Novelty Assessment Report

Paper: Trained on Tokens, Calibrated on Concepts: The Emergence of Semantic Calibration in LLMs
PDF URL: https://openreview.net/pdf?id=0sCyk9Tr5J
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Large Language Models (LLMs) often lack meaningful confidence estimates for the semantic content of their outputs. While base LLMs are known to exhibit next-token calibration, it remains unclear whether they can assess confidence in the actual meaning of their responses beyond the token level. We find that, when using a certain sampling-based notion of semantic calibration, base LLMs are remarkably well-calibrated: they can meaningfully assess confidence in various open-ended question-answering tasks, despite training only on next-token prediction. To formalize this phenomenon, we introduce "B-calibration," a notion of calibration parameterized by the choice of equivalence classes. Our main theoretical contribution establishes a mechanism for why semantic calibration emerges in base LLMs, leveraging a recent connection between calibration and local loss optimality. This theoretical mechanism leads to a testable prediction: base LLMs will be semantically calibrated when they can easily predict their own distribution over semantic answer classes before generating a response. We state three implications of this prediction, which we validate through experiments: (1) Base LLMs are semantically calibrated across question-answering tasks, (2) instruction-tuning procedures systematically break this calibration, and (3) chain-of-thought reasoning breaks calibration (intuitively because models cannot predict their final answers before completing their generation). To our knowledge, our work provides the first principled explanation of when and why semantic calibration emerges in LLMs.

Core Task Landscape

This paper addresses: **semantic confidence calibration in large language models**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Confidence Estimation Methods and Frameworks**
- **Calibration via Fine-Tuning and Optimization**
- **Theoretical Foundations and Empirical Analysis**
- **Domain-Specific Applications**
- **Adaptive Inference and Selective Prediction**
- **Uncertainty Taxonomy and Semantic Distinctions**

Complete Taxonomy Tree

- semantic confidence calibration in large language models Survey Taxonomy
- Confidence Estimation Methods and Frameworks
  - Black-Box Uncertainty Quantification
  - Consistency-Based Methods (3 papers)
    - [3] Generating with confidence: Uncertainty quantification for black-box large language models (Lin Zhen, 2023) View paper
    - [32] Think Twice Before Assure: Confidence Estimation for Large Language Models through Reflection on Multiple Answers (Li, 2024) View paper
    - [45] Calibrating large language models with sample consistency (Lyu Qing, 2025) View paper
  - Semantic Similarity and Clustering (6 papers)
    - [18] Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space (Risto Miikkulainen, 2024) View paper
    - [19] Detecting hallucinations in large language models using semantic entropy (Sebastian Farquhar, 2024) View paper
    - [24] Semantically diverse language generation for uncertainty estimation in language models (Lukas Aichberger, 2024) View paper
    - [28] Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities (Nikitin Alexander, 2024) View paper
    - [30] Improving uncertainty quantification in large language models via semantic embeddings (Bonilla, 2024) View paper
    - [38] Beyond Semantic Entropy: Boosting LLM Uncertainty Quantification with Pairwise Semantic Similarity (Nguyen-Dang, 2025) View paper
  - Verbalized Confidence Elicitation (5 papers)
    - [11] Confidence under the hood: An investigation into the confidence-probability alignment in large language models (Kumar Abhishek, 2024) View paper
    - [14] Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models (Valdenegro-Toro, 2024) View paper
    - [17] The calibration gap between model and human confidence in large language models (Mark Steyvers, 2024) View paper
    - [40] Calibrating the confidence of large language models by eliciting fidelity (Zhang Mozhi, 2024) View paper
    - [49] Uncertainty distillation: Teaching language models to express semantic confidence (Mueller, 2025) View paper

- White-Box and Hybrid Approaches
- Token Probability-Based Methods (3 papers)
  - [9] Revisiting Uncertainty Estimation and Calibration of Large Language Models (Tao Linwei, 2025) View paper
  - [33] Probabilities Are All You Need: A Probability-Only Approach to Uncertainty Estimation in Large Language Models (Manh Nguyen, 2025) View paper
  - [34] Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs (Yaldiz, 2024) View paper
- Intermediate Representation Analysis (1 papers)
  - [22] Beyond the Final Layer: Intermediate Representations for Better Multilingual Calibration in Large Language Models (Zhang, 2025) View paper
- Auxiliary Model Training (2 papers)
  - [23] Graph-based Confidence Calibration for Large Language Models (Li, 2024) View paper
  - [31] Calibrating large language models using their generations only (Ulmer, 2024) View paper
- Comprehensive Surveys and Taxonomies (3 papers)
- [1] A survey of confidence estimation and calibration in large language models (Jiahui Geng, 2024) View paper
- [5] Uncertainty quantification and confidence calibration in large language models: A survey (Xiaoou Liu, 2025) View paper
- [12] Look before you leap: An exploratory study of uncertainty measurement for large language models (Huang Yuheng, 2023) View paper

- Calibration via Fine-Tuning and Optimization
  - Temperature Scaling and Post-Training Calibration (2 papers)
  - [6] Semantic-Level Confidence Calibration of Language Models via Temperature Scaling (TA Lamb, 2025) View paper
  - [15] Task calibration: Calibrating large language models on inference tasks (Yingjie Li, 2025) View paper
  - Reinforcement Learning and Listener-Aware Fine-Tuning (3 papers)
  - [20] Prompt4Trust: A Reinforcement Learning Prompt Augmentation Framework for Clinically-Aligned Confidence Calibration in Multimodal Large Language Models (Shen Xing, 2025) View paper
  - [37] SEED-GRPO: Semantic Entropy Enhanced GRPO for Uncertainty-Aware Policy Optimization (Chen, 2025) View paper
  - [48] LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models (Stengel-Eskin, 2024) View paper
  - Uncertainty-Aware Training Objectives (1 papers)
  - [47] Enhancing trust in large language models with uncertainty-aware fine-tuning (Krishnan, 2024) View paper
- Theoretical Foundations and Empirical Analysis
  - Emergence and Mechanisms of Calibration ★ (1 papers)
  - [0] Trained on Tokens, Calibrated on Concepts: The Emergence of Semantic Calibration in LLMs (Anon et al., 2026) View paper
  - Empirical Calibration Studies (3 papers)
  - [4] A close look into the calibration of pre-trained language models (Chen Yangyi, 2023) View paper
  - [7] Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models (Chhikara, 2025) View paper
  - [27] How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering (Zhengbao Jiang, 2021) View paper
  - Confidence-Probability Alignment Analysis (1 papers)
  - [16] Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback (Tian, 2023) View paper
  - Decoding Strategy Effects on Uncertainty (1 papers)
  - [46] Semantic uncertainty in advanced decoding methods for LLM generation (Fan Si-min, 2025) View paper
- Domain-Specific Applications
  - Medical and Healthcare Applications (2 papers)
  - [8] The challenge of uncertainty quantification of large language models in medicine (Safavi-Naini, 2025) View paper
  - [29] Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment (Thomas Savage, 2025) View paper
  - Structured Data Generation (3 papers)
  - [2] Confidence Estimation for Text-to-SQL in Large Language Models (Pourreza, 2025) View paper
  - [21] LLM4Tag: Automatic Tagging System for Information Retrieval via Large Language Models (Ruiming Tang, 2025) View paper
  - [36] Confidence-Aware Sub-Structure Beam Search (CABS): Mitigating Hallucination in Structured Data Generation with Large Language Models (Wei, 2024) View paper
  - Code Generation and Summarization (2 papers)
  - [35] L2CEval: Evaluating Language-to-Code Generation Capabilities of Large Language Models (Ansong Ni, 2023) View paper
  - [44] Calibration of Large Language Models on Code Summarization (Devanbu, 2025) View paper
  - Multilingual Calibration (1 papers)
  - [41] MlingConf: A Comprehensive Study of Multilingual Confidence Estimation on Large Language Models (Xue, 2024) View paper
  - Vision-Language Models (1 papers)
  - [13] Confidence is key: Uncertainty estimation in large language models and vision language models (Groot, 2024) View paper
- Adaptive Inference and Selective Prediction
  - Adaptive Computation and Early Exiting (1 papers)
  - [10] Confident adaptive language modeling (Schuster, 2022) View paper
  - Selective Answering and Refusal (1 papers)
  - [39] Trusted Uncertainty in Large Language Models: A Unified Framework for Confidence Calibration and Risk-Controlled Refusal (Conti Giulia, 2025) View paper
  - Multi-Agent Deliberation (1 papers)
  - [42] Confidence Calibration and Rationalization for LLMs via Multi-Agent Deliberation (Yang Rui-xin, 2024) View paper
- Uncertainty Taxonomy and Semantic Distinctions (4 papers)
  - [25] Semantic volume: Quantifying and detecting both external and internal uncertainty in llms (Li Xiaomin, 2025) View paper
  - [26] Confidence Calibration of Large Language Models (LLMC) (Bien, 2025) View paper
  - [43] Confidence Calibration in Large Language Model-Based Entity Matching (Tashu, 2025) View paper
  - [50] Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models (Zhao Xin-ran, 2024) View paper

Narrative

Core task: semantic confidence calibration in large language models. The field addresses how well a model's expressed confidence aligns with the true correctness or semantic consistency of its outputs. The taxonomy organizes research into six main branches: Confidence Estimation Methods and Frameworks develop techniques to measure uncertainty from model outputs, often using semantic clustering or consistency checks (e.g., Semantic Entropy[19], Semantic Density[18]); Calibration via Fine-Tuning and Optimization explores training-time interventions to improve alignment between confidence and accuracy (e.g., Uncertainty-Aware Fine-tuning[47], SEED-GRPO[37]); Theoretical Foundations and Empirical Analysis investigates the underlying mechanisms and emergence of calibration properties; Domain-Specific Applications tailors calibration methods to specialized settings like medical question answering (Medical Uncertainty Challenge[8]) or text-to-SQL (Text-to-SQL Confidence[2]); Adaptive Inference and Selective Prediction leverages confidence scores to decide when to abstain or escalate; and Uncertainty Taxonomy and Semantic Distinctions clarifies different notions of uncertainty, distinguishing epistemic from aleatoric sources and token-level from semantic-level measures.

A particularly active line of work contrasts token-probability-based methods with semantic-level approaches: while traditional calibration often relies on softmax probabilities or temperature scaling (Temperature Scaling[6]), many recent studies argue that semantic consistency across paraphrases or sampled outputs better captures true uncertainty (Generating with Confidence[3], Semantic Embeddings[30]). Another key tension involves whether to estimate confidence from generations alone (Generations Only[31]) or to incorporate internal model states and intermediate representations (Intermediate Representations[22]). The original paper, Semantic Calibration[0], sits squarely within the Theoretical Foundations and Empirical Analysis branch, focusing on the emergence and mechanisms of calibration. It complements works like Calibration Pre-trained Models[4] and Revisiting Calibration[9] by examining how and why semantic-level confidence signals arise during pretraining and scaling, offering insights that inform both estimation frameworks and fine-tuning strategies across the taxonomy.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf 'Emergence and Mechanisms of Calibration' focuses on theoretical explanations for why and when semantic calibration emerges in LLMs, distinguishing itself from purely empirical work. The sibling subtopics cover complementary aspects: confidence-probability alignment examines the relationship between internal and expressed confidence, decoding strategy effects analyze how generation methods impact uncertainty, and empirical calibration studies provide large-scale evaluations without theoretical contributions. Together, these categories separate theoretical mechanisms from empirical observations and specific technical factors affecting calibration.

Similarities: - All subtopics address aspects of semantic calibration and uncertainty in LLMs - All exclude calibration improvement methods and adaptive inference techniques - All focus on understanding or measuring calibration properties rather than optimizing them - All deal with the relationship between model confidence and actual performance

Differences: - The original leaf emphasizes theoretical analysis and mechanistic explanations, while Empirical Calibration Studies explicitly excludes theoretical contributions - Confidence-Probability Alignment Analysis focuses specifically on the internal-external confidence relationship, whereas the original leaf addresses broader emergence phenomena - Decoding Strategy Effects examines a specific technical factor (generation methods), while the original leaf seeks general principles of when/why calibration emerges - Empirical Calibration Studies covers large-scale evaluations across multiple dimensions (models, tasks, training stages), while the original leaf targets explanatory frameworks - The original leaf's scope is mechanistic understanding, while siblings either measure specific relationships or provide descriptive evaluations

Suggested Search Directions: - Theoretical frameworks explaining calibration emergence during pretraining or scaling - Mechanistic interpretability studies of confidence formation in transformer architectures - Mathematical models of how model capacity or training dynamics affect calibration properties - Causal analysis of factors driving calibration behavior in different model families

Sibling Subtopics

- **Confidence-Probability Alignment Analysis** (leaves: 1, papers: 1)
- Scope: Research examining alignment between internal model probabilities and externally expressed confidence.
- Exclude: Excludes calibration improvement methods; see Calibration via Fine-Tuning and Optimization.
- **Decoding Strategy Effects on Uncertainty** (leaves: 1, papers: 1)
- Scope: Studies analyzing how different decoding methods affect semantic uncertainty and calibration.
- Exclude: Excludes adaptive inference methods; see Adaptive Inference and Selective Prediction.
- **Empirical Calibration Studies** (leaves: 1, papers: 3)
- Scope: Large-scale empirical evaluations of calibration properties across models, tasks, or training stages.
- Exclude: Excludes theoretical explanations or method proposals; see sibling categories.

Contributions Analysis

Overall novelty summary. The paper introduces B-calibration, a parameterized framework for semantic calibration in base LLMs, and provides a theoretical mechanism linking semantic calibration to local loss optimality. It resides in the 'Emergence and Mechanisms of Calibration' leaf under 'Theoretical Foundations and Empirical Analysis,' where it is currently the sole paper. This leaf focuses on explaining why semantic calibration emerges through theoretical analysis, distinguishing it from purely empirical evaluations or method proposals. The sparse population of this leaf suggests that theoretical explanations of calibration emergence remain underexplored in the literature, positioning the work in a relatively open research direction.

The taxonomy reveals substantial activity in neighboring areas: the sibling leaf 'Empirical Calibration Studies' contains three papers examining calibration properties across models and tasks, while 'Confidence Estimation Methods and Frameworks' encompasses multiple leaves with 20+ papers developing black-box and white-box uncertainty quantification techniques. The parent branch 'Theoretical Foundations and Empirical Analysis' also includes work on confidence-probability alignment and decoding strategy effects. The paper's theoretical focus on emergence mechanisms differentiates it from these empirical and methodological neighbors, though it shares conceptual ground with studies analyzing when and why calibration properties manifest during training or scaling.

Among 20 candidates examined across three contributions, no clearly refuting prior work was identified. The B-calibration framework examined 10 candidates with zero refutable matches, suggesting novelty in formalizing semantic calibration via equivalence classes. The theoretical mechanism linking calibration to local loss optimality examined only 1 candidate, reflecting limited prior theoretical work in this specific direction. The testable predictions contribution examined 9 candidates, again with no refutations, indicating that the predictive framework and its experimental validation appear distinct from existing empirical studies. The limited search scope (20 candidates total) means these findings reflect top semantic matches rather than exhaustive coverage.

Given the sparse theoretical landscape and the absence of refuting work among examined candidates, the paper appears to occupy a relatively novel position within its immediate research context. However, the small search scope and the single-paper status of its taxonomy leaf suggest caution: while no overlapping prior work surfaced in top-20 semantic matches, a broader literature review might

reveal related theoretical analyses not captured here. The work's novelty seems strongest in its formal B-calibration framework and mechanistic explanation, with empirical validation building on established evaluation paradigms from neighboring leaves.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: B-calibration framework for semantic calibration in LLMs

Description: The authors introduce B-calibration, a formal framework that generalizes calibration to arbitrary equivalence classes defined by a collapsing function B. This framework enables rigorous analysis of semantic calibration by treating the LLM as inducing a classifier over semantic classes.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. QA-Calibration of Language Model Confidence Scores

URL: View paper

Brief Assessment

QA-Calibration[64] focuses on calibration across question-and-answer groups in QA systems, not on semantic equivalence classes defined by collapsing functions. The frameworks address different calibration objectives.

---

2. FOCoOp: Enhancing Out-of-Distribution Robustness in Federated Prompt Learning for Vision-Language Models

URL: View paper

Brief Assessment

FOCoOp[60] addresses federated prompt learning for vision-language models with a focus on out-of-distribution robustness through prompt optimization, not calibration frameworks for language model outputs or equivalence classes in LLMs.

---

3. Task calibration: Calibrating large language models on inference tasks

URL: View paper

Brief Assessment

Task Calibration[15] addresses calibration in inference tasks through task reformulation using mutual information, focusing on mitigating context preference bias. This differs from the ORIGINAL paper's B-calibration framework, which generalizes calibration to arbitrary equivalence classes defined by collapsing functions.

---

4. InfAlign: Inference-aware language model alignment

URL: View paper

Brief Assessment

InfAlign[66] focuses on reward calibration for inference-time alignment methods (best-of-n, controlled decoding), not on semantic calibration frameworks for LLM outputs or equivalence class-based calibration.

---

5. Examining the efficacy of generative artificial intelligence in item generation: comparative analysis of human-developed and AI-generated reading tests

URL: View paper

Brief Assessment

AI-Generated Tests[65] focuses on psychometric calibration methods for educational test items using Rasch models, not on semantic calibration frameworks for language model outputs or equivalence classes in LLMs.

---

6. Linguistic calibration of long-form generations

URL: View paper

Brief Assessment

Linguistic Calibration[62] focuses on calibrating long-form generations through downstream user forecasts in decision-making contexts, not on defining equivalence class-based calibration frameworks like B-calibration.

---

7. Calibrating long-form generations from large language models

URL: View paper

Brief Assessment

Long-form Generations[58] focuses on calibrating long-form text generation where correctness is continuous (0-1 scores), not on semantic equivalence classes defined by collapsing functions. The candidate addresses graded correctness distributions rather than semantic class-based calibration frameworks.

---

8. The Geometry of Creative Variability: How Credal Sets Expose Calibration Gaps in Language Models

URL: View paper

Brief Assessment

Creative Variability Geometry[59] focuses on uncertainty quantification in creative text generation using credal sets to measure human-AI alignment, not on semantic calibration frameworks for equivalence classes in language models.

---

9. Self-Calibrated Listwise Reranking with Large Language Models

URL: View paper

Brief Assessment

Self-Calibrated Listwise[61] focuses on calibrating relevance scores for document reranking tasks, not on semantic calibration of LLM outputs or confidence estimation in question-answering.

---

10. OstQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting

URL: View paper

Brief Assessment

OstQuant[63] focuses on post-training quantization techniques for compressing LLMs, not on calibration frameworks for equivalence classes or semantic calibration in language model outputs.

Contribution 2: Theoretical mechanism linking semantic calibration to local loss optimality

Description: The authors establish a theoretical mechanism explaining emergent semantic calibration in base LLMs by connecting B-calibration to local loss optimality. They prove that B-calibration is equivalent to local loss optimality with respect to a corresponding perturbation family, and show when such perturbations are easy for autoregressive models to implement.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Self-modulated gradient diffusion for large language model internal consistency calibration

URL: View paper

Brief Assessment

Self-modulated Gradient[67] focuses on internal consistency calibration through gradient diffusion mechanisms in autoregressive systems, addressing temporal stability and energy accumulation in local minima. This is technically distinct from the original paper's theoretical framework connecting semantic calibration emergence to local loss optimality in base LLMs through B-calibration equivalence.

Contribution 3: Testable predictions about when semantic calibration emerges

Description: The authors derive testable predictions from their theory, stating that base LLMs exhibit semantic calibration when they can predict their own semantic class distribution before generation. They validate three specific implications: base LLMs are semantically calibrated on question-answering tasks, instruction-tuning breaks this calibration, and chain-of-thought reasoning breaks calibration.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Semantically diverse language generation for uncertainty estimation in language models

URL: View paper

Brief Assessment

Semantically Diverse Generation[24] focuses on improving uncertainty estimation through diverse sampling methods for language generation, not on predicting when semantic calibration emerges in base LLMs or analyzing the effects of instruction-tuning and chain-of-thought on calibration properties.

2. Calibrating Verbalized Probabilities for Large Language Models

URL: View paper

Brief Assessment

Verbalized Probabilities[57] focuses on calibrating verbalized confidence scores for discriminative tasks using temperature scaling on inverted probabilities. It does not address predicting semantic class distributions before generation or the emergence of semantic calibration in base LLMs through next-token prediction training.

3. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space

URL: View paper

Brief Assessment

Semantic Density[18] focuses on uncertainty quantification through semantic space analysis and does not address predicting semantic class distributions before generation or the specific conditions (base LLMs, instruction-tuning effects, chain-of-thought impacts) that the original paper investigates.

4. Semantic probabilistic control of language models

URL: View paper

Brief Assessment

Semantic Probabilistic Control[51] focuses on steering LM generations toward semantic constraints using verifier gradients, not on predicting semantic class distributions before generation or analyzing when calibration emerges in base LLMs.

5. Improving uncertainty quantification in large language models via semantic embeddings

URL: View paper

Brief Assessment

Semantic Embeddings[30] focuses on uncertainty quantification using semantic embeddings and cosine similarity, not on predicting semantic class distributions before generation or deriving testable predictions about calibration emergence in base LLMs.

6. Estimating the Probabilities of Rare Outputs in Language Models

URL: View paper

Brief Assessment

Rare Outputs[54] focuses on estimating probabilities of rare model outputs through importance sampling and activation extrapolation methods, not on predicting semantic class distributions before generation or semantic calibration emergence in LLMs.

7. Next Semantic Scale Prediction via Hierarchical Diffusion Language Models

URL: View paper

Brief Assessment

Hierarchical Diffusion[53] focuses on hierarchical discrete diffusion for language generation with semantic scale prediction, not on predicting semantic class distributions before generation for calibration purposes in LLMs.

8. A visionâ□□language model-based traffic sign detection method for high-resolution drone images: A case study in Guyuan, China

URL: View paper

Brief Assessment

Traffic Sign Detection[56] focuses on vision-language models for traffic sign detection in drone images, not on semantic calibration or confidence estimation in language model outputs. The domains are entirely different.

9. Exploiting latent semantic information in statistical language modeling

URL: View paper

Brief Assessment

Latent Semantic Information[55] focuses on statistical language modeling using latent semantic analysis for speech recognition, not on semantic calibration or confidence estimation in LLMs. The candidate paper addresses predicting semantic class distributions in the context of document-word co-occurrences for language modeling, not predicting a model's own semantic output distribution before generation as required by the original contribution.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Trained on Tokens, Calibrated on Concepts: The Emergence of Semantic Calibration in LLMs View paper
- [1] A survey of confidence estimation and calibration in large language models View paper
- [2] Confidence Estimation for Text-to-SQL in Large Language Models View paper
- [3] Generating with confidence: Uncertainty quantification for black-box large language models View paper
- [4] A close look into the calibration of pre-trained language models View paper
- [5] Uncertainty quantification and confidence calibration in large language models: A survey View paper
- [6] Semantic-Level Confidence Calibration of Language Models via Temperature Scaling View paper
- [7] Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models View paper
- [8] The challenge of uncertainty quantification of large language models in medicine View paper
- [9] Revisiting Uncertainty Estimation and Calibration of Large Language Models View paper
- [10] Confident adaptive language modeling View paper
- [11] Confidence under the hood: An investigation into the confidence-probability alignment in large language models View paper
- [12] Look before you leap: An exploratory study of uncertainty measurement for large language models View paper
- [13] Confidence is key: Uncertainty estimation in large language models and vision language models View paper
- [14] Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models View paper
- [15] Task calibration: Calibrating large language models on inference tasks View paper
- [16] Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback View paper
- [17] The calibration gap between model and human confidence in large language models View paper
- [18] Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space View paper
- [19] Detecting hallucinations in large language models using semantic entropy View paper
- [20] Prompt4Trust: A Reinforcement Learning Prompt Augmentation Framework for Clinically-Aligned Confidence Calibration in Multimodal Large Language Models View paper
- [21] LLM4Tag: Automatic Tagging System for Information Retrieval via Large Language Models View paper
- [22] Beyond the Final Layer: Intermediate Representations for Better Multilingual Calibration in Large Language Models View paper
- [23] Graph-based Confidence Calibration for Large Language Models View paper
- [24] Semantically diverse language generation for uncertainty estimation in language models View paper
- [25] Semantic volume: Quantifying and detecting both external and internal uncertainty in llms View paper
- [26] Confidence Calibration of Large Language Models (LLMC) View paper
- [27] How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering View paper
- [28] Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities View paper
- [29] Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment View paper
- [30] Improving uncertainty quantification in large language models via semantic embeddings View paper
- [31] Calibrating large language models using their generations only View paper
- [32] Think Twice Before Assure: Confidence Estimation for Large Language Models through Reflection on Multiple Answers View paper
- [33] Probabilities Are All You Need: A Probability-Only Approach to Uncertainty Estimation in Large Language Models View paper
- [34] Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs View paper
- [35] L2CEval: Evaluating Language-to-Code Generation Capabilities of Large Language Models View paper
- [36] Confidence-Aware Sub-Structure Beam Search (CABS): Mitigating Hallucination in Structured Data Generation with Large Language Models View paper
- [37] SEED-GRPO: Semantic Entropy Enhanced GRPO for Uncertainty-Aware Policy Optimization View paper
- [38] Beyond Semantic Entropy: Boosting LLM Uncertainty Quantification with Pairwise Semantic Similarity View paper
- [39] Trusted Uncertainty in Large Language Models: A Unified Framework for Confidence Calibration and Risk-Controlled Refusal View paper
- [40] Calibrating the confidence of large language models by eliciting fidelity View paper
- [41] MlingConf: A Comprehensive Study of Multilingual Confidence Estimation on Large Language Models View paper
- [42] Confidence Calibration and Rationalization for LLMs via Multi-Agent Deliberation View paper
- [43] Confidence Calibration in Large Language Model-Based Entity Matching View paper
- [44] Calibration of Large Language Models on Code Summarization View paper
- [45] Calibrating large language models with sample consistency View paper
- [46] Semantic uncertainty in advanced decoding methods for LLM generation View paper
- [47] Enhancing trust in large language models with uncertainty-aware fine-tuning View paper
- [48] LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models View paper
- [49] Uncertainty distillation: Teaching language models to express semantic confidence View paper
- [50] Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models View paper
- [51] Semantic probabilistic control of language models View paper
- [52] Tcp: Textual-based class-aware prompt tuning for visual-language model View paper
- [53] Next Semantic Scale Prediction via Hierarchical Diffusion Language Models View paper
- [54] Estimating the Probabilities of Rare Outputs in Language Models View paper
- [55] Exploiting latent semantic information in statistical language modeling View paper

- [56] A visionâlanguage model-based traffic sign detection method for high-resolution drone images: A case study in Guyuan, China View paper
- [57] Calibrating Verbalized Probabilities for Large Language Models View paper
- [58] Calibrating long-form generations from large language models View paper
- [59] The Geometry of Creative Variability: How Credal Sets Expose Calibration Gaps in Language Models View paper
- [60] FOCoOp: Enhancing Out-of-Distribution Robustness in Federated Prompt Learning for Vision-Language Models View paper
- [61] Self-Calibrated Listwise Reranking with Large Language Models View paper
- [62] Linguistic calibration of long-form generations View paper
- [63] OstQuant: Refining Large Language Model Quantization with Orthogonal and Scaling Transformations for Better Distribution Fitting View paper
- [64] QA-Calibration of Language Model Confidence Scores View paper
- [65] Examining the efficacy of generative artificial intelligence in item generation: comparative analysis of human-developed and AI-generated reading tests View paper
- [66] InfAlign: Inference-aware language model alignment View paper
- [67] Self-modulated gradient diffusion for large language model internal consistency calibration View paper