Learning to Route LLMs with Confidence Tokens

Paper Presentation & Analysis

Presenters

Shiduo Li & Qianyu Zheng

Georgia Institute of Technology

ICML 2026 Paper

CS 8803: LLM (SPRING 2026)

Problem Statement: Confidence in LLMs

CONTEXT

- LLMs are deployed in high-stakes settings
  - E.g. medicine, education, customer support
- Systems need to know *when* an LLM answer can be trusted
  - I.e. the LLM should report inconfidence in inference stage
- No reliable confidence signal = forced tradeoff:
  - Always use the cheap small model (risky)
  - Always use the expensive large model (costly)
  - No principled way to route or abstain

EXISTING APPROACHES

On LLM Uncertainty Quantification:

- Confidence from internal feature (e.g. logits)
- Verbalizing confidence (i.e. predict with 0 - 1)
- Iterative Prompting (consistency as confidence)
- Instruction Learning / Pretraining for Calibration
  - [IDK] token [5]

On Routing/Rejection:

- Extra Router / Classifier Models for routing
- LLM Cascades
- Rejection knowledge injection with finetuning

Proposed Solution: Confidence Token

GOAL

- Train a local LLM to reliably signal whether its own answer is correct or not

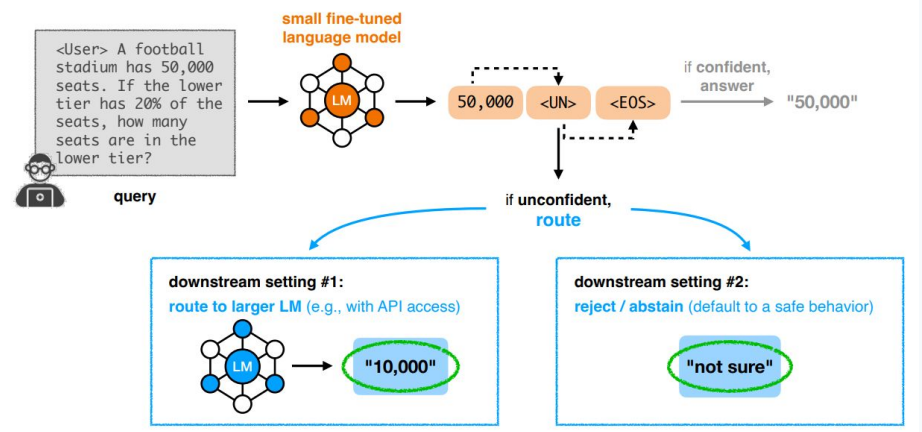


Figure 1: Self-REF attaches a confidence token to each prediction. Routing proceeds based on the confidence.

APPROACH

Two special tokens are introduced:

- <CN> *(Confident)* — generated when the model believes its answer is correct
- <UN> *(Unconfident)* — generated when the model believes its answer is wrong

Key insight:

- Finetune the model to output one of the confidence tokens at the end of the sentence before EOS.
- Confidence is assessed *after* the model generates its answer — not on the query alone.

Finetune an LLM with the Confidence Token

SFT DATASET CURATION

Starting Point:

- Run the **base model M** on each training sample to generate predictions.

Labeling:

- **Correct** ($\hat{y} = y$) → append <CN> → forms the *confident sample set*
- **Incorrect** ($\hat{y} \neq y$) → append <UN> → forms the *unconfident sample set*

SUPPRESS INCORRECT SAMPLES

Gradient Masking:

- The **incorrect answer tokens are masked** during training
  - The model only learns to predict <UN> after a wrong answer — not to reproduce the wrong answer itself

Down-sampling:

- Unconfident samples are **downsampled** by a tunable ratio $\alpha \in [0,1]$

Algorithm 1 Data Augmentation for Self-REF

input Base LLM $\mathcal{M}(\cdot)$ and original training data $\mathcal{D}_{\text{train}} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N_{\text{train}}}$.

output Augmented data $\mathcal{D}'_{\text{train}}$ with unconfident samples and confident samples. Use $\mathcal{M}(\cdot)$ to make predictions $\hat{y}^{(i)}$ for the given input query $x^{(i)}$.

1: Create a set of unconfident samples:

$$\mathcal{D}'_{\text{train,<UN>}} = \{(x^{(i)}, \hat{y}^{(i)}\text{<UN>}) : y^{(i)} \neq \hat{y}^{(i)}\}_{i=1}^{N_{\text{train}}}.$$

2: Create a set of confident samples:

$$\mathcal{D}'_{\text{train,<CN>}} = \{(x^{(i)}, \hat{y}^{(i)}\text{<CN>}) : y^{(i)} = \hat{y}^{(i)}\}_{i=1}^{N_{\text{train}}}.$$

3: Mix the unconfident and confident data to form the combined augmented dataset, subsampling the unconfident samples with tunable proportion $\alpha \in [0, 1]$:

$$\mathcal{D}'_{\text{train}} = \text{subsample}_{\alpha}(\mathcal{D}'_{\text{train,<UN>}}) \cup \mathcal{D}'_{\text{train,<CN>}}.$$

Confidence-based Routing & Rejection

CONFIDENCE SCORE

Final confidence score is computed as:

$$c_{\mathcal{M}}(x^{(i)}, \hat{y}^{(i)}) = \frac{P(\text{<CN>})}{P(\text{<UN>}) + P(\text{<CN>})}$$

If confidence > threshold t:

- Directly answer

Otherwise:

- Route to a larger LLM
- Abstain from answering

ROUTING

Redirect the question to a larger, more capable LLM:

- Maintains high accuracy without always paying the cost of a large LLM
- Routing incurs network transmission costs; the larger LLM may also be wrong in answering the query.

REJECTION

Model abstains and returns a safe default (e.g., "not sure" or "none of the above")

- Prevents unsafe or incorrect answers in high-stakes settings
- Abstaining provides no answer — tradeoff for accuracy

The Evaluation Setup: Defining the Playing Field

DATASETS

A rigorous evaluation across diverse knowledge domains:

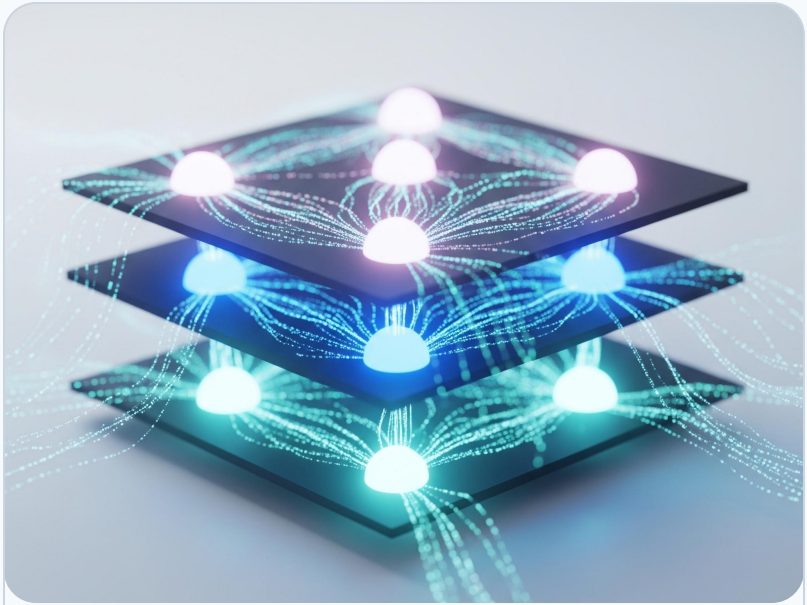
- **MMLU:** 57 subjects (Humanities, STEM, etc.)
- **OpenbookQA:** Elementary science reasoning
- **GSM8K:** Grade-school math word problems
- **MedQA:** Professional medical board exams

BASELINES

Compared against state-of-the-art estimators:

- **Zero-shot & Fine-tuned Logits:** Token probabilities
- **Verbalizing Uncertainty:** Direct LLM estimation
- **Verbalizing Yes/No Tokens:** Confidence check

All methods normalized to a 0–1 continuous score.



Result 1: The Routing Trade-off

THE SETUP

Infrastructure: Local Llama3-8B routing to API Llama3-70B.

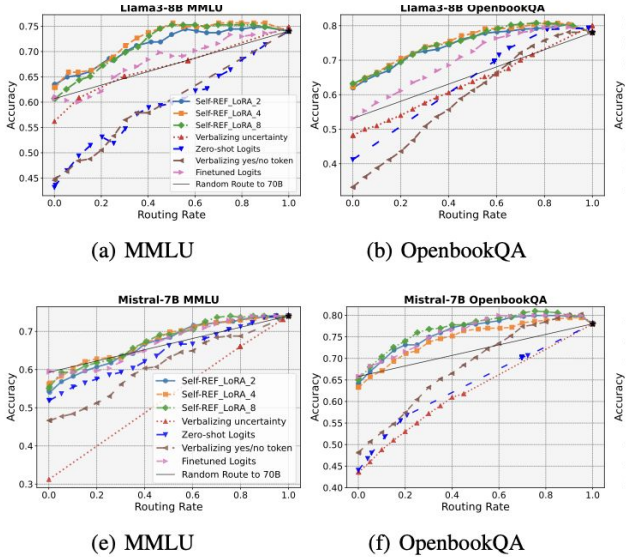
Decision Logic: If 8B confidence $c < t$, route to 70B.

HIGHLIGHT

Self-REF achieved the best accuracy-to-routing-rate trade-off across all datasets.

Concrete ROI (MMLU Dataset):

- **39% Routing Rate:** Same accuracy as 100% routing.
- **2.03x Efficiency:** Direct reduction in per-token latency.



(a) MMLU (b) OpenbookQA

(e) MMLU (f) OpenbookQA

Routing is from local LLMs Llama3-8B and Mistral-7B to Llama3-70B on MMLU, OpenbookQA, GSM8K, and MedQA.

Result 2: Prioritizing Safety via Rejection

THE SETUP

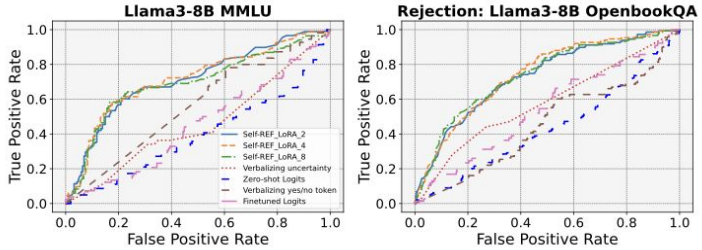
Evaluation Set Design:

50% of queries had the correct answer choice deliberately removed from the list to test abstention behavior.

THE RESULT

Self-REF accurately identifies "None of the Above" scenarios.

- **Token Trigger:** Correctly appends <UN> when no valid choice exists.
- **Superior Detection:** Outperformed all baselines in rejection accuracy.
- **Reliable Calibration:** High correlation between confidence tokens and correctness.



(a) MMLU (b) OpenbookQA

Figure 3: ROC curves for rejection learning.

The true positive rate (recall) indicates the proportion of samples where "none of the above" is the ground truth that is correctly rejected.

Across all settings, Self-REF outperforms other baselines in learning to reject.

The Calibration Paradox

METRICS

Three metrics were used to assess the alignment of predicted vs. actual probabilities:

- **ECE:** Expected Calibration Error
- **BS:** Brier Score
- **CE:** Cross-Entropy Score

THE INSIGHT

Calibration does not equal Utility.

Self-REF did not always achieve the best ECE or Brier Scores in comparison to baselines. However, it consistently delivered the highest ROI in actual routing and rejection tasks.

Llama3-8B-Instruct	MMLU				OpenbookQA				GSM8K				MedQA			
	Route	ECE	BS	CE	Route	ECE	BS	CE	Route	ECE	BS	CE	Route	ECE	BS	CE
Verbalizing Uncertainty	100%	0.217	0.333	7.383	75%	0.148	0.311	7.688	92%	0.047	0.307	7.999	100%	0.069	0.331	1.206
Verbalizing Yes/No Token	100%	0.466	0.397	10.281	95%	0.291	0.415	5.319	100%	0.107	0.384	8.369	100%	0.090	0.470	11.276
Zero-shot Logits	100%	0.347	0.418	1.733*	90%	0.225	0.496	1.705	90%	0.027	0.292*	3.312	100%	0.108	0.736	3.767
Finetuned Logits	90%	0.081	**0.206**	**0.602**	70%	0.095	**0.238**	**0.676**	70%	0.055	**0.256**	**0.930**	80%	**0.059**	**0.265**	**0.844**
Self-REF - Llama3-8B-Instruct	39%	**0.040**	0.320*	11.534	49%	**0.090**	0.254*	1.118*	65%	**0.019**	0.400	3.088*	40%	0.066*	0.278*	1.085*

Mistral-7B-Instruct	MMLU				OpenbookQA				GSM8K				MedQA			
	Route	ECE	BS	CE	Route	ECE	BS	CE	Route	ECE	BS	CE	Route	ECE	BS	CE
Verbalizing Uncertainty	100%	0.126	0.313	8.725	100%	0.221	0.434	10.846	100%	0.163	0.669	3.822	100%	0.087	0.564	2.480
Verbalizing Yes/No Token	100%	0.297	0.345	4.380	100%	0.252	0.394	5.863	100%	0.146	0.821	23.98	100%	0.090	0.579	17.14
Zero-shot Logits	95%	0.311*	0.353	1.643	100%	0.232	0.442	2.498	100%	**0.046**	0.398	1.957	100%	0.087	0.507*	2.409
Finetuned Logits	95%	**0.173**	**0.263**	**0.751**	50%	**0.143**	0.207*	**0.627**	100%	0.050*	0.323	1.236*	90%	**0.023**	**0.226**	**0.633**
Self-REF - Mistral-7B-Instruct	70%	0.369	0.293*	1.021*	40%	0.157*	**0.193**	0.869*	75%	0.068	**0.273**	**0.968**	70%	0.037*	**0.226**	0.653*

Note: Lower scores are better for ECE and BS.

Takeaway: Downstream utility is more valuable than perfect probability alignment.

Result 3: Zero-Shot Transferability

THE TEST

Generalizing Self-Reflection

LoRA weights trained exclusively on **OpenbookQA** were deployed for inference on the **MMLU dataset** to test for conceptual learning over memorization.

THE RESULT

Transferred weights maintain competitive utility:

- Outperformed all baselines on MMLU despite never seeing the data during training.
- **Mistral-7B:** Achieved parity with 70B model at a 75% routing rate.
- **Llama3-8B:** Required 70% routing rate to match the 70B model's accuracy.

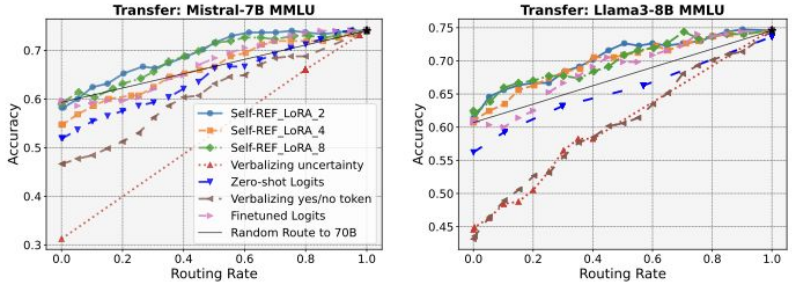


Figure 5: Transferability of Self-REF, assessed on the routing task. Local LLMs are fine-tuned with Self-REF on OpenbookQA and inference is run on the MMLU dataset.

Local LLMs are fine-tuned with Self-REF on OpenbookQA and inference is run on the MMLU dataset.

Self-REF: Proving the model learned the concept of self-reflection, not just dataset memorization.

Result 4: When Routing Saves the Day

EXAMPLE 1

8B + Self-REF > 70B (Base)

Case: Complex moral dispute

Prompt: *"...versions of the ticking bomb hypothetical she discusses are 'stunningly stupid,' but she claims this is actually evidence of?"*
Base 70B: `[A] the stupidity of most traditional philosophical examples.` (Incorrect)
Self-REF 8B: `[D] the readiness on the part of many intelligent people... <CN>` (Correct + Confident) Seeing the literal `<CN>` and `<UN>` tokens in action makes the architecture feel real.

Result: 8B gets it right + **<CN>**

EXAMPLE 2

Appropriate Routing

Case: Abstract Algebra question

Prompt: *"Statement 1: Every function from a finite set onto itself must be one to one. Statement 2: Every subgroup of an abelian group is abelian."*
8B Response (Self-REF): `[C] True, False. <UN>` → *(Incorrect + Unconfident → Triggers Route)*
70B Response (Expert): `[A] True, True.` → *(Correct!)*

Result: 8B identifies failure **<UN>** 70B Corrects

Self-REF: Tangible accuracy gains through intelligent model collaboration

Result 5: Understanding Model Failure Modes

OVERCONFIDENCE

False Positives in Technical Domains

The model often generates `<CN>` for incorrect answers in highly structured fields.

Affected Domains:

- College Computer Science
- Conceptual Physics
- Jurisprudence

The Blind Spot: Model recognizes technical syntax and patterns but misses nuanced, edge-case reasoning.

UNDERCONFIDENCE

False Negatives in Ambiguous Contexts

Model becomes unnecessarily hesitant, outputting `<UN>` despite being correct.

Affected Domains:

- High School European History
- Human Sexuality
- Interpretive Subjects

The Blind Spot: Hesitation occurs on context-heavy, ambiguous topics requiring complex qualitative interpretation.

Self-REF: Honestly presenting the model's blind spots to improve safety-critical routing

Conclusion: The Self-REF Advantage

FINAL SUMMARY

Granular Control

Precise thresholds for **cost**, **latency**, and **safety**.

Confidence scores enable customizable routing logic.

Zero Degradation

Learning confidence tokens does **not** harm base LLM performance.

Preserves original reasoning capabilities.

Effective Filtering

Operates as a high-fidelity filter for **Routing** and **Abstention**.

Significant ROI in real-world tasks.

Self-REF provides the missing link between raw model output and reliable system-level utility.

Critiques on the Method

DRAWBACK OF THRESHOLD-BASED METHODS

- Global confidence threshold
  - Acceptable for multiple choice questions in this paper, but not generalizable to more open-ended cases [1].
  - Sensitive to distribution shift & fine tuning dataset.
  - No adaptation to different risk thresholds [2].
- Threshold-based Methods for UQ are generally questionable:
  - Benchmark against other methods is hard [3]
  - Single threshold conflates different sources of uncertainty (i.e. Aleatoric vs Epistemic uncertainty) [4]
    - So far UQ methods are benchmarked on zero-aleatoric datasets
  - No difference for different reasons of uncertainty - route or abstain?

References

[1] Rubashevskii, A., Piatrashyn, D., Nakov, P., & Panov, M. (2026). Adaptive Conformal Prediction for Improving Factuality of Generations by Large Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2604.13991
[2] Wu, S., Gustafsson, F. K., Phillips, E., Gao, B., Thakur, A., & Clifton, D. A. (2026). BAS: A Decision-Theoretic Approach to Evaluating Large Language Model Confidence. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2604.03216
[3] Vashurin, R., Fadeeva, E., Vazhentsev, A., Rvanova, L., Vasilev, D., Tsvigun, A., … Shelmanov, A. (03 2025). Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. Transactions of the Association for Computational Linguistics, 13, 220–248. doi:10.1162/tacl_a_00737
[4] Walha, N., Gruber, S. G., Decker, T., Yang, Y., Javanmardi, A., Hüllermeier, E., & Buettner, F. (2025). Fine-Grained Uncertainty Decomposition in Large Language Models: A Spectral Approach. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2509.22272
[5] Cohen, R., Dobler, K., Biran, E., & de Melo, G. (2024). I Don't Know: Explicit Modeling of Uncertainty with an [IDK] Token. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2412.06676