LLM Beliefs Are in Their Heads

Alessandro Corona Mendozza* **Anders Søgaard**
University of Copenhagen

Abstract

We investigate belief-like representations in decoder-only autoregressive LLMs using linear controlled probes on residual stream activations and single attention heads. Following Herrmann and Levinstein's (2025) criteria (Accuracy, Use, Coherence, and Uniformity), we find that large models exhibit strong truth-sensitivity (Accuracy), and steering activations along probe directions reliably changes downstream behavior (Use). Coherence, measured via calibrated probes and cross-dataset probing, is moderate across models, while training on diverse data yields domain-consistent truth directions (Uniformity). The results are particularly encouraging at the head level and align with *some* standard philosophical accounts of belief, e.g., minimal functionalism, supporting the view that LLMs can maintain propositional attitudes under such theoretical frameworks.

1 Introduction

Large Language Model (LLM) intermediate representations have been shown to be truth-sensitive: they seem to linearly encode different statements according to their truth value (Burns et al., 2023). This finding is noteworthy, because our human notion of belief refers to the non-verbal attitude we take "whenever we take something to be the case or regard it as true" (Schwitzgebel, 2024). As a result, interpretability researchers and philosophers alike have begun to consider these representations a serious candidate for the role of LLMs' underlying beliefs (Christiano et al., 2021; Chalmers, 2025).

Recently, Herrmann and Levinstein (2025) operationalized a set of additional standards that representations should satisfy to count as underlying beliefs. In this study, we follow their lead by developing four different experiments on decoder-only autoregressive transformers from six families – and consider how they fare against these standards through probing (Hewitt and Liang, 2019; Belinkov, 2022), activation steering (Turner et al., 2024), model confidence estimation (Shorinwa et al., 2025), and cross-domain validation. We investigate all layers and show results on two different degrees of granularity by testing the residual stream and the single attention heads.

Our research questions are exploratory and can be articulated as follows:

- **(RQ1)**: To what extent do LLM internal representations of truth exhibit separability, causal involvement, implicit coherence, and cross-domain robustness?
- **(RQ2)**: Which architectural components (residual stream vs. attention heads) serve as the most effective substrate for activations with respect to the criteria in RQ1?
- **(RQ3)**: Under what conditions can these activations be interpreted as computational proxies for belief-like states?

We find that our models **(1)** consistently surpass baselines across all experiments, with larger and instructed models outperforming smaller and base models. Crucially, we find that **(2)** *attention heads* tend to show activations which are particularly well-separable, responsive to steering, and uniform across domains. Whereas (Levinstein and Herrmann, 2025; Bürger et al., 2024) report a failure of coherence, we observe moderate scores from linear probes trained on rich datasets. While our results are generally positive for the standards that we employed, we conclude by **(3)** clarifying the sense in which intermediate representations should be interpreted as proxies for beliefs: we argue that our findings provide non-trivial support for the more permissive functionalist theories of mental states (Lewis, 1972). We close our discussion by claiming

*Correspondence to: alessandro.corona.m@gmail.com
Code: https://github.com/Supersheep7/beliefs_llms

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41033–41067
July 2-7, 2026

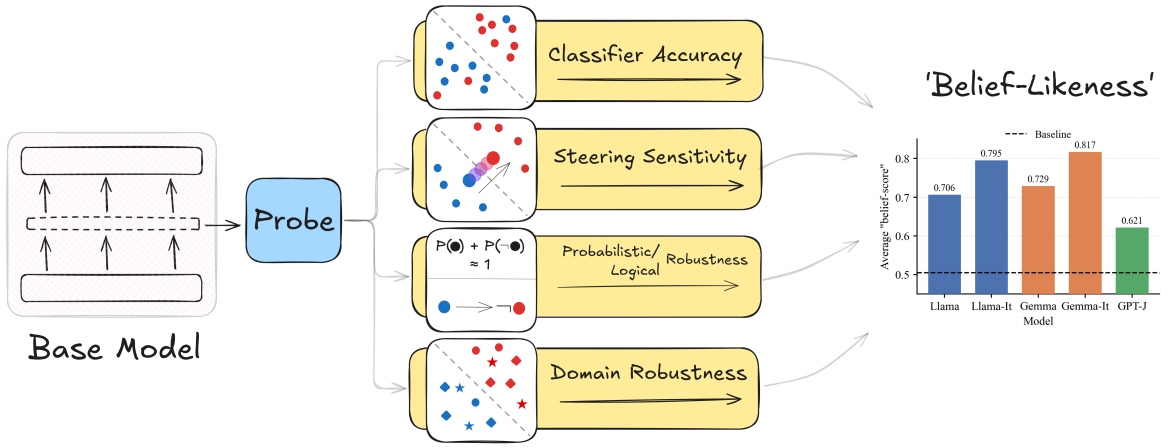


Figure 1: *Standards for Belief Representations* (Herrmann and Levinstein, 2025). To take the role of beliefs, intermediate activations should be distinctly separable by a classifier (*Accuracy*), causally implied in logit computation (*Use*), robust logically and probabilistically (*Coherence*), and uniform across different domains (*Uniformity*).

that, *if* we accept this liberal framework on cognition, activations that are truth-sensitive, minimally coherent, and robustly drive logit-level truth assessments can be seen as beliefs about the sequences they process in LLMs.

2 Background

The claim that LLMs encode belief-like or knowledge-like states originates in work on parametric information implicitly stored in model weights. In this tradition, propositions "known" or "believed" by a model are elicited via cloze-style prompts (Petroni et al., 2019, 2020; De Cao et al., 2021; Akyurek et al., 2022; Hase et al., 2023, 2024). On the other hand, representation-level studies show that LLM latent spaces encode structured information about properties such as color (Abdou et al., 2021), space and time (Gurnee and Tegmark, 2024), or game states (Li et al., 2023a; Nanda et al., 2023) that closely mirror their real-world counterparts. This line of inference-time representational interpretability (Zou et al., 2025) deploys a mix of probing (Alain and Bengio, 2018; Belinkov, 2022), activation interventions (Turner et al., 2024), latent space and circuit analysis (Olah et al., 2020).

Several studies have investigated factual truth representations in LLMs using these methods. Burns et al. (2023) found that unsupervised probes can reveal a task-agnostic latent dimension representing truth. Supervised approaches have used shallow neural networks (Azaria and Mitchell, 2023) or linear classifiers such as logistic regression or mass-mean probes (Marks and Tegmark, 2024). While most work focuses on the residual stream, Li et al. (2023b) showed that individual attention heads can also track truth-related features.

Logical robustness remains a challenge. Levinstein and Herrmann (2025) showed that most probes fail to generalize to negated statements, indicating limited coherence. Bürger et al. (2024) addressed this by adding a polarity direction to their classifiers, providing evidence for a two-dimensional subspace generalizing across logical paraphrases. Building on this, Herrmann and Levinstein (2025) proposed formal standards for intermediate truth-sensitive representations to count as genuinely belief-like, informed by formal Bayesian epistemology (Ramsey, 1931; Bovens and Hartmann, 2003) and decision theory (Jeffrey, 1965). No study has yet applied their framework to assess the belief-likeness of different LLMs.

Finally, ways of extracting meaningful probability measures from LLMs' intermediate representations have been partially inspected in epistemic uncertainty quantification studies (Shorinwa et al., 2025). We will follow common UQ methods in LLMs by analyzing the self-reported answers from the models (Krause et al., 2023; Tang et al., 2024) and the post-logit probability distribution (Kadavath et al., 2022; Ling et al., 2024). Using probabilities extracted from latent spaces for assessing model uncertainty has not yet been explored.

3 Requirements for belief

We tested our set of models on the standards for belief representations described by Herrmann and Levinstein (2025). They identify four dimensions: *Accuracy*, *Use*, *Coherence* and *Uniformity*. On this

framework, LLM activations will be more or less belief-like depending on their implicit score in each of these dimensions.

- ***Accuracy*** measures how well intermediate activations separate true from false inputs, reflecting whether the model encodes truth-sensitive representations. It is the basic requirement for activations to be considered belief-like and has been the main metric in prior work (Azaria and Mitchell, 2023; Burns et al., 2023). Higher probe classification success indicates stronger model performance on this dimension.

- A key aspect of beliefs is their functional role in the production of behavior (Lewis, 1972; Block, 1978), particularly in assertions (Zimmerman, 2018). The desideratum of ***Use*** stems from this theoretical backdrop and can be tested through aimed intervention on the activation space. Specifically, it is expected that a model with a crisp truth direction will change its logit distribution accordingly when the intermediate activations are steered across that direction.

- A minimal form of ***Coherence*** is often required for some mental representation to count as a belief (Davidson, 1973; Jeffrey, 1965; Joyce, 1998). This means, for example, that if an agent believes that p, they should not believe that $\neg p$. A minimal coherence requirement is whether activations for logical paraphrases fall in the same truth partition and satisfy probabilistic coherence constraints when decoded by probes.

- The truth direction should be as uniform as possible across different domains. This ***Uniformity*** requirement is justified by the fact that beliefs should hinge on a direction of general truth, while domain-dependent directions of truth would lead to representations that are harder to track and diagnose for safety reasons.

We deploy four experiments for testing the desiderata.

4 Methods

Each experiment employs distinct methods, detailed in its respective section, while sharing a common set of models, probes, and datasets.

4.1 Models

We focused on decoder-only autoregressive transformers from 6 different families and across various parameter counts: GPT-2 (Radford et al., 2019), Llama-3.1-8B (Grattafiori et al., 2024), Gemma-2-9B (Team, 2024), Yi-6B (0.1AI, 2025), Pythia-6.9B (Biderman et al., 2023) and GPT-J-6B (Wang and Komatsuzaki, 2021). When available, both base and variants trained through instruction fine-tuning (IFT) were included.

The study on Accuracy was performed on all models. For Use, Coherence, and Uniformity, we focused on the best models (Top-3 probing accuracy on residual $> 80\%$) since we did not expect meaningful effects for activations with low truth-sensitivity. The restricted set counts the models from the Llama and Gemma families, as well as GPT-J-6B, often listed as the competitive baseline. All layers and heads were analyzed in the first two experiments, but for Coherence and Uniformity we only used the best-performing layer/head to extract probabilities and accuracy, assuming the most separable activations yield the strongest results.

4.2 Probes

We probed the models' activations through linear classifiers to find directions that can also be used for steering. For the Accuracy and Uniformity experiments, we focused on logistic regression probes; for the Use experiment, we shifted to raw mass-mean probes, as they have been shown to be more effective for steering (Li et al., 2023b). During the Coherence experiment, both classifiers were calibrated through isotonic regression and compared as pseudoprobability estimators.

4.3 Datasets

Following Marks and Tegmark (2024), we use their True/False dataset of factual, non-misleading statements across multiple domains, restricting the number of statements per domain for balance. For the Coherence experiment, we extended this dataset with negations, conjunctions, and disjunctions derived from the original statements. We split and validate the dataset according to experiment requirements and stratifying domains and labels across folds, except for leave-one-group-out settings in Coherence and Uniformity experiments; full details and examples can be found in appendix A.3.

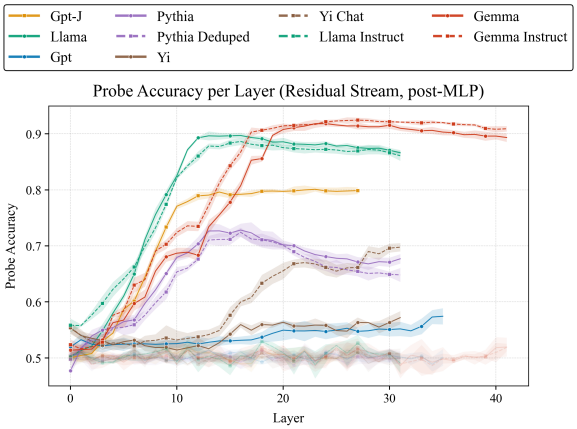


Figure 3: Probe accuracy ± std on the residual streams, averaged over 5 folds. Truth-sensitivity emerges in the middle layers (10-20) in all cases; GPT-J, as well as Llama and Gemma models, are the only LLMs that show high probe accuracy. Control probes (shown in faint lines) remain at chance level on all layers.

5 Experiments

5.1 Accuracy

We report probing results for all models using linear classifiers, reporting absolute accuracy rather than selectivity (Hewitt and Liang, 2019), since probes perform at chance on control tasks. We probe the final tokens and analyze each layer using two activations: (i) the post-MLP residual stream, capturing coarse-grained patterns (Figure 3), and (ii) attention heads, capturing more localized structure (Figure 2, Appendix A.4).

Smaller and task-specific models show low average accuracy overall (Top-3 residual probe accuracy: GPT-2-Large 0.567; Yi 0.630; Pythia 0.721), whereas the largest models exhibit stable truth-sensitive encodings from mid to pre-unembedding layers (Llama 0.889; Gemma 0.920). Similar trends appear at the head level, with average probe accuracy reaching 0.931 on Gemma Instruct.

Attention heads exhibit more localized behavior than the residual stream: in base models, truth-sensitive heads emerge in middle layers and fade later, a pattern absent in IFT models. This suggests IFT promotes finer-grained truth tracking in later layers (Figure 2). Head-level representations are comparably sensitive to residuals, indicating that the dimensionality of head subspaces is adequate for capturing the level of abstraction of truth features.

While prior work shows that single-domain datasets encode truth and falsity in the principal components of the residual stream (Marks and Tegmark, 2024), we do not observe the same effect in a cross-domain setting. However, a subset of highly truth-sensitive *heads* (Figure 4, Table 1) can encode generalized truth along their principal components, suggesting a structural specialization for the feature.

5.2 Use

In this experiment, we intervene on the models' latent spaces by steering the most truth-sensitive activations found in the Accuracy experiment.

Methods We steer activations along mass-mean class difference directions $\theta_{act} = \mu^{+}_{act} - \mu^{-}_{act}$, where μ^{+}_{act} and μ^{-}_{act} are the means for the true and

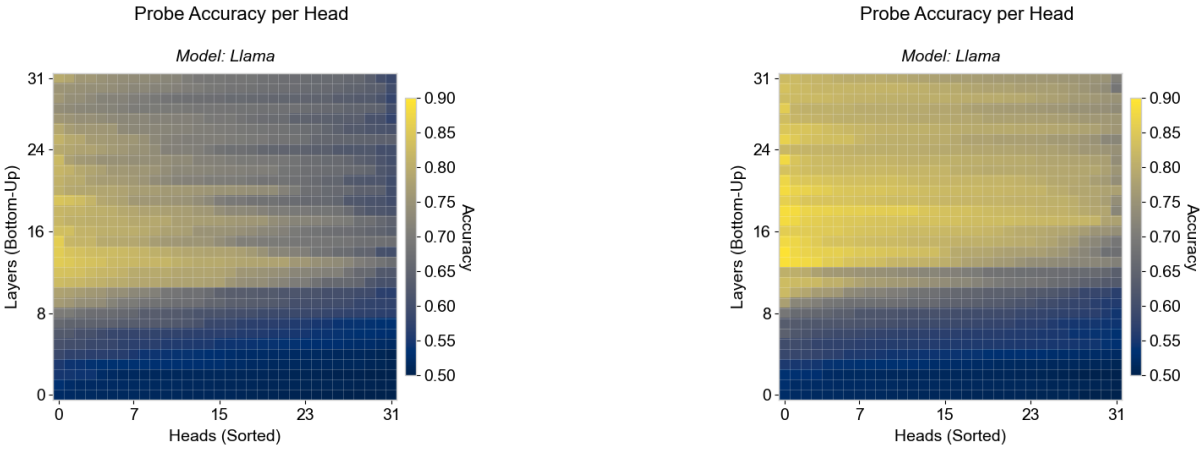


Figure 2: Probe accuracy on pre-output heads, sorted, averaged over 5 folds. Average std: .01 (Llama); .00 (Llama Instruct). Layer behavior is consistent with probe accuracy on residual (Figure 3). However, sweeping probes on the heads reveals higher locality of the truth feature and different patterns in IFT models, where more heads contribute to representation in higher-level layers.

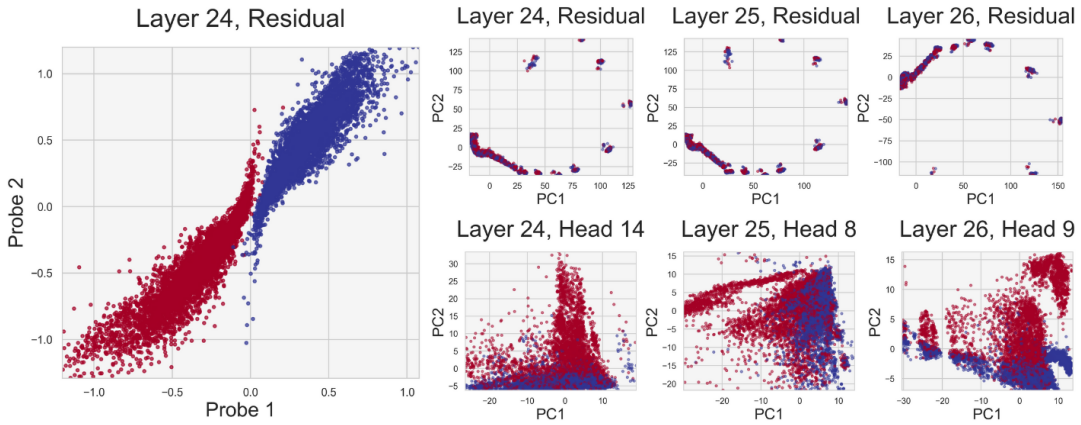


Figure 4: Projections of the residual stream on directions extracted from two orthogonal probes (Roger, 2023) (left); PCA for notably truth-sensitive heads (bottom-right) and their corresponding residual streams (upper-right). Results collected from Gemma Instruct.

Accuracy	Residual	Heads
> 0.80	-	**2**
0.75-0.80	-	**5**
0.60-0.75	-	44
0.55-0.60	3	83
< 0.55	39	538

Table 1: Subset of Gemma Instruct activations most predictive of truth under linear probe tested on PC projections; while attention heads are highly localized, the probe fails to track truth in residual projections.

false classes for a given activation space. Our objective is to elicit *incorrect* truth value answers from the model. Single-layer or single-head interventions are often insufficient, so following Li et al. (2023b), we apply two hyperparameters determined through grid search: K, the number of the top-K most truth-sensitive activations collected from the first experiment, and α, the intervention strength. Because a head's contribution can reverse through the attention output, we determine its effect on the residual stream and apply steering with the appropriate sign; see Figure 5 for intervention effects across a parameter sweep. Our residual-level intervention is defined as such:

$$\mathbf{r}^{(\ell)\prime} = \mathbf{r}^{(\ell)} + \alpha\, \theta_\ell \cdot \mathbf{1}_{[\ell \in \text{Top-K}]} \tag{1}$$

Where $\ell = 1, \ldots, L$ denotes the residual extracted from the respective layer, Top-K $= \{\ell_1, ..., \ell_K\}$ denotes the K residuals with highest truth-sensitivity across the entire model, and $\mathbf{1}_{[\cdot]}$ is the indicator function.

Similarly, we define head-level interventions:

$$\mathbf{h}^{(\ell,i)\prime} = \mathbf{h}^{(\ell,i)} + \tilde{\alpha}\, \theta_{(\ell,i)} \cdot \mathbf{1}_{[(\ell,i) \in \text{Top-K}]} \tag{2}$$

Where (ℓ, i) denotes head i at layer ℓ, with $\ell = 1, \ldots, L$ and $i = 1, \ldots, H$, Top-K $= \{(\ell_1, i_1), ..., (\ell_K, i_K)\}$ denotes the K heads with highest truth-sensitivity, and $\tilde{\alpha}$ is the signed steering scalar. We use zero-shot prompts (see Appendix A.1) and report results for both steering directions (false $\rightarrow$ true and true $\rightarrow$ false), applying negative α and $\tilde{\alpha}$ for the latter.

Metrics We define our metrics as the absolute effect of intervention E and the direction-specific effect $\underset{\rightarrow}{E}$ compared against a random direction sampled from a standard normal distribution applied to the same activations:

$$\begin{aligned} E &= \left(P^{\times}_{\text{post}} - P^{\checkmark}_{\text{post}}\right) - \left(P^{\times}_{\text{pre}} - P^{\checkmark}_{\text{pre}}\right) \\ \underset{\rightarrow}{E} &= \frac{E/S}{E_{\text{rand}}/S_{\text{rand}}} \end{aligned} \tag{3}$$

Where $P^{\times}$ and $P^{\checkmark}$ are the probabilities of the incorrect and correct truth values, collected from pre (P_{pre}) and post-intervention (P_{post}) models. Both E and E_{rand} were clipped at a minimum of .05 to ensure numerical stability. For the direction-specific effect, we compute the strength of intervention S, defined as:

$$S = \sqrt{|\alpha|\, K\, \frac{D_{\text{act}}}{D_{\text{model}}}} \tag{4}$$

Here D_{act} is the dimension of the intervened activation and D_{model} is the dimension of the residual

Model		False → True					True → False				
		$\lvert\alpha\rvert$	K	S	E	$\underset{\rightarrow}{E}$	$\lvert\alpha\rvert$	K	S	E	$\underset{\rightarrow}{E}$
Llama	Residual	9	1	3.00	0.21	3.02±1.98	6	1	2.44	0.05	1.00
	Heads	3	40	1.94	0.29	3.90±1.79	9	1	0.53	0.06	0.90±0.39
Llama_it	Residual	3	7	4.58	1.23	10.41±4.27	5	2	3.16	0.90	32.50±22.15
	Heads	2	20	1.12	**1.23**	6.31±1.31	3	10	0.96	1.23	20.51±16.07
Gemma	Residual	31	7	14.73	0.12	2.45	33	22	26.94	0.35	6.57±0.79
	Heads	12	15	3.58	0.29	5.67±0.17	10	60	6.54	0.31	6.25±0.01
Gemma_it	Residual	30	16	21.90	0.54	9.81±2.03	30	26	27.92	1.07±0.10	19.94±2.96
	Heads	10	65	6.81	0.87	**12.91±5.44**	14	25	5.00	**1.48**	**26.84±5.43**
GPT-J	Residual	15	4	7.74	0.05	1.00	30	15	21.21	0.20	2.52±1.26
	Heads	2	5	0.79	0.05	1.00	10	50	5.59	0.09	1.38±0.47

Table 2: Full steering results, averaged across 5 seeds. Llama and Gemma models are generally successful, with differences between steering setups. Steering interventions on attention heads produce stronger and cheaper effects on truth-value computations than residual stream interventions, suggesting a more direct causal role. Variance in $\underset{\rightarrow}{E}$ reflects variability in the random baseline rather than effect instability.

stream. Intuitively, E captures raw impact on output probabilities and $\underset{\rightarrow}{E}$ quantifies how effective a direction-specific intervention is relative to random steering, adjusted for the strength of intervention.

This is an important distinction: an intervention showing high E and low $\underset{\rightarrow}{E}$ points to an effect which is significant, but not exclusive to the mass-mean direction; an intervention showing low E and high $\underset{\rightarrow}{E}$ denotes a specificity effect which is scaled by very small effects on the baseline.

Results We report the full results on Table 2, collected by selecting K/α pairs in order to maximize E/S. We also sweep baseline hyperparameters; if no meaningful effect is observed, we keep hyperparameters from the clean-runs search, otherwise we select the best-performing configuration from the search. GPT-J-6B is generally unresponsive to steering and will serve as our competitive baseline for the experiment. Interventions on models from the Llama and Gemma families are generally successful. We have three additional takeaways:

1. The results are influenced by model output bias. This is clearly visible in two cases: (1) when we compare the strength of the intervention to the effect exhibited in the Llama and Gemma family, and (2) when we compare the steering directions within models, which exhibit a 'True/False'-response bias.

2. Consistent with previous observations, IFT models show greater absolute steering sensitivity than base models (avg. E ratio: 6.35). Gemma models also exhibit higher sensitivity than Llama models (avg. $\underset{\rightarrow}{E}$ ratio: 1.15).

3. Attention heads are consistently more steerable than the residual stream, producing larger effects at lower strength (avg. E ratio: 1.28). In Llama Instruct models, random head perturbations can also be effective, yielding lower and less stable direction-specific scores.

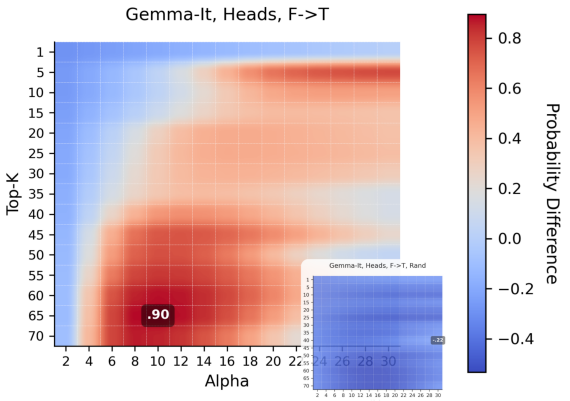


Figure 5: Steering grid search on Gemma. While the effect is clear on the mass-mean direction, there is no effect on a random direction (grid on the bottom-right). The mass-mean direction is causally implied in downstream computation, satisfying Use. The set of hyperparameters swept is model-dependent, to account for model family bias (see also Appendix A.4).

5.3 Coherence

We assess agreement between intermediate representations both absolutely and probabilistically. In the first case, by testing probes on datasets containing logical paraphrases differing from the training set. For the probabilistic experiment, we develop metrics and estimators to extract the probabilities from the latent space. We focused again on the best models selected during the Accuracy experiments.

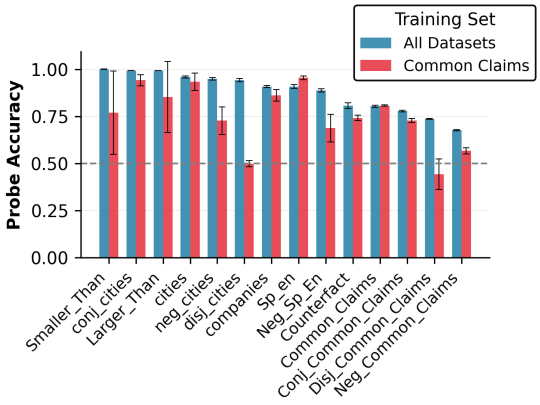


Figure 6: Probe accuracy ± std on logically diverse datasets with corresponding training sets, Gemma Instruct, layer 25, head 8, averaged over 5 seeds. When trained on all datasets, the probe is able to classify datasets across logical paraphrases, with a moderate drop in accuracy for 'Common Claims' paraphrases.

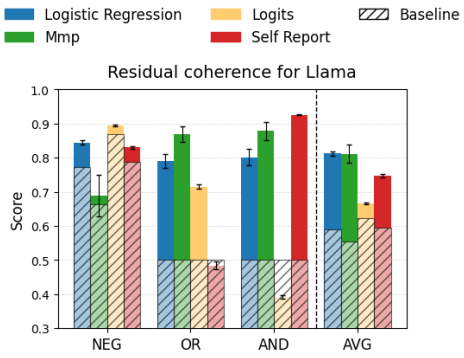


Figure 7: Probabilistic coherence scores ± std, Llama, residual, layer 13, averaged over 5 seeds. Calibrated linear classifiers show better results than logits and self-reporting on conjunction and disjunction. Calibrated logistic regression shows the best coherence on negation, accounting for the baseline. On an aggregate level, probes perform better than the other methods.

Metrics We evaluate probabilistic coherence using tests from different logical paraphrases, following synchronic norms of rationality from Bayesian epistemology (Lin, 2024). Using negation, a rational agent must satisfy finite additivity and the complement rule:

$$\hat{P}(\phi \cup \neg\phi) = \hat{P}(\phi) + \hat{P}(\neg\phi) \approx 1 \qquad (5)$$

Where $\hat{P}(\phi)$ is the model's subjective probability for statement ϕ. Conjunction and disjunction have to follow monotonicity:

$$\hat{P}(\phi \cap \psi) \leq \hat{P}(\phi) \leq \hat{P}(\phi \cup \psi) \qquad (6)$$

Since $(\phi \cap \psi) \subseteq \phi \subseteq (\phi \cup \psi)$. Subjective probabilities should respect this ordering. To measure how closely the model outputs approximate ideal-agent norms, we compute an inverse-MSE score for negation and an accuracy score for conjunction and disjunction. Full details on the metrics appear in Appendix A.2.

Methods We extract subjective probabilities from the latent space to test the model's coherence. Both logistic and mass-mean probes are calibrated using isotonic regression on the post-sigmoid outputs, after which we read out the pseudoprobabilities (Zadrozny and Elkan, 2002). Intuitively, activations that project further along the latent truth direction should represent higher estimated probabilities that a given statement is true. For comparison, we use two alternative estimators: (1) self-reported assessments produced via few-shot prompting, and (2) probabilities for the model's True and False tokens obtained from inference. See Appendix A.1 for more details.

Results In the logic task (Figure 6), we observe moderate overall coherence on paraphrases, with the negated representations in the diverse 'Common Claims' set yielding the lowest results from linear probes. This is also true for probes trained on datasets aggregating multiple logical paraphrases (Average accuracy on negated common claims: Llama, 0.51; Llama Instruct, 0.59; Gemma, 0.56; Gemma Instruct, 0.67; GPT-J, 0.44). Except for this set, probes can generalize well when trained on datasets that include logical paraphrases (cf. full figures in Appendix A.4).

In the probabilistic experiment (Figure 7), logistic regression probes tend to retrieve subjective probabilities that are stably coherent across three tasks and broadly surpass baselines. Additionally, in several configurations, we report probabilities that are more coherent than the models' self-reported assessments and extracted logits (Llama: all tasks; Llama Instruct: conjunction; Gemma: negation; Gemma Instruct: negation and conjunction: see also Appendix A.4 for aggregate results).

While results on probabilistic coherence are promising, the logic task suggests that generalized negation may be too strong a confounder for linear probes trained on representations extracted from smaller base models. This supports the hypothesis that in models of this size, linear truth directions are dependent on the logical operators seen in the prompt (see also Bürger et al.'s (2024) work for

a non-linear probe showing reliable truth-tracking for logical paraphrases).

5.4 Uniformity

We assess probe generalization capabilities by testing cross-domain uniformity on the best models selected during the Accuracy experiment. Probes are trained on (i) generic statements from 'Common Claims', (ii) single-domain statements (e.g., 'Cities', 'Companies'), and (iii) both; each of them is tested on one of the domains of the True/False dataset. Following the Coherence experiment, we exclude the datasets including logical paraphrases, to avoid results confounded by failures across negation. We focus again on the best layer/head found in each model. Example train-test matrices are shown in Figure 8, with full results in Appendix A.4.

Uniformity results are strong, both at the head and the residual level, especially with large multi-domain training sets (average/worst test accuracies: Llama 0.87/0.74; Llama Instruct 0.87/0.72; Gemma 0.88/0.68; Gemma Instruct 0.90/0.74). Probes trained on the naturally multi-domain 'common-claims' dataset partially generalize to unseen domains, particularly in IFT models (average/worst test accuracies: Llama 0.64/0.57; Llama Instruct 0.75/0.58; Gemma 0.75/0.66; Gemma Instruct 0.77/0.50). Rich, diverse datasets, thus, enable probes to converge on directions that robustly track truth across domains.

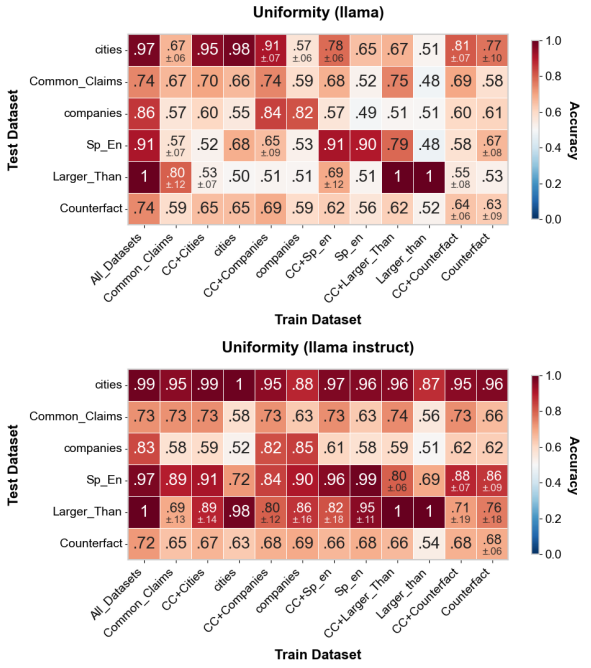


Figure 8: Uniformity results on residual for Llama layer 13 and Llama Instruct, layer 15, averaged over 10 seeds. Std $>$ .05 reported for unstable train/test pairs. IFT models show more uniform activations across domains. Training on diverse sets leads to uniformity - an effect that is also present at the head level (see Appendix A.4).

6 Discussion

6.1 Takeaways

Figure 9 shows a summary of the models' scores. We have three main empirical takeaways to report.

1. We observed generally strong evidence of truth-tracking across all experiments, with three model families (Gpt-J, Llama, Gemma) exhibiting clear signals under our standards. Nonetheless, the moderate results on logical coherence suggest that more expressive probes may be needed to reliably track truth across negation in complex domains. Truth-sensitivity emerges at middle layers, and IFT tends to expose it further in later layers.

2. The most important factors that covary with crisp truth-tracking representations seem to be model size and the use of IFT, which appears to enhance the quality of the representations. This was especially evident in the experiments on Use and Uniformity, where IFT models showed higher sensitivity and cross-domain consistency than their base versions.

3. ***Attention heads*** seem to be the best locus for model beliefs: they show strong truth-sensitivity (even on PCs), exert causal influence on outputs, and exhibit coherence comparable to the residual stream. Unlike the higher-dimensional residual stream, where features are more superposed and fragile under intervention, attention heads provide a more interpretable substrate for truth representation.

6.2 The sense in which models believe

Our tests are grounded in formal epistemology, which models beliefs as abstract entities to capture cognitive behavior (Weisberg, 2021). Although this abstraction supports measurable standards, we conclude by clarifying to what extent these representations may count as genuine instances of AI cognition (cf. Goldstein and Levinstein, 2024). Beliefs are ubiquitous in our common-sense psychology (Hutto and Ravenscroft, 2021), but there is no unified mechanistic story about what consti-

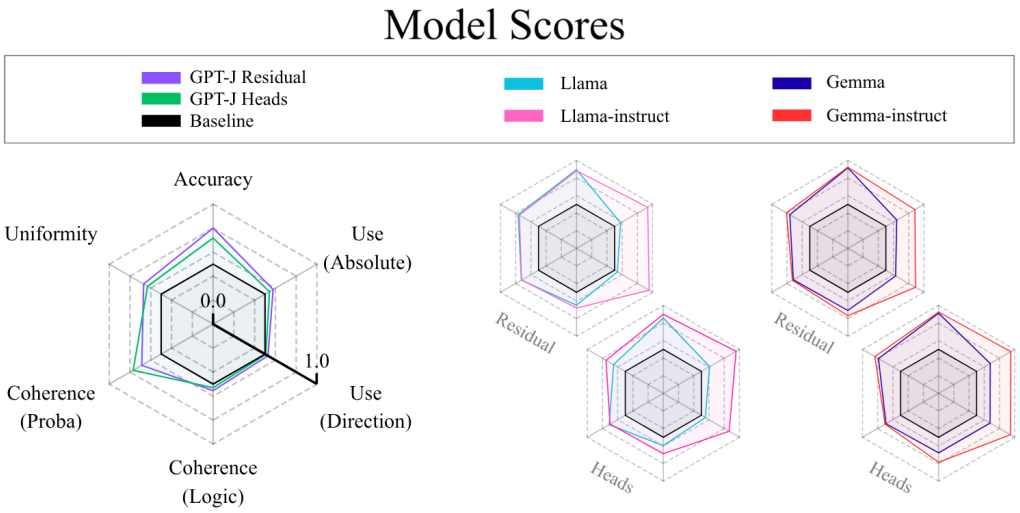


Figure 9: General scores for models in the restricted set. See Appendix A.2 for formal score definitions. The head level is comparable to the residual stream and sometimes can carry crisper and more causal-effective representations.

tutes these mental states – and some philosophers have even questioned whether such things exist in biological agents (Churchland, 1981; Stich, 1983).

Standard accounts of belief in philosophy and cognitive science can be broadly divided into *deflationary* and more *realist* views. On deflationary accounts, beliefs are fictions or abstract patterns invoked to interpret behavior, rather than genuine mental states (Dennett, 1971, 1981; Toon, 2023). On this view, attributing beliefs to LLMs is trivially licensed by their linguistic behavior, and mechanistic analysis adds no further justification (cf. Cappelen and Dever, forthcoming-a; forthcoming-b). Realist accounts, by contrast, treat beliefs as precise mental states with properties such as compositionality, biological realization, phenomenal character, or semantic content (Fodor, 1975; Searle, 1990; Schwitzgebel, 2002; Poth and Schuster, 2026). Our tests offer no evidence that LLMs process belief-like states in this substantive sense.

One brand of realism understands beliefs in functionalist terms, on which mental states are individuated by *the roles* they play within a system (Armstrong, 1968; Block, 1978). This is an influential approach that, most importantly, allows for multiple realizability: the possibility that cognitive entities can be instantiated by distinct physical systems. What counts as the relevant functional role varies across functionalist theories. For example, Millikan (1984) asks for a function to be biologically grounded in the system's evolutionary history, and Bratman (1987) characterizes beliefs with respect to their connection with other mental states such as desires or intentions. However, our tests target a shared core of functionalism (Schwitzgebel, 2024) by examining whether LLM activations can realize central and broadly uncontroversial belief-roles. These roles map to our experiments: tracking truth (Accuracy), supporting assertion (Use), and properly integrating with other beliefs (Coherence). For proponents of a minimal version of functionalism, our results are directly philosophically relevant.

Conclusion

We evaluated intermediate representations in decoder-only LLMs against standards of belief-likeness: Accuracy, Use, Coherence, and Uniformity. Across six model families, larger models, especially when instructed, develop latent representations that are truth-sensitive across domains, causally impactful on outputs, and moderately coherent. We found that attention heads provide a particularly effective substrate for truth-tracking.

Our observations are relevant for AI safety, model editing, and uncertainty quantification: localizing truth-related features in specific heads offers a target for diagnosing factual inconsistency, pre-decoding hallucinations, and may enable local interventions in weight-editing approaches. Moreover, if probe-based pseudoprobabilities prove more coherent than their competitors, future research may explore the best way to characterize a model's 'credences' from their internal states.

Ultimately, our results suggest that, for the more liberal functionalists, LLMs can encode representations that are non-trivially belief-like.

Limitations

Our study has various limitations.

- Our work focused on a limited pool of models, spanning parameter sizes up to approximately 9 billion. While this is justified by the size and design of our experiments, we note that previous studies analyzed models up to approximately 70 billion parameters (Marks and Tegmark, 2024).
- Coherence results have been shown to be improvable by non-linear probes trained on truth *and* polarity (Bürger et al., 2024). We did not compare our results with non-linear probes, leaving this possibility for further research.
- The True/False dataset that we employed is diverse, but lacks cross-lingual statements except for Spanish and is intentionally simple. The results may generalize in different ways according to model size for more complex and diverse datasets.
- Our experiments face the standard limitations for probing. Most importantly, since the True/False dataset is underdetermined with respect to the feature it tracks, we cannot be sure that the represented feature is *truth* or a feature that heavily correlates with truth (as pointed out by Levinstein and Herrmann, 2025). This study, however, is not limited to reporting probe-derived correlations due to our steering experiment, which verifies counterfactual robustness for the directions.
- The experiments only show that LLMs instantiate beliefs in a minimal functionalist sense, but not in a more substantive, 'Searlean' sense (§6.2). Furthermore, LLMs trivially have beliefs in the deflationary sense. The results, then, are of philosophical interest for a specific, albeit popular (Bourget and Chalmers, 2023) view on mental states.

Ethical Considerations

Two main ethical considerations arise from our study.

- Our results are directly relevant to AI safety. Since vector steering can significantly alter the behavior of a language model, findings using this methodology - especially on tasks related to factual truth - can be exploited for malicious intent, such as pushing chatbots to lie or, more generally, polluting the model's answers. On the other hand, advancements in inference-time lie detection supported by our experiments can have positive impacts for alignment.
- While our findings are philosophically relevant, it is important to note that the idea that models may have belief-like representations is strongly related to claims about richer cognitive states or consciousness and sentience. These claims can contribute to spread serious problems related to undue anthropomorphization; see (Deroy, 2023) for a thorough argument and (Corona Mendozza, 2025) specifically for beliefs. To avoid that, we remark that such interpretations should be made cautiously, and our results should not be taken as evidence of a substantive mental life in the models.

Acknowledgments

We would like to thank the Meetings of the Content-Centered Computing group at the University of Turin and the audience in the Center for Language Technology at the University of Copenhagen for the useful feedback on early versions of this study. This work was funded by a Carlsberg Semper Ardens Advance grant for the Center for Philosophy of AI.

References

0.1AI. 2025. Yi: Open foundation models by 01.ai.

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, page 109–132, Online. Association for Computational Linguistics.

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

David M. Armstrong. 1968. *A Materialist Theory of the Mind*. Routledge Kegan Paul.

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*, page 967–976, Singapore. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Honolulu, Hawaii, USA. JMLR.org.

Ned Block. 1978. Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9:261–325.

David Bourget and David J. Chalmers. 2023. Philosophers on philosophy: The 2020 philpapers survey. *Philosophers' Imprint*, 23(11).

Luc Bovens and Stephan Hartmann. 2003. *Bayesian Epistemology*. Oxford University Press, Oxford.

Michael Bratman. 1987. *Intention, Plans, and Practical Reason*. Harvard University Press.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*.

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*, volume 37, page 138393–138431. Curran Associates, Inc.

Herman Cappelen and Josh Dever. forthcoming-a. *A Hyper-Externalist Manifesto for LLMs*. Oxford University Press.

Herman Cappelen and Josh Dever. forthcoming-b. *Going Whole Hog: A Philosophical Defense of AI Cognition*. Oxford University Press.

David J. Chalmers. 2025. Propositional interpretability in artificial intelligence.

Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you (report). Technical report, Alignment Research Center.

Paul M. Churchland. 1981. Eliminative materialism and the propositional attitudes. *Journal of Philosophy*, 78(2):67–90.

Alessandro Corona Mendozza. 2025. Don't believe the belief hype! In *Mechanistic Interpretability Workshop at NeurIPS 2025*.

Donald Davidson. 1973. Radical interpretation. *Dialectica*, 27(1):313–328.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel C. Dennett. 1971. Intentional systems. *Journal of Philosophy*, 68(February):87–106.

Daniel C. Dennett. 1981. *The Intentional Stance*. MIT Press.

Ophelia Deroy. 2023. The ethics of terminology: Can we use human terms to describe ai? *Topoi*, 42(3):881–889.

Jerry Fodor. 1975. *The Language of Thought*. Harvard University Press.

Simon Goldstein and Benjamin A. Levinstein. 2024. Does chatgpt have a mind?

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.

Wes Gurnee and Max Tegmark. 2024. Language models represent space and time. In *International Conference on Representation Learning*, volume 2024, page 2483–2503.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, page 2714–2731, Dubrovnik, Croatia. Association for Computational Linguistics.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. 2024. Fundamental problems with model editing: How should rational belief revision work in llms? *Transactions on Machine Learning Research*.

Daniel A. Herrmann and Benjamin A. Levinstein. 2025. Standards for belief representations in llms. *Minds and Machines*, 35(1):5.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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), page 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Daniel Hutto and Ian Ravenscroft. 2021. *Folk Psychology as a Theory*, fall 2021 edition. Metaphysics Research Lab, Stanford University.

Richard C. Jeffrey. 1965. *The Logic of Decision*. University of Chicago Press, New York, NY, USA.

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)*, page 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

James M. Joyce. 1998. A nonpragmatic vindication of probabilism. *Philosophy of Science*, 65(4):575–603.

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

Lea Krause, Wondimagegnhue Tufa, Selene Baez Santamaria, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: Exploring the calibration and expression of (un)certainty of large language models in a multilingual setting. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, page 1–9, Prague, Czech Republic. Association for Computational Linguistics.

Benjamin A. Levinstein and Daniel A. Herrmann. 2025. Still no lie detector for language models: probing empirical and conceptual roadblocks. *Philosophical Studies*, 182:1539–1565.

David K. Lewis. 1972. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3):249–258.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc. Event-place: New Orleans, LA, USA.

Hanti Lin. 2024. *Bayesian Epistemology*, summer 2024 edition. Metaphysics Research Lab, Stanford University.

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)*, page 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets.

Ruth G. Millikan. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Neel Nanda, Andrew Lee, and Martin Wattenberg. 2023. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, page 16–30, Singapore. Association for Computational Linguistics.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

OpenAI. 2025. Introducing gpt-4.1 in the api.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, page 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poth and Annika Schuster. 2026. Mental, scientific, and artificial representations. *Philosophy and the Mind Sciences*, 7(1).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Frank Ramsey. 1931. *Truth and Probability*, page 156–98. Routledge Kegan Paul.

Fabien Roger. 2023. What discovering latent knowledge did and did not find.

Eric Schwitzgebel. 2002. A phenomenal, dispositional account of belief. *Noûs*, 36(2):249–275.

Eric Schwitzgebel. 2024. *Belief*, spring 2024 edition. Metaphysics Research Lab, Stanford University.

John R. Searle. 1990. Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association*, 64(3):21–37. American Philosophical Association.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3).

Stephen P. Stich. 1983. *From folk psychology to cognitive science: The case against belief.* MIT Press, Cambridge, MA, US.

Zhisheng Tang, Ke Shen, and Mayank Kejriwal. 2024. An evaluation of estimative uncertainty in large language models. *Preprint*, arXiv:2405.15185.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Adam Toon. 2023. *Mind as Metaphor: A Defence of Mental Fictionalism*, 1 edition. Oxford University PressOxford.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering.

Ben Wang and Aran Komatsuzaki. 2021. Gpt-j-6b: A 6 billion parameter autoregressive language model.

Jonathan Weisberg. 2021. *Formal Epistemology*, spring 2021 edition. Metaphysics Research Lab, Stanford University.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 694–699, New York, NY, USA. Association for Computing Machinery. Event-place: Edmonton, Alberta, Canada.

Aaron Z. Zimmerman. 2018. *Belief: A Pragmatic Picture*. Oxford University Press, Oxford.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency.

A Appendices

A.1 Prompting strategies

We used several contexts across the experiments for our statements `STMT`.

A.1.1 Accuracy & Uniformity

Each model was prompted to register the last position logits during the Accuracy experiment. The prompt was zero-shot, with no context except:

```
"{STMT} This sentence is:"
```

We designed the context with the aim of tracking directions that would prove useful in the second experiment, too, where the models are asked to assess the truth value of each statement under interventions. The same formats were used for extracting the projections in Figure 4.

A.1.2 Use

We used zero-shot prompting with light instructions in the context to prevent logits from collapsing to true/false tokens. The prompts were in this form:

```
"Determine whether the following
statement is factually correct.
Respond with exactly one of:
True, False, Unknown. Answer
Unknown unless you are certain.
\n{STMT}\nAnswer:"
```

The 'Unknown' option was added as a confounder for tracking model uncertainty and avoiding biased defaulting to one of the truth values. Since each model presents different biases and tokenizers, we had to slightly modify the prompts according to the families, but kept our changes on the whitespaces to preserve comparability across models.

A.1.3 Coherence

For the probabilistic coherence task, we had two prompts. The first one was fed to the models to extract probe-based probabilities and logit-based probabilities, and it had this form:

```
"The sky is blue. This
statement is: True
\n\nThe earth is flat. This
statement is: False
 \n\n{STMT} This statement is:"
```

This time, we used two-shot prompting to encourage the model to collapse most of its uncertainty to the true and false tokens in order to track the competing probabilities. For the self-reporting, we had to resort to few-shot prompting together with a stronger context to teach the task to the models:

```
"I am a fact-checking AI. For each
statement, I rate the probability
that the statement is true on a
scale from 0 to 1.
\n\nStatement: Paris is the
capital of France.
\nP(True): 0.95
\n\nStatement: The largest bear
in the world
is currently in Italy.
\nP(True): 0.25
\n\nStatement: Milan is the
capital of Italy.
\nP(True): 0.05
\n\nStatement: Humans have five senses.
\nP(True): 0.65
\n\nStatement: {STMT}
\nP(True):"
```

The shots were built to cover four different cases of true/false uncertainty.

A.2 Scoring

A.2.1 Probabilistic Coherence

The scores for probabilistic coherence were tailored to how negation, conjunction, and disjunction should behave on degrees of belief (Section 5.3). More specifically, the inverse MSE score for negation was defined as:

$$\text{NegScore} = \frac{1}{1 - \text{MSE}\big(\hat{\mathbf{p}}(\phi) + \hat{\mathbf{p}}(\neg\phi), \overline{1}\big)} \tag{7}$$

Where $\hat{\mathbf{p}}(\phi), \hat{\mathbf{p}}(\neg\phi) \in [0,1]^n$ are the vectors of predicted probabilities for statement pairs in ϕ and $\neg\phi$ form, and:

$$\overline{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n \tag{8}$$

Is the vector of ones of the same dimension as $\hat{\mathbf{p}}(\phi)$.

The scores for conjunction and negation were defined as a pair:

$$\text{ConjScore} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{P}_i(\phi \cap \psi) \leq \hat{P}_i(\phi)\right] \tag{9}$$

$$\text{DisjScore} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{P}_i(\phi \cup \psi) \geq \hat{P}_i(\phi)\right] \tag{10}$$

Where $\phi \cap \psi$ and $\phi \cup \psi$ are statements that pair ϕ with a new proposition ψ under conjunction and disjunction respectively; $\mathbf{1}[\cdot]$ is the indicator function.

The baselines for conjunction and disjunction are set to 0.5. For negation, rather than using a fixed value, the baseline is constructed by randomly permuting the estimator's predictions for positive and negative statements before summing them. This preserves the distributional properties of the predictions while destroying the pairing between complementary statement pairs.

A.2.2 Final scoring

To enable meaningful comparison across the different dimensions of our study, we extracted scores from our experiments that fall in the range $[0, 1]$, with random baselines around 0.5. These scores illustrate high-level differences in belief-likeness across dimensions, but are not intended as a competitive benchmark, and each aggregation is bound to bring some distortion for the actual results, which are reported in the rest of the study.

Accuracy For the Accuracy score we averaged the top-5 activations for each of the tested models, resulting in the following scores: **Gpt-J**, 0.799 (Residual) - 0.718 (Heads); **Llama**, 0.895 (Residual) - 0.857 (Heads); **Llama Instruct**, 0.882 (Residual) - 0.899 (Heads); **Gemma**, 0.916 (Residual) - 0.911 (Heads); **Gemma Instruct**, 0.923 (Residual) - 0.925 (Heads).

Use (Absolute) For the Use (Absolute) score, we collected the absolute effect E for each configuration and mapped the values through a logistic function with a fixed $k = 2.5$ slope applied for spread:

$$\sigma(x) = \frac{1}{1 + e^{-kx}} \tag{11}$$

Our baseline is the unintervented model, which trivially shows an effect of 0 leading to a 0.5 baseline. The resulting scores are: **Gpt-J**, 0.578 (Residual) - 0.543 (Heads); **Llama**, 0.579 (Residual) - 0.608 (Heads); **Llama Instruct**, 0.935 (Residual) - 0.956 (Heads); **Gemma**, 0.643 (Residual) - 0.679 (Heads); **Gemma Instruct**, 0.882 (Residual) - 0.949 (Heads).

Use (Direction) The Use (Direction) score is computed similarly to Use (Absolute), except for an additional shift in the logistic function applied to map the random baseline ($\underset{\rightarrow}{E} = 1$) to 0.5:

$$\sigma(x) = \frac{1}{1 + e^{-k(x-1)}} \tag{12}$$

we fix $k = 0.15$ considering the average higher spread of values for $\underset{\rightarrow}{E}$. The resulting scores are: **Gpt-J**, 0.529 (Residual) - 0.507 (Heads); **Llama**, 0.538 (Residual) - 0.552 (Heads); **Llama Instruct**, 0.956 (Residual) - 0.865 (Heads); **Gemma**, 0.629 (Residual) - 0.678 (Heads); **Gemma Instruct**, 0.889 (Residual) - 0.944 (Heads).

Coherence (Logic) For the Coherence (Logic) score, we compute the average probe performance on the best residual/head when trained on all datasets and penalize variability, using mean minus standard deviation. This favors representations that are both high-performing and consistent after exposure to all cross-logic datasets. The resulting scores are: **Gpt-J**, 0.555 (Residual) - 0.530 (Heads); **Llama**, 0.648 (Residual) - 0.596 (Heads); **Llama Instruct**, 0.679 (Residual) - 0.687 (Heads); **Gemma**, 0.708 (Residual) - 0.678 (Heads); **Gemma Instruct**, 0.768 (Residual) - 0.784 (Heads).

Coherence (Probabilities) For the Coherence (Probabilities) score, we average the probabilistic performance of logistic regression estimators trained and tested on the best residuals/heads across negation, conjunction, and disjunction tasks. Each score is baseline-corrected and shifted by +0.5, such that baseline performance always corresponds to 0.5. The resulting scores are: **Gpt-J**, 0.687 (Residual) - 0.770 (Heads); **Llama**, 0.726 (Residual) - 0.712 (Heads); **Llama Instruct**, 0.722 (Residual) - 0.701 (Heads); **Gemma**, 0.716 (Residual) - 0.691 (Heads); **Gemma Instruct**, 0.728 (Residual) - 0.703 (Heads).

Uniformity As for Coherence (Logic), we average the probe performance on the best residual/head when the classifier has been exposed to all cross-domain datasets prior to the accuracy report and penalize variability. The resulting scores are: **Gpt-J**, 0.669 (Residual) - 0.631 (Heads); **Llama**, 0.768 (Residual) - 0.654 (Heads); **Llama Instruct**, 0.752 (Residual) - 0.753 (Heads); **Gemma**, 0.763 (Residual) - 0.790 (Heads); **Gemma Instruct**, 0.807 (Residual) - 0.838 (Heads).

A.3 Dataset

We employed and extended the True/False dataset built by Marks and Tegmark (2024), chosen to avoid overly misleading or hard statements - as, for example, in TruthfulQA (Lin et al., 2022) or TriviaQA (Joshi et al., 2017).

A.3.1 Dataset Composition

The original dataset includes statements across multiple domains: Cities, Companies, Larger, Smaller, Spanish-English Translation (Sp_En_Trans), Counterfact, and Common Claims. It already includes negated statements for Cities and Sp_En_Trans, as well as conjunction and disjunction statements for Cities. We extended the dataset by creating logical paraphrases for two domains: Companies (Negation only) and Common Claims (all paraphrases).

We used string manipulation to generate the conjunction and disjunction paraphrases for Common Claims. For the negated versions of Common Claims, we used GPT-4.1-mini via OpenAI's API (OpenAI, 2025) to ensure natural phrasing. Table 3 shows representative examples from each domain and paraphrase type.

A.3.2 Data Splits and Validation

Due to the varied experimental setups, we organized the data splits as follows:

Accuracy: We first performed a 50/50 stratified split (by domain) to reserve half the dataset for the Use experiment. The Accuracy probes were then trained and tested using an 80/20 split on the first half. We employed 5-fold cross-validation for statistical robustness.

Use: Based on results from the Coherence experiments, we excluded logical paraphrases from this experiment. We used the held-out 50% of the data, and performed exploratory grid searches on a random subset of 100 statements. The results displayed in Table 2 are collected on the full partition of the data.

Coherence (Logic) and Uniformity: We used a 75/25 split in a cross-domain setup to test the relative properties of these measures. Since no cross-validation scheme was applicable here, the split was repeated over 5 (logic) and 10 (uniformity) random seeds to ensure statistical robustness in both experiments.

Coherence (Probabilistic): We used a 70/30 split on the full dataset, with logically paraphrased versions paired in the same rows to prevent data leakage. The split was repeated over 5 random seeds for each estimator. For testing, we analyzed paraphrased versions from each domain and added paraphrases from Counterfact and Companies.

The results reported in the Appendix were obtained from a single run to save computational resources.

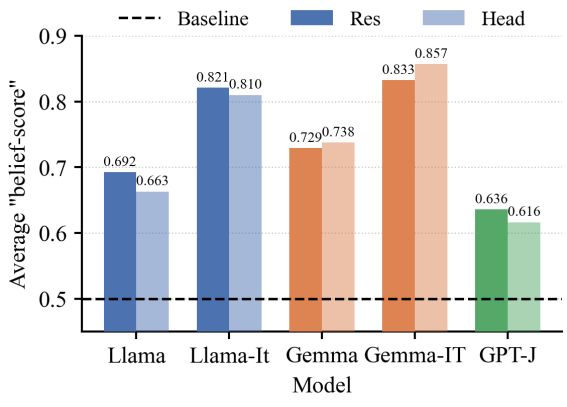


Figure 10: Average model scores across all experiments

Statement	Label	Domain/Logic	N
The city of Hangzhou is in China.	True	Cities	1496
The city of Casablanca is in Russia.	False	Cities	
The city of Hangzhou is not in China.	True	Cities (Neg)	1496
The city of Casablanca is not in Russia.	False	Cities (Neg)	
It is the case both that the city of Nasiriyah is in Iraq and that the city of Tlalpan is in Mexico.	True	Cities (Conj)	1496
It is the case both that the city of Coimbatore is in Uzbekistan and that the city of Daqing is in China.	False	Cities (Conj)	
It is the case either that the city of Rajkot is in India or that the city of Guankou is in Saudi Arabia.	True	Cities (Disj)	1496
It is the case either that the city of Jeddah is in India or that the city of Feira de Santana is in China.	False	Cities (Disj)	
Wild monkeys groom each other's hair as a social activity.	True	Common Claims	4450
More vitamin D makes some people bleed	False	Common Claims	
More vitamin D does not make some people bleed.	True	Common Claims (Neg)	4450
Wild monkeys do not groom each other's hair as a social activity	False	Common Claims (Neg)	
It is both the case that a seed is an embryonic plant enclosed in a protective outer covering and some Indians worship oxen	True	Common Claims (Conj)	4450
It is both the case that More vitamin D makes some people bleed and Lemons produce more electricity than batteries	False	Common Claims (Conj)	
It is either the case that Wild monkeys groom each other's hair as a social activity or Female crows have been known to gather around and help new mothers with childcare	True	Common Claims (Disj)	4450
It is either the case that More vitamin D makes some people bleed or Lemons produce more electricity than batteries.	False	Common Claims (Disj)	
American Express operates in the industry of diversified financials.	True	Companies	1200
Barclays has headquarters in Russia.	False	Companies	
Fifty-six is larger than fifty-three.	True	Larger	1980
Sixty-five is larger than eighty-seven.	False	Larger	
Sixty-five is smaller than eighty-seven.	True	Smaller	1980
Fifty-six is smaller than fifty-three.	False	Smaller	
The Spanish word 'madre' means 'mother'.	True	Sp_En_Trans	354
The Spanish word 'insecto' means 'then'.	False	Sp_En_Trans	
The Spanish word 'insecto' does not mean 'then'.	True	Sp_En_Trans (Neg)	354
The Spanish word 'madre' does not mean 'mother'.	False	Sp_En_Trans (Neg)	
The Big Bang Theory debuted on CBS.	True	Counterfact	2000
Klemens von Metternich's profession is an actor.	False	Counterfact	

Table 3: Statements extracted from each domain of the dataset.

A.4 Full results by model[1]

A.4.1 Gpt-2-Large

Accuracy We report a full probe sweep on the model's attention heads.

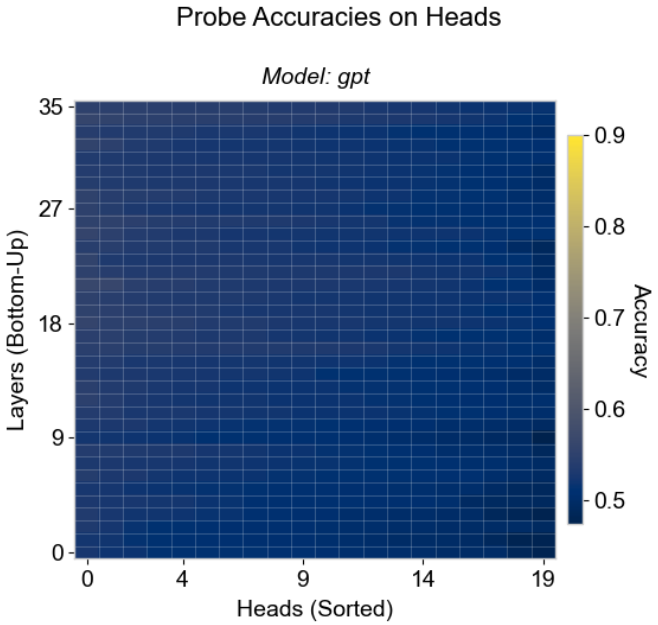


Figure 11: Probe sweep on heads, Gpt-2-Large

A.4.2 Pythia-6.9B

Accuracy We report a full probe sweep on the model's attention heads.

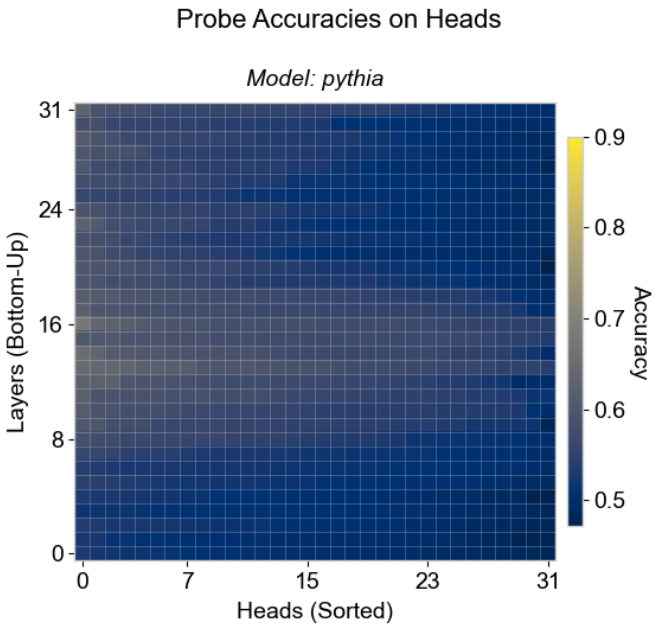


Figure 12: Probe sweep on heads, Pythia-6.9B

[1]For the full model-by-model analysis, we collected single-run results to save computational resources. While intervention sweeps are useful for visualizing how model sensitivity changes according to intervention strength, we refer back to Table 2 for full results on that experiment.

A.4.3 Pythia-6.9B-Deduped

Accuracy We report a full probe sweep on the model's attention heads.

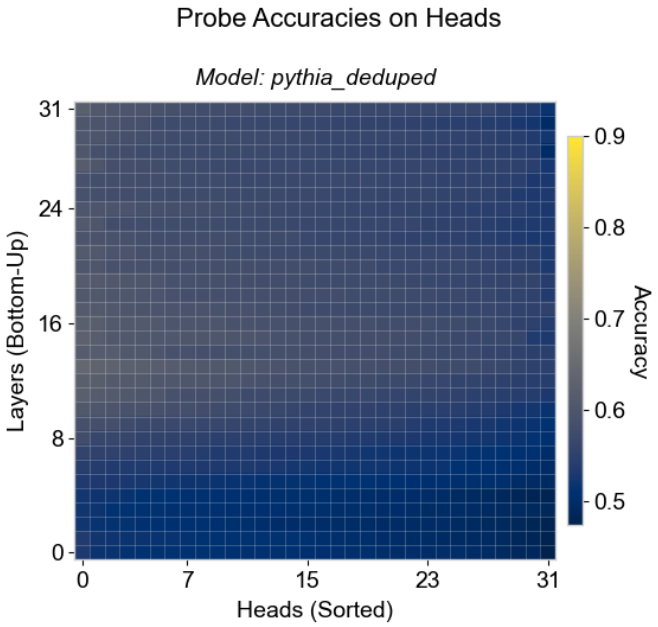


Figure 13: Probe sweep on heads, Pythia-6.9B-Deduped

A.4.4 Yi-6B

Accuracy We report a full probe sweep on the model's attention heads.

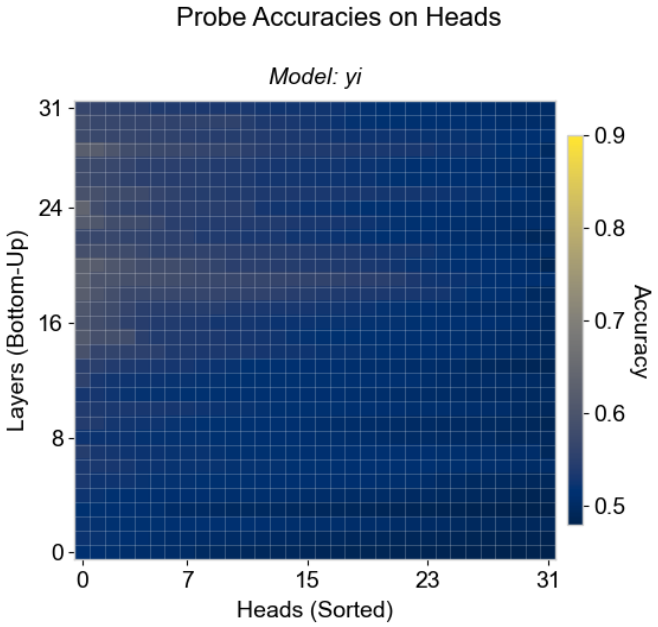


Figure 14: Probe sweep on heads, Yi-6B

A.4.5 Yi-6B-Chat

Accuracy We report a full probe sweep on the model's attention heads.

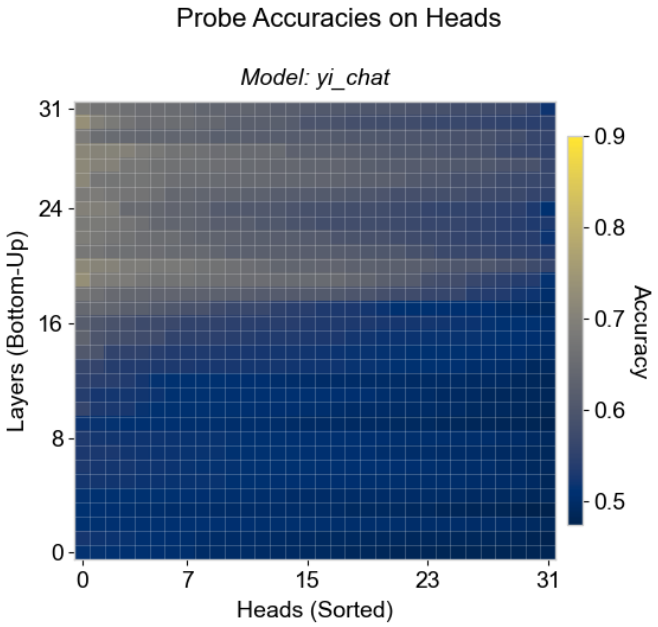


Figure 15: Probe sweep on heads, Yi-6B-Chat

A.4.6 GPT-J-6B

Accuracy We report a full probe sweep on the model's attention heads.

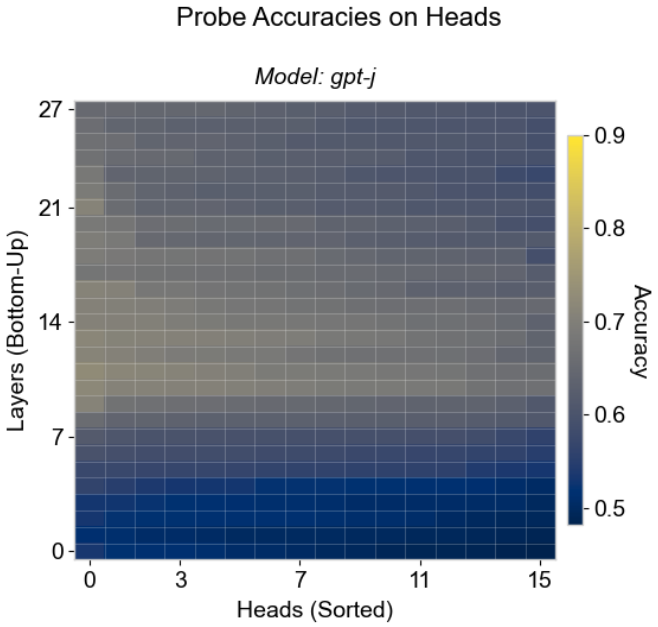


Figure 16: Probe sweep on heads, GPT-J-6B

Use Example sweeps, in both directions, plus control for reference.
Residual steering:

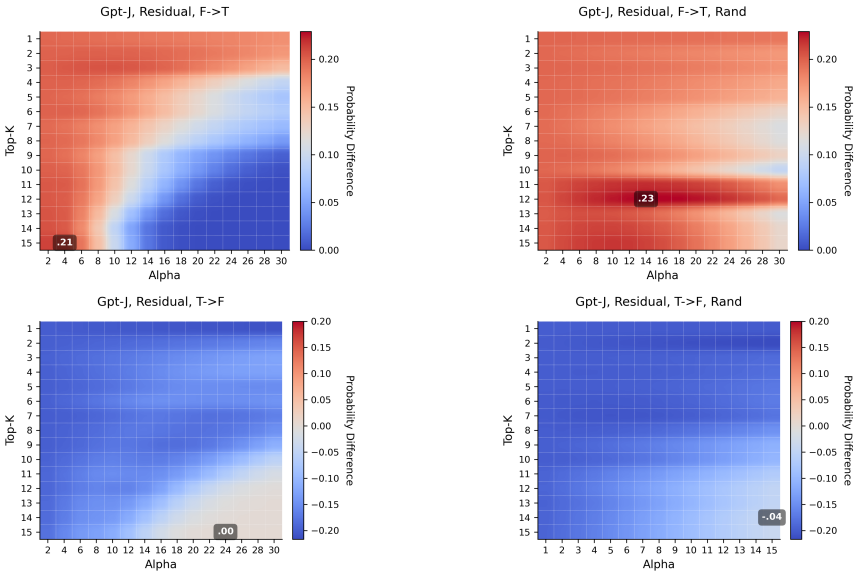


Figure 17: Grid search K, α, GPT-J-6B, Residual. Control sweeps on the right.

Head steering:

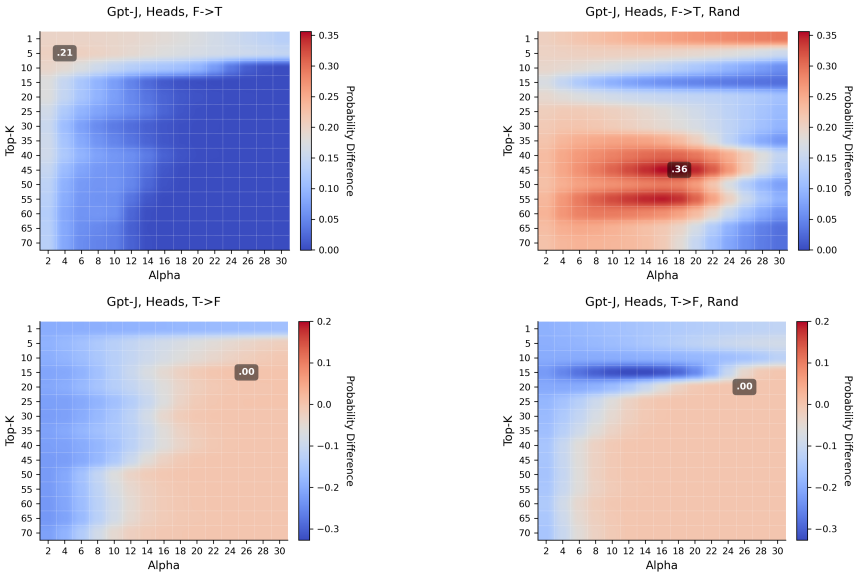
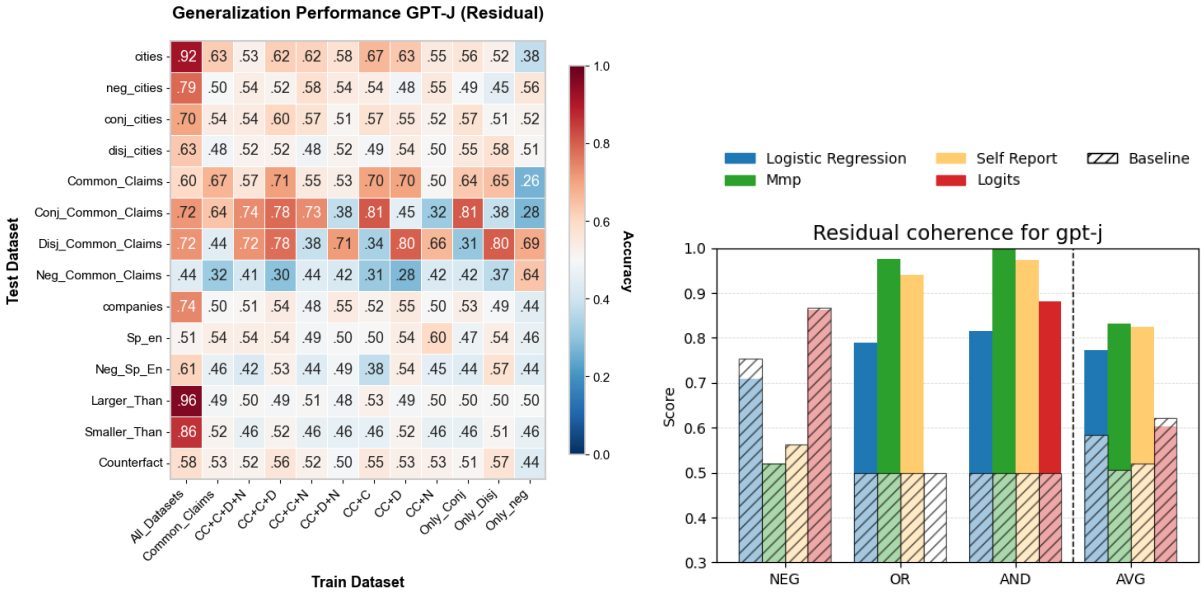


Figure 18: Grid search K, α, GPT-J-6B, Heads. Control sweeps are on the right.

Coherence We report a full cross-dataset probing experiment (left) and full results for the probabilistic experiment (right) for the model. Results on the residual stream:

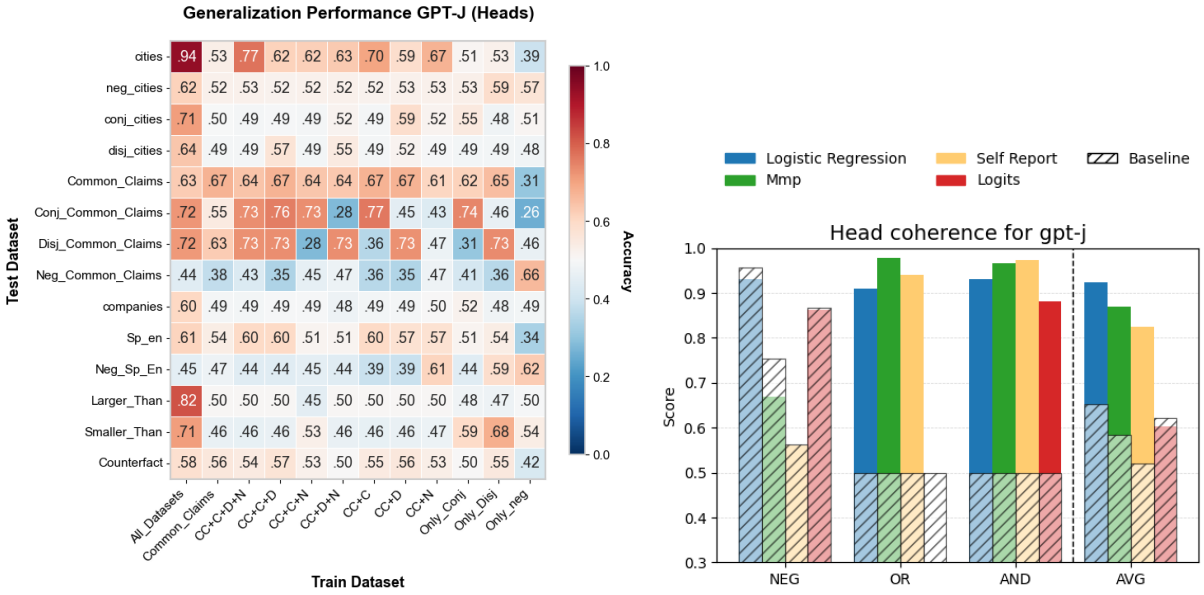


(a) Logical Coherence (Residual)

(b) Probabilistic Coherence (Residual)

Figure 19: Coherence results

Results on the attention heads:

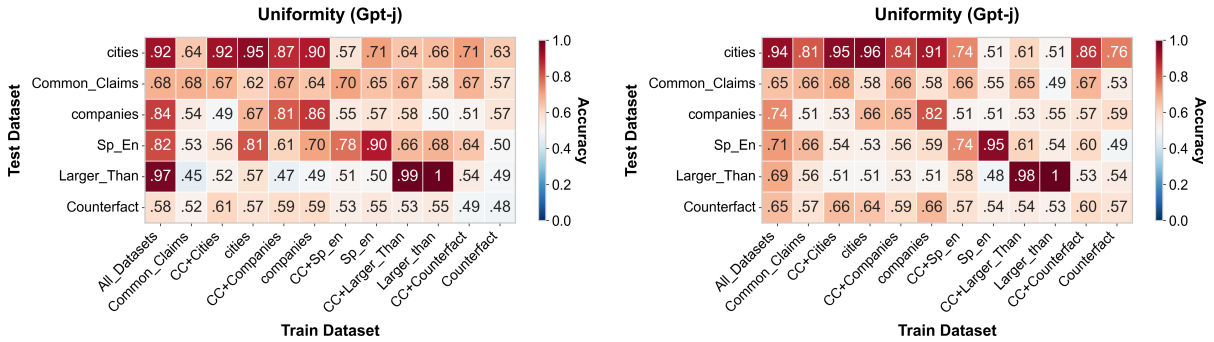


(a) Logical Coherence (Heads)

(b) Probabilistic Coherence (Heads)

Figure 20: Coherence results

Uniformity We report the full cross-domain uniformity experiment for the model, results extracted from the residual stream (left) and from the attention heads (right).



(a) Cross-domain Uniformity (Residual)

(b) Cross-domain Uniformity (Heads)

Figure 21: Coherence results

A.4.7 Llama-3.1-8B

Accuracy We report a full probe sweep on the model's attention heads (cf. Figure 2).

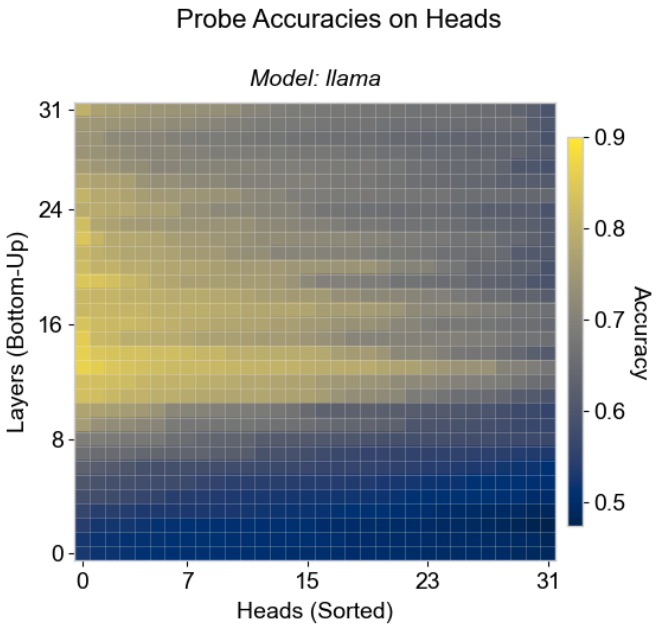


Figure 22: Probe sweep on heads, Llama-3.1-8B

Use Example sweeps, in both directions, plus control for reference.
Residual steering:

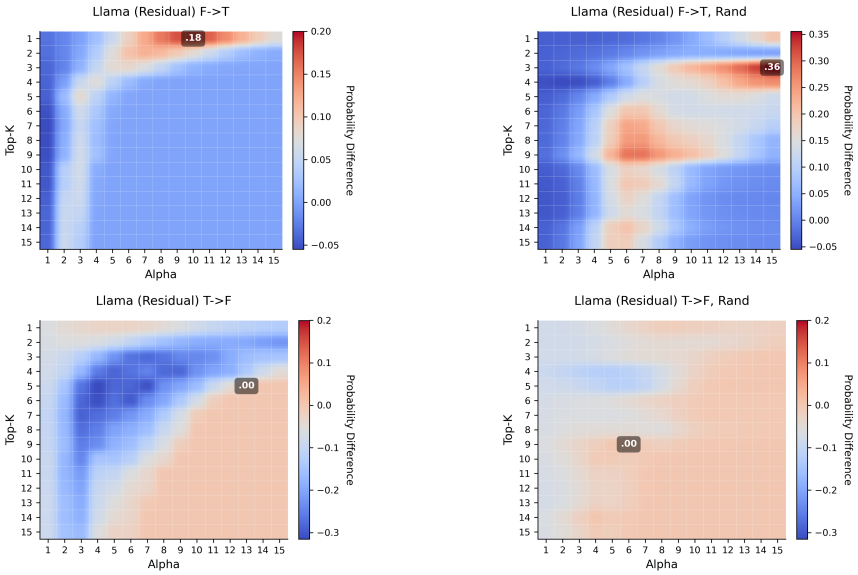


Figure 23: Grid search K, α, Llama-3.1-8B. Random direction effect on F->T (upper right) and no effect on any T->F task (lower) suggest a strong bias of the model towards 'True' tokens.

Head steering:

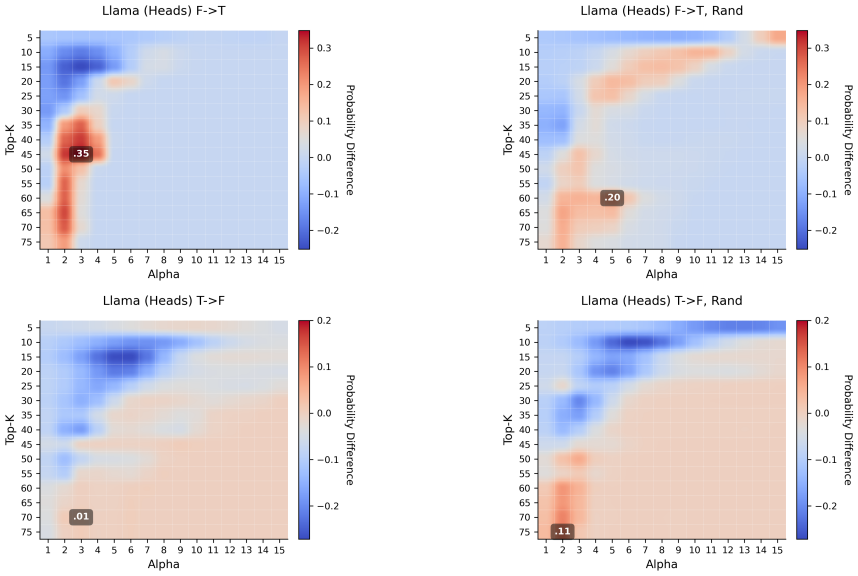
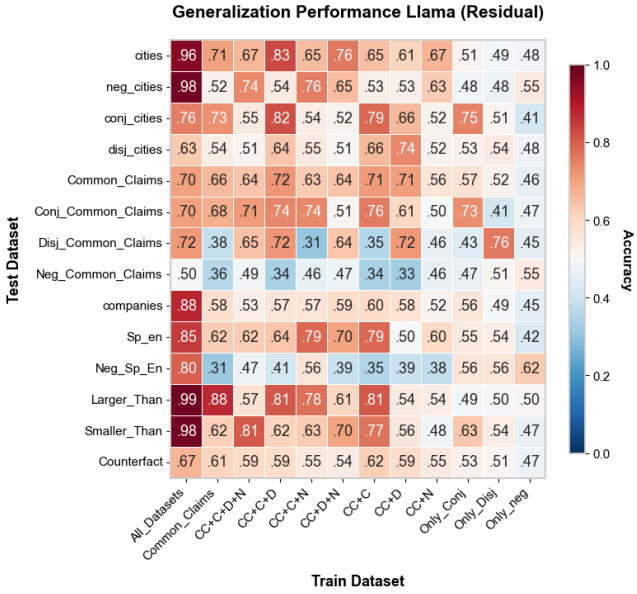


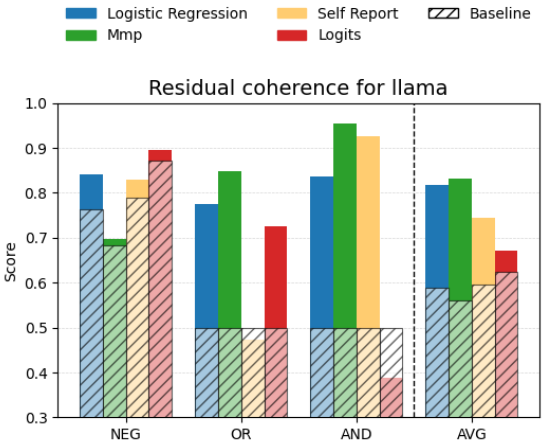
Figure 24: Grid search K, α, Llama-3.1-8B, Heads. See remarks on Figure 23

Coherence We report a full cross-dataset probing experiment (left) and full results for the probabilistic experiment (right) for the model.

Results on the residual stream:



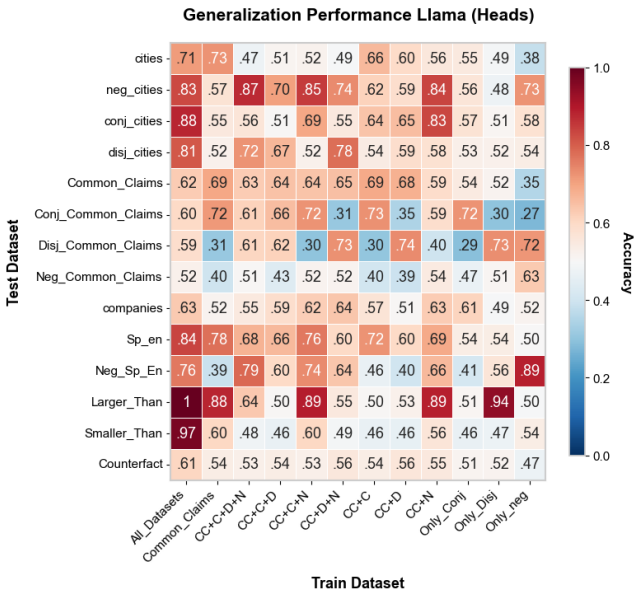
(a) Logical Coherence (Residual)



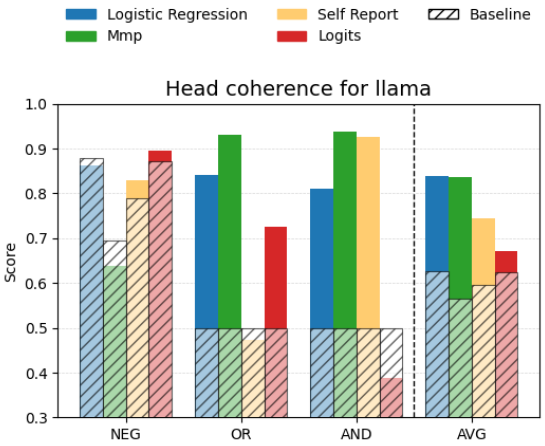
(b) Probabilistic Coherence (Residual)

Figure 25: Coherence results

Results on the attention heads:



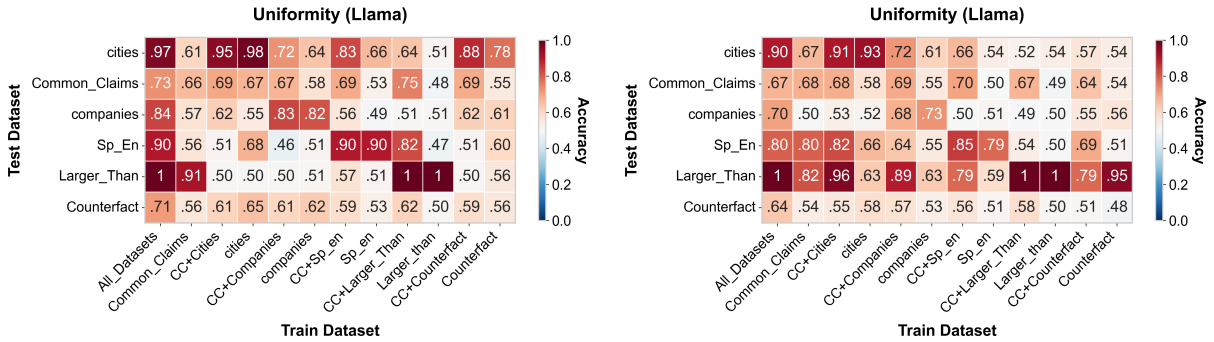
(a) Logical Coherence (Heads)



(b) Probabilistic Coherence (Heads)

Figure 26: Coherence results

Uniformity We report the full cross-domain Uniformity experiment for the model, results extracted from the residual stream (left) and from the attention heads (right).



(a) Cross-domain Uniformity (Residual)

(b) Cross-domain Uniformity (Heads)

Figure 27: Uniformity results

A.4.8 Llama-3.1-8B-instruct

Accuracy We report a full probe sweep on the model's attention heads (cf. Figure 2).

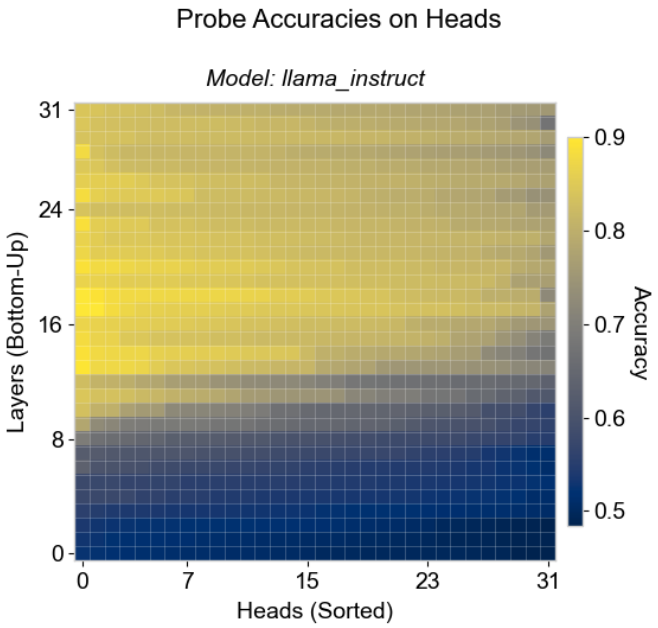


Figure 28: Probe sweep on heads, Llama-3.1-8B-instruct

Use Example sweeps we performed on the model, in both directions, and showing sweeps performed on random directions for reference.
Residual steering:

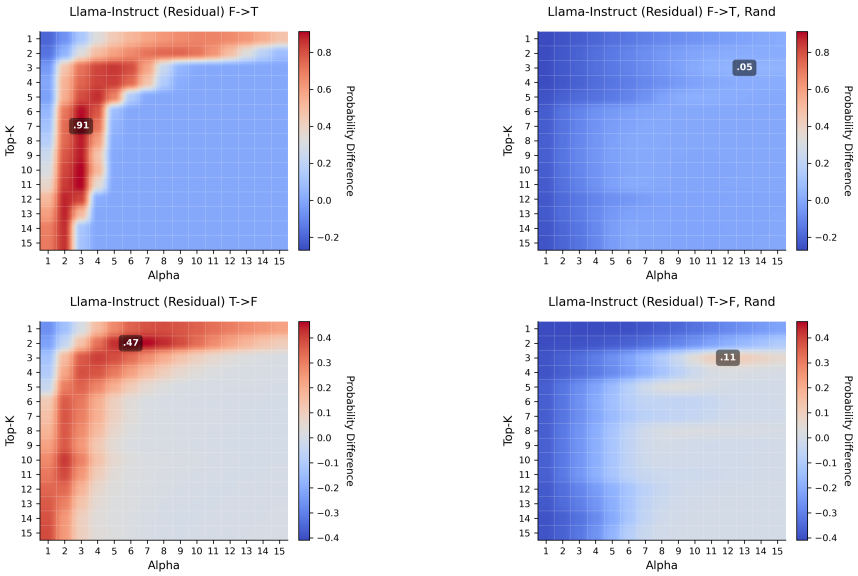


Figure 29: Grid search K, α, Llama-3.1-8B-Instruct, Residual. Control sweeps on the right.

Head steering:

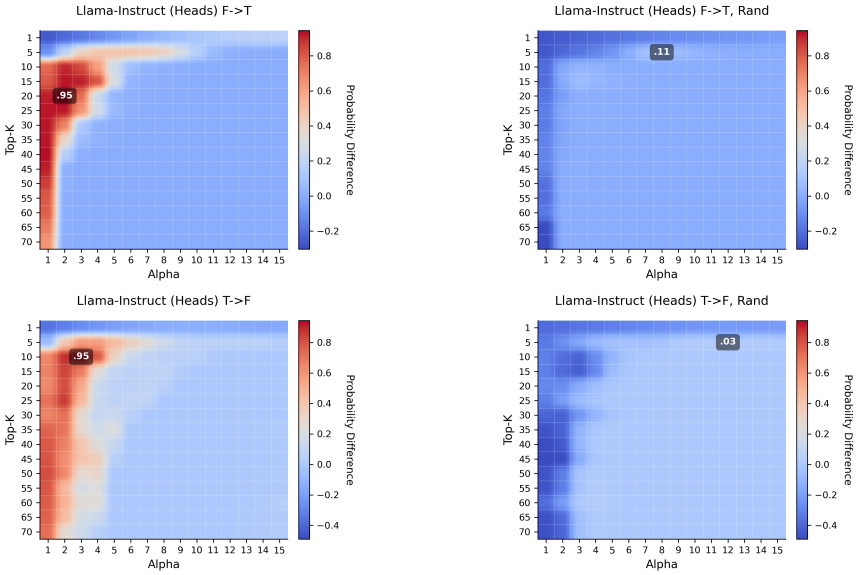
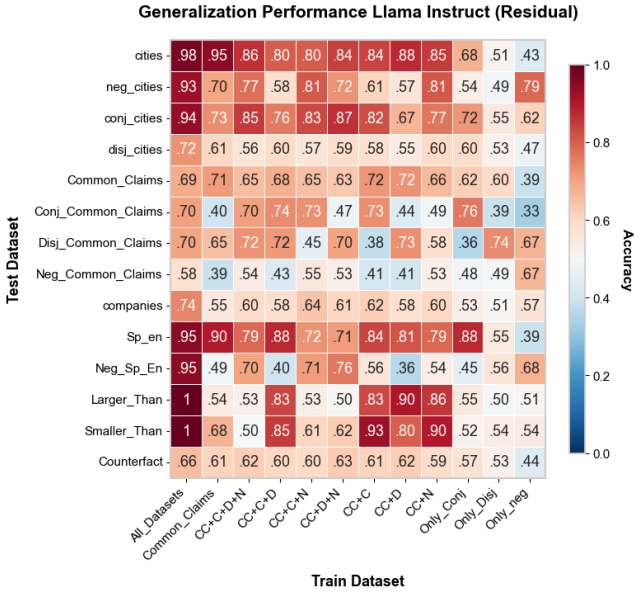


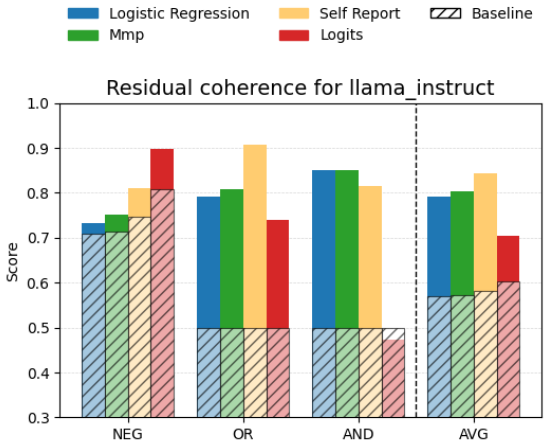
Figure 30: Grid search K, α, Llama-3.1-8B-Instruct, Heads. Control sweeps on the right.

Coherence We report a full cross-dataset probing experiment (left) and full results for the probabilistic experiment (right) for the model.

Results on the residual stream:



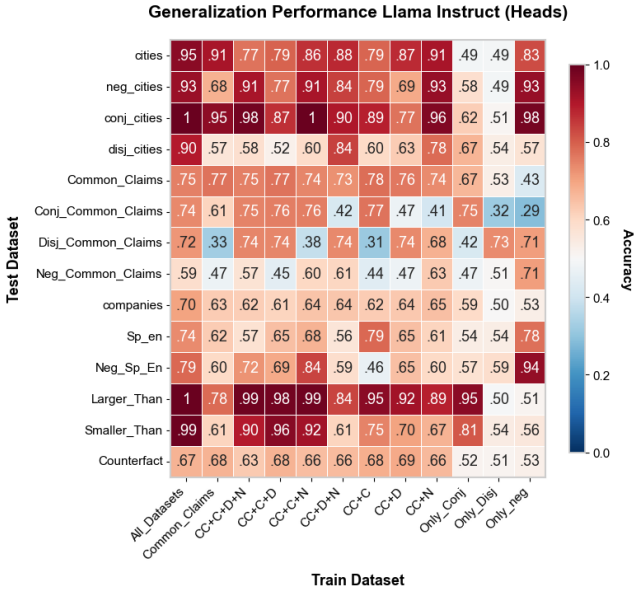
(a) Logical Coherence (Residual)



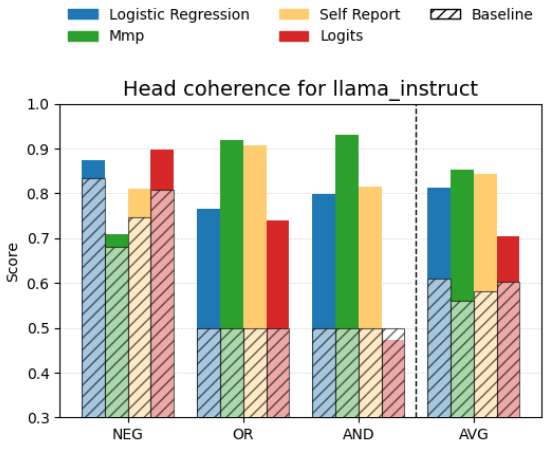
(b) Probabilistic Coherence (Residual)

Figure 31: Coherence results

Results on the attention heads:



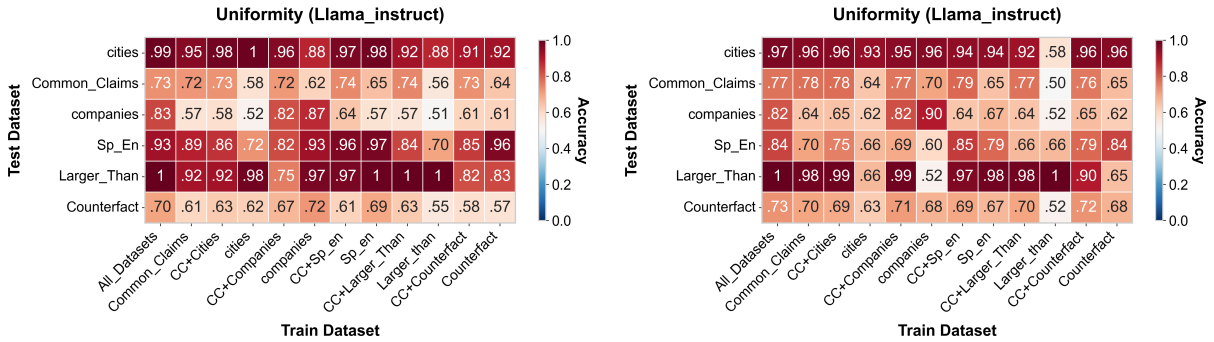
(a) Logical Coherence (Heads)



(b) Probabilistic Coherence (Heads)

Figure 32: Coherence results

Uniformity We report the full cross-domain Uniformity experiment for the model, results extracted from the residual stream (left) and from the attention heads (right).



(a) Cross-domain Uniformity (Residual)

(b) Cross-domain Uniformity (Heads)

Figure 33: Uniformity results

A.4.9 Gemma-2-9B

Accuracy We report a full probe sweep on the model's attention heads.

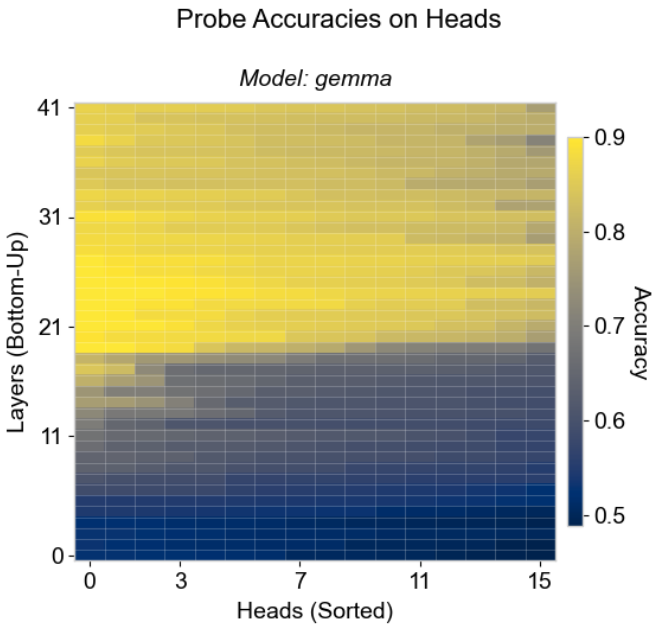


Figure 34: Probe sweep on heads, Gemma-2-9B

Use Example sweeps we performed on the model, in both directions, and showing sweeps performed on random directions for reference.
Residual steering:

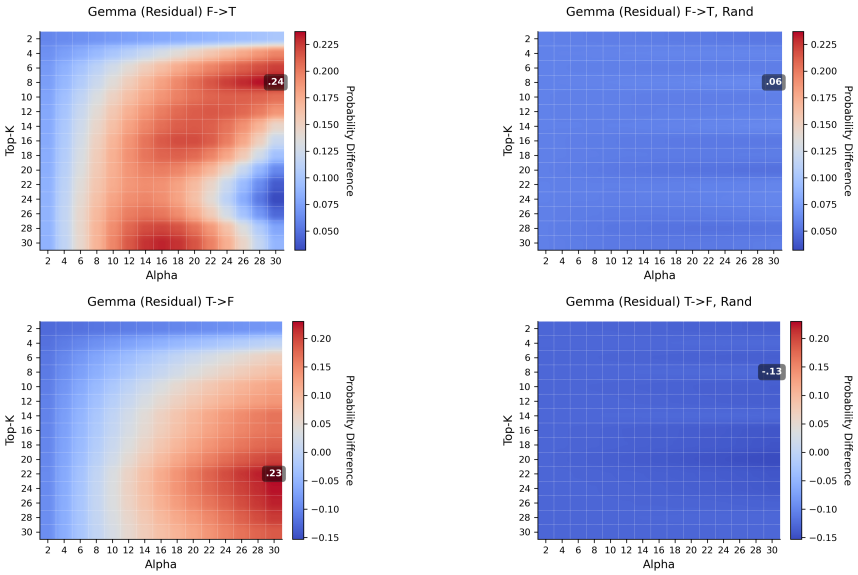


Figure 35: Grid search K, α, Gemma-2-9B, Residual. Control sweeps on the right.

Head steering:

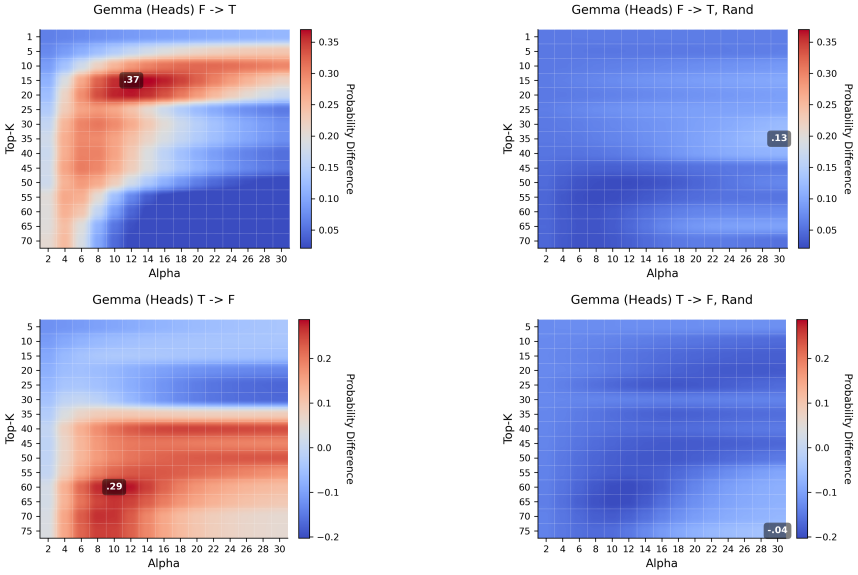
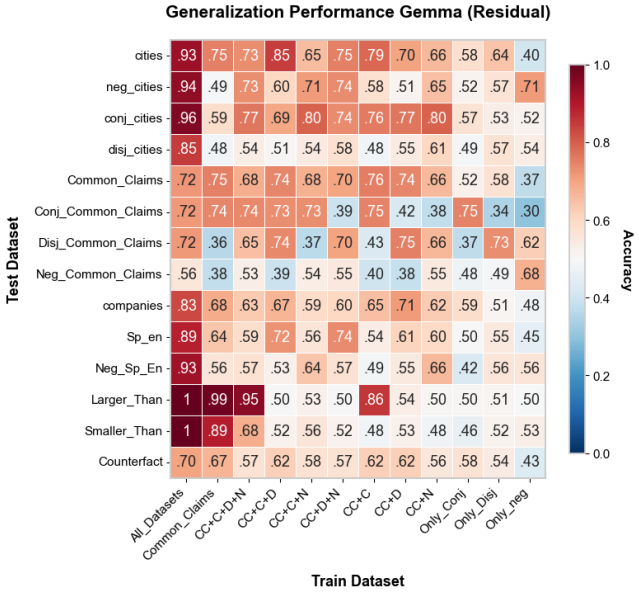


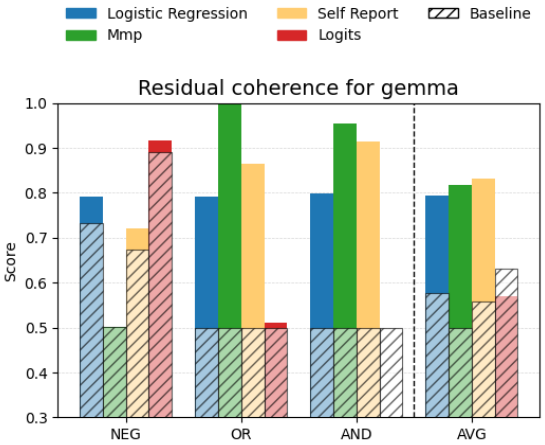
Figure 36: Grid search K, α, Gemma-2-9B, Heads. Control sweeps on the right.

Coherence We report a full cross-dataset probing experiment (left) and full results for the probabilistic experiment (right) for the model.

Results on the residual stream:



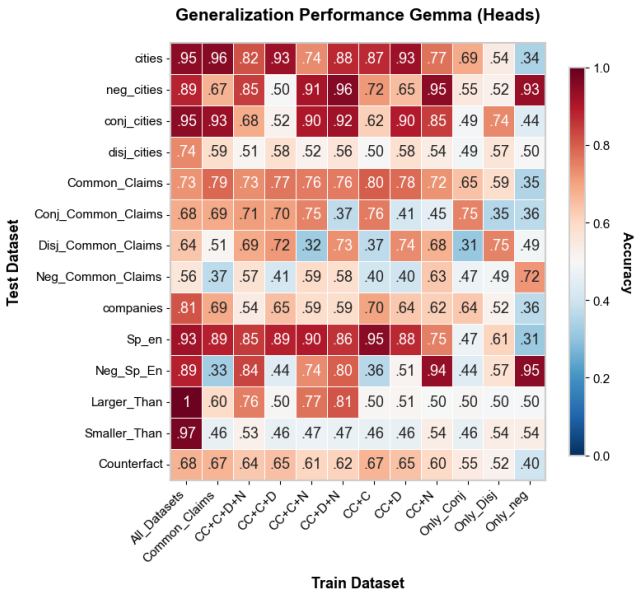
(a) Logical Coherence (Residual)



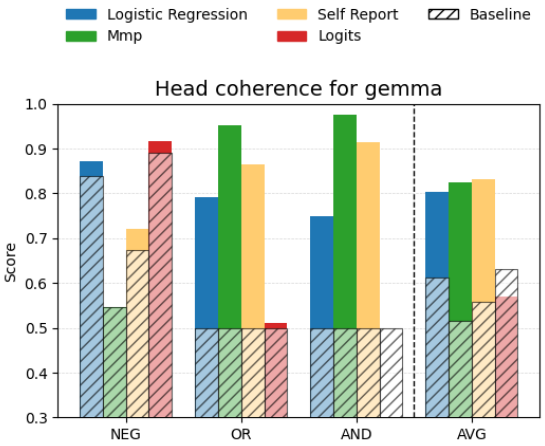
(b) Probabilistic Coherence (Residual)

Figure 37: Coherence results

Results on the attention heads:



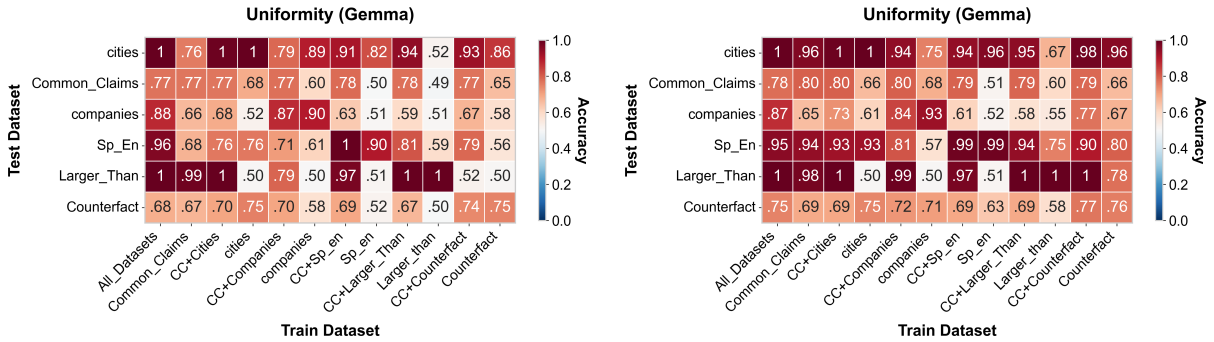
(a) Logical Coherence (Heads)



(b) Probabilistic Coherence (Heads)

Figure 38: Coherence results

Uniformity We report the full cross-domain Uniformity experiment for the model, results extracted from the residual stream (left) and from the attention heads (right).



(a) Cross-domain Uniformity (Residual)

(b) Cross-domain Uniformity (Heads)

Figure 39: Uniformity results

A.4.10 Gemma-2-9B-instruct

We report a full probe sweep on the model's attention heads.

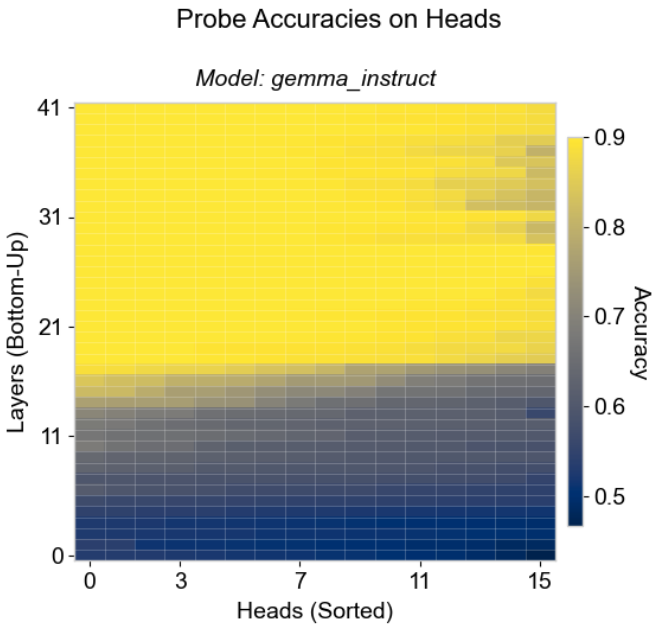


Figure 40: Probe sweep on heads, Gemma-2-9B-instruct

Use Example sweeps we performed on the model, in both directions, and showing sweeps performed on random directions for reference.
Residual steering:

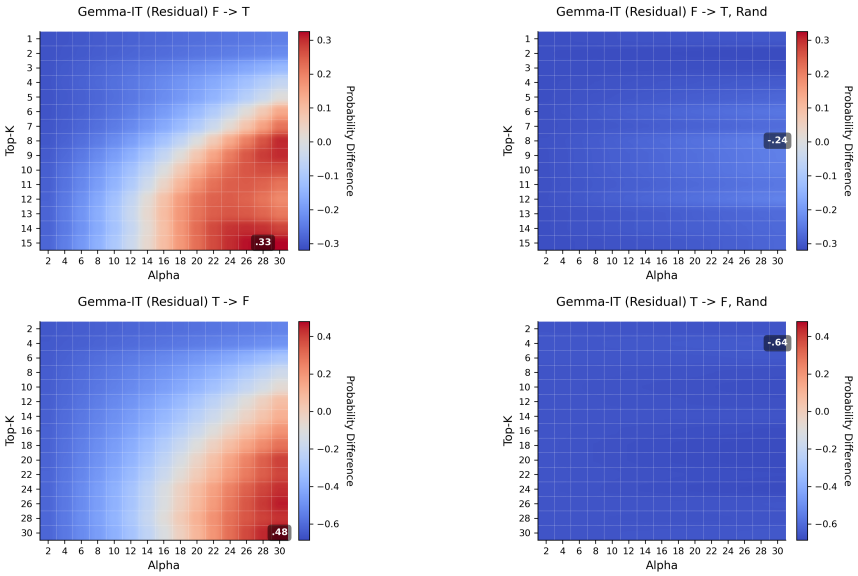


Figure 41: Grid search K, α, Gemma-2-9B-Instruct, Residual. Control sweeps on the right.

Head steering:

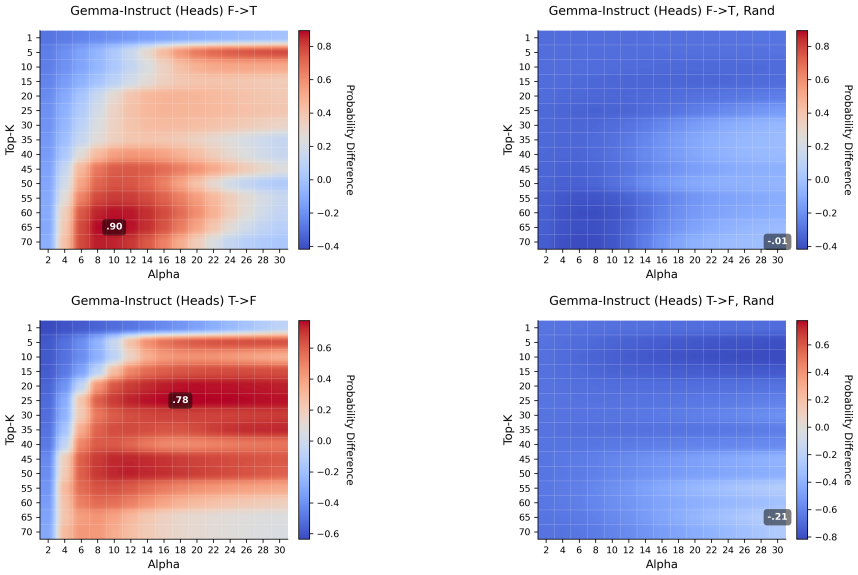
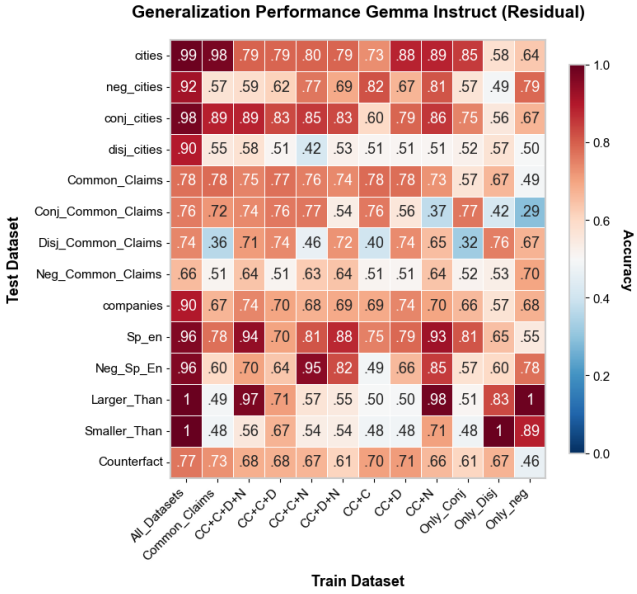


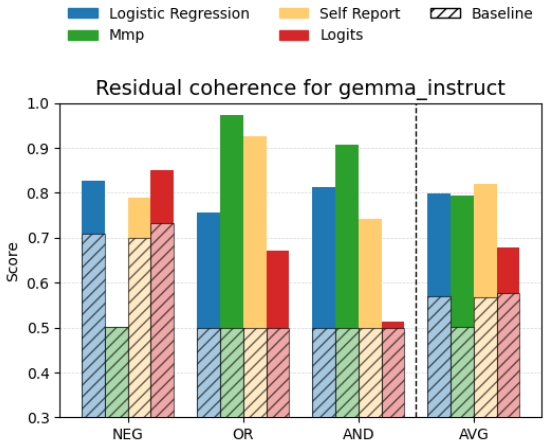
Figure 42: Grid search K, α, Gemma-2-9B-Instruct, Heads. Control sweeps on the right.

Coherence We report a full cross-dataset probing experiment (left) and full results for the probabilistic experiment (right) for the model.

Results on the residual stream:



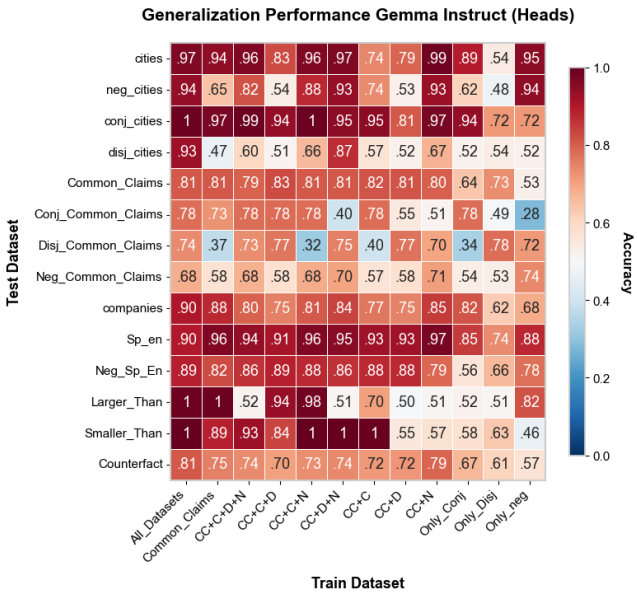
(a) Logical Coherence (Residual)



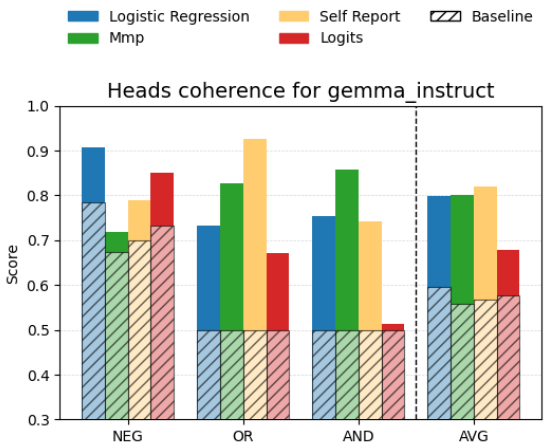
(b) Probabilistic Coherence (Residual)

Figure 43: Coherence results

Results on the attention heads:



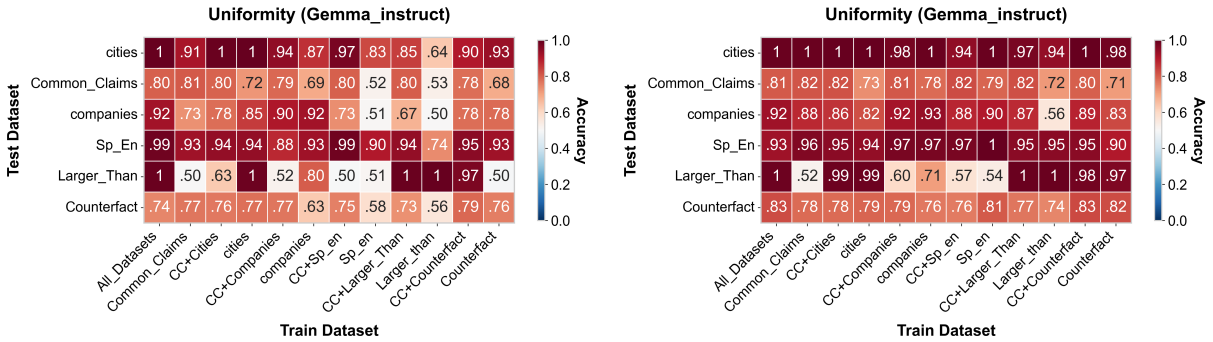
(a) Logical Coherence (Heads)



(b) Probabilistic Coherence (Heads)

Figure 44: Coherence results

Uniformity We report the full cross-domain Uniformity experiment for the model, results extracted from the residual stream (left) and from the attention heads (right).



(a) Cross-domain Uniformity (Residual)

(b) Cross-domain Uniformity (Heads)

Figure 45: Uniformity results