Arizona State University

Position: Uncertainty Quantification in LLMs is Just Unsupervised Clustering

Tiejin Chen, Longchao Da, Xiaoou Liu, Hua Wei

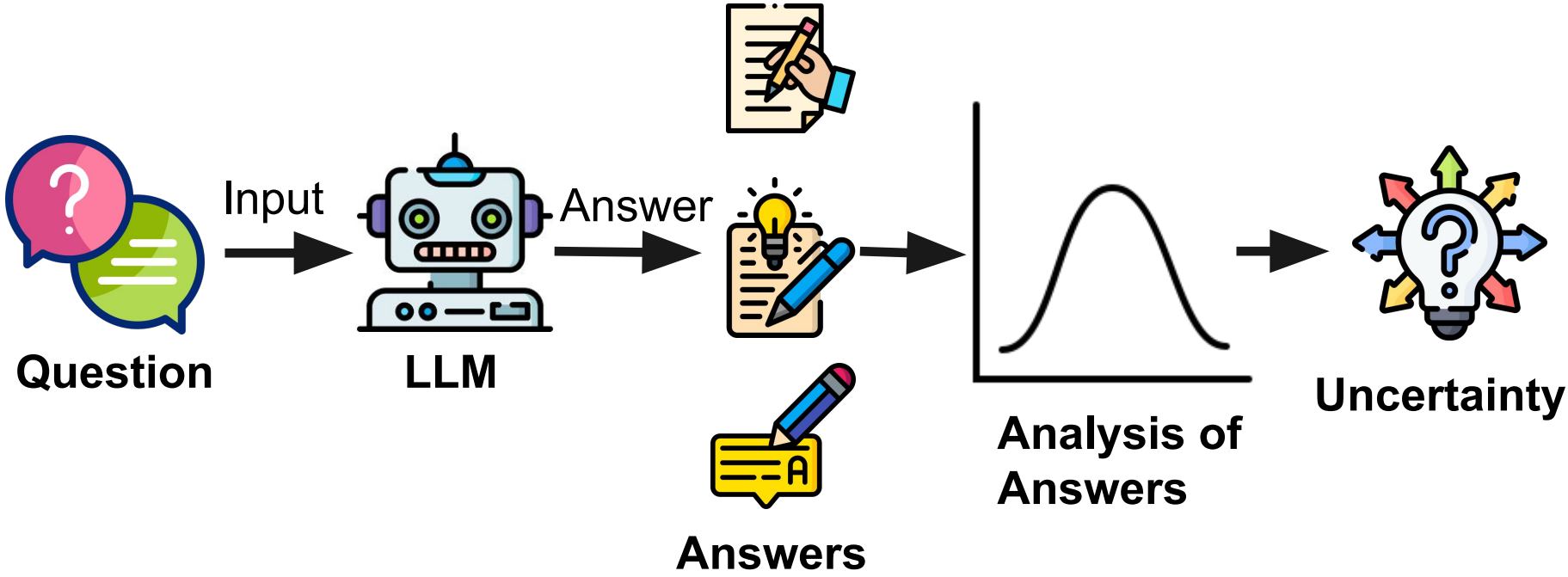
Arizona State University

ICML International Conference On Machine Learning

Copyright © 2026 Arizona Board of Regents

Background

What is Mainstream Uncertainty Quantification in LLMs?



For evaluation, uncertainty is used as the binary classifier to detect whether the answer is correct or not. A higher AUROC indicates a better uncertainty.

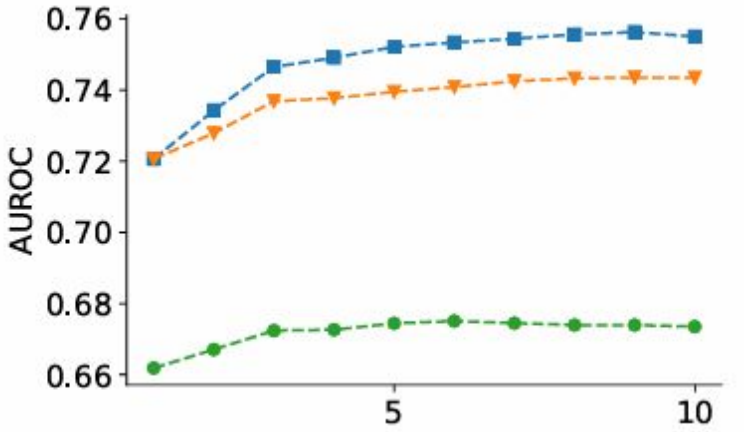
Liu, Xiaoou, et al. "Uncertainty quantification and confidence calibration in large language models: A survey." Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025.

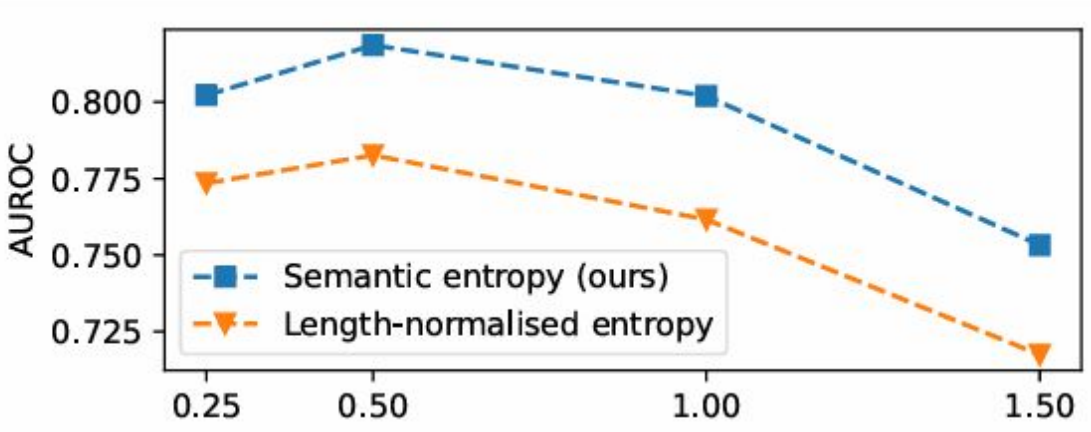
Copyright © 2026 Arizona Board of Regents

Background

What is Mainstream Uncertainty Quantification in LLMs?

Uncertainty Tutorial





The sampling-based method performs much better than entropy.

Kuhn, Lorenz, Yarin Gal, and Sebastian Farquhar. "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation." The Eleventh International Conference on Learning Representations.

Copyright © 2026 Arizona Board of Regents

Position

Mainstream UQ in LLMs is basically just an unsupervised clustering problem. It only measures internal consistency and fails to act as a real safeguard.

Copyright © 2026 Arizona Board of Regents

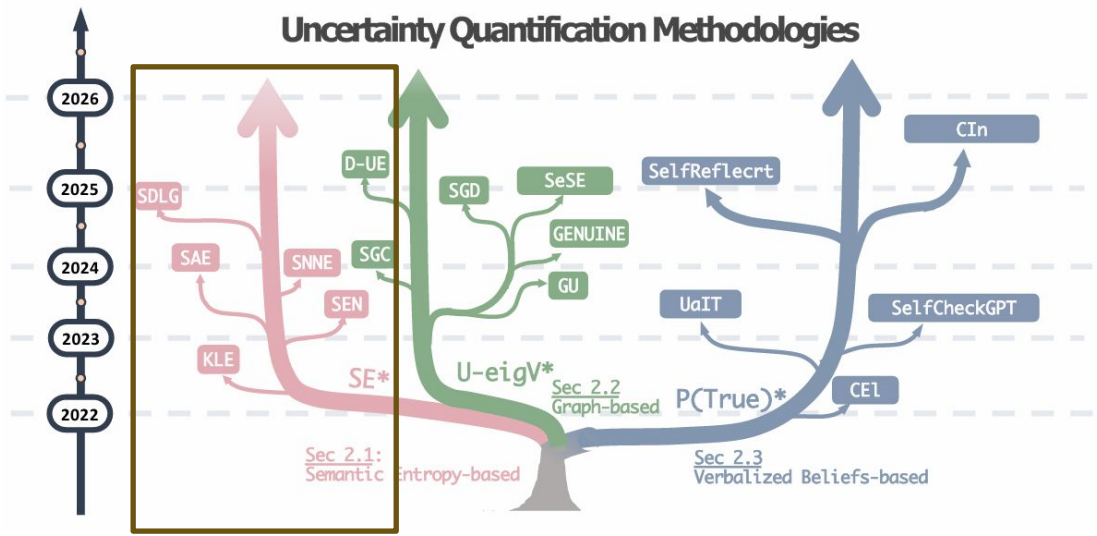
Why is it Clustering?

- **Explicit Clustering**: Semantic Entropy and Its Variants
- **Implicit Clustering:** Graph-based Methods
- **Latent Confidence Clustering:** P(True)

Copyright © 2026 Arizona Board of Regents

Why is it Clustering?

Explicit Clustering



Kuhn, Lorenz, Yarin Gal, and Sebastian Farquhar. "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation." The Eleventh International Conference on Learning Representations.

McCabe, Lucas H., et al. "Estimating Semantic Alphabet Size for LLM Uncertainty Quantification." The Fourteenth International Conference on Learning Representations.

Ma, Huan, et al. "Semantic energy: Detecting llm hallucination beyond entropy." arXiv preprint arXiv:2508.14496 (2025).

Nikitin, Alexander, et al. "Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities." Advances in Neural Information Processing Systems 37 (2024): 8901-8929.

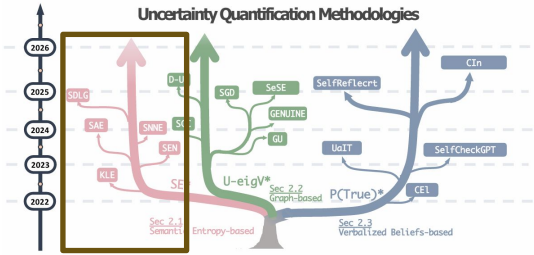
Aichberger, Lukas, et al. "Improving uncertainty estimation through semantically diverse language generation." International Conference on Learning Representations. Vol. 2025. 2025.

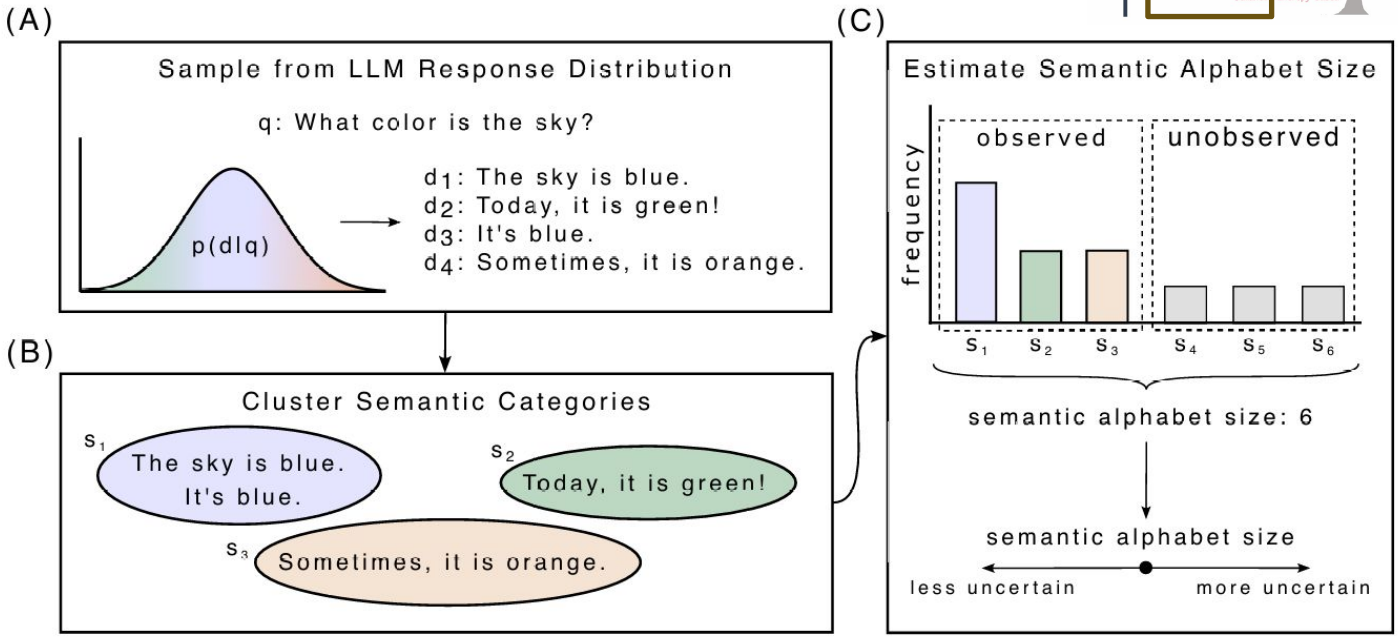
Vashurin, Roman, et al. "Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in llms." Advances in Neural Information Processing Systems 38 (2026): 106236-106281.

Copyright © 2026 Arizona Board of Regents

Why is it Clustering?

Explicit Clustering





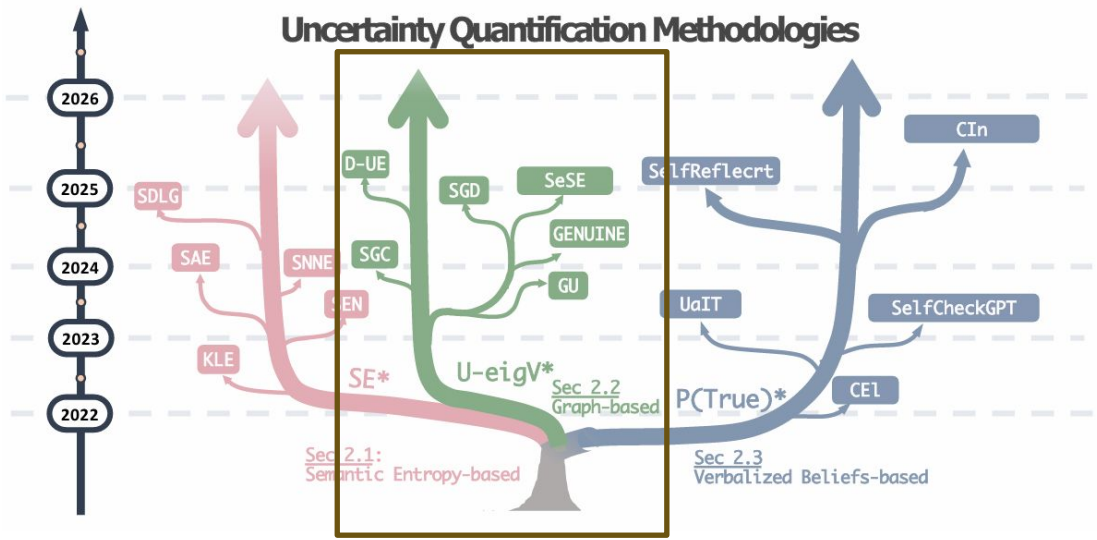
- Explicitly conduct clustering based on semantics.
- The clustering of semantics will influence the final uncertainty.

McCabe, Lucas H., et al. "Estimating Semantic Alphabet Size for LLM Uncertainty Quantification." The Fourteenth International Conference on Learning Representations.

Copyright © 2026 Arizona Board of Regents

Why is it Clustering?

Implicit Clustering



Lin, Zhen, Shubhendu Trivedi, and Jimeng Sun. "Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models." Transactions on Machine Learning Research.

Da, Longchao, et al. "Llm uncertainty quantification through directional entailment graph and claim level response augmentation." arXiv preprint arXiv:2407.00994 (2024).

Da, Longchao, et al. "Understanding the Uncertainty of LLM Explanations: A Perspective Based on Reasoning Topology." Second Conference on Language Modeling.

Liu, Xiaoou, et al. "Diagnosing Multi-step Reasoning Failures in Black-box LLMs via Stepwise Confidence Attribution.“ International Conference on Machine Learning. 2026.

Li, Zhaoye, et al. "Uncertainty Quantification for Black-Box LLMs via Star Graphs Connectivity: Exploring Alternatives for Semantic Density." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer Nature Switzerland, 2025.

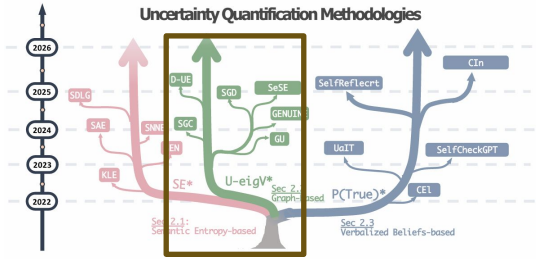
Jiang, Mingjian, et al. "Graph-based uncertainty metrics for long-form language model generations." Advances in Neural Information Processing Systems 37 (2024): 32980-33006.

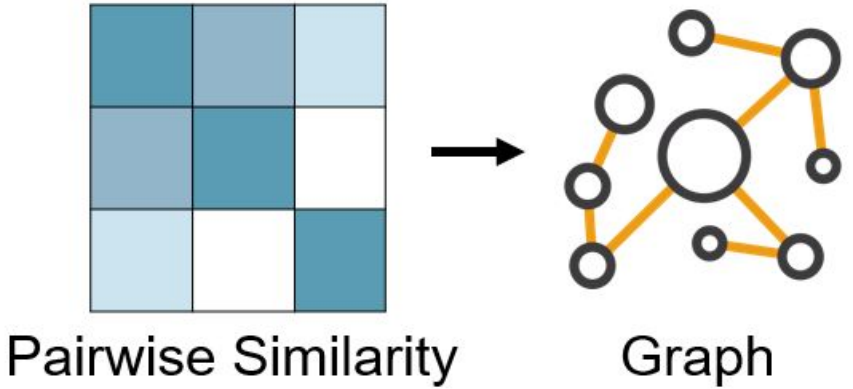
Wang, Tuo, et al. "GENUINE: Graph Enhanced Multi-level Uncertainty Estimation for Large Language Models." Findings of the Association for Computational Linguistics: EMNLP 2025. 2025.

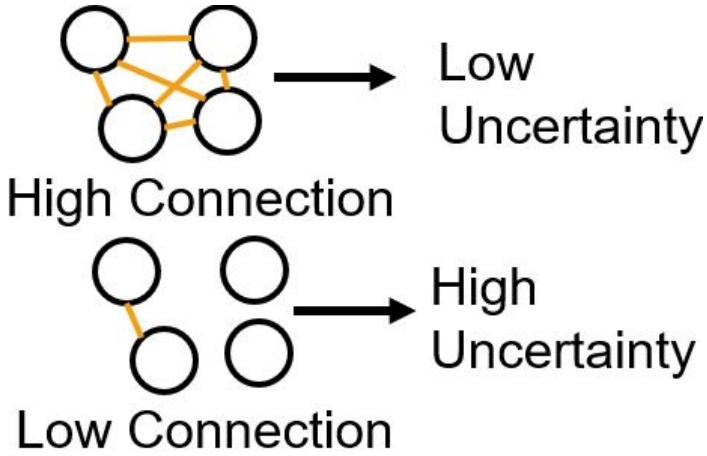
Copyright © 2026 Arizona Board of Regents

Why is it Clustering?

Implicit Clustering





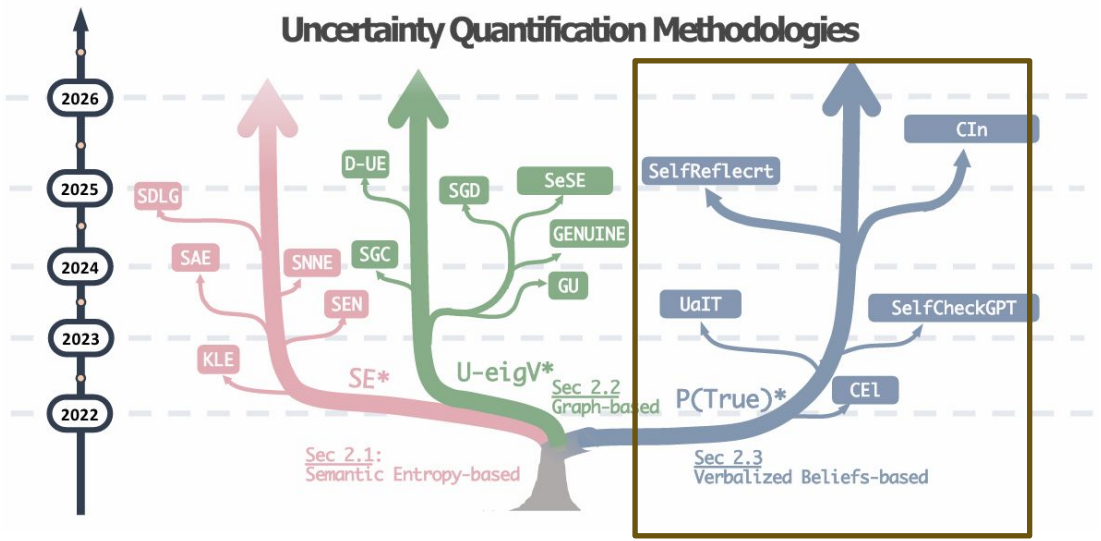


- Each node is a different answer.
- It can be viewed as clustering because these methods normally mathematically estimate how many distinct, coherent groups the model's responses divide into.

Copyright © 2026 Arizona Board of Regents

Why is it Clustering?

Explicit Clustering

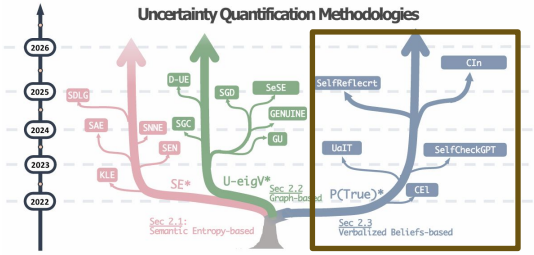


Kadavath, Saurav, et al. "Language models (mostly) know what they know." arXiv preprint arXiv:2207.05221 (2022).
Xiong, Miao, et al. "Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms." International Conference on Learning Representations. Vol. 2024. 2024.
Xi, Tengxiao, Chen Wang, and Jiajun Zhang. "Confidence Introspection: A Self-reflection Method for Reliable and Helpful Large Language Models." IEEE Transactions on Audio, Speech and Language Processing (2026).

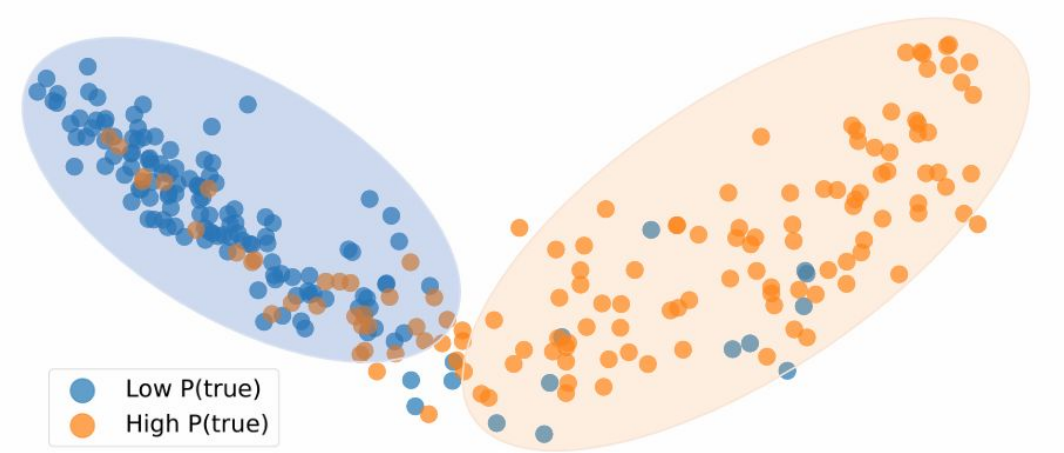
Copyright © 2026 Arizona Board of Regents

Why is it Clustering?

Latent Confidence Clustering



- Directly ask LLMs whether the answer is correct or not.
- It has a different clustering mechanism because it does not cluster on the semantic level, but can be viewed as the clustering inside LLMs confidence region. LLMs cluster all answers into high uncertainty and low uncertainty.



We can see it more clearly using the hidden state from LLM (Qwen2.5-32b-Instruct). It shows that we can cluster the samples with high and low P(true) using hidden state.

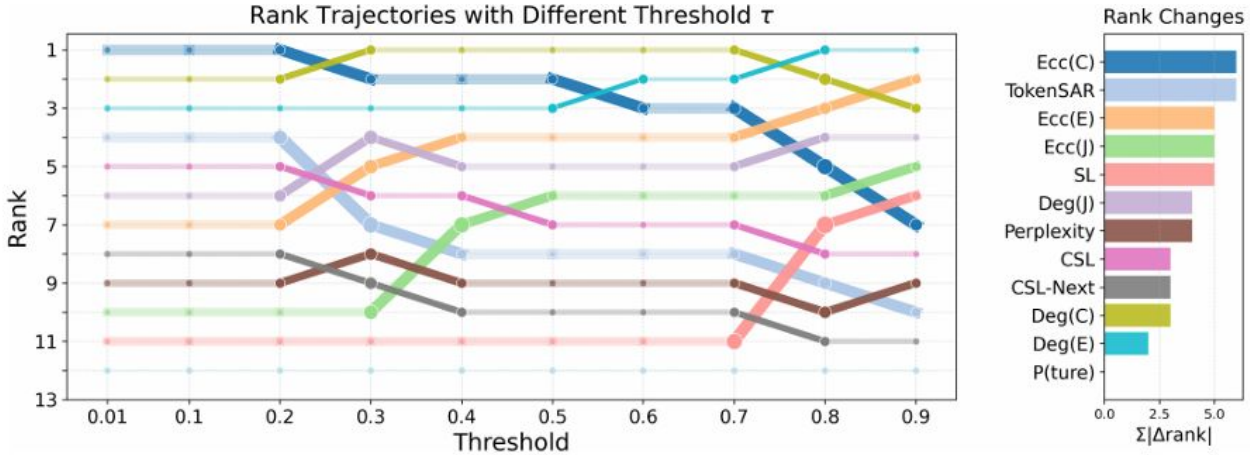
Kadavath, Saurav, et al. "Language models (mostly) know what they know." arXiv preprint arXiv:2207.05221 (2022).

Copyright © 2026 Arizona Board of Regents

Why Clustering Fails?

The Parameter Sensitivity Crisis

- **The Parameter Sensitivity Crisis:** The performance of clustering is highly sensitive to the parameter.



Liu, Xiaoou, et al. "MCQA-Eval: Efficient Confidence Evaluation in NLG with Gold-Standard Correctness Labels." arXiv preprint arXiv:2502.14268 (2025).

Copyright © 2026 Arizona Board of Regents

Discussion

The Parameter Sensitivity Crisis

Alternative: Sensitivity is a Feature for Uncertainty

Our Response:

- Instability is a flaw.
- It ruins fair evaluation and fakes progress.

Future Direction:

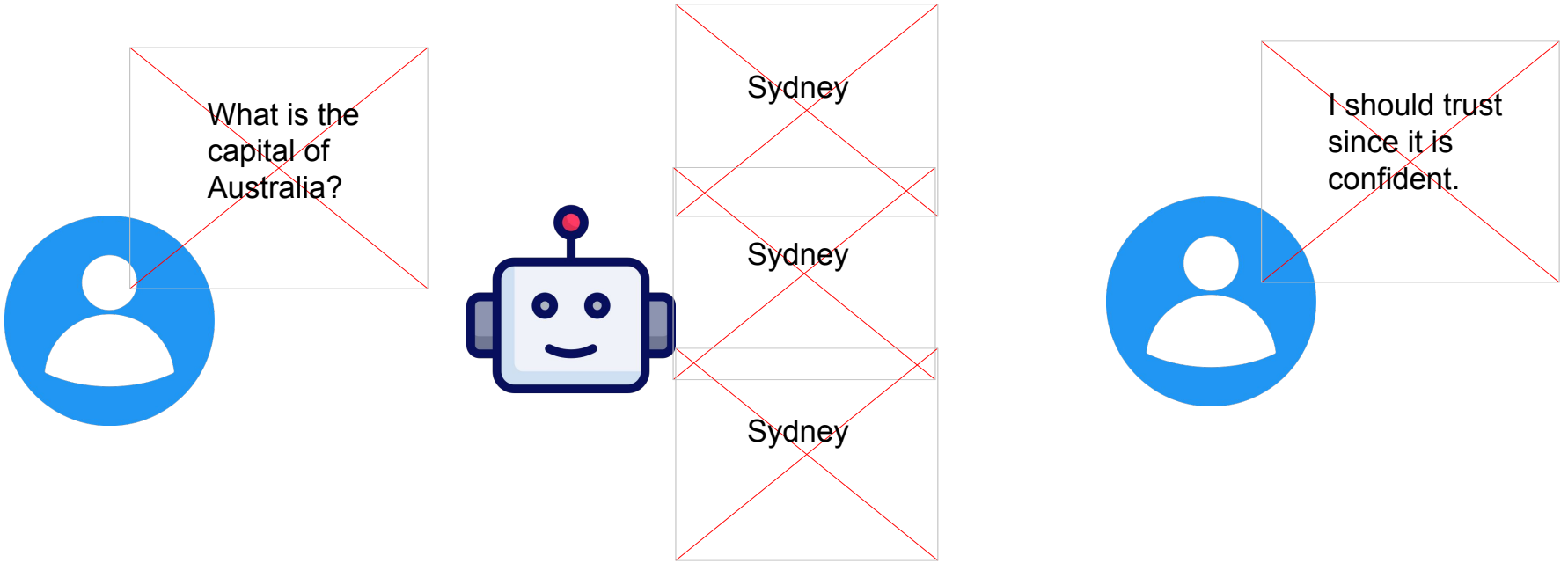
- Tail-risk Evaluation: detect errors when the false positive rate is low.
- Sensitivity Reporting of Parameters: report metrics such as AUROC with different parameters.

Copyright © 2026 Arizona Board of Regents

Why Clustering Fails?

The “Internal Evaluation” Trap

- **The “Internal Evaluation” Trap:** Current methods evaluate consistency, but fail to detect confident errors.



Copyright © 2026 Arizona Board of Regents

Discussion

The “Internal Evaluation” Trap

Alternative: Uncertainty Measures Belief, Not Truth.

Our Response:

- Internal belief is valuable for research.
- Trustworthy deployment requires a focus on practical safety.
- The field already evaluates UQ based on truth.

Future Direction:

- Integrate the Uncertainty in Application: use uncertainty in application such as Conformal Prediction.
- Training for Native Uncertainty: use accuracy as the label.

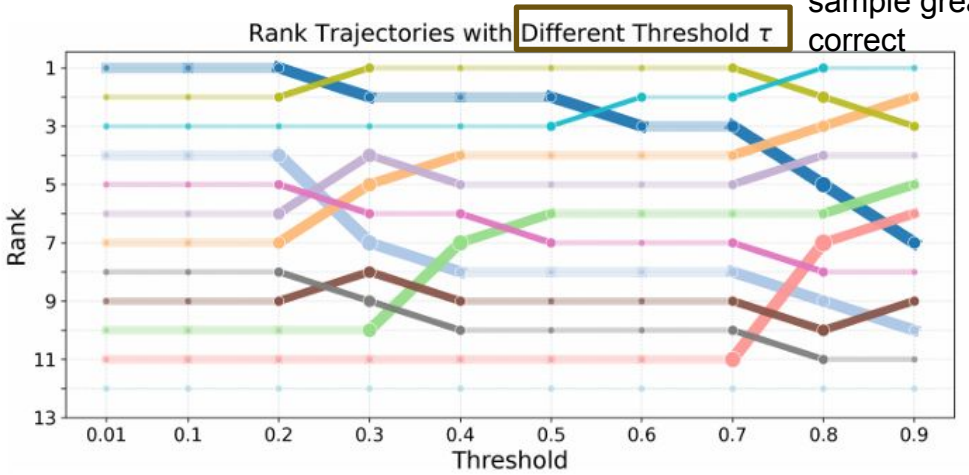
Copyright © 2026 Arizona Board of Regents

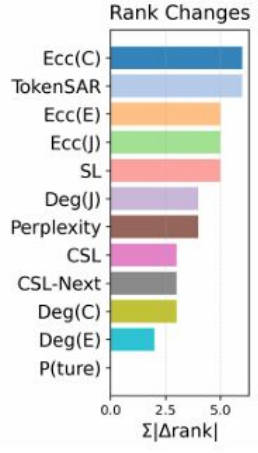
Why Clustering Fails?

The Lack of Ground Truth

- **The Lack of Ground Truth:** There is no true uncertainty, so UQ evaluations rely on noisy and biased 'correctness' functions that distort results.

Accuracy Threshold! Use LLM-as-a-judge and consider the sample greater than this threshold as correct





Liu, Xiaoou, et al. "MCQA-Eval: Efficient Confidence Evaluation in NLG with Gold-Standard Correctness Labels." arXiv preprint arXiv:2502.14268 (2025).

Copyright © 2026 Arizona Board of Regents

Discussion

The Lack of Ground Truth

Alternative: Ground Truth is Intractable for Generative Tasks

Our Response:

- High-stakes facts are strictly objective.
- Difficulty is not an excuse.

Future Direction:

- Mandatory Unit Testing and Atomic Fact Verification: use tools that help get correct metrics when encountering open-ended generation

Copyright © 2026 Arizona Board of Regents

Thank you!

Paper

Copyright © 2026 Arizona Board of Regents