers. The short-form generations have 2.1 words on average per answer, compared to 37.5 for the long-form generations.

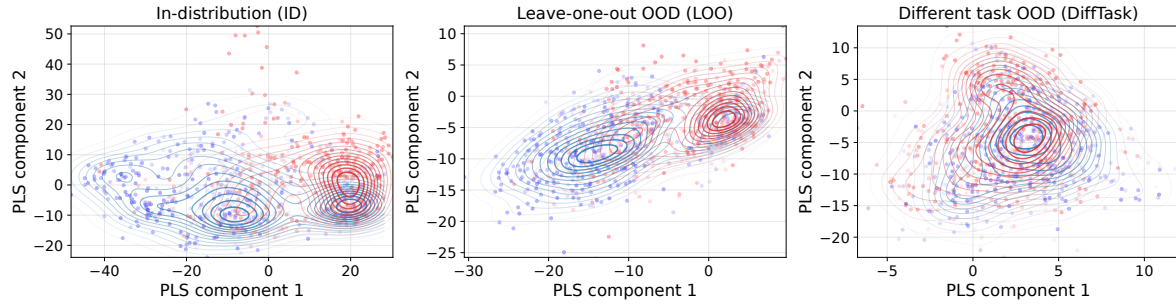


Figure 4: Visualisation of Llama 3.1-8B hidden states (layer 16) evaluated on SciQ, either for in-distribution, leave-one-out or different task OOD settings. Hidden states are down-projected to 2d subspaces using PLS models fitted to predict AlignScore from the training examples. As OOD-ness increases, hidden states of more correct generations (blue) and less correct generations (red) become less separable. Kernel density contour lines show the density of generations above and below the median AlignScore.

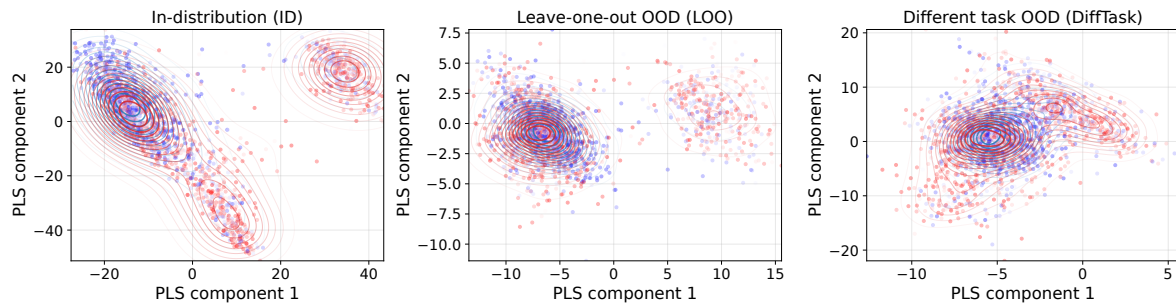


Figure 5: Visualisation of Llama 3.1-8B hidden states (layer 16) evaluated on PubmedQA, either for in-distribution, leave-one-out or different task OOD settings. Hidden states are down-projected to 2d subspaces using PLS models fitted to predict AlignScore from the training examples. As OOD-ness increases, hidden states of more correct generations (blue) and less correct generations (red) become less separable. Kernel density contour lines show the density of generations above and below the median AlignScore.

QA LLM-as-a-Judge prompt:
<p>Evaluate the following answers to questions. For each question you will be given a model answer and the correct answer. You must rate how correct the model answer is on a scale from 0.0 to 1.0, where:</p> <ul style="list-style-type: none"> - 1.0 means the model answer is completely correct - 0.0 means the model answer is completely incorrect or wrong - Values in between reflect answers that are partially correct but incomplete or imprecise <p>Only respond with a number between 0.0 to 1.0. Do not write any explanation.</p> <p>For example:</p> <p>Question: who is the young guitarist who played with buddy guy? Ground Truth: Quinn Sullivan Model Answer: Ronnie Earl Score: 0.0</p> <p>Question: name of the first episode of stranger things Ground Truth: Chapter One: The Vanishing of Will Byers Model Answer: The disappearance of Will Byers Score: 0.0</p> <p>Question: What are the symptoms of diabetes? Ground Truth: Common symptoms include increased thirst, frequent urination, fatigue, and blurred vision. Model Answer: Symptoms of diabetes include increased thirst and frequent urination. Score: 0.6</p> <p>Question: What is the capital of Australia? Ground Truth: Canberra Model Answer: Canberra, which is located in the Australian Capital Territory Score: 1.0</p> <p>{input_question} Ground Truth: {label} Model Answer: {model_answer} Score:</p>
Summarisation LLM-as-a-Judge prompt:
<p>Only respond with a number between 0.0 to 1.0. Do not write any explanation.</p> <p>The task below is a text summarisation task. You will see a Text, a Ground Truth Summary, and a Model Summary. Score how well the Model Summary matches the Ground Truth Summary from 0.0 to 1.0, where 1.0 means it conveys the same information, 0.0 means it is completely different or irrelevant, and scores in between reflect partial overlap in the key points covered.</p> <p>{input_question} Ground Truth Summary: {label} Model Summary: {model_answer} Score:</p>

Figure 6: LLM prompts used for question answering and summarisation tasks.

Evaluation data	ID	LOO	ID-SameTask	DiffTask	1D-DiffTask
QA evaluation data					
SciQ (1000)	SciQ (1800)	SamSum (200), XSum (200), CNN-Dailymail (200), COQA (200), TriviaQA (200), MMLU (200), PubmedQA (200), MedQUAD (200), TruthfulQA (200)	MedQUAD (1800)	SamSum (600), XSum (600), CNN-Dailymail (600)	SamSum (1800)
TriviaQA (2000)	TriviaQA (1800)	SamSum (200), XSum (200), CNN-Dailymail (200), COQA (200), SciQ (200), MMLU (200), PubmedQA (200), MedQUAD (200), TruthfulQA (200)	MedQUAD (1800)	SamSum (600), XSum (600), CNN-Dailymail (600)	SamSum (1800)
PubmedQA (2000)	PubmedQA (1800)	SamSum (200), XSum (200), CNN-Dailymail (200), COQA (200), SciQ (200), MMLU (200), TriviaQA (200), MedQUAD (200), TruthfulQA (200)	MedQUAD (1800)	SamSum (600), XSum (600), CNN-Dailymail (600)	SamSum (1800)
COQA (2000)	COQA (1800)	SamSum (200), XSum (200), CNN-Dailymail (200), PubmedQA (200), SciQ (200), MMLU (200), TriviaQA (200), MedQUAD (200), TruthfulQA (200)	MedQUAD (1800)	SamSum (600), XSum (600), CNN-Dailymail (600)	SamSum (1800)
Summarisation evaluation data					
XSum (2000)	XSum (1800)	SamSum (200), SciQ (200), CNN-Dailymail (200), COQA (200), TriviaQA (200), MMLU (200), PubmedQA (200), MedQUAD (200), TruthfulQA (200)	SamSum (1800)	SciQ (300), MMLU (300), TriviaQA (300), COQA (300), TruthfulQA (300), MedQUAD (300)	MedQUAD (1800)
CNN-Dailymail (2000)	CNN-Dailymail (1800)	SamSum (200), SciQ (200), XSum (200), COQA (200), TriviaQA (200), MMLU (200), PubmedQA (200), MedQUAD (200), TruthfulQA (200)	SamSum (1800)	SciQ (300), MMLU (300), TriviaQA (300), COQA (300), TruthfulQA (300), MedQUAD (300)	MedQUAD (1800)

Table 22: For each dataset used for evaluation, the different datasets used for training are provided for each OOD setting. In each case, the number of training examples are provided next to each dataset (the total is always 1,800 for each experiment). The total amount of evaluation data used for each dataset is also provided in the first column (we sample at most 2000 test instances).

Method	Short-form			Avg. Short	Long-form			Avg. Long
	SciQ	COQA	TriviaQA		XSum	CNN	PubmedQA	
Llama 3.1-8B results								
MSP	0.58	0.45	0.69	0.57	-0.13	0.10	-0.01	-0.01
Perplexity	0.20	0.45	0.69	0.45	-0.11	0.24	-0.04	0.03
DegMat NLI Score Ent	0.55	0.43	0.72	0.57	0.06	0.16	0.07	0.10
Eccentricity NLI Score Ent	0.52	0.44	0.65	0.54	0.02	0.05	0.01	0.03
Eig ValLaplacian NLI Score Ent	0.52	0.47	0.71	0.57	0.06	0.16	0.07	0.10
Lexical Similarity ROUGE-L	0.49	0.42	0.71	0.54	0.02	0.09	0.05	0.05
SAR	0.45	0.45	0.76	0.55	0.01	0.13	0.02	0.05
Semantic Entropy	0.52	0.44	0.56	0.51	0.02	0.08	0.02	0.04
SentenceSAR	0.55	0.46	0.73	0.58	-0.04	0.06	-0.03	-0.00

Table 23: Performance of unsupervised methods using Llama-3.1-8B on our chosen datasets, using PRR with AlignScore for evaluation.

Method	Short-form				Long-form			
	SciQ	COQA	TriviaQA	Avg. Short	XSum	CNN	PubmedQA	Avg. Long
Gemma-2-9B results								
MSP	0.61	0.53	0.77	0.64	-0.14	0.05	-0.21	-0.10
Perplexity	0.18	0.52	0.73	0.48	-0.19	0.01	0.20	0.01
DegMat NLI Score Ent	0.51	0.42	0.74	0.56	0.09	0.10	0.02	0.07
Eccentricity NLI Score Ent	0.50	0.48	0.72	0.57	0.02	0.06	-0.00	0.03
Eig ValLaplacian NLI Score Ent	0.48	0.45	0.75	0.56	0.09	0.10	0.03	0.07
Lexical Similarity ROUGE-L	0.47	0.44	0.74	0.55	0.06	0.10	0.05	0.07
SAR	0.44	0.48	0.75	0.56	0.03	0.11	0.03	0.06
Semantic Entropy	0.49	0.49	0.60	0.53	0.07	0.08	0.02	0.06
SentenceSAR	0.58	0.53	0.75	0.62	-0.00	0.06	-0.06	-0.00

Table 24: Performance of unsupervised methods using Gemma-2-9B on our chosen datasets, using PRR with AlignScore for evaluation.

Method	Short-form			Avg. Short	Long-form			Avg. Long
	SciQ	COQA	TriviaQA		XSum	CNN	PubmedQA	
Llama 3.1-8B results								
SAPLMA (ID)	0.63	0.52	0.74	0.63	0.32	0.24	0.39	0.32
SAPLMA (LOO)	0.57	0.32	0.48	0.46	0.07	-0.03	0.04	0.03
SAPLMA (1D-SameTask)	0.27	0.03	0.46	0.26	0.13	0.02	0.06	0.07
SAPLMA (DiffTask)	0.35	0.33	0.36	0.35	0.12	0.05	0.19	0.12
SAPLMA (1D-DiffTask)	0.35	0.27	0.33	0.32	0.09	-0.04	0.08	0.04
Obeso et al. (2025)	0.50	0.26	0.37	0.38	0.01	-0.01	-0.05	-0.02

Table 25: We compare the robustness of the method proposed by [Obeso et al. \(2025\)](#), showing no better robustness on long-form tasks compared to the SAPLMA probe (using the middle-layer hidden states).

Layer	AlignScore			ROUGE-L			Accuracy		
	ID	LOO	DiffTask	ID	LOO	DiffTask	ID	LOO	DiffTask
Llama-3.1-8B									
middle-12	0.39	0.03	0.20	0.47	0.07	0.25	0.50	0.39	0.34
middle-8	0.49	0.21	0.17	0.57	0.21	0.15	0.59	0.35	0.38
middle-4	0.58	0.21	0.18	0.63	0.26	0.14	0.64	0.46	0.21
middle	0.59	0.36	0.30	0.65	0.32	0.24	0.68	0.55	0.38
middle+4	0.55	0.31	0.24	0.59	0.27	0.19	0.62	0.50	0.37
middle+8	0.51	0.28	0.09	0.56	0.26	0.12	0.60	0.46	0.24
middle+12	0.51	0.28	0.06	0.57	0.23	0.10	0.61	0.47	0.26
last	0.51	0.17	-0.10	0.57	0.12	-0.08	0.61	0.36	0.12
Llama-3.1-8B-IT									
middle-12	0.45	0.10	-0.02	0.34	0.11	0.02	0.47	0.31	0.20
middle-8	0.55	0.09	0.05	0.45	0.11	0.05	0.54	0.30	0.26
middle-4	0.62	0.31	0.06	0.55	0.25	0.08	0.62	0.44	0.28
middle	0.64	0.29	0.11	0.56	0.28	0.13	0.64	0.48	0.33
middle+4	0.62	0.31	0.06	0.51	0.28	0.05	0.63	0.46	0.24
middle+8	0.59	0.24	0.01	0.50	0.23	0.06	0.60	0.43	0.26
middle+12	0.59	0.22	0.02	0.48	0.22	0.07	0.60	0.41	0.27
last	0.55	0.25	0.01	0.46	0.26	0.05	0.58	0.43	0.26
Gemma-2-9B									
middle-16	0.35	0.06	0.11	0.40	0.01	0.07	0.46	0.33	0.23
middle-12	0.44	0.02	0.07	0.54	-0.07	-0.03	0.50	0.28	0.26
middle-8	0.50	0.15	0.15	0.56	0.07	-0.01	0.57	0.40	0.23
middle-4	0.56	0.17	0.25	0.64	0.12	0.28	0.60	0.47	0.29
middle	0.58	0.22	0.20	0.64	0.15	0.09	0.62	0.48	0.38
middle+4	0.59	0.29	0.27	0.69	0.22	0.17	0.64	0.53	0.44
middle+8	0.60	0.26	0.21	0.69	0.18	0.16	0.63	0.52	0.37
middle+12	0.55	0.26	0.17	0.63	0.19	0.12	0.60	0.52	0.32
middle+16	0.53	0.24	0.17	0.64	0.16	0.12	0.60	0.49	0.31
last	0.54	0.17	0.01	0.65	0.12	0.03	0.62	0.44	0.14
OLMo-2-7B									
middle-12	0.33	0.10	0.08	0.32	0.08	0.07	0.28	0.15	0.13
middle-8	0.42	0.15	0.10	0.42	0.13	0.08	0.41	0.18	0.16
middle-4	0.52	0.28	0.18	0.49	0.23	0.19	0.48	0.24	0.27
middle	0.58	0.42	0.37	0.56	0.43	0.37	0.55	0.47	0.31
middle+4	0.54	0.35	0.33	0.53	0.36	0.34	0.51	0.38	0.37
middle+8	0.52	0.34	0.26	0.49	0.30	0.22	0.49	0.36	0.23
middle+12	0.51	0.33	0.14	0.46	0.29	0.13	0.48	0.30	0.13
last	0.48	0.31	0.16	0.44	0.29	0.16	0.47	0.35	0.14
Falcon-7B									
middle-12	0.46	0.03	0.16	0.54	-0.10	0.28	0.52	0.27	0.26
middle-8	0.50	0.10	0.16	0.60	-0.08	0.25	0.56	0.29	0.26
middle-4	0.56	0.23	0.17	0.61	0.09	0.05	0.60	0.44	0.37
middle	0.56	0.25	0.20	0.65	0.11	0.15	0.61	0.44	0.30
middle+4	0.54	0.20	0.18	0.64	0.10	0.16	0.62	0.44	0.25
middle+8	0.54	0.25	0.26	0.66	0.13	0.31	0.62	0.45	0.31
middle+12	0.52	0.21	0.22	0.63	0.10	0.17	0.59	0.45	0.28
last	0.51	0.16	0.10	0.61	0.05	-0.01	0.57	0.39	0.22
Qwen-2.5-3B-IT									
middle-12	0.46	0.06	-0.00	0.55	0.04	-0.04	0.49	0.32	0.24
middle-8	0.48	0.09	0.01	0.56	0.17	-0.19	0.49	0.28	0.14
middle-4	0.54	0.14	0.08	0.61	0.19	-0.01	0.54	0.36	0.20
middle	0.61	0.09	0.03	0.70	0.11	-0.18	0.62	0.36	0.12
middle+4	0.64	0.21	0.13	0.72	0.20	0.12	0.65	0.45	0.25
middle+8	0.67	0.34	0.20	0.76	0.37	-0.01	0.68	0.55	0.29
middle+12	0.65	0.23	0.07	0.74	0.23	-0.06	0.65	0.45	0.19
last	0.59	0.22	0.11	0.68	0.17	-0.02	0.60	0.44	0.20
Average (all models)									
middle-12	0.42	0.06	0.08	0.46	0.02	0.09	0.46	0.29	0.24
middle-8	0.49	0.13	0.11	0.53	0.10	0.05	0.53	0.30	0.24
middle-4	0.56	0.22	0.15	0.59	0.19	0.12	0.58	0.40	0.27
middle	0.59	0.27	0.20	0.63	0.23	0.13	0.62	0.46	0.30
middle+4	0.58	0.28	0.20	0.61	0.24	0.17	0.61	0.46	0.32
middle+8	0.57	0.29	0.17	0.61	0.25	0.15	0.61	0.46	0.28
middle+12	0.55	0.26	0.11	0.58	0.21	0.09	0.59	0.43	0.24
last	0.53	0.21	0.05	0.57	0.17	0.02	0.57	0.40	0.18

Table 26: PRR results (higher is better) by representation layer across three correctness functions and six LLMs, evaluated on ProbeDriftLight. Bold indicates the best value per column within each model section. We consider the following models: Llama-3.1-8B (Grattafiori et al., 2024), Llama 3.1-8B-IT, Gemma-2-9B (Team et al., 2024), OLMo-2-7B (OLMo et al., 2025), Falcon-7B (Almazrouei et al., 2023) and Qwen-2.5-3B-IT (Qwen et al., 2025).

Model	Method	ID	LOO	DiffTask
Llama-3.1-8B	Avg response tokens	0.59	0.36	0.30
	Last response token	0.58	0.36	0.16
Llama-3.1-8B-IT	Avg response tokens	0.64	0.29	0.11
	Last response token	0.61	0.21	-0.06
Gemma-2-9B	Avg response tokens	0.58	0.22	0.20
	Last response token	0.57	0.25	0.02
OLMo-2-7B	Avg response tokens	0.58	0.42	0.37
	Last response token	0.53	0.40	0.28
Falcon-7B	Avg response tokens	0.56	0.25	0.20
	Last response token	0.56	0.26	0.19
Qwen-2.5-3B-IT	Avg response tokens	0.61	0.09	0.03
	Last response token	0.56	0.17	-0.09
Average	Avg response tokens	0.59	0.27	0.20
	Last response token	0.57	0.27	0.08

Table 27: PRR results for SAPLMA using AlignScore, comparing average vs last response token aggregation, averaged across SciQ, TriviaQA, and PubmedQA. We consider the following models: Llama-3.1-8B (Grattafiori et al., 2024), Llama 3.1-8B-IT, Gemma-2-9B (Team et al., 2024), OLMo-2-7B (OLMo et al., 2025), Falcon-7B (Almazrouei et al., 2023) and Qwen-2.5-3B-IT (Qwen et al., 2025).

Aggregation Method	ID	LOO	DiffTask
AlignScore			
Average of response tokens	0.59	0.36	0.30
Last response token	0.58	0.36	0.16
Random response token	0.41	0.27	0.24
Average of all tokens	0.38	0.05	0.17
Average of context tokens	0.32	0.07	0.16
Last context token	0.46	0.25	0.13
ROUGE-L			
Average of response tokens	0.65	0.32	0.24
Last response token	0.63	0.35	0.21
Random response token	0.46	0.22	0.16
Average of all tokens	0.47	0.02	0.04
Average of context tokens	0.29	0.03	0.05
Last context token	0.42	0.19	0.02
Accuracy			
Average of response tokens	0.68	0.55	0.38
Last response token	0.66	0.51	0.33
Random response token	0.54	0.45	0.32
Average of all tokens	0.51	0.22	0.19
Average of context tokens	0.37	0.20	0.20
Last context token	0.52	0.36	0.18
LLM-as-a-judge			
Average of response tokens	0.61	0.51	0.35
Last response token	0.56	0.50	0.17
Random response token	0.45	0.35	0.28
Average of all tokens	0.40	0.19	0.36
Average of context tokens	0.36	0.15	0.33
Last context token	0.50	0.43	0.23

Table 28: PRR results (higher is better) for SAPLMA (middle) across token aggregation strategies and four evaluation metrics, for Llama-3.1-8B. Results are averaged across SciQ, TriviaQA, and PubmedQA.

	ID	LOO	DiffTask
Representation layer			
middle-12	0.35	-0.16	0.02
middle-8	0.36	-0.05	0.01
middle-4	0.42	-0.09	0.29
middle	0.39	-0.09	0.00
middle+4	0.38	-0.12	-0.02
middle+8	0.38	-0.12	0.07
middle+12	0.38	-0.13	0.06
last	0.33	-0.13	0.05
Token aggregation			
Avg. response tokens	0.39	-0.09	0.01
Last response token	0.38	-0.07	-0.02
Random response token	0.21	-0.12	-0.01
Avg. all tokens	0.22	-0.08	0.00
Avg. context tokens	0.12	-0.08	-0.01
Last context token	0.23	0.01	0.00
Probe architecture			
SAPLMA (4 layers)	0.39	-0.09	0.01
Uhead (1 layer)	0.36	-0.14	0.14
Uhead (2 layers)	0.34	-0.15	-0.04
SAPLMA (2 layers)	0.39	-0.15	0.10
SAPLMA (3 layers)	0.38	-0.11	0.01
SAPLMA (5 layers)	0.42	-0.09	0.07

Table 29: PRR results (higher is better) for PubmedQA (long-form), averaged across Llama-3.1-8B and Gemma-2-9B. Each section varies one design dimension while holding others fixed. The layer section uses the SAPLMA architecture with average response token aggregation; the aggregation section uses the middle layer; the architecture section uses the middle layer with average response token aggregation.

Method	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask
MSP	0.57	0.57	0.57	0.57	0.57
SAPLMA (mid)	0.63	0.46	0.26	0.35	0.32
HBO (paper version)	0.66	0.61	0.57	0.57	0.57
HBO (simplified)	0.63	0.55	0.57	0.57	0.57

Table 30: PRR results (higher is better) comparing the HBO configuration used in the paper against a simplified configuration ($UQ_{\text{usv}} = R$, $UQ_{\text{sv}} = 1 - R$), averaged across all three short-form datasets (SciQ, TriviaQA, CoQA) for Llama 3.1-8B.

Method	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask
MSP	0.58	0.58	0.58	0.58	0.58
SAPLMA (mid)	0.63	0.57	0.27	0.35	0.35
HBO (paper version)	0.68	0.64	0.58	0.58	0.58
HBO (simplified)	0.63	0.62	0.58	0.59	0.58

Table 31: PRR results (higher is better) comparing the HBO configuration used in the paper against a simplified configuration, for SciQ only – the dataset used as a validation set to select between the two configurations – for Llama 3.1-8B.

Method	ID	LOO	1D-SameTask	DiffTask	1D-DiffTask
MSP	0.57	0.57	0.57	0.57	0.57
SAPLMA (mid)	0.63	0.40	0.25	0.35	0.30
HBO (paper version)	0.66	0.59	0.57	0.57	0.57
HBO (simplified)	0.63	0.51	0.57	0.56	0.57

Table 32: PRR results (higher is better) comparing the HBO configuration used in the paper against a simplified configuration, averaged across the short-form datasets excluding SciQ (TriviaQA, CoQA) for Llama 3.1-8B.

Method	Degrees of OOD eval				Breadth of eval		Strong OOD baselines		Claims
	None	LOO	1D-SameTask	DiffTask	3+ Eval	Nat.	Unsup.	Sup.	+ve claim
Vazhentsev et al. (2025b)		✓			✓	✓	✓		✓
Shelmanov et al. (2025)			✓		✓		✓		✓
Obeso et al. (2025)			✓	✓	✓	✓	✓		✓
Chuang et al. (2024)			✓	✓	✓	✓		✓	✓
McKenzie et al. (2025)				✓	✓	✓		✓	✓
He et al. (2024)		✓			✓			✓	✓
Kapoor et al. (2024)		✓		✓		✓			✓
Vazhentsev et al. (2025a)		✓		✓	✓	✓	✓		
Orgad et al. (2025)			✓	✓	✓	✓	✓		
Tong et al. (2025)			✓		✓	✓	✓		
Kossen et al. (2024)		✓			✓	✓	✓	✓	
Bürger et al. (2024)		✓		✓	✓			✓	
Zhang et al. (2025)			✓	✓	✓	✓			
Huang et al. (2025)	✓								
Su et al. (2024)	✓								
Bhattacharjya et al. (2025)	✓								
Ours (HBO)		✓	✓	✓	✓	✓	✓	✓	✓

Table 33: Supervised UQ work which includes: 1) no OOD experimentation, 2) testing robustness in a leave-one-out setting (**LOO**), 3) training on a single dataset for the same task (**SameTask**), 4) training only on different tasks (**DiffTask**), 5) containing OOD experimentation beyond synthetically generated data (**Nat.**), 6) testing on 3 or more OOD datasets (**3+ eval**), 7) using a logit-based baseline such as MSP or entropy (**Unsup.**), and 8) including a strong supervised baseline where linear or MLP probes are trained only on the middle hidden state of an LLM (**Sup.**). **+ve claim** refers to methods that on average outperform each of the baselines they were compared to when tested OOD.