

UNIFYING LATENT UNCERTAINTY SIGNALS IN LARGE LANGUAGE MODELS FOR IMPROVED FACTUAL PRECISION

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) have emerged as powerful tools for knowledge-intensive tasks, yet their tendency to generate factually incorrect or misleading outputs—commonly referred to as hallucinations—poses a fundamental challenge to their reliability. While uncertainty estimation is critical for mitigating such errors, LLMs are not explicitly trained to represent or express uncertainty. In this work, we investigate whether and how uncertainty is implicitly encoded within pretrained models. Through a probing-based analysis, we demonstrate that LLMs internalize multiple distinct and dataset-specific uncertainty signals, which can be extracted as linear directions in their latent space. These signals are most pronounced in intermediate layers, exhibit limited cross-task generalization, and are substantially enhanced by instruction tuning and [IDK]-token training. Building on these findings, we propose Linear Uncertainty Alignment, a novel framework that leverages a unified uncertainty direction to train LLMs to classify their own correctness. Our experiments show that this significantly improves factual precision and reduces hallucination rates under zero-shot evaluation. Together, these results provide new insights into the internal structure of uncertainty in LLMs and introduce a practical method for aligning models toward more trustworthy behavior.

1 INTRODUCTION

Large Language Models (LLMs) are trained on vast corpora of text data (Brown et al., 2020; Raffel et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Le Scao et al., 2023; Jiang et al., 2023a), enabling them to comprehend and generate human language. These training datasets encompass a wide range of written human knowledge, including books, news articles, Wikipedia, and scientific publications. Through this extensive pretraining, LLMs retain significant portions of the information they are exposed to, effectively embedding real-world knowledge within their parameters such that they are able to serve as knowledge repositories (Petrone et al., 2019; Roberts et al., 2020; Cohen et al., 2023a; Pan et al., 2023). This capability allows LLMs to be leveraged in tasks that depend on such knowledge, such as closed-book question answering (Brown et al., 2020; Roberts et al., 2020) and information retrieval (Tay et al., 2022).

Despite their widespread adoption, LLMs are widely known to suffer from ‘hallucinations’—a predisposition towards producing outputs that are false or misleading—which significantly undermines their accuracy and trustworthiness (Ji et al., 2023; Manduchi et al., 2024). Hallucinations may manifest in various forms, including factually incorrect statements (Maynez et al., 2020; Devaraj et al., 2022; Tam et al., 2023), internal inconsistencies (Elazar et al., 2021; Mündler et al., 2023), contradictions (Cohen et al., 2024a), or statements lacking clear sources or attribution (Bohnet et al., 2022; Rashkin et al., 2023; Yue et al., 2023).

Uncertainty, however, is a concept that LLMs are not generally known to capture (Yin et al., 2023; Kapoor et al., 2024). At the very least, they are generally not explicitly trained on it. This lack of competency regarding uncertainty, however, often results in misinformation generation, which can be harmful and misleading (Maynez et al., 2020; Devaraj et al., 2022; Tam et al., 2023), as LLMs have a hard time expressing a lack of knowledge both verbally and through their output distribution.

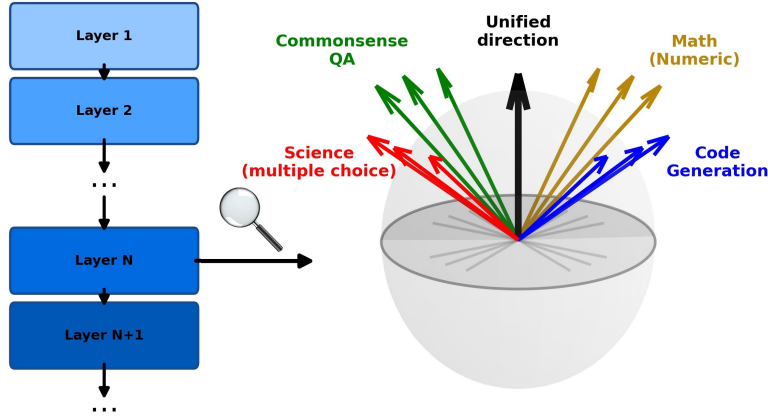


Figure 1: Multiple data-specific linear uncertainty vectors identified at the end of each transformer layer. Vectors are near-orthogonal across topics and align within a topic; a unified vector correlates positively with all.

Some more advanced methods such as instruction tuning (Ouyang et al., 2022; Zhang et al., 2023) during post-training and [IDK] tuning (Cohen et al., 2024b) during pretraining aim, inter alia, to align LLMs to more efficiently express their uncertainty and refrain from misinformation generation. While instruction tuning more generally aligns LLMs with human intent by fine-tuning them on task-specific instructions and corresponding outputs, the model is often also encouraged to refrain from answering questions when the specific answer is not known to it.

In this work, we first propose an analysis mechanism to study the uncertainty captured by large language models (LLMs). Using this mechanism, we show that LLMs internalize a notion of uncertainty during pretraining, which can be extracted via linear probes from their latent representations. Specifically, we identify linear uncertainty vectors—directions in hidden space—that correlate with generation correctness across multiple models and datasets, even without further weight training. This suggests that uncertainty is a learnable, linearly separable concept within LLMs. Our analysis further shows that LLMs do not encode a single unified notion of uncertainty, but rather multiple distinct vectors tied to different datasets or knowledge types. While these vectors are often nearly independent, partial transfer exists—e.g., across mathematics benchmarks. These findings suggest new ways of mitigating hallucination, since inconsistencies between uncertainty signals may underlie unreliable outputs.

Building on these insights, we introduce *Linear Uncertainty Alignment*, a method for aligning models with their internal uncertainty signals by training them to exploit the identified classifiers to predict their own correctness. This alignment substantially improves factual precision in zero-shot evaluation and reinforces that the linear directions correspond to meaningful representations of uncertainty.

In summary, our contributions are twofold: (1) we provide a systematic analysis showing that LLMs implicitly encode multiple forms of uncertainty that can be isolated through linear probes (Section 2), and (2) we present a training framework that leverages these signals to improve factual precision and reduce hallucination (Section 3).

2 IDENTIFYING UNCERTAINTY PREDICTORS

2.1 FRAMEWORK

In this work, we assume that uncertainty is represented within an LLM’s latent space across layers. Specifically, let $\mathbf{h}_i(x)$ denote the hidden state at the end of the i -th layer, taken at the last vector (created from the last input token), given input prompt x . For each hidden state, we search for a linear vector $\mathbf{u}_i$ such that the classifier $C(x, i) = \mathbf{u}_i^\top \mathbf{h}_i(x) + b_i$ predicts the model’s subsequent answer correctness with accuracy significantly above chance. Intuitively, this identifies a linear concept encoding the model’s uncertainty about its own outputs.

Model	OpenBookQA	PopQA	Qampari	ROMQA	SVAMP	StrategyQA	TriviaQA	TruthfulQA
Llama-3.2-1B	0.534	0.857	0.634	0.750	0.729	0.689	0.716	0.737
Llama-3.2-3B	0.590	0.793	0.734	0.583	0.750	0.608	0.742	0.600
Llama-3.1-8B	0.644	0.757	0.630	0.763	0.711	0.684	0.757	0.722
Llama-3.1-8B-Instruct	0.694	0.768	0.679	0.750	0.767	0.639	0.776	0.719
Mistral-7B-v0.1	0.597	0.747	0.727	0.750	0.687	0.643	0.760	0.673
IDK-tuned-Mistral-7B-v0.1	0.611	0.829	0.789	0.667	0.628	0.547	0.693	0.725
Qwen2.5-7B	0.678	0.817	0.697	0.615	0.696	0.698	0.717	0.678
Qwen3-14B	0.743	0.833	0.630	0.596	0.789	0.561	0.782	0.699
Qwen3-14B-Instruct	0.619	0.771	0.861	0.655	0.767	0.711	0.756	0.726

Table 1: Correctness prediction accuracy across 8 of the evaluation datasets (see Table 5 for additional results).

Linear Uncertainty Search Let M be a language model and $\mathcal{D} = \{(q_j, a_j)\}_{j=1}^n$ a dataset of questions and answers. To find $\mathbf{u}_i$ for layer i , we train a linear classifier to predict whether M ’s answer to q_j is correct. Using a training split $\mathcal{D}_{\text{TRAIN}} = \{(q_j, a_j)\}_{j=1}^m$, $m < n$, we consider the model’s predictions and define labels as

$$L(q_j) = \begin{cases} 1 & \text{if } M(q_j) = a_j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

This yields $\hat{\mathcal{D}}_{\text{TRAIN}} = \{(q_j, L(q_j))\}_{j=1}^m$. A classifier is then trained on the hidden states $\mathbf{h}_i(q_j)$ to predict $L(q_j)$. With a linear classifier, this corresponds to identifying a direction $\mathbf{u}_i(\mathcal{D})$ in the latent space—referred to as the uncertainty direction for dataset $\mathcal{D}$ —along with its learned bias b_i .

Uncertainty Vector as a Predictor We evaluate the uncertainty predictor $\mathbf{u}_i(\mathcal{D})$ on held-out test sets from $\mathcal{D}$. Given input x , $\mathbf{h}_i(x)$ is the hidden state at layer i , and the classifier

$$C_{\mathbf{u}_i(\mathcal{D})}(x) = \begin{cases} \text{INCORRECT} & \text{if } [\mathbf{u}_i(\mathcal{D})]^\top \mathbf{h}_i(x) + b_i > 0 \\ \text{CORRECT} & \text{otherwise} \end{cases} \quad (2)$$

predicts whether M ’s next-token choice is correct. Comparing predictions against the ground truth allows us to compute accuracy, precision, recall, and related metrics.

2.2 EXPERIMENTAL SETUP FOR ANALYSIS

To evaluate and utilize our uncertainty identification framework, we consider a series of experiments, for which we first introduce the experimental setup.

Foundation Models. In order to reach general conclusions that are not specific to any particular LLM, in this work we study three different families of models: the Llama family of models (Touvron et al., 2023; Dubey et al., 2024), Mistral (Jiang et al., 2023b), and Qwen (Bai et al., 2023; Yang et al., 2024). Specifically, for Llama, we study Llama-3.2-1B, Llama-3.2-3B, and Llama-3.1-8B, for Mistral, we consider Mistral-7B-v0.1, and finally for Qwen, we rely on Qwen2.5-7B and Qwen3-14B.

Advanced Models. For evaluating the effects of different types of training on the linear uncertainty encodings, we exploit three particular additional models in our experiments. To capture the effect of instruction tuning (Ouyang et al., 2022; Zhang et al., 2023), we use Llama-3.1-8B-Instruct and Qwen3-14B-Instruct, which both were post-trained using instruction tuning. Furthermore, we follow Cohen et al. (2024b) and use the IDK-tuned-Mistral-7B-v0.1 model in our experiments to evaluate the effect of [IDK] tuning—a method that essentially adds a new uncertainty token to the model’s vocabulary and teaches the model to use it during pretraining by adapting its loss to consider the new token.

Datasets and Benchmarks. We utilize 16 QA datasets and benchmarks in both our linear uncertainty search (Section 2.1) and the induced classifier evaluation (Section 2.1). We group them into six thematic categories:

- **Commonsense QA:** *CommonsenseQA* (Talmor et al., 2019), *StrategyQA* (Geva et al., 2021a). These include questions that assess the model’s ability to apply everyday reasoning and background knowledge to answer questions beyond surface-level facts.
- **Fact-Lookup and Adversarial QA:** *GranolaEntityQuestions* (Yona et al., 2024), *Natural Questions* (Kwiatkowski et al., 2019), *PopQA* (Mallen et al., 2022), *TriviaQA* (Joshi et al., 2017), *TruthfulQA* (Lin et al., 2021). These consist of questions that test the model’s factual recall and resilience to misleading or adversarial question phrasing.
- **List-Output QA:** *QAMPARI* (Amouyal et al., 2023), *RoMQA* (Zhong et al., 2022). Both evaluate whether models can produce comprehensive sets of correct answers, challenging their ability to recall multiple relevant facts simultaneously
- **Science QA (K-12):** *ARC-Easy* (Clark et al., 2018), *OpenBookQA* (Mihaylov et al., 2018). These focus on elementary school and high-school level science, requiring models to combine factual knowledge with basic reasoning.
- **Math Word Problems:** *GSM8K* (Cobbe et al., 2021), *ASDiv-A* (Miao et al., 2020), *SVAMP* (Patel et al., 2021). These include queries that test models on arithmetic and algebraic reasoning through natural language mathematical problems.
- **Code Generation:** *HumanEval-X* (Zheng et al., 2023), *MBPP* (Austin et al., 2021). We use these datasets to evaluate the ability of models to generate correct and functional software code given natural language programming prompts.

Notably, for each of these, we create a fixed training split, which will be used to derive our uncertainty vectors, and a test split, which will be used to evaluate their performance.

Linear Uncertainty Search Details. For every model M , transformer layer i , and evaluation dataset $\mathcal{D}$, we fit a logistic regression probe on the hidden states $\mathbf{h}_i(x)$ and obtain a single weight vector,

$$\mathbf{u}_i(\mathcal{D}),$$

which serves as the *linear uncertainty direction* for that (layer, dataset) pair.

To obtain a dataset-agnostic baseline, we also train an additional probe on the **concatenation of all datasets**. The resulting vector is denoted as

$$\mathbf{u}_i(\mathcal{D}_{\text{UNIFIED}}).$$

Evaluation. We evaluate the ability of our identified uncertainty linear vectors to predict the correctness of the model’s generation. For this, we consider the following metrics: (i) **Accuracy:** the ratio of correct predictions by the classifier that is induced by the uncertainty linear vector, (ii) **Precision:** the ratio of actually wrong completions by the model among those that the induced classifier predicted to be wrong.

2.3 ANALYSIS RESULTS

A Linear Representation of Uncertainty is Learned during Pretraining. Tables 1 and 5 report the performance of correctness classifiers, derived from linear uncertainty vectors, across models and datasets (best-performing layers only; full layer-wise analysis in a later section). Despite keeping model weights frozen, we identify linear directions in latent space that yield meaningful correctness predictions, with accuracy well above the 0.5 random baseline. The strongest signals consistently emerge from *upper-intermediate* layers (roughly 2/3 depth), while performance declines in final layers (Figures 8 and 9). This concentration suggests that uncertainty becomes most explicit once knowledge has been consolidated but before it is fully transformed into generation-specific representations. These findings provide strong evidence that uncertainty is encoded in a linearly separable form, concentrated in the middle of the network.

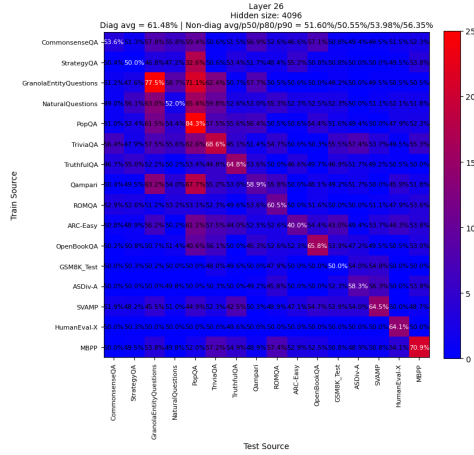


Figure 2: Correctness prediction accuracy results of the classifier induced by $u_{26}(\mathcal{D})$, for datasets $\mathcal{D}$ given on the y -axis, using Llama-3.1-8B, while testing on the test set for datasets given on the x -axis.

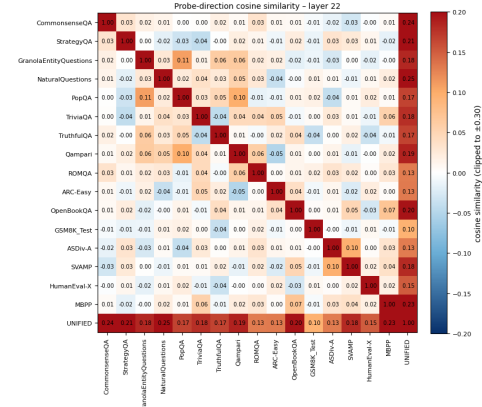


Figure 3: Cosine similarity results across all linear uncertainty vectors at layer number 22 of Llama-3.1-8B

LLMs Learn Multiple Different Linear Uncertainty Vectors. Linear uncertainty vectors appear across layers but are typically dataset-specific. A classifier trained on $\mathcal{D}_1$ often fails on a different $\mathcal{D}_2$ ($\mathcal{D}_1 \neq \mathcal{D}_2$), and cosine similarity between vectors is often near-zero (Figure 3). Nonetheless, exceptions exist: in domains like **Math Word Problems**, classifiers transfer effectively across datasets such as GSM8K, ASDiv, and SVAMP (Figure 7), sometimes matching or exceeding in-domain accuracy. Thus, while uncertainty is largely dataset-specific, certain domains exhibit shared structures that support generalization.

A Unified Uncertainty Axis with Universal Positive Alignment. The *unified* probe $u_i(\mathcal{D}_{\text{UNIFIED}})$, trained on all datasets, aligns positively with each dataset-specific vector, despite the near-orthogonality of those vectors. One would ordinarily anticipate a mix of positive and negative associations, so this universal positivity reveals a shared component across tasks. Conceptually, this indicates a generalized axis of uncertainty (Figure 1), offering both theoretical insight and a methodological tool for surfacing common uncertainty structure.

Fine-Tuning, Not Scale, Enhances Uncertainty Representation. Scaling does little to improve uncertainty prediction: larger Llama models perform only marginally better than smaller ones (Figure 9). In contrast, specialized training substantially boosts performance. Instruction tuning improves the accuracy and shifts peak performance to earlier layers, while [IDK] tuning not only increases accuracy but also reduces overconfidence, yielding more reliable predictors. Both enhance cross-dataset transfer, underscoring that fine-tuning—not scale—is key to strengthening uncertainty representations. These results suggest that tuning strategies can explicitly guide models to align latent uncertainty signals with observable correctness, whereas sheer model size does not.

3 LINEAR UNCERTAINTY ALIGNMENT

As a consequence of our previous analysis (Section 2), we propose a novel tuning mechanism designed to explicitly align language models with their own internal uncertainty representations. The key idea is to leverage the dataset-specific uncertainty vectors $u_i(\mathcal{D})$ identified in Section 2.1 and integrate them into the training objective. By doing so, the model is encouraged not only to predict correct answers but also to recognize and classify the correctness of its own generations.

3.1 DUAL-OBJECTIVE TRAINING

Our method tunes the model with two complementary objectives:

Standard Cross-Entropy Loss. As in conventional fine-tuning, the model is trained to maximize the likelihood of gold answers. Given a dataset $\mathcal{D} = \{(q_j, a_j)\}_{j=1}^n$, the cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n} \sum_{j=1}^n \log P_{\theta}(a_j | q_j), \quad (3)$$

where $P_{\theta}(a_j | q_j)$ denotes the probability assigned by the model with parameters θ to the correct answer a_j given question q_j .

Uncertainty Classification Loss. To align the model with its own uncertainty, we introduce a secondary objective that encourages hidden states to reflect correctness predictions. For each input q_j , let $\mathbf{h}_i(q_j)$ denote the hidden state of layer i , and let $L(q_j) \in \{0, 1\}$ be the correctness label defined in Equation 1. Using the fixed uncertainty vector $\mathbf{u}_i(\mathcal{D})$ and bias b_i , the predicted correctness is given by

$$\hat{L}(q_j) = \sigma(\mathbf{u}_i(\mathcal{D})^{\top} \mathbf{h}_i(q_j) + b_i), \quad (4)$$

where $\sigma(\cdot)$ is the logistic sigmoid function. The corresponding classification loss is

$$\mathcal{L}_{\text{U}} = -\frac{1}{n} \sum_{j=1}^n \left[L(q_j) \log \hat{L}(q_j) + (1 - L(q_j)) \log(1 - \hat{L}(q_j)) \right]. \quad (5)$$

3.2 UNIFIED OBJECTIVE

The final training objective combines both components as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{U}}, \quad (6)$$

where λ is a tunable hyperparameter controlling the trade-off between improving factual correctness and aligning with the model’s internal uncertainty representation.

This dual-objective formulation encourages the model to (a) produce accurate answers where possible, while (b) learning to reflect and calibrate its uncertainty in line with the fixed, data-derived uncertainty vectors. As we show in Section 4, models tuned with this method exhibit improved factual precision and reduced hallucination rates under zero-shot evaluation.

4 EXPERIMENTS AND RESULTS

We proceed by detailing the setup underlying the experimental evaluation of our proposed *Linear Uncertainty Alignment* method, followed by a discussion of the corresponding results and conclusions.

4.1 EXPERIMENTAL SETUP

Our experimental design builds directly on the models, datasets, and evaluation metrics introduced in Section 2. Here, we describe the details specific to assessing our alignment method.

Instruction-Based Evaluation Protocol. To probe both factual precision and calibrated abstention behavior, we employ an instruction-driven evaluation format. The instruction is designed to explicitly invite models to abstain when uncertain, thereby testing their ability to leverage internal uncertainty representations. Specifically, each test instance is presented in the form:

“Please answer the following question. Please answer with *I don’t know the answer* in cases where you’re not certain in your answer. The question is: ...”

Dataset	Llama-3.2-1B						Llama-3.2-3B						Llama-3.1-8B					
	Tuned			+Aligned			Tuned			+Aligned			Tuned			+Aligned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CommonsenseQA	40.5	25.0	30.9	50.7	23.9	32.5	38.9	28.7	33.0	50.5	28.1	36.2	45.3	34.0	38.9	57.2	32.7	41.6
Natural Questions	27.0	18.9	22.2	36.1	18.1	24.1	28.5	21.7	24.7	36.9	20.4	26.0	31.7	25.5	28.3	40.4	24.0	30.1
TriviaQA	56.2	47.0	51.2	65.9	46.2	54.4	58.5	52.7	55.5	66.7	50.9	57.8	59.8	58.9	59.3	68.1	57.8	62.5
PopQA	45.1	32.0	37.4	55.0	30.8	39.6	45.7	36.2	40.4	56.9	34.9	43.2	48.4	42.5	45.3	60.1	39.9	48.0
TruthfulQA	35.6	27.9	31.3	44.8	26.8	33.4	35.5	31.6	33.4	45.0	30.1	36.1	39.6	35.0	37.1	49.8	34.2	40.7
GSM8K	46.9	39.7	43.0	54.7	39.2	45.8	48.6	42.5	45.4	56.3	41.6	47.9	53.0	46.8	49.7	61.8	46.3	53.0
Average	41.9	31.7	36.0	51.2	30.8	38.3	42.6	35.6	38.7	52.0	34.3	41.2	46.3	40.5	43.1	56.2	39.2	46.0

Table 2: Comparison of precision (P), recall (R), and F1-score between tuned and uncertainty-aligned LLaMA models across evaluation datasets. For each dataset and metric, the better result is highlighted in bold.

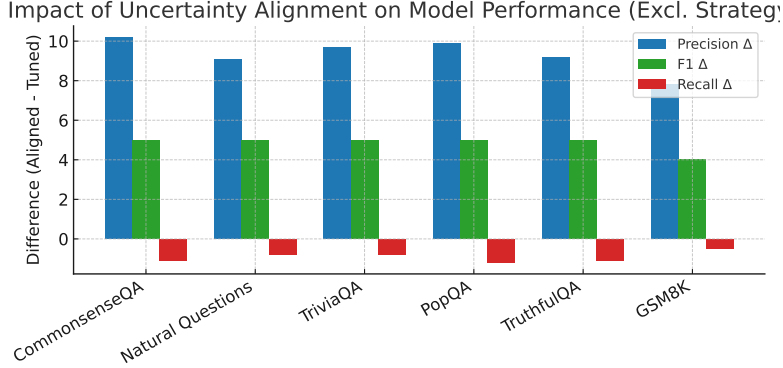


Figure 4: Comparison of precision, recall, and F1-score between tuned and uncertainty-aligned LLaMA models (Llama-3.2-1B, Llama-3.2-3B, Llama-3.1-8B). Alignment consistently improves **precision** and **F1**, with only a minor decrease in recall.

Tuning Variants. For each foundation model considered, we fine-tune the model under the dual-objective formulation from Section 3. The tuning is conducted separately with respect to:

1. each layer-specific uncertainty vector $\mathbf{u}_i(\mathcal{D})$ identified in Section 2;
2. the unified uncertainty direction $\mathbf{u}_{\text{uni}}$ obtained by aggregating datasets (Section 2.3).

These two variants allow us to directly compare the impact of alignment to individual vectors versus the global, unified representation.

Evaluation. Each tuned model variant is evaluated across all benchmark datasets used in our earlier analysis, following the same correctness, precision, and hallucination metrics. We thus obtain a comprehensive picture of how uncertainty alignment impacts model behavior across diverse tasks. In particular, we measure improvements in factual precision, the reduction of spurious hallucinations, and the calibrated use of abstentions (*I don't know*) as a function of the tuning vector employed.

4.2 EVALUATION RESULTS

Per-Dataset Evaluation. Our first evaluation examines the effect of the proposed method on factual precision in a dataset-specific setup. Concretely, for each dataset, we train a separate model using its training split and then evaluate performance on the corresponding held-out test split. Importantly, we choose the fixed uncertainty vector to be the one corresponding to the layer for which the corresponding uncertainty vector produces the best classification results out of all the model layers. Across all model sizes and datasets, we observe a consistent pattern: alignment markedly boosts precision, with average gains of nearly 10 points, while recall is only moderately affected (Tables 2 and 6, Figure 4). This balance results in substantial improvements in F1, confirming that the method not only reduces factual errors but also preserves the model's ability to retrieve correct answers. For example, on CommonsenseQA, precision improves from 45.3 to 57.2 for the 8B model, with recall remains at a competitive 32.7, leading to an F1 increase from 38.9 to 41.6. Similar trends are observed across Natural Questions, TriviaQA, PopQA, TruthfulQA, and GSM8K, where precision gains consistently outweigh the relatively small recall drops, yielding stronger overall F1 scores.

Dataset	LLaMA-1B						LLaMA-3B						LLaMA-8B					
	Tuned			+Aligned			Tuned			+Aligned			Tuned			+Aligned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CommonsenseQA	36.1	25.7	30.0	43.8	23.5	31.0	33.4	27.0	29.9	41.5	26.8	32.6	41.0	31.3	35.5	47.6	32.5	38.6
StrategyQA	35.0	20.3	25.7	32.2	24.1	27.7	34.6	26.5	30.0	36.0	24.9	29.5	33.9	29.0	31.3	38.0	25.1	30.3
Natural Questions	24.6	17.1	20.2	29.6	18.7	23.0	25.4	19.9	22.3	31.0	19.8	24.4	32.0	23.2	26.9	34.4	24.8	28.8
TriviaQA	56.0	46.1	50.6	60.7	44.9	51.6	58.3	53.8	56.0	61.1	50.3	56.0	58.3	58.9	58.6	65.0	56.1	60.4
PopQA	47.3	32.3	38.4	48.2	30.1	37.5	43.9	33.9	38.3	48.9	34.3	40.3	50.2	43.6	46.7	51.4	40.0	45.0
TruthfulQA	33.7	28.2	30.7	38.1	26.0	31.1	37.3	32.5	34.7	38.7	29.4	33.4	39.1	33.9	36.3	42.5	32.5	36.6
GSM8K	41.7	39.9	40.8	50.4	38.0	43.6	45.3	41.7	43.4	52.8	40.6	46.0	50.4	38.0	43.3	57.0	44.5	50.1

Table 3: Performance comparison between the baseline tuned LLaMA models and our **Aligned** variants trained with the unified uncertainty vector as the classification signal in the auxiliary loss. Results are reported on all evaluation datasets in terms of precision (P), recall (R), and F1-score. For each dataset and metric, the better score between **Tuned** and **+Aligned** is highlighted in **bold**.

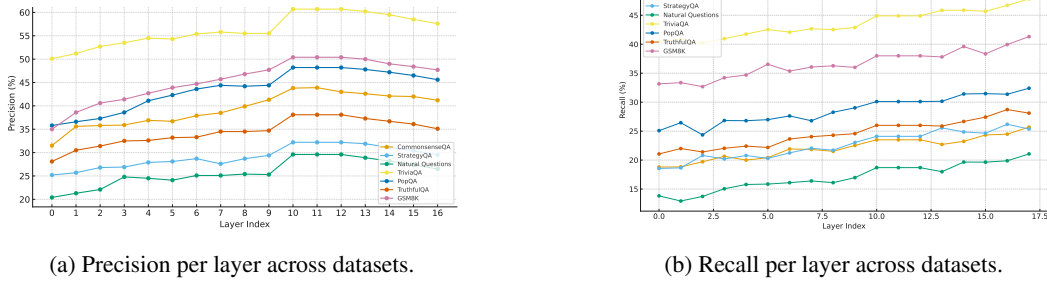


Figure 5: Layer-wise evaluation of LLaMA-3.2-1B models aligned using the different layer-specific uncertainty vectors. Results are shown across all datasets in terms of (a) precision and (b) recall. The curves demonstrate consistent improvements near the alignment layers, indicating that the extracted vectors capture meaningful uncertainty signals.

These findings provide strong empirical support for our central hypothesis: the linear uncertainty direction we identified indeed captures meaningful uncertainty in the model’s latent space. By aligning generations along this axis, the model becomes more calibrated with respect to factual correctness, which manifests as higher precision and better overall reliability.

Unified Vector results. In contrast to our above per-dataset alignment experiments, where a separate model was tuned for each dataset-specific uncertainty vector, here we trained a single model by leveraging the unified uncertainty vector as the auxiliary classification signal. For both the baseline tuned models and our aligned variants, the training set was constructed by randomly sampling 10% of each dataset’s original training split used in the previous experiments, ensuring a fair and consistent comparison. As shown in Tables 3 and 7, the unified alignment consistently improves precision and F1 scores relative to the tuned baseline across nearly all datasets, while recall remains competitive. Importantly, the fact that a single unified vector—learned once and applied across heterogeneous tasks—leads to systematic gains provides strong evidence that this vector indeed captures and represents the general concept of uncertainty within the model’s latent space.

4.3 FURTHER ANALYSIS

Layer Analysis. Figure 5 shows per-layer trends in precision and recall across all datasets. Performance is consistently lowest at the embedding layer (layer 0), increases through intermediate layers, and peaks around layers 10–12. Beyond this range, precision and recall either plateau or decline slightly, suggesting that mid-layer representations capture the most informative uncertainty signal for alignment. These findings are compatible with our earlier analysis results, further reinforcing the conclusion that the unified uncertainty vector captures the underlying concept effectively (see Appendix C).

Uncertainty Classification Loss Only. In this ablation study, we trained models using only the auxiliary classification loss based on the uncertainty vector, while omitting the standard cross-entropy objective. We then compared their performance to our full +Aligned method across all datasets. The results show that removing cross-entropy leads to a substantial drop in recall, in some cases exceeding

Dataset	LLaMA-1B						LLaMA-3B						LLaMA-8B					
	+Aligned			Ablation			+Aligned			Ablation			+Aligned			Ablation		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CommonsenseQA	43.8	23.5	31.0	44.2	17.9	25.3	41.5	26.8	32.6	41.2	19.7	26.5	47.6	32.5	38.6	47.0	24.8	32.3
StrategyQA	32.2	24.1	27.7	33.0	16.0	21.2	36.0	24.9	29.5	35.6	17.8	23.2	38.0	25.1	30.3	37.7	19.4	25.2
Natural Questions	29.6	18.7	23.0	30.1	13.5	18.6	31.0	19.8	24.4	30.6	14.2	19.6	34.4	24.8	28.8	33.9	17.1	22.4
TriviaQA	60.7	44.9	51.6	61.2	32.5	40.6	61.1	50.3	56.0	60.9	36.4	45.2	65.0	56.1	60.4	64.7	41.2	49.9
PopQA	48.2	30.1	37.5	47.8	19.7	27.9	48.9	34.3	40.3	49.2	21.8	29.8	51.4	40.0	45.0	51.0	28.4	36.6
TruthfulQA	38.1	26.0	31.1	38.5	17.9	24.2	38.7	29.4	33.4	39.0	21.0	27.1	42.5	32.5	36.6	42.2	23.7	29.9
GSM8K	50.4	38.0	43.6	50.7	25.7	32.6	52.8	40.6	46.0	52.3	27.1	35.4	57.0	44.5	50.1	56.6	29.4	38.3

Table 4: Ablation study: comparison between our full **+Aligned** method (with cross-entropy and classification loss) and the **Ablation** variant trained only with the classification loss. Results are reported on all evaluation datasets in terms of precision (P), recall (R), and F1-score. The better score for each dataset and metric is highlighted in **bold**.

25%, while precision is occasionally comparable or slightly higher. However, the overall F1 scores consistently favor the full method, highlighting that the cross-entropy loss is crucial for maintaining balanced predictions and preventing the model from becoming overly conservative.

5 RELATED WORK

Model Calibration. Our analysis is closely related to the challenge of model calibration (Guo et al., 2017): providing a measure of the probability that a prediction is incorrect alongside the prediction itself. Factual error detection can be viewed as a variation of calibration, where instead of a continuous probability, we output a binary judgment of correctness. Common approaches include transformations of model logits (Desai & Durrett, 2020; Jiang et al., 2021) and uncertainty-based methods (e.g., see Kuhn et al., 2023). Recent work explores supervised calibration with LMs, using fine-tuning (Kadavath et al., 2022; Lin et al., 2022), in-context learning (Cohen et al., 2023a; Alivanistos et al., 2022), zero-shot instruction methods (Cohen et al., 2023b), and consistency sampling (Yoran et al., 2023). Other studies leverage internal states for certainty classification (Azaria & Mitchell, 2023), introduce special tokens for unanswerable inputs (Lu et al., 2022), or design datasets for refusal tuning (Zhang et al., 2024). Our work instead analyzes the dynamics of uncertainty encoding in pretrained and calibrated models.

Mechanistic Interpretability Recent work has been aiming to identify circuits and features within models that correspond to interpretable concepts such as factual recall, syntax, or positional reasoning (Olsson et al., 2022; Yu et al., 2023). For instance, tools such as SAE (Sparse Autoencoders) have been used to isolate human-interpretable features from residual stream activations (Meng et al., 2022). Other studies explore how knowledge is stored and manipulated across layers, such as tracing factual associations or memorized content to specific directions in the latent space (Geva et al., 2021b; Gurnee et al., 2023; Geva et al., 2023; Yu et al., 2024). Despite promising progress, full mechanistic understanding remains an open challenge due to the scale and complexity of modern models.

6 CONCLUSION

Our systematic analysis of uncertainty representation in Large Language Models reveals four key findings: (1) linear uncertainty representations are learned during pretraining, with probing classifiers achieving above-chance accuracy in predicting answer correctness, most pronounced in upper-intermediate layers around two-thirds of model depth; (2) models acquire multiple, dataset-specific uncertainty vectors that are largely orthogonal yet show meaningful cross-dataset generalization within task families; (3) despite this near-orthogonality, a unified uncertainty axis aligns positively with all dataset-specific directions, revealing a shared component across tasks; and (4) specialized training strategies—instruction tuning and [IDK] tuning—substantially enhance uncertainty representation, while scaling alone provides only marginal gains.

Building on the third finding, we introduce Linear Uncertainty Alignment, a training method that teaches models to classify their own correctness through dual objectives. This approach improves factual precision while maintaining competitive recall, showing that uncertainty is not only learnable and accessible but also deliberately alignable for building more trustworthy language models.

REFERENCES

- Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. Prompting as probing: Using language models for knowledge base construction. *arXiv preprint arXiv:2208.11057*, 2022.
- Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. QAMPARI: A benchmark for open-domain questions with many answers. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz (eds.), *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 97–110, Singapore, December 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.gem-1.9/>.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it’s lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL <https://aclanthology.org/2023.findings-emnlp.68>.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *ArXiv*, abs/2309.16609, 2023. URL <https://api.semanticscholar.org/CorpusID:263134555>.
- Bernd Bohnet, Vinh Q Tran, Pat Verga, Roei Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafford. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Roi Cohen, Mor Geva, Jonathan Berant, and Amir Globerson. Crawling the internal knowledge-base of language models. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1856–1869, Dubrovnik, Croatia, May 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.139. URL <https://aclanthology.org/2023.findings-eacl.139>.

- Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. LM vs LM: Detecting factual errors via cross examination. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12621–12640, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.778. URL <https://aclanthology.org/2023.emnlp-main.778>.
- Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024a.
- Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I Don’t Know: Explicit modeling of uncertainty with an [IDK] token. *Advances in Neural Information Processing Systems*, 37: 10935–10958, 2024b.
- Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.21. URL <https://aclanthology.org/2020.emnlp-main.21>.
- Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. Evaluating factuality in text simplification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7331–7345, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.506. URL <https://aclanthology.org/2022.acl-long.506/>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL <https://aclanthology.org/2021.tacl-1.60/>.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021a. doi: 10.1162/tacl_a_00370. URL <https://aclanthology.org/2021.tacl-1.21/>.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL <https://aclanthology.org/2021.emnlp-main.446/>.
- Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL <https://aclanthology.org/2023.emnlp-main.751/>.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/guo17a.html>.

- Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL <https://doi.org/10.1145/3571730>.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.
- Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L’elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023b. URL <https://api.semanticscholar.org/CorpusID:263830494>.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL <https://aclanthology.org/2021.tacl-1.57/>.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislaw Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don’t know. *arXiv preprint arXiv:2406.08391*, 2024.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL <https://aclanthology.org/Q19-1026/>.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2023.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.
- Stephanie C. Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL <https://api.semanticscholar.org/CorpusID:237532606>.
- Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. On controlling fallback responses for grounded dialogue generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2591–2601,

- Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.204. URL <https://aclanthology.org/2022.findings-acl.204/>.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.
- Laura Manduchi, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, Marius Kloft, Yingzhen Li, Christoph Lippert, Gerard de Melo, Eric Nalisnick, Björn Ommer, Rajesh Ranganath, Maja Rudolph, Karen Ullrich, Guy Van den Broeck, Julia E Vogt, Yixin Wang, Florian Wenzel, Frank Wood, Stephan Mandt, and Vincent Fortuin. On the challenges and opportunities in Generative AI. *arXiv preprint arXiv:2403.00025*, 2024. URL <https://arxiv.org/abs/2403.00025>.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL <https://aclanthology.org/2020.acl-main.173/>.
- Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=-h6WAS6eE4>.
- Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing English math word problem solvers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 975–984, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.92. URL <https://aclanthology.org/2020.acl-main.92/>.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.
- Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeljanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38, 2023. doi: 10.4230/TGDK.1.1.2. URL <https://drops.dagstuhl.de/entities/document/10.4230/TGDK.1.1.2>.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL <https://aclanthology.org/2021.naacl-main.168/>.

- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.
- Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421/>.
- Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through news summarization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5220–5255, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.322. URL <https://aclanthology.org/2023.findings-acl.322>.
- Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843, 2022.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don’t know? *arXiv preprint arXiv:2305.18153*, 2023.
- Gal Yona, Roei Aharoni, and Mor Geva. Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6737–6751, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.365. URL <https://aclanthology.org/2024.acl-long.365/>.
- Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. Answering questions by meta-reasoning over multiple chains of thought. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5942–5966, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.364. URL <https://aclanthology.org/2023.emnlp-main.364>.
- Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. *arXiv preprint arXiv:2403.18167*, 2024.

- Qinan Yu, Jack Merullo, and Ellie Pavlick. Characterizing mechanisms for factual recall in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9924–9959, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.615. URL <https://aclanthology.org/2023.emnlp-main.615/>.
- Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.307. URL <https://aclanthology.org/2023.findings-emnlp.307>.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say ‘I don’t know’. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394. URL <https://aclanthology.org/2024.naacl-long.394/>.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.
- Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, et al. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5673–5684, 2023.
- Victor Zhong, Weijia Shi, Wen tau Yih, and Luke Zettlemoyer. RoMQA: A benchmark for robust, multi-evidence, multi-answer question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2022. URL <https://api.semanticscholar.org/CorpusID:253116788>.

A LIMITATIONS

While our analysis provides compelling evidence for the existence of linearly accessible uncertainty representations in LLMs, it is limited to linear probes and does not explore more complex, nonlinear structures that may further explain model behavior. Our evaluation focuses on a fixed set of models and datasets, which, although diverse, may not capture the full variability seen in real-world applications or domain-specific tasks. Additionally, correctness is treated as a proxy for uncertainty, which may not fully align with how uncertainty manifests in open-ended or ambiguous generation scenarios. Finally, the performance of our classifiers may also be influenced by dataset-specific biases, potentially limiting generalizability.

B COMPUTATIONAL RESOURCES

In our experiments we use one NVIDIA A100 80G GPU.

C ADDITIONAL RESULTS

Here we provide additional noteworthy results from our experiments.

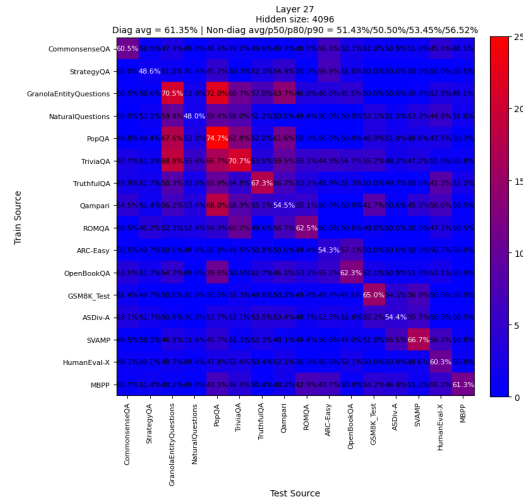


Figure 6: Correctness prediction accuracy results of the classifier induced by $u_{27}(\mathcal{D})$, for datasets $\mathcal{D}$ given on the y -axis, using Mistral-7B-v0.1, while testing on the test set for datasets given on the x -axis.

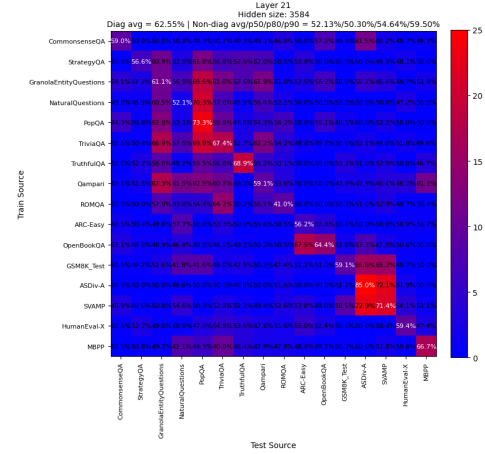
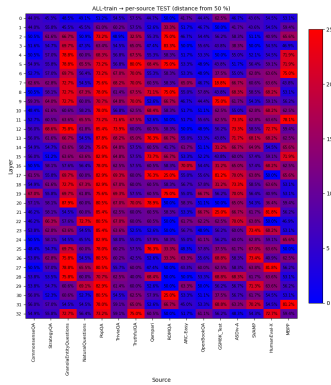


Figure 7: Correctness prediction accuracy results of the classifier induced by $u_{21}(\mathcal{D})$, for datasets $\mathcal{D}$ given on the y -axis, using Qwen2.5-7B, while testing on the test set for datasets given on the x -axis..



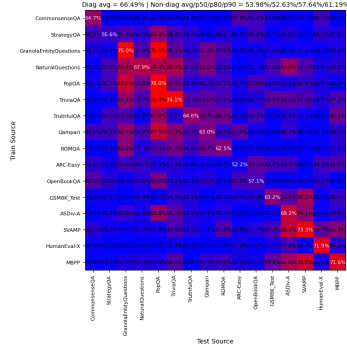


Figure 10: Correctness prediction accuracy results of the classifier induced by $u_{15}(\mathcal{D})$, for datasets $\mathcal{D}$ given on the y -axis, using Llama-3.1-8B-Instruct, while testing on the test set for datasets given on the x -axis.

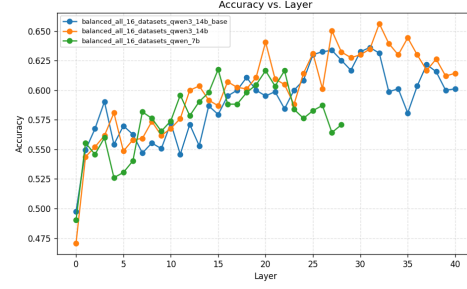


Figure 11: Correctness prediction accuracy averaged over all datasets of the induced classifier, considering the Qwen family: Qwen2.5-7B, Qwen3-14B, and Qwen3-14B-Instruct

Model	ARC-Easy	ASDiv-A	CommonsenseQA	GSM8K	GranolaEntityQuestions	HumanEval-X	MBPP	NaturalQuestions
Llama-3.2-1B	0.535	0.670	0.625	0.444	0.789	0.708	0.769	0.600
Llama-3.2-3B	0.710	0.648	0.598	0.688	0.790	0.732	0.641	0.675
Llama-3.1-8B	0.657	0.667	0.649	0.577	0.763	0.692	0.722	0.590
Llama-3.1-8B-Instruct	0.652	0.885	0.667	0.737	0.705	0.781	0.728	0.655
Mistral-7B-v0.1	0.657	0.691	0.709	0.550	0.782	0.707	0.707	0.630
IDK-tuned-Mistral-7B-v0.1	0.600	0.750	0.571	0.688	0.758	0.545	0.688	0.673
Qwen2.5-7B	0.750	0.800	0.718	0.682	0.704	0.578	0.648	0.750
Qwen3-14B	0.727	0.786	0.655	0.878	0.738	0.800	0.694	0.651
Qwen3-14B-Instruct	0.800	0.750	0.638	0.702	0.770	0.688	0.625	0.674

Table 5: Correctness prediction accuracy across the remaining datasets, complementing the results given in Table 1.

Dataset	LLaMA-1B						LLaMA-3B						LLaMA-8B					
	Tuned			+Aligned			Tuned			+Aligned			Tuned			+Aligned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GranolaEntityQ	33.1	22.4	26.7	34.7	21.0	26.6	32.8	25.1	28.5	35.2	24.3	28.9	36.0	28.7	32.0	38.1	27.1	31.7
QAMPARI	29.4	19.8	23.6	31.2	18.7	23.7	28.5	22.6	25.2	32.0	21.8	26.0	33.1	27.2	29.9	35.0	26.1	29.8
RoMQA	26.8	16.5	20.4	28.3	15.1	19.7	27.4	19.7	22.8	29.1	18.9	22.8	30.6	24.5	27.1	32.8	23.3	27.1
ARC-Easy	42.2	31.6	36.1	44.0	30.4	36.1	45.1	37.2	40.7	47.0	35.8	40.6	48.2	41.0	44.3	49.9	39.6	44.2
OpenBookQA	39.0	27.3	32.1	40.5	26.0	31.6	38.7	31.8	34.9	41.2	30.7	35.2	42.5	35.4	38.6	44.3	34.0	38.4
ASDiv-A	34.8	28.1	31.1	37.1	26.5	30.9	36.5	31.2	33.6	39.0	30.1	34.0	40.2	34.0	36.9	42.5	32.7	36.8
SVAMP	32.7	26.4	29.2	34.5	25.0	29.1	33.9	29.3	31.4	36.1	28.1	31.9	37.5	32.1	34.6	39.8	30.5	34.6
HumanEval-X	21.0	12.4	15.5	23.2	11.5	15.9	22.6	15.8	18.6	24.0	14.9	18.4	25.8	19.7	22.3	27.6	18.4	22.1
MBPP	26.5	18.1	21.5	27.3	17.2	21.3	27.1	20.6	23.5	28.8	19.8	23.6	29.9	24.0	26.6	31.5	22.7	26.4

Table 6: Performance comparison on the remaining benchmarks not included in Table 3. Consistent with earlier results, the **+Aligned** models typically improve recall and F1 across datasets, while tuned models sometimes retain slightly higher precision.

Dataset	LLaMA-1B						LLaMA-3B						LLaMA-8B					
	Tuned			+Aligned			Tuned			+Aligned			Tuned			+Aligned		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GranolaEntityQ	30.5	18.6	23.1	34.8	20.1	25.5	33.0	22.0	26.4	36.7	23.5	28.7	37.9	26.0	31.0	41.0	27.1	32.7
QAMPARI	25.2	14.7	18.6	28.5	15.9	20.4	27.1	17.5	21.3	30.2	19.0	23.2	30.5	22.0	25.6	34.1	23.5	27.9
RoMQA	21.0	13.2	16.2	24.7	14.0	18.0	22.4	15.9	18.6	26.3	16.8	20.6	27.0	18.3	21.9	30.8	19.6	23.9
ARC-Easy	43.2	34.5	38.4	46.0	33.1	38.7	45.0	36.8	40.5	49.1	37.2	42.3	50.3	41.0	45.2	53.0	42.1	46.9
OpenBookQA	39.5	26.3	31.6	41.7	27.5	33.0	40.9	28.0	33.2	44.3	29.6	35.4	46.0	33.1	38.6	48.5	34.4	40.3
ASDiv-A	37.8	29.4	33.1	42.2	30.7	35.6	39.4	31.8	35.1	44.0	33.0	37.9	45.5	36.5	40.5	49.0	37.6	42.5
SVAMP	34.0	25.0	29.0	39.1	26.5	31.5	35.8	28.2	31.6	41.3	29.7	34.7	43.0	33.0	37.4	47.2	34.5	39.9
HumanEval-X	20.5	12.0	15.1	23.8	13.1	17.0	21.7	13.7	16.7	25.5	14.9	19.2	26.1	16.5	20.2	29.0	17.6	21.8
MBPP	28.7	19.2	23.0	32.0	20.1	24.8	30.1	21.5	25.2	34.4	22.8	27.2	35.0	24.2	28.7	38.1	25.6	30.6

Table 7: Performance comparison on the additional datasets not shown in Table 3. Same setup as before, results are reported in terms of precision (P), recall (R), and F1-score, with better scores highlighted in **bold**.