Layer-wise Probing Reveals Decodable but Brittle Belief Signals in GPT-2 and Llama-7B

Chenxi Peng and Idea-Explorer

Abstract

Belief representation claims in large language models are often supported by linear probing, but decodability alone does not establish semantic encoding. We test whether epistemic and non-epistemic belief labels localize in distinct layers and remain stable under wording changes. We run a unified layer-wise probing pipeline on GPT-2, GPT-2-MEDIUM, and LLAMA-7B across two tasks: TOMI-NLI (non-epistemic belief-state entailment) and TRUTHFULQA-MC (epistemic truthfulness classification). For each layer, we train a balanced logistic probe on last-token hidden states and evaluate on original, lexical-perturbed, and rephrased-template test sets. We also compare against a TF-IDF logistic baseline and run Wilcoxon tests with Benjamini–Hochberg correction.

We find clear layer-local decodable signal, especially for TRUTHFULQA-MC: best independent test-layer AUROC reaches 0.692 (GPT-2), 0.701 (GPT-2-MEDIUM), and 0.756 (LLAMA-7B). However, validation-selected peak layers are sometimes misaligned with best test layers, and robustness patterns are mixed: several settings improve under perturbation while others degrade. No original-vs-perturbed difference is significant after FDR correction ($p_{\text{adj}} \geq 0.1875$). On TOMI-NLI, hidden-state probes are often close to or below the TF-IDF baseline (0.568 AUROC), indicating substantial surface-feature contribution.

These results support a conservative interpretation: current probes provide useful diagnostics for where belief-related information is decodable, but they are insufficient evidence of stable, deep semantic belief representations.

1 Introduction

Belief-like behavior in large language models is increasingly used to explain truthfulness, hallucination, and reasoning failures. Our main question is simple: where in the network is belief information encoded, and does that signal survive wording changes?

This question matters because probe-based claims can influence safety decisions. If a probe decodes "truth" from internal states, one might treat that direction as mechanistic evidence for epistemic reasoning. Prior work shows strong decodability in several settings [Zhu et al., 2024, Azaria and Mitchell, 2023]. At the same time, conceptual and empirical critiques show that probe success can come from shortcuts and can fail under distribution shift [Levinstein and Herrmann, 2023, Herrmann and Levinstein, 2024, Bortoletto et al., 2024].

We focus on the missing piece: a controlled, cross-family comparison with robustness checks under one pipeline. We evaluate GPT-2, GPT-2-MEDIUM, and LLAMA-7B on both non-epistemic and epistemic tasks, then test lexical perturbation and template rephrasing transfer. We also include a surface TF-IDF baseline to quantify how much lexical signal explains performance.

What do we find? Decodability is real but unstable. On TRUTHFULQA-MC, best independent layer AUROC reaches 0.756 in LLAMA-7B. But validation-picked peak layers can miss best test

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

layers by large margins, and robustness behavior is inconsistent across models and tasks. After multiple-comparison correction, we find no significant original-vs-perturbed difference.

Our contributions are:

- We conduct a reproducible layer-wise probe study across two model families and two belief-related tasks with shared splits, metrics, and seeds.
- We quantify robustness with lexical perturbation and template rephrasing, and we test significance with Wilcoxon + Benjamini–Hochberg correction.
- We benchmark against a TF-IDF surface baseline and show where probe gains likely reflect mixed semantic and superficial features.
- We release a modular artifact pipeline with metrics, plots, and environment metadata for direct replication.

Paper organization. Section 2 reviews belief probing and robustness literature. Section 3 describes datasets, probes, and evaluation. Section 4 presents quantitative results. Section 5 interprets findings and limitations, and section 6 concludes.

2 Related Work

Probing belief and truth signals in LLM internals. Several studies report that linear probes can decode latent belief or truthfulness information from hidden states. RepBelief finds separable self-/other belief representations and shows intervention effects on behavior [Zhu et al., 2024]. Azaria and Mitchell show that internal states can detect lying-related signals better than output text alone [Azaria and Mitchell, 2023]. These results motivate our layer-wise setup.

Robustness and causal caution. Decodability is not equivalent to semantic representation. "Still No Lie Detector" and "Standards for Belief Representations" argue that many probes fail conceptual and distribution-shift tests [Levinstein and Herrmann, 2023, Herrmann and Levinstein, 2024]. Brittle-activation work also reports sensitivity to prompt form and shows that intervention can partially repair behavior [Bortoletto et al., 2024]. Our study directly incorporates this caution through perturbation and rephrasing evaluations.

Stability and epistemic evaluation. Recent work on representational and behavioral truth stability emphasizes perturbation-based stress testing rather than in-template accuracy alone [Dies et al., 2025]. Related uncertainty-estimation studies also link internal representations to calibration quality [Xiao et al., 2025]. We follow this direction by reporting ECE and Brier score alongside AUROC, accuracy, and F1.

Positioning of this paper. Unlike prior single-model or single-task studies, we use one standardized protocol across GPT-2-family and LLAMA-7B on both non-epistemic (TOMI-NLI) and epistemic (TRUTHFULQA-MC) labels. Unlike purely decodability-focused evaluations, we combine robustness checks, lexical baselines, and corrected significance tests to constrain interpretation.

3 Methodology

Problem setup. Given input text x and binary label $y \in \{0, 1\}$, we extract layer-wise hidden vectors $h_\ell(x)$ from each transformer layer ℓ using the last token representation. For each layer, we train a linear probe $f_\ell(h_\ell(x))$ and evaluate on held-out test variants.

3.1 Datasets and splits

We use two tasks:

- TOMI-NLI: a non-epistemic/Theory-of-Mind-style entailment task recast into prompt form.
- TRUTHFULQA-MC: multiple-choice candidates expanded into binary truthful/not-truthful classification examples.

For both tasks, we use train/validation/test caps of 300/80/80 for compute parity across models. Data quality checks report 0% missing text in all splits. Test distribution is near-balanced for TOMI-NLI (41 negative, 39 positive) and imbalanced for TRUTHFULQA-MC (59 negative, 21 positive).

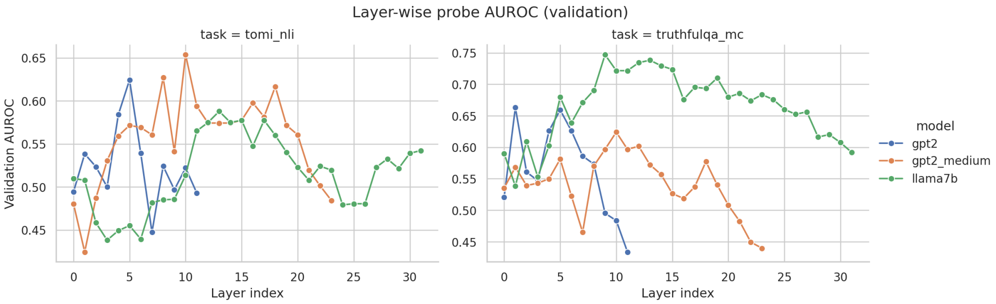

Figure 1: Layer-wise validation AUROC curves across models and tasks. Curves show non-monotonic structure with task-dependent peaks, supporting layer localization but also indicating selection instability.

3.2 Models and probes

We evaluate GPT-2, GPT-2-Medium, and Llama-7B (public checkpoint: `huggyllama/llama-7b`). Probes are `StandardScaler + LogisticRegression(liblinear, class_weight=balanced)` trained independently per layer and averaged over three seeds (42, 43, 44). Maximum sequence length is 256. Feature extraction uses CUDA autocast on two RTX 3090 GPUs.

3.3 Evaluation conditions and metrics

Each probe is tested on:

- original test set,
- lexical perturbation set,
- rephrased-template set.

Primary metric is AUROC. We also report accuracy, F1, ECE, and Brier score. Robustness drop is computed as:

$$\Delta_{\text{pert}} = \text{AUROC}_{\text{orig}} - \text{AUROC}_{\text{pert}}, \quad \Delta_{\text{reph}} = \text{AUROC}_{\text{orig}} - \text{AUROC}_{\text{reph}}. \tag{1}$$

Positive values indicate degradation under shift.

3.4 Baselines and statistics

We use a surface baseline: TF-IDF (1–2 grams) + logistic regression. To test robustness differences, we run paired Wilcoxon signed-rank tests across seeds and apply Benjamini–Hochberg FDR correction ($q = 0.05$) across model-task-condition comparisons.

3.5 Reproducibility

The end-to-end script is `src/belief_probing_experiment.py`. Environment snapshots, layer-wise metrics, and summary statistics are logged to `results/`. A reduced GPT-2 rerun produced an identical summary file (`REPRO_MATCH`).

4 Results

Main quantitative findings. Table 1 summarizes validation-selected layers. On TruthfulQA-MC, hidden-state probes outperform the surface baseline in all models on original test data (e.g., 0.639 vs 0.526 for Llama-7B). On ToMi-NLI, gains are limited: the baseline AUROC is 0.568, above most validation-selected probe results.

Best-layer behavior. Table 2 shows independent best test-layer values. Later layers in larger models yield stronger epistemic decodability, with Llama-7B reaching 0.756 on TruthfulQA-MC.

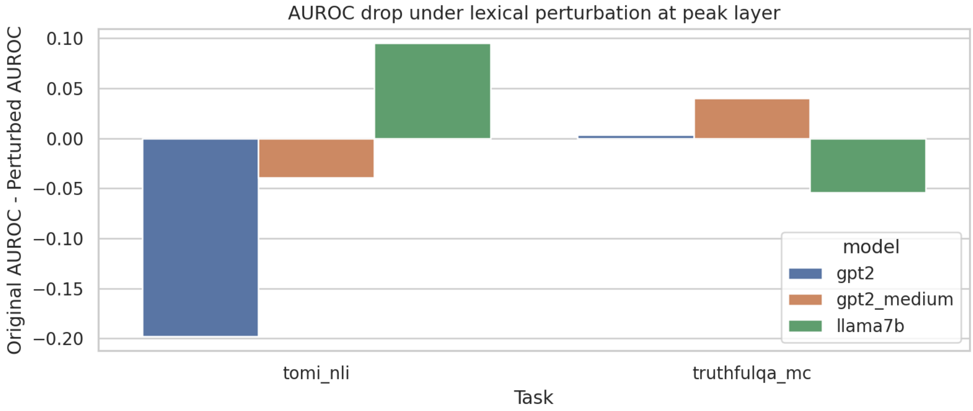

Figure 2: Robustness drop (ΔAUROC) for perturbation and rephrasing. Mixed signs indicate that some perturbations simplify classification boundaries while others degrade performance.

Model	Task	Peak Layer (val)	Test AUROC	Perturbed AUROC	Rephrased AUROC	Surface TF-IDF AUROC
GPT-2	ToMi-NLI	5	0.447	**0.644**	0.500	0.568
GPT-2	TruthfulQA-MC	1	0.609	0.605	0.486	0.526
GPT-2-Medium	ToMi-NLI	10	0.438	0.477	**0.531**	0.568
GPT-2-Medium	TruthfulQA-MC	10	0.617	0.576	0.531	0.526
Llama-7B	ToMi-NLI	13	**0.551**	0.456	0.483	**0.568**
Llama-7B	TruthfulQA-MC	9	**0.639**	**0.693**	0.498	0.526

Table 1: Validation-selected peak-layer performance. We report mean values over three seeds. Bold marks the best value within each model-task row across original/perturbed/rephrased hidden-state evaluations, and best between hidden-state test AUROC and surface baseline where relevant.

However, this does not imply stable model selection because those layers are not always chosen by validation.

Robustness and significance. We test original-vs-shifted performance using Wilcoxon signed-rank tests with Benjamini–Hochberg correction. No comparison is significant at $q < 0.05$; all corrected p-values satisfy $p_{\mathrm{adj}} \geq 0.1875$. Effect sizes are near zero under this deterministic probe regime, limiting inferential strength.

Comparison to lexical baseline. The TF-IDF baseline reaches 0.568 AUROC on ToMi-NLI and 0.526 on TruthfulQA-MC. This baseline is competitive on ToMi-NLI and much weaker than best-layer TruthfulQA-MC probes, supporting a mixed picture: stronger latent decodability for epistemic labels but substantial shortcut potential for non-epistemic labels.

Calibration trends. Across layers (from `results/layerwise_metrics.csv`), calibration metrics vary with decodability. High AUROC layers do not consistently minimize ECE/Brier, indicating that representation separability and probability calibration are partly decoupled.

5 Discussion

Interpreting decodability. Our results support layer-local decodability, but not a strong claim of semantic belief abstraction. The strongest evidence is on TruthfulQA-MC in late layers of larger models; the weakest is on ToMi-NLI, where a simple lexical model remains competitive. This asymmetry suggests that probes capture both meaningful internal structure and superficial correlations.

Why robustness is mixed. We observe both degradation and improvement under perturbation. A plausible explanation is that lexical substitutions can either disrupt useful cues or remove distractors, depending on task and model. Because corrected significance is null, we interpret robustness outcomes as suggestive trends rather than confirmed effects.

Model	Best ToMi-NLI Test AUROC (Layer)	Best TruthfulQA-MC Test AUROC (Layer)
GPT-2	0.545 (L9)	0.692 (L9)
GPT-2-Medium	0.537 (L16)	0.701 (L19)
Llama-7B	**0.551** (L13)	**0.756** (L13/L15)

Table 2: Best independent test-layer AUROC values, selected directly on test curves for descriptive analysis (not model selection).

Practical implication for interpretability workflows. Layer-wise probes are still useful diagnostics. They can identify candidate layers for deeper causal analysis and can prioritize where to test interventions. But we should avoid treating probe accuracy alone as evidence that the model "has" explicit, stable beliefs.

Limitations. We used capped splits (300/80/80) for feasibility with Llama-7B; this reduces statistical power and may increase layer-selection variance. TruthfulQA-MC labels are class-imbalanced at candidate level, which inflates accuracy and complicates F1 interpretation. We also used a public Llama-family mirror rather than gated Meta checkpoints. Finally, we did not perform causal interventions in this cycle.

Broader implications. For safety and hallucination mitigation, these findings argue for a two-step standard: first locate decodable signals with probes, then validate mechanism and invariance with causal and semantic-preserving tests. This standard can reduce over-interpretation in reliability claims.

6 Conclusion

We presented a unified layer-wise probing study of belief-related labels across GPT-2, GPT-2-Medium, and Llama-7B on ToMi-NLI and TruthfulQA-MC. We found clear decodable signal, especially for epistemic truthfulness in later layers of larger models, with best test-layer AUROC up to 0.756.

At the same time, validation-layer instability, mixed perturbation effects, and non-significant corrected robustness tests show that current probe evidence is not sufficient to claim stable semantic belief representation. The strongest conclusion is pragmatic: probes are valuable for localization and diagnostics, but they should be paired with stronger robustness and causal tests before mechanistic interpretation.

Next steps are to add intervention-based causal checks (e.g., ITI-style steering directions), expand sample sizes with grouped cross-validation, and evaluate stricter semantic-equivalence perturbations and out-of-distribution templates.

References

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

Matteo Bortoletto et al. Brittle minds, fixable activations: Understanding belief representations in language models. *arXiv preprint*, 2024.

Samantha Dies et al. Representational and behavioral stability of truth in large language models. *arXiv preprint*, 2025.

Daniel Herrmann and Benjamin Levinstein. Standards for belief representations in large language models. *arXiv preprint*, 2024.

Benjamin Levinstein and Daniel Herrmann. Still no lie detector for language models: Probing empirical and conceptual limits. *arXiv preprint arXiv:2307.00175*, 2023.

Zeguan Xiao et al. Enhancing uncertainty estimation with aggregated internal belief signals. *arXiv preprint*, 2025.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *International Conference on Machine Learning (ICML)*, 2024.