Uncertainty Beneath the Surface: A Study of How Language Models Enocde Linguistic Uncertainty

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly being deployed in high-stakes domains where the ability to reason under uncertainty is critical. Despite recent progress in evaluating factuality and calibration, little is known about how LLMs internally represent epistemic modality: linguistic cues that signal speaker uncertainty (e.g., "might", "probably"). To our knowledge, this work presents one of the first systematic investigations into *whether and how LLMs encode sensitivity to epistemic modality in their activation space*. We curate a contrastive multiple choice dataset of *3114* sentence pairs that vary in epistemic certainty and introduce a probing framework to quantify activation-level differences between certain and uncertain prompts. We further propose Model Sensitivity to Uncertainty (MSU), a layerwise metric that captures representational shifts attributable to epistemic cues. Our findings suggest that LLMs exhibit measurable and layer-specific sensitivity to epistemic modality, raising implications for their deployment in sensitive decision-making contexts.Code can be accessed at https://anonymous.4open.science/r/Uncertainty-Beneath-the-Surface-11D1.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in high-stakes domains such as law, medicine, and public policy—settings that demand not only accurate outputs but also the ability to reason responsibly under uncertainty. While previous research has emphasized calibrating model confidence and evaluating output probabilities, comparatively little is known about how LLMs internally represent input-side uncertainty—particularly linguistic uncertainty expressed through epistemic modality (e.g., might, could, probably). Figure 1 shows that even minimal shifts in modality, such as replacing 'should' with 'could', lead to consistent differences in model generations, despite all other input remaining constant. These variations are not artifacts of sampling noise; rather, they indicate a systematic and grounded sensitivity to uncertainty cues. This raises important questions: How are such epistemic signals encoded across the layers of a model? Are these representations stable across model variants? Understanding the internal treatment of linguistic uncertainty is critical for building trustworthy, transparent systems, especially in applications where appropriately responding to hedged or speculative language can directly affect outcomes and user trust.

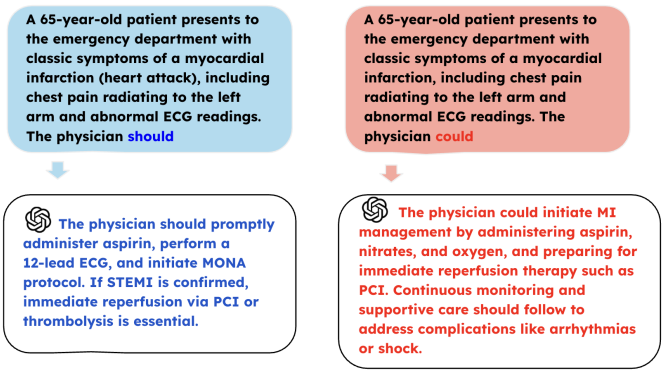


Figure 1: Although the prompt pairs differ only in epistemic modality (*should vs could*), the responses vary: those prompted with *could* tend to offer a broader range of medical possibilities and are more open-ended compared to those ending with *should*. This could imply how the model interprets linguistic uncertainty.

1.1 Uncertainty in LLM Outputs

Much of the recent work on LLM uncertainty has focused on model outputs, through calibration techniques (Desai and Durrett, 2020), truthfulness under uncertainty (Lin et al., 2022), or confidence alignment with ground-truth (Ghafouri et al., 2024). While these approaches provide useful diagnostics on model predictions, they do not investigate how input uncertainty is internally represented or

whether models distinguish between certain and uncertain prompts at a representational level.

1.2 Epistemics and Modal Reasoning

Several recent studies have examined the role of modal verbs in model reasoning. (Holliday et al., 2024)) show that LLMs often struggle with logical tasks involving modal operators, suggesting a lack of systematic reasoning with modality.

(Zhou et al., 2023) analyze use of epistemic markers in LLM-generated text, showing large effects on accuracy depending on whether uncertainty or certainty markers are used, although they do not study how these markers are internally represented in neural activations. Similarly, (Lee et al., 2025) show that LLM-based evaluators are systematically biased against responses containing expressions of uncertainty.

Our work complements these efforts by offering a mechanistic perspective on how epistemic modality is encoded inside models, using probing over activation spaces. In contrast to prior work focusing on usage or output alignment, we provide empirical evidence of internal sensitivity to epistemic variation.

1.3 Probing Internal Representations

A growing body of mechanistic interpretability research seeks to understand LLM internals by intervening in the *activation space*. One core technique is activation patching, also known as causal mediation or causal tracing, which substitutes hidden activations from a clean forward pass into a corrupted run, thereby identifying components causally responsible for specific behaviors (Vig et al., 2020). Building on this, path patching refines the approach by restricting interventions to specific computational paths, enabling finer-grained localization of functional subcircuits (Goldowsky-Dill et al., 2023). Complementary methods include causal scrubbing (Chan et al., 2022), which tests whether abstract circuits maintain function across structural perturbations, and automated circuit discovery frameworks such as ACDC (Conmy et al., 2023).

In this work, we pose the question – *Are large language models sensitive to epistemic modality in their input?* This question is investigated by contrasting representations elicited by semantically similar prompts that differ only in the degree of certainty they convey. As depicted in Figure 2, we curate a multiple-choice dataset of 3,114 sentence pairs that differ in their expression of certainty and examine how these variations influence the model's internal activation space.

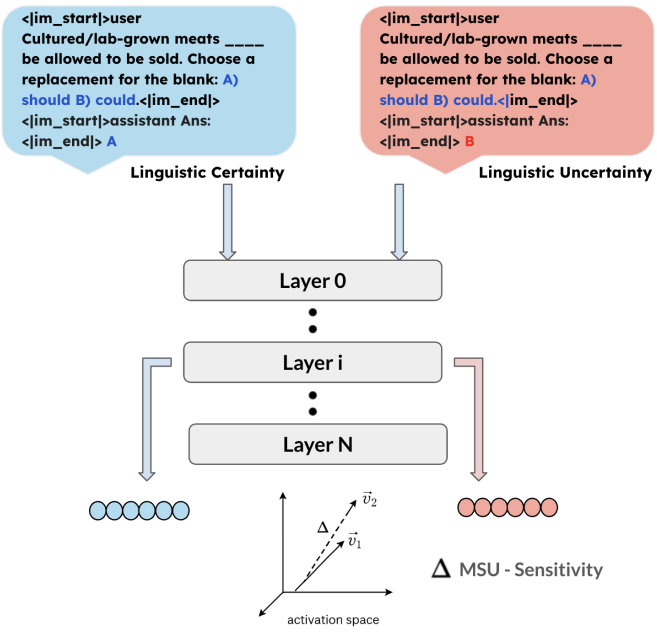


Figure 2: Paired inputs with linguistic certainty and uncertainty are passed through the model. Layerwise activation differences (Δ) are used to compute MSU, capturing the model's sensitivity to uncertainty across layers.

To quantify such effects, we introduce a novel metric, **Model Sensitivity to Uncertainty (MSU)**, which captures the representational shift induced by epistemic cues at each layer of the model. Through this lens, we assess whether and where in the model epistemic modality is internally encoded. Our key contributions are as follows:

- We propose a probing framework to assess whether LLMs encode epistemic modality through changes in their activation space.
- We introduce **MSU**, a layerwise metric for quantifying model sensitivity to linguistic uncertainty.
- We release the dataset of 3,114 sentence pairs depicting linguistic certainty and uncertainty.

2 Dataset

Unlike contrastive datasets used in prior work (Rimsky et al., 2024), data was not constructed to be semantically opposite or adversarial. Instead, they

vary along a fine-grained linguistic axis of uncertainty, making them well-suited for probing representational sensitivity to uncertainty.

The dataset used in this work is derived from claims in the `Anthropic/Persuasion` (Durmus et al., 2024) corpus. Sentences containing modal verbs such as "should" and "must" were identified and programmatically masked using the `pandas` (Wes McKinney, 2010), `NumPy` (Harris et al., 2020), and `NLTK` (Bird and Loper, 2004) libraries. These masked positions were then filled with controlled multiple-choice options representing either certain or uncertain linguistic modality (e.g., "should" vs. "could").

For example, consider a sample from the `Anthropic/Persuasion` dataset (Durmus et al., 2024):

Example Prompt

Original: *"Governments and technology companies must do more to protect online privacy and security."*
Modified prompt:

```
<|im_start|>user
Governments and technology companies
[MASK] do more to protect online
privacy and security.
Choose a replacement for the MASK.
A) Must B) Might
<|im_end|>
```

One of the pair would be appended with A, and the other with B, as shown in Figure 2 to form the linguistic certain-uncertain pair. The changes made to the dataset ensure the following:

Controlled Variation: The only systematic difference across each pair is the use of uncertainty markers (e.g., "might", "possibly", "maybe"), ensuring minimal lexical confounds.

Semantic Stability: As the core semantics is preserved, any variation in activation vectors can be more confidently attributed to modality, not content drift. For examples of these some of these changes, refer Appendix A.4.

Each sentence was paired with two variants, one expressing certainty and one expressing uncertainty by appending the starting letter of the option at the end of the instruction, resulting in 3,114 samples per condition, and a total of 6,228 examples.

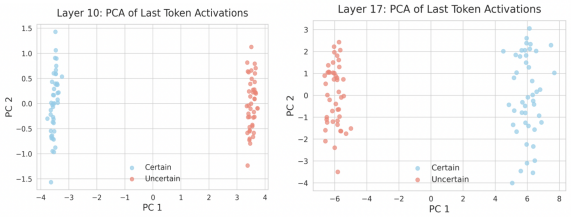


Figure 3: PCA plots of the last token activations of layers 10 and 17 for Qwen2.5-0.5B-Instruct model. A geometric inversion can be observed in the Projections for the Uncertain and Certain input activations.

3 Evaluation Setup

We conduct our analysis on three small-scale language models: Qwen2.5-0.5B-Instruct (Team, 2024), Qwen1.5-0.5B-Chat (Bai et al., 2023), and LLaMA-3.2-1B-Instruct(Grattafiori et al., 2024). Details on the number of parameters, layers, model sources, and the reason for selecting those are provided in Appendix A.1.

Can Linguistic Uncertainty be probed in the Activation Space?

Principal Component Analysis (PCA) is applied layer by layer to the model activation vectors to examine whether the uncertainty is encoded linearly separable. For each layer, activations of certain and uncertain examples are projected onto the two main components using Scikit-learn (Pedregosa et al., 2011), allowing visualization of potential clustering patterns. This analysis serves as a diagnostic to assess the validity of the data set to study linguistic uncertainty (refer to appendix A.3).

Results

In all three models, we observe clear clustering patterns that validate the separability of epistemic modalities in the model's internal representation space. Interestingly, in the later layers of the Instruct model (Layer 17 and 23), as well as in Layers 13 through 15 of the Chat model, we detect a geometric inversion in the PCA projections: the cluster corresponding to uncertain statements flips position relative to that of certain statements along the primary axes. *This inversion suggests a deeper semantic reorganization in the latent space, potentially signaling a transition from syntactic or lexical representation toward task-relevant abstractions of linguistic uncertainty.* These structured shifts reinforce our hypothesis that uncertainty is not only linearly encoded but is also semantically

recontextualized in deeper layers, underscoring the interpretability and representational richness of the models under investigation.

How does Sensitivity to Linguistic Uncertainty change across layers?

We extract activation vectors from all transformer layers using the TransformerLens library (Nanda and Bloom, 2022), which allows access to cached internal states without modifying the model architecture. Specifically, for each input pair, consisting of a certain and an epistemically uncertain variant, we record the residual stream activations at the final token position, where model output is most strongly influenced.

To quantify the representational shift induced by linguistic uncertainty, we introduce a metric called *Model Sensitivity to Uncertainty (MSU)*. This metric captures the average distance between the representations of the certain and uncertain variants of each input sentence pair.

Formally, for a given model layer ℓ, we define MSU as:

$$\text{MSU}^{(\ell)} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{h}_i^{(\ell,\text{certain})} - \mathbf{h}_i^{(\ell,\text{uncertain})} \right\|_2 \tag{1}$$

where $\mathbf{h}_i^{(\ell,\cdot)}$ denotes the activation vector obtained from layer ℓ for the i-th input in its certain or uncertain form, and N is the total number of input pairs.

MSU provides a quantitative estimate of how much linguistic uncertainty perturbs the model's internal representations. Larger MSU values indicate greater sensitivity to uncertainty at that layer.

Results

Across three models, we observe a strikingly consistent trend: the MSU scores increase monotonically with depth, indicating that sensitivity to epistemic uncertainty is a progressively emerging phenomenon in the transformer stack (Figure 5). Later layers exhibit substantially higher sensitivity to epistemic modals than early layers, indicating that semantic distinctions introduced by modality are progressively amplified across depth. This aligns with prior work indicating that deeper layers are responsible for encoding abstract, compositional semantics and final decision-making(Zhao et al., 2024). Our results suggest that epistemic uncertainty is treated as a high-level semantic feature.

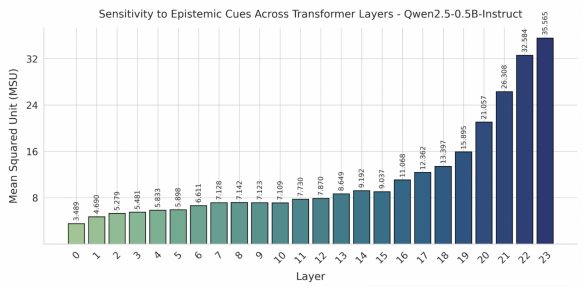


Figure 4: Layer-wise MSU scores for Qwen2.5-0.5B-Instruct, indicate progressively increasing scores across layers suggesting that later layers are responsible for encoding uncertainties in language

Model	Average MSU
LLaMA 3.2–1B	9.361
Qwen1.5–0.5B-Chat	**16.968**
Qwen2.5–0.5B-Instruct	11.520

Table 1: Average MSU values across transformer layers for each model. Qwen1.5-0.5B-Chat comes out to be the most sensitive to linguistic uncertainty which is then followed by Qwen2.5–0.5B-Instruct and LLaMA 3.2–1B, among the 3 tested models.

4 Conclusion

Our investigation reveals that the encoding of epistemic uncertainty is a distributed and emergent property of deep transformer architectures. Rather than residing in isolated layers, epistemic modality unfolds progressively, peaking in the final layers—across models and variants alike. By proposing the Mean Sensitivity to Uncertainty (MSU) metric, we provide a targeted lens into this phenomenon. This structural consistency underscores a deeper semantic organization within LLMs and opens new pathways for designing models that are not only more interpretable, but also more epistemically aware.

Limitations

This study takes an initial step toward understanding how language models encode linguistic uncertainty. Our findings are based on a limited set of instruction-tuned models, so their generality across architectures, sizes, and pretraining paradigms remains uncertain.

Future work should expand to include diverse model types, like non-instruct, multilingual, domain-specific and inputs namely, varied modal verbs, syntax, discourse). A key direction is localizing uncertainty to specific neurons or attention

heads, and examining how internal uncertainty signals (e.g., logits, entropy) relate to output confidence and calibration.

References

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

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing: a method for rigorously testing interpretability hypotheses.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the persuasiveness of language models.

Bijean Ghafouri, Shahrad Mohammadzadeh, James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, Mayank Goel, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2024. Epistemic integrity in large language models. In *Neurips Safe Generative AI Workshop 2024*.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *Preprint*, arXiv:2304.05969.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, ..., and 1 others. 2024. The llama 3 herd of models. https://arxiv.org/abs/2407.21783. ArXiv:2407.21783.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and 7 others. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Wesley H. Holliday, Matthew Mandelkern, and Cedegao E. Zhang. 2024. Conditional and modal reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3800–3821, Miami, Florida, USA. Association for Computational Linguistics.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of the 2025 NAACL-Long Papers*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Zheng Zhao, Yftah Ziser, and Shay B Cohen. 2024. Layer by layer: Uncovering where multi-task learning happens in instruction-tuned large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15195–15214, Miami, Florida, USA. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. https://arxiv.org/abs/2302.13439. ArXiv:2302.13439.

A Appendix

A.1 Model Specifications

These models were selected to cover a range of instruction-tuned and chat-oriented variants with varying parameter counts while maintaining manageable computational overhead. All models provide access to internal activations, which is crucial for our layer-wise representation analysis. ageable computational overhead. Table 2 details the model sizes and number of layers for the language models used in our experiments. All internal activations were accessed using the `TransformerLens` (Nanda and Bloom, 2022) library.

Model	Parameters	Layers
Qwen2.5-0.5B-Instruct	391M	24
Qwen1.5-0.5B-Chat	308M	24
Llama-3.2-1B-Instruct	1.1B	16

Table 2: Model sizes, number of layers, and sources for LLM variants used in our analysis.

A.2 Layer-wise Sensitivity to Linguistic Uncertainty

Transformer-based language models adopt a deep, autoregressive architecture in which representations are refined layer by layer—gradually building from lexical cues to nuanced semantic abstraction. This layered structure raises a key question: at which depth is linguistic uncertainty, especially that conveyed by epistemic modals (e.g., must, might, should, could), most saliently represented? To investigate this, we analyze the Mean Sensitivity to Uncertainty (MSU) across transformer layers for three Qwen2.5-0.5B variants—Chat, Instruct, and Base—as well as the LLaMA 3.2-1B-Instruct model.

Across all models, a consistent pattern emerges: early layers (e.g., Layers 0–5) demonstrate low sensitivity to uncertainty (MSU $\approx$ 2.5–6), reflecting a focus on surface-level representations. In contrast, later layers (e.g., Layers 13–23) show a sharp rise in MSU, often surpassing 30, suggesting that deeper layers increasingly encode and amplify epistemic cues. For instance, in LLaMA 3.2-1B-Instruct, MSU climbs from 2.67 at Layer 0 to 31.96 at Layer 15. Among the Qwen variants, the Chat model displays the highest overall sensitivity, with elevated MSU values throughout the stack, indicating a heightened responsiveness to modal uncertainty. These findings support the hypothesis that the semantic abstraction required to capture epistemic nuance is a late-stage phenomenon within the transformer hierarchy.

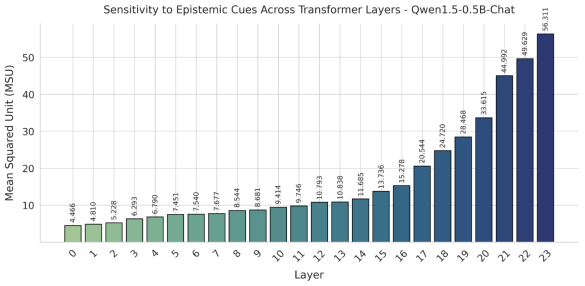


Figure 5: Layer-wise MSU scores for Qwen2.5-0.5B-Chat. Among all variants, the Chat model exhibits the highest sensitivity to linguistic uncertainty, with a steep increase in MSU across deeper layers, reaching a score of .

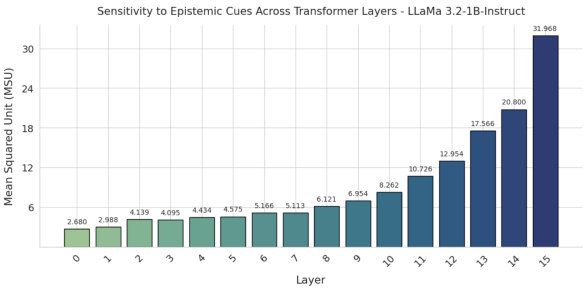


Figure 6: Layer-wise MSU scores for LLaMA 3.2-1B-Instruct, illustrating a steady increase in sensitivity to epistemic uncertainty across layers. The scores begin modestly in early layers (e.g., 2.68 at Layer 0) and rise sharply in deeper layers, peaking at 31.96 in Layer 15. This trend supports the hypothesis that later layers in autoregressive transformers are more attuned to modeling linguistic uncertainty.

A.3 PCA-analysis

We observe a noticeable shift in the principal components of internal representations in the deeper layers of all models A.1. (Figures: 8a) Specifically, while earlier and mid-level layers exhibit stable projection patterns, the final layers display a reorientation in the direction of PC1 and PC2. This structural transition likely reflects a late-stage reorganization of semantic or epistemic features, where uncertainty-related signals become more linearly separable or concentrated. Such emergent behavior may indicate that LLMs progressively

consolidate abstract modality cues toward the final layers, where decision-critical information is encoded. This suggests that the geometry of representations—not just their magnitude—may carry functional signals related to epistemic reasoning.

A.4 Dataset Examples

Example Prompt

Original: *"Social media companies should be required to label AI-generated content"* Modified prompt - Certain:

```
<|im_start|>user Social media
companies [MASK] be required
to label AI-generated content.
Choose a replacement. A) should
B) could.<|im_end|>
<|im_start|>assistant
Ans: <|im_end|>(A
```

Modified prompt - Uncertain:

```
<|im_start|>user
Social media companies [MASK] be required
to label AI-generated content.
Choose a replacement. A) should
B) could.<|im_end|>
<|im_start|>assistant
Ans: <|im_end|>(B
```

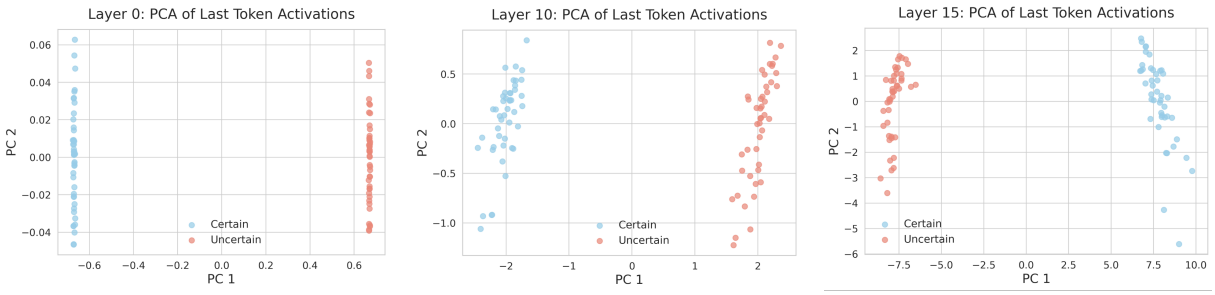
Example Prompt

Original: *"Individuals must take responsibility for online privacy without excessive government mandates."* Modified prompt - Certain:

```
<|im_start|>user
Individuals ____ take responsibility
for online privacy without excessive government ma
replacement. A) must B) might.
<|im_end|>
<|im_start|>assistant
Ans: <|im_end|>(A
```

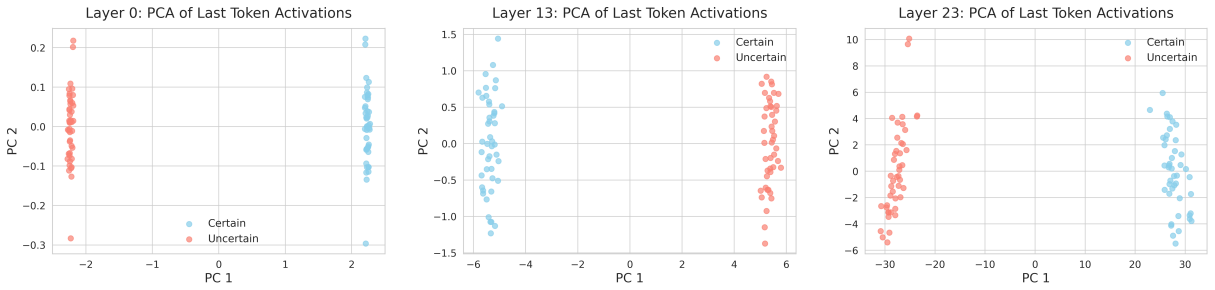
Modified prompt - Uncertain:

```
<|im_start|>user
Individuals ____ take responsibility
for online privacy without excessive government ma
replacement. A) must B) might.
<|im_end|>
<|im_start|>assistant
Ans: <|im_end|>(B
```



(a) Layer 0 (b) Layer 10 (c) Layer 15

Figure 7: PCA projections of internal representations at different layers of the LLaMA 3.2–1B model. While early and mid layers exhibit relatively stable clustering patterns, the final layer shows a notable shift in the orientation of the principal components, suggesting a reorganization of the representational space. This directional change in PC1 vs. PC2 may reflect the model's transition from encoding general contextual features to more task-specific or decision-relevant information.



(a) Layer 0 (b) Layer 10 (c) Layer 15

Figure 8: PCA projections of internal representations at different layers of the Qwen1.5-0.5B-Chat model. Most layers display a consistent structure in the representational space; however, a marked shift in the direction of PC1 vs. PC2 emerges between layers 13 and 15. This transition suggests that epistemic information becomes reorganized or amplified in deeper layers, aligning with the model's increasing sensitivity to uncertainty.