LLMs are Overconfident: Evaluating Confidence Interval Calibration with FermiEval

Elliot L. Epstein John Winnicki Thanawat Sornwanee Rajat Dwaraknath

AIR-FM Workshop@AAAI 2026
Assessing and Improving Reliability of Foundation Models in the Real World

Why Confidence Intervals for LLM Estimates?

Motivation: LLMs are strong at numerical estimation, but we also need reliable uncertainty.

- Many real-world tasks need a range (risk assessment, planning, decision support).
- A nominal 95% confidence interval should contain the truth about 95% of the time.
- We study: Do LLMs' confidence intervals match their stated coverage?

Key finding (from our experiments): models are systematically overconfident.

FermiEval Benchmark

Task: Fermi-style questions with order-of-magnitude ground truth.

- Dataset source: Science Olympiad Fermi questions.
- Labels are base-10 exponents: $y \in \mathbb{Z}$ corresponds to 10^y.
- Split: 500 train / 500 test, filtered to 10^{-100} to 10^{100}.

Prompted output: an interval in exponent space, e.g. $[10^L, 10^U]$ with integer $L \leq U$.

Example: Measuring Calibration

Example question (from the benchmark):

How many pennies would it take to cover the state of Pennsylvania?

Ground truth label: $y = 13$ (meaning the answer is on the order of 10^{13}).

Model output format: for a target level p (e.g., 95%), the model returns integers (L, U) defining $[10^L, 10^U]$.

Coverage check (per item): count it as "covered" if $y \in [L, U]$. Over the whole dataset, observed coverage is the fraction of items covered, and we plot observed vs. nominal p.

Calibration plot:

for targets $p \in \{0.90, 0.95, 0.99, 0.998\}$, compare

$$\text{nominal } p \quad \text{vs. observed coverage } \widehat{\Pr}(y \in [L, U]).$$

Method: Conformal Calibration

Idea: treat LLM intervals as base intervals, then learn a single safety margin from a held-out calibration set.

- **Train vs. test:** use the train split to form a calibration subset (to learn q), then report coverage on the untouched test split (with q fixed).
- For each calibration question, measure how far the truth falls outside the base interval:

$$s_i = 0 \text{ if } y_i \in [L_i, U_i], \quad \text{else } s_i = \text{distance to the interval.}$$

- Pick q so that about a $(1-\alpha)$ fraction of calibration examples have $s_i \leq q$.
- Then widen every future interval by the same amount q to fix under-coverage.

$$s_i = \max\{L(x_i) - y_i,\, y_i - U(x_i)\}, \qquad q_{1-\alpha} = (1-\alpha)\text{-quantile of } \{s_i\}_{i=1}^{n}$$

$$\mathrm{CI}^{\mathrm{conf}}(x) = [\, L(x) - q_{1-\alpha},\, U(x) + q_{1-\alpha}\,].$$

Guarantee: under exchangeability, $\Pr\{y \in \mathrm{CI}^{\mathrm{conf}}(x)\} \geq 1 - \alpha$.

Result: Confidence Intervals Are Overconfident

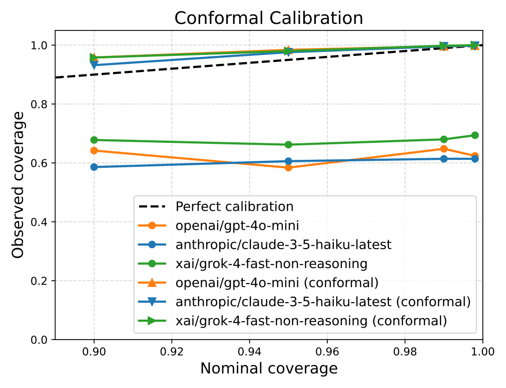

Figure 1: Calibration curves for representative models. Dashed line is perfect calibration ($y = x$).

- Coverage is below nominal: overconfident intervals.
- Coverage plateaus as nominal p increases.
- Nominal 99% covers only $\sim$ 65% on average.
- Same overconfidence appears for several open-weight models (appendix).
- Other fixes: log-probability elicitation (appendix) and multi-query quantile aggregation.

$\Rightarrow$ **Conformal calibration lifts observed coverage from $\sim$ 65% at nominal 99% to $\sim$ 99% (nominal).**

Interval Quality: Winkler Score

Beyond coverage: we also want intervals to be sharp (not overly wide). **Winkler score** (lower is better) combines width + a penalty when the truth is outside:

$$\mathrm{WS} = (U - L) + \frac{2}{\alpha}\left|y - \mathrm{proj}_{[L,U]}(y)\right| \qquad (\alpha = 1 - p).$$

Observed: for $p = 0.99$, conformal calibration reduces the average Winkler score by 54%.

Takeaways

1. We propose FermiEval: a benchmark for confidence-interval calibration on Fermi-style estimation.
2. We find significant overconfidence: observed coverage is far below nominal and plateaus for large nominal levels.
3. We propose an efficient conformal method that brings coverage back to nominal levels.
4. We propose a perception-tunnel hypothesis explaining why LLMs under-represent tails.