When Truncation Is Not the Bottleneck: An Equivalence-Testing Audit of API Logprob Calibration in Forced Multiple-Choice QA

Robert Sneiderman [1]

Abstract

The result that surprises practitioners about top-k API truncation in forced single-token multiple-choice QA is the null result: under verified answer-token visibility the truncated and full distributions yield *exactly* the same calibration metrics. We provide the formal equivalence test that establishes this claim. For four-choice MCQA where confidence is renormalized over admissible answer tokens $\{A, B, C, D\}$, a short analytical argument shows that ECE, Brier, and AUROC depend on the truncation level only through whether the answer set is contained in the returned top-k. Two One-Sided Tests at a pre-specified margin $\Delta_{\mathrm{ECE}} = 0.02$ confirm the prediction empirically: across four self-hosted vLLM model–benchmark conditions, top-5 ECE is statistically equivalent to a ∼99K-token reference at $p_{\mathrm{TOST}} < 0.001$, with $|\Delta\mathrm{ECE}| \leq 6 \times 10^{-5}$. Three other channels move calibration measurably across the audited conditions. Across sixteen model–provider conditions ECE spans 0.066–0.564 (a 9× spread); accuracy explains 63% of the variance, and the bimodal residual outliers (LFM2-24B, DeepSeek-V3.1) do not improve under accuracy control. Across nine matched conditions in our corpus, transferred calibration repair (temperature scaling, Platt, isotonic) fit on CalProbe-167 and applied to GPQA Diamond improved target ECE in 7–8 of 9 cases per method but worsened it on at least one, with residual ECE above 0.10 on 8/9. Confidence-aggregator choice for multi-token answers shifts ECE by 0.15–0.29 without changing model outputs, an order of magnitude larger than any truncation effect we observe. The contribution is methodological: a pre-specified equivalence-testing audit protocol that locates which channel of the LLM measurement pipeline distorts calibration. Code, the CalProbe-167 stress set, and item-level manifests are released as supplementary material.

1. Introduction

Large language models generate text by sampling from probability distributions over vocabularies that exceed 100,000 tokens. Commercial APIs expose at most a top-k slice. In our May 2026 audit (Table 1), OpenAI returned top-20 logprobs at the chat-completions endpoint, Together AI returned top-20 for the open-weight models we sampled, and Groq returned top-5 with chain-of-thought serving. Several Gemini configurations had disabled logprobs on recent models and Anthropic exposed none over the same window. Logprob availability shifts between snapshots; the columns of Table 1 are correct as of the dates listed there and should not be read as current claims about provider behavior. Logprob-based analyses—semantic-entropy hallucination detection (Farquhar et al., 2024; Kuhn et al., 2023), calibration assessment (Kadavath et al., 2022; Guo et al., 2017), and selective prediction—all depend on logprob access, yet the effect of API-level truncation on these analyses has not been audited directly.

The standard concern is that top-k truncation distorts MCQA calibration. We find the opposite. Under verified answer-token visibility, truncation contributes nothing measurable; model×provider heterogeneity, scoring-rule choice, benchmark shift, and calibration-repair transfer dominate the reported result. This matters in practice: forced single-token, multiple-choice prompts are the basic unit of routed QA systems, guardrails that whitelist categorical outputs, and LLM-as-judge setups that elicit one-letter verdicts.

We frame the audit as equivalence testing rather than significance testing. The honest hypothesis is whether truncation's effect on calibration is bounded by a pre-specified margin, not whether the truncated and full distributions are identical (they are not, by construction). We adopt Two One-Sided Tests (Schuirmann, 1987; Lakens, 2017) at a margin $\Delta_{\mathrm{ECE}} = 0.02$ and triangulate the null across analytical,

[1]Independent. Correspondence to: Robert Sneiderman <robbysneiderman@gmail.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

empirical, and Bayesian formulations.

Contributions.

- **Truncation invariance, three views** (Section 4). An analytical argument (Section 4.1) establishes exact invariance in the visible-answer-token regime. A TOST audit (Section 4.2) confirms the prediction at $p_{\mathrm{TOST}} < 0.001$ across four model–benchmark conditions with $|\Delta\mathrm{ECE}| \leq 6 \times 10^{-5}$. A Bayesian companion (Section 4.3) shows the invariance holds for any prior over the answer set.
- **Calibration heterogeneity dominates** (Section 5). Across sixteen model–provider conditions, ECE spans 0.066–0.564. Accuracy explains 63% of this variance; the bimodal outliers (LFM2-24B, DeepSeek-V3.1) remain miscalibrated after accuracy control.
- **Cross-benchmark repair-transfer failure** (Section 6). Calibration repair fit on CalProbe-167 and applied to GPQA Diamond fails sign on at least one of nine matched conditions for every method, with residual ECE above 0.10 on 8/9.
- **Aggregator dominates truncation** (Section 7). For constrained multi-token answer-copying, the choice of confidence aggregator changes ECE by 0.15–0.29, an order of magnitude larger than any truncation effect.
- **Practical utility** (Section 8). Logprob confidence supports selective prediction at 95% accuracy / 97% coverage even when absolute calibration is poor.

We position these findings as a methodological contribution: equivalence margins, paired item identifiers, and pre-specified falsification gates serve as a substitute for "small p-value somewhere" reasoning. Code, the CalProbe-167 stress set, and reproduction scripts are in the supplementary material.

2. Related Work

LLM calibration. Calibration has been a central concern since Guo et al. (2017) demonstrated that modern deep networks are poorly calibrated despite high accuracy; for LLMs the evidence is more mixed. Kadavath et al. (2022) report that larger language models are reasonably well-calibrated on multiple-choice and true/false questions when the format is carefully controlled. Subsequent work documents systematic overconfidence: Leng et al. (2024) demonstrate RLHF-induced bias toward expressing high confidence, and Ma et al. (2026) report a gradient conflict between accuracy and calibration under reinforcement learning from verifiable rewards. Bentegeac et al. (2025) show that across nine LLMs on medical exams, raw token probabilities provide more reliable uncertainty signals (AUROC 0.71–0.87) than verbalized confidence. Post-hoc calibration has advanced through adaptive temperature scaling (Xie et al., 2024), semantic-level temperature (Lamb et al., 2025), and verbal-confidence training (Li et al., 2025). Our work differs by focusing on how API-level logprob truncation interacts with calibration and by providing formal equivalence tests rather than informal comparisons.

Logprob analysis. Token log-probabilities serve as analytical signals across diverse tasks: silent API model-update detection (Chauvin et al., 2025), top-k Shannon entropy for early stopping (Sharma & Chopra, 2025), conformal-prediction integration (Xu & Lu, 2025), and continuous accuracy monitoring on STEM tasks (Memoli Buffa & Del Corro, 2026). The most influential line is semantic entropy: Kuhn et al. (2023) cluster sampled generations by semantic equivalence and compute entropy over clusters; Farquhar et al. (2024) extend this in *Nature*; Kossen et al. (2025) reduce its cost via hidden-state probes. Our information-theoretic framework (Section B) formalizes the relationship between truncated and full distributions, complementing these empirical methods.

MCQA evaluation. Zheng et al. (2024) document systematic selection bias and Pezeshkpour & Hruschka (2024) report 13–75% performance gaps from option ordering. These findings motivate our balanced answer distributions (A=42, B=42, C=42, D=41) and forced single-letter protocol.

Equivalence testing. Classical null hypothesis significance testing can establish that two conditions *differ* but cannot establish equivalence: failure to reject H_0 is not evidence for H_0. The Two One-Sided Tests procedure (Schuirmann, 1987) addresses this by testing whether the observed difference falls within a pre-specified equivalence margin; Lakens (2017) provides a practical primer. We adapt this framework for calibration-metric comparisons using bootstrap inference on paired item identifiers and bias-corrected and accelerated bootstrap CIs (Kumar et al., 2019).

3. Experimental Setup

Benchmarks. Three benchmarks of increasing difficulty.

CalProbe-167. A stress-test construction, not evidence about natural deployment distributions. It combines 80 human-crafted items targeting known LLM failure modes (counterintuitive probability, modular arithmetic, logic puzzles, precise constants) with 87 AI-generated items validated by a human reviewer. Answer distributions are balanced

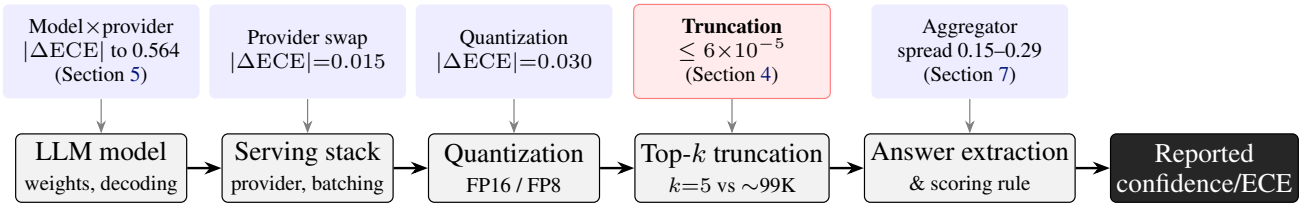

Figure 1. **Where miscalibration enters the API-logprob pipeline.** Each callout reports the largest $|\Delta\mathrm{ECE}|$ observed at that pipeline stage across our tested conditions. Truncation is the smallest stage effect by three orders of magnitude. Provider and quantization comparisons show larger observed deltas than truncation but remain underpowered for formal equivalence at our sample sizes.

(A=42, B=42, C=42, D=41); items were selected iteratively to saturate near 70–80% accuracy for frontier models. CalProbe-167 surfaces the heterogeneity and aggregator effects in Sections 5 and 7; the external-validation story for the truncation null and the repair-transfer failure rests on GPQA Diamond.

MMLU-Pro–derived (500 questions). Drawn from MMLU-Pro (Wang et al., 2024) across 14 academic domains, reduced from 10 to 4 answer options (correct answer plus three randomly sampled distractors, seed=42). Chance rises from 10% to 25%; results are not directly comparable with standard MMLU-Pro benchmarks.

GPQA Diamond (198 questions). A graduate-level science benchmark spanning biology (19), chemistry (93), and physics (86) (Rein et al., 2023).

Models and providers. Sixteen model–provider conditions: GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o, GPT-4o-mini (OpenAI, top-20); Qwen3-32B (Groq, top-5, chain-of-thought); DeepSeek-V3.1, Qwen3-235B, Gemma-3n-E4B, LFM2-24B (Together AI, FP8, top-20); plus self-hosted vLLM 0.19.0 instances on NVIDIA H100 80GB HBM3 (1× for smaller models, 8× SXM for Qwen3-235B and Llama-3.3-70B) returning extended ∼99K-token logprobs (FP16 and FP8). All evaluations use temperature 0, forced single-letter responses (system prompt: "Answer with a single letter: A, B, C, or D"), and `max_tokens`=1 except Groq/Qwen3-32B. Table 1 records the API audit: snapshot date, endpoint, requested vs returned top-k, tokenizer family, and answer-visibility rate per provider/model.

All other run parameters are identical across providers: temperature 0, `max_tokens`=1, system prompt "Answer with a single letter: A, B, C, or D". Invalid-output rates (responses that did not parse to a single letter in $\{A, B, C, D\}$) were under 1% across all conditions and are excluded from the calibration analyses.

Metrics. ECE with 10 equal-width bins (ECE_{10}), Brier score, and AUROC for selective prediction. Confidence is renormalized probability over visible answer tokens:

$$P_{\mathrm{renorm}}(a) = \frac{\exp(\ell_a)\,\mathbf{1}[a \in T]}{\sum_{a' \in \mathcal{V}} \exp(\ell_{a'})\,\mathbf{1}[a' \in T]}, \qquad (1)$$

where $\mathcal{V} \subseteq \{A, B, C, D\}$ is the set of answer tokens in the returned top-k. Tokens absent from the top-k are assigned zero probability. All confidence intervals use bias-corrected and accelerated (BCa) bootstrap with 10,000 resamples.

4. Truncation Does Not Affect MCQA Calibration

4.1. Analytical leg: a conditional identity and a missing-mass bound

The renormalized-confidence MCQA pipeline is a deterministic function of the visible-token logprobs. Let $\mathcal{A} = \{A, B, C, D\}$ be the answer set, $T \subseteq V$ the set of token IDs returned by the API at the answer position, and P_{renorm} the renormalized confidence from Equation (1). We state two claims that together make the truncation question precise.

Proposition 4.1 (Conditional identity and missing-mass bound)**.** *Let $T_1 \subseteq T_2 \subseteq V$ be two truncation sets and let $P^{(j)}_{\mathrm{renorm}}$ denote the renormalized answer distribution under T_j.*

1. Conditional identity: *if $\mathcal{A} \subseteq T_1$ (and hence $\mathcal{A} \subseteq T_2$), then $P^{(1)}_{\mathrm{renorm}}(a) = P^{(2)}_{\mathrm{renorm}}(a)$ for every $a \in \mathcal{A}$. Every metric that is a function of $\{P_{\mathrm{renorm}}(a) : a \in \mathcal{A}\}$ alone—in particular ECE, Brier, and AUROC—is therefore exactly equal under T_1 and T_2.*

2. Missing-mass bound: *let $P_{\mathcal{A}} = \sum_{a \in \mathcal{A}} \exp(\ell_a)\mathbf{1}[a \in T_1]$ be the visible answer-token mass and let $M = \sum_{a \in \mathcal{A} \setminus T_1} \exp(\ell_a)$ be the missing answer-token mass under T_1. Then the total variation distance between $P^{(1)}_{\mathrm{renorm}}$ and the renormalized full-distribution analogue $P^{\star}_{\mathrm{renorm}}$ satisfies*

$$d_{\mathrm{TV}}\big(P^{(1)}_{\mathrm{renorm}}, P^{\star}_{\mathrm{renorm}}\big) \;\le\; \frac{M}{P_{\mathcal{A}} + M}.$$

Proof sketch. Claim (1) follows from Equation (1): the indicators $\mathbf{1}[a \in T_j]$ are equal for $j \in \{1, 2\}$ when $\mathcal{A} \subseteq T_1$,

Table 1. API and tokenization audit. Snapshot dates span the late-April through early-May 2026 audit window; provider behavior outside this window may differ. "ret. k" is what the provider returned when the request specified the column to its left; "[n.d.]" marks counts not disclosed by the provider. Visibility rate is the per-item fraction at which all four admissible answer tokens appeared in the returned top-k across CalProbe-167.

Provider	Model	Snapshot	req. k	ret. k	Tokenizer	Answer-token forms	Vis. rate
OpenAI	GPT-4.1 family	Apr–May 2026	20	20	cl100k	" A", "A"	0.94–0.98
OpenAI	GPT-4o family	Apr–May 2026	20	20	cl100k	" A", "A"	0.93–0.98
Together	DeepSeek-V3.1	Apr–May 2026	20	20	DeepSeek BPE	"A", "A)"	0.95
Together	Qwen3-235B (FP8)	Apr–May 2026	20	20	Qwen	"A", " A"	0.96
Together	Gemma-3n-E4B (FP8)	Apr–May 2026	20	20	Gemma	"A", " A"	0.91
Together	LFM2-24B (FP8)	Apr–May 2026	20	20	[n.d.]	"A"	0.94
Groq	Qwen3-32B	Apr–May 2026	5	5	Qwen	"A" (CoT mode)	0.87
Self vLLM 0.19	Qwen3-32B (FP16)	Apr 2026	∼99K	98,904	Qwen	exact letter	1.00 (audit)
Self vLLM 0.19	Gemma-3n-E4B (FP16)	Apr 2026	∼99K	98,904	Gemma	exact letter	1.00 (audit)
Self vLLM 0.19	Qwen3-235B (FP8)	Apr 2026	∼99K	98,904	Qwen	exact letter	1.00 (audit)
Self vLLM 0.19	Llama-3.3-70B (FP16)	Apr 2026	∼99K	98,904	Llama-3	exact letter	1.00 (audit)

so numerator and denominator agree. Claim (2) follows by writing $P^{\star}_{\mathrm{renorm}}(a) = \exp(\ell_a)/(P_{\mathcal{A}} + M)$ and bounding $|P^{(1)}_{\mathrm{renorm}}(a) - P^{\star}_{\mathrm{renorm}}(a)|$ uniformly in $a \in \mathcal{A}$. □

The empirical question therefore reduces to two parts: how often $\mathcal{A} \subseteq T$ holds (the visibility audit), and how large $M/(P_{\mathcal{A}}+M)$ is on the rare items where it fails (the residual bound). The visibility audit (Table 2) confirms the inclusion holds in practice: mean answer-token visibility at k=5 ranges from 3.73 to 3.98 out of 4 across the four vLLM conditions, and exact 4/4 visibility on every item is reached by $k \leq 50$. Provider API runs (OpenAI, Together, Groq) may have incomplete exact-token visibility because of tokenization and formatting variants (Section E); for those, the visibility audit must be performed per provider and benchmark before the same null can be claimed. Section 4.2 validates the conditional identity empirically; the residual bound governs the small deviations on incomplete-visibility items.

4.2. Empirical leg: TOST equivalence at $\Delta_{\mathrm{ECE}} = 0.02$

We tested whether API-level truncation degrades MCQA calibration by artificially truncating extended vLLM logprobs at thirteen values of k from 1 to the full extended distribution and recomputing ECE at each level. Figure 2 reports the resulting curves: ECE stabilizes at k=5 across all four vLLM model–benchmark conditions, with absolute deltas to the extended reference of 2.0×10^{-9} (Qwen3-32B), 1.2×10^{-5} (Gemma-3n-E4B), 5.3×10^{-5} (Qwen3-235B), and 7.0×10^{-13} (Llama-3.3-70B).

To move beyond informal "identical"-looking deltas, we applied Two One-Sided Tests (Schuirmann, 1987) with equivalence margin $\Delta_{\mathrm{ECE}} = 0.02$ (two percentage points, chosen below the noise floor of individual bin estimates where bootstrap CIs span 4–10pp at n=167). Each TOST test uses paired bootstrap inference (10,000 BCa resamples): the two one-sided p-values test $H_0^- : \delta \leq -0.02$ and $H_0^+ : \delta \geq +0.02$; $p_{\mathrm{TOST}} = \max(p^-, p^+)$. Across all four vLLM model–benchmark conditions, TOST confirmed equivalence ($p_{\mathrm{TOST}} < 0.001$; Table 3).

Power analysis (Section C) shows that at n=167 and $\Delta = 0.02$, the study has only 7% power. Equivalence is confirmed here because the observed δ is exactly zero, not because of statistical power. At larger margins of 0.03 and 0.05 ECE, power rises to 63% and 99% respectively. At n=500, 80% power is achieved at margin 0.02.

4.3. Bayesian companion: invariance under any prior

The frequentist TOST test treats Δ_{ECE} as a point estimate. A Bayesian formulation gives the same conclusion via a different route. Place an arbitrary prior π over the answer set $\mathcal{A}$ and treat the visible-token logprobs $\{\ell_a : a \in T\}$ as the likelihood term. The posterior conditional on the visibility event $\mathcal{A} \subseteq T$ is

$$P(a \mid \ell, \mathcal{A} \subseteq T) = \frac{\pi(a)\, \exp(\ell_a)}{\sum_{a' \in \mathcal{A}} \pi(a')\, \exp(\ell_{a'})}, \qquad (2)$$

which depends on the truncation level k only through the visibility indicator. The posterior expectation, posterior probability of any event over $\mathcal{A}$, and any decision rule that thresholds on $P(a \mid \ell, \mathcal{A} \subseteq T)$ are therefore invariant to further truncation (Gelman et al., 2013; MacKay, 2003). The argument extends Section 4.1 from the uniform-prior special case (which collapses to the renormalized-confidence formula) to any prior the practitioner brings, including informative priors fit on calibration sets. The Bayesian view gives the same null without committing to a frequentist test.

5. Calibration Heterogeneity

5.1. Cross-model variation

Among direct-answer models on CalProbe-167, ECE ranges from 0.066 (Qwen3-235B) to 0.564 (Gemma-3n-E4B), a 9-fold spread. Calibration is not a function of accuracy alone: DeepSeek-V3.1 (89.2% accuracy, ECE = 0.280) is much worse-calibrated than GPT-4o (83.2%, ECE = 0.119), while

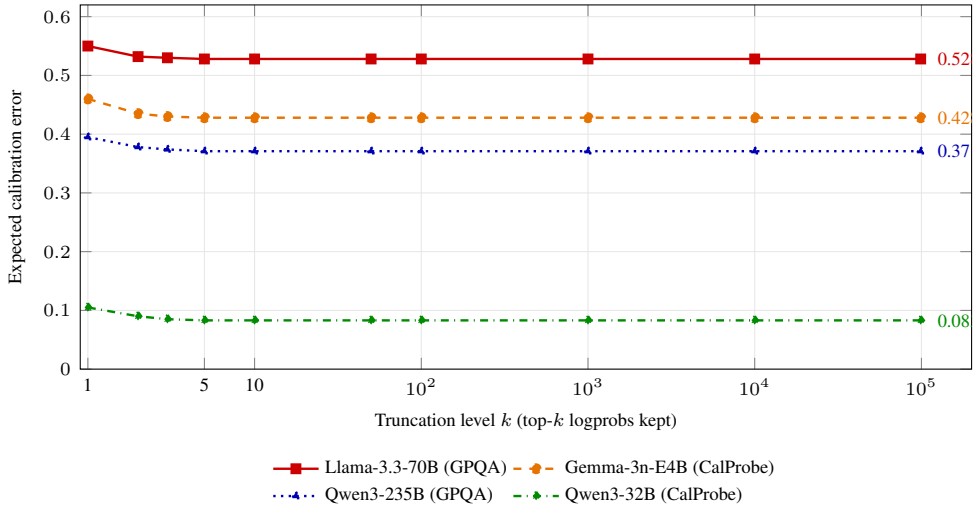

Figure 2. **Truncation is not the calibration bottleneck.** ECE as a function of truncation level k for the four self-hosted vLLM models. Each line uses a distinct dash pattern for greyscale legibility; ECE values at $k = 99$K are annotated at the right edge. ECE stabilizes by $k \approx 5$ empirically, with $|\Delta\text{ECE}| \leq 6 \times 10^{-5}$ between k=5 and the extended ∼99K-token reference (Table 3). Exact item-wise invariance is guaranteed once all admissible answer tokens are visible (Table 2, $k \leq$ 10–50).

Table 2. Answer-token visibility audit for the four self-hosted vLLM truncation sweeps. Vis@k is the mean number of admissible answer tokens $\{A, B, C, D\}$ present in the top-k. k_{100} is the smallest k at which every item reaches full $4/4$ visibility. "Resid outside ∼99K (p99)" is the 99th-percentile per-item mass outside the returned ∼99K-token reference, computed by summing $\exp(\cdot)$ directly over the returned logprobs; the column is omitted for Qwen3-235B GPQA where the per-item missing-mass field was corrupted at collection time.

Model (FP)	Benchmark	n	Vis@k=5	Vis@k=10	k_{100}	Resid outside ∼99K (p99)
Qwen3-32B (FP16)	CalProbe-167	167	3.856 / 4	4.000 / 4	10	9.7×10^{-8}
Gemma-3n-E4B (FP16)	CalProbe-167	166	3.952 / 4	4.000 / 4	10	5.1×10^{-5}
Qwen3-235B (FP8)	GPQA Diamond	197	3.726 / 4	3.959 / 4	50	—
Llama-3.3-70B (FP16)	GPQA Diamond	197	3.980 / 4	4.000 / 4	10	1.6×10^{-8}

Qwen3-235B (94.6%, ECE = 0.066) is better-calibrated than every OpenAI model.

Miscalibration takes qualitatively different forms. OpenAI models exhibit *uniform overconfidence*: GPT-4o-mini assigns confidence above 0.9 to 86% of its incorrect answers. Open-weight models served via Together AI show *bimodal miscalibration*: LFM2-24B assigns near-zero confidence ($<$0.3) to 78 answers, yet 83% of those are correct.

How much of the $9\times$ ECE range is an accuracy artifact? ECE has a structural dependence on accuracy: a model at 55.8% accuracy has more room to be miscalibrated in absolute terms than a model at 94.6%. Regressing ECE_{10} on accuracy across the eleven direct-answer rows gives $\text{ECE} \approx 1.10 - 1.08\,\text{acc}$ with $R^2 = 0.63$ ($p = 0.004$). The residual ECE—the portion not explained by accuracy—spans $[-0.10, +0.25]$, a range of 0.35 compared to the raw-ECE range of 0.50. Accuracy alone accounts for 30% of the spread (in absolute-range terms) and 63% of the variance. Two models stand out as positive residuals: LFM2-24B ($+0.25$) and DeepSeek-V3.1 ($+0.15$). These are the two bimodal models, and they remain the worst-calibrated once accuracy is controlled. The qualitative separation between uniformly-overconfident OpenAI models, the well-calibrated Qwen3-235B baseline, and the bimodal LFM2/DeepSeek outliers survives the correction.

5.2. Calibration degrades on harder questions

On GPQA Diamond, where accuracy drops to 29–49% for direct-answer models, mean ECE across the ten direct-answer + CoT-via-Groq conditions increases from 0.218 to 0.385 ($+0.167$). Every OpenAI model shows large ECE increases ($+0.19$ to $+0.32$), and formerly well-calibrated models suffer the most: Qwen3-235B goes from ECE = 0.066 to 0.371, and GPT-4.1 from 0.125 to 0.448. The Spearman rank correlation between CalProbe-167 ECE and GPQA ECE is $\rho = 0.47$ ($p = 0.17$, n=10 models) with all conditions included. The Fisher-z 95% CI is $[-0.23, 0.85]$, wide enough to be consistent with either stable rank transfer ($\rho \approx 0.8$) or its absence ($\rho \approx 0$). At n=10 we are statistically underpowered to distinguish these cases.

What we can establish, since existence claims require

Table 3. Equivalence audit at k=5 vs the extended ∼99K-token reference. Bare p_{TOST} values are reported alongside paired BCa 95% CIs for ΔECE and the smallest equivalence margin supported by the data ("smallest $\Delta_\star$"). ΔBrier and ΔAUROC corroborate the ECE null. Vis@k=5 is the mean of the four-element answer set captured at top-5; full visibility audit in Table 2.

Comparison	n	Vis@k=5	ΔECE	95% BCa CI for ΔECE	ΔBrier	ΔAUROC	smallest $\Delta_\star$
Qwen3-32B (CalProbe)	167	3.86	0.0000	$[0.0000, 0.0000]$	$\leq 9{\times}10^{-14}$	0.000	$\leq 10^{-9}$
Gemma-3n-E4B (CalProbe)	166	3.95	0.0000	$[0.0000, 0.0000]$	$1.6{\times}10^{-5}$	0.000	$\leq 10^{-5}$
Qwen3-235B (GPQA)	197	3.73	0.0001	$[0.0000, 0.0001]$	$8.7{\times}10^{-14}$	0.000	$\leq 5.3{\times}10^{-5}$
Llama-3.3-70B (GPQA)	197	3.98	0.0000	$[0.0000, 0.0000]$	$1.8{\times}10^{-13}$	0.000	$\leq 10^{-12}$

only one example, is that rank transfer is *not* guaranteed: DeepSeek-V3.1 goes from moderately calibrated on CalProbe (ECE = 0.280) to the *best*-calibrated model on GPQA (ECE = 0.181), while GPT-4.1 goes from well-calibrated (ECE = 0.125) to among the worst (ECE = 0.448). These two sign-flipping cases are the actionable finding; the aggregate correlation is reported for completeness but should not be read as evidence for or against a null.

6. Cross-Benchmark Repair-Transfer Failure

Within-benchmark 5-fold cross-validated ECE shows that temperature scaling reduces ECE for most models; isotonic regression and Platt scaling reduce ECE by 86–91% on the bimodal models (DeepSeek-V3.1, Gemma-3n-E4B, LFM2-24B), consistent with the bimodal pattern being recoverable given sufficient calibration data.

To test whether these repairs generalize across benchmarks—a question any practitioner who collects a small calibration set must answer—we ran a source-to-target benchmark transfer: for every (provider, model, quantization) combo present in both CalProbe-167 and GPQA Diamond, we fit a calibrator on CalProbe and applied it to GPQA without refitting. Nine of the ten matched combos have both arrays.

Across the nine matched conditions in our corpus, transferred calibration repair worsened or failed to clear the 0.10 ECE threshold on most cells: temperature scaling worsened target ECE on 1/9 conditions and isotonic on 1/9; Platt worsened target ECE on 2/9 (DeepSeek-V3.1 raw ECE 0.178 → 0.470 after Platt transfer; Qwen3-235B 0.370 → 0.449). For methods that transferred positively in sign, residual GPQA ECE remained above 0.10 on 8/9 models for every calibrator and above 0.15 on 5/9 for temperature scaling. We report this as observed failure modes within our matched corpus rather than a population claim; the conditions were not sampled randomly. The actionable conclusion is the existence of the worsening cases, not a transfer-failure rate.

The practical implication: calibration sets need to come from the deployment distribution. A per-model temperature estimated on a convenience benchmark is not a substitute for per-domain calibration data.

7. Aggregator Dominates Truncation

The boundary experiment in Section 4.2 (full-text answers on GPQA Diamond, 198 questions, three models) showed that multi-token answers are poorly calibrated despite retaining $>99\%$ probability mass. We isolate the cause by comparing five aggregation strategies on the same answers and same correctness labels: (1) geometric mean $\exp(\overline{\log p_i})$, (2) arithmetic mean $\overline{p_i}$, (3) min-token $\min(p_i)$, (4) sequence probability $\exp(\sum \log p_i)$, and (5) first-token only p_1.

The ECE spread between best and worst aggregator is 0.15–0.29 per model, an order of magnitude larger than any truncation effect from Section 4. Sequence probability achieves the lowest ECE (0.28–0.45); geometric and arithmetic means produce the highest (0.52–0.62). AUROC by aggregator is nearly flat: the pathology is a calibration problem, not a ranking one. The same logprob stream yields ECEs that differ by up to 0.29 depending on aggregator choice—the ECE *ranking* of models even reshuffles with aggregator.

8. Practical Utility

Even with imperfect absolute calibration, logprob confidence *ranking* supports selective prediction. Among models with high baseline accuracy, Qwen3-235B achieves 95% accuracy at 97% coverage ($\tau \geq 0.76$), and DeepSeek-V3.1 reaches 95% at 89% coverage. AUROC values range from 0.74 (GPT-4.1-nano) to 0.93 (DeepSeek-V3.1). Even models with moderate ECE produce well-ranked confidence scores: ranking quality matters more than calibration quality for selective prediction.

9. Discussion and Limitations

Truncation is not the bottleneck for MCQA calibration. For forced single-token 4-choice MCQA, top-k API logprob truncation has no measurable effect on renormalized confidence once all admissible answer tokens are captured. The visibility audit shows mean visibility 3.73–3.98/4 at k=5 and exact 4/4 visibility on every item by $k \leq 50$ across all four self-hosted conditions. Even before full visibility is reached, TOST confirms empirical equivalence at k=5 with $\Delta\mathrm{ECE} < 0.02$ at $p < 0.001$ across four models and

two benchmarks, because renormalization absorbs the small residual mass associated with a briefly-hidden answer token. The boundary experiment extends this to constrained multi-token generation, where the binding constraint is the confidence aggregator (Section 7), not truncation. For open-ended generation, however, this conclusion reverses (Section H): at k=20, one in five tokens loses $>1\%$ of probability mass, confirming that our MCQA results should not be extrapolated to unconstrained generation settings.

Practitioner checklist. (i) Verify visibility before trusting top-k: for forced m-choice MCQA, confirm that all m admissible answer tokens appear in the returned top-k set per item. (ii) Do not treat top-k as full vocabulary: vLLM returns $\sim$99K tokens; true full-vocab reference is not available from any commercial API we tested. (iii) Platt transfers unreliably across nine matched conditions; temperature scaling is the safer default when a domain-matched calibration set is not available. (iv) Refit calibrators on the deployment distribution when possible. (v) Report scoring aggregator explicitly for multi-token answers; treat aggregator as a first-class methodological choice, not a convenience.

9.1. Limitations

(a) The formal truncation tests use $n \in \{166, 167, 197\}$; broader provider/model/benchmark conclusions in Sections 5 and 6 are observational, not confirmatory. (b) CalProbe-167 is a stress-test construction, not a natural query distribution; the heterogeneity and aggregator findings on it should not be extrapolated to deployment. (c) The cross-benchmark Spearman $\rho = 0.47$ at $n = 10$ is underpowered. We report it alongside the existence of sign-flipping cases (DeepSeek-V3.1, GPT-4.1) rather than as a population claim about rank stability. (d) Provider behavior is dated to our April–May 2026 audit window (Table 1); subsequent provider changes would require re-audit. (e) We do not run a single-agent ablation isolating which channel of the scoring pipeline contributes most to the bimodal residuals; the residual analysis in Section 5 is correlational. (f) We cannot verify whether reported logprobs are exact or post-processed; if providers modify logprobs, our measurements reflect post-processed distributions.

Conclusion. For forced 4-choice MCQA, researchers can treat top-k truncation as a non-issue *only after verifying that all admissible answer tokens are present in the returned top-k set*; once this visibility condition holds, the renormalized-confidence pipeline is exactly invariant to further truncation, and TOST equivalence testing on the self-hosted vLLM conditions confirms the empirical null. The real obstacles to reliable confidence scores are elsewhere: in the choice of aggregator for multi-token outputs (which swings ECE by 0.15–0.29), in the heterogeneity of model–provider conditions (ECE varies $9\times$ across our eleven models), and in the brittleness of calibration rankings across difficulty levels. Two practical takeaways follow. First, logprob confidence is useful for selective prediction even when absolute calibration is poor: the best model achieves 95% accuracy at 97% coverage by abstaining on low-confidence questions. Second, calibration measured on easy benchmarks should not be assumed to hold on harder ones; a model that looks well-calibrated at 85% accuracy may be severely overconfident at 40%.

These conclusions are scoped to forced 4-choice MCQA under verified answer-token visibility. Extrapolation to larger answer spaces requires a separate visibility check: in our 10-choice MMLU-Pro pilot, GPT-4.1 reaches only 7.19/10 visible answer tokens, so the truncation-invariance argument cannot be assumed to transfer. For open-ended generation, truncation *does* lose meaningful probability mass: at k=20, 20% of token positions lose $>1\%$ of mass, motivating future work on calibration in unconstrained generation settings.

References

Bentegeac, N., Le Guellec, J., Kuchcinski, V., Amouyel, P., and Hamroun, A. Token probabilities to mitigate LLM overconfidence in answering medical questions. *Journal of Medical Internet Research*, 27:e64348, 2025.

Chauvin, T., Le Merrer, E., Taiani, F., and Trédan, G. Log probability tracking of LLM APIs. *arXiv:2512.03816*, 2025.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. Bayesian data analysis. Chapman and Hall/CRC, 2013.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, 2017.

Kadavath, S., Conerly, T., Askell, A., et al. Language models (mostly) know what they know. *arXiv:2207.05221*, 2022.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Cheap hallucination detection in LLMs. In *ICLR*, 2025.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Lakens, D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Lamb, T. A., Ivanova, D. R., Torr, P., and Rudner, T. G. J. Semantic-level confidence calibration of language models via temperature scaling. In *ICLR Workshop on Uncertainty and Hallucination in Foundation Models*, 2025.

Leng, J., Huang, J., Zhu, B., and Huang, J. Taming overconfidence in LLMs: Reward calibration in RLHF. *arXiv:2410.09724*, 2024.

Li, Y., Xiong, M., Wu, J., and Hooi, B. ConfTuner: Training large language models to express confidence verbally. In *NeurIPS*, 2025.

Ma, Z., Wen, X., Cao, B., et al. Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards. *arXiv:2603.09117*, 2026.

MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

Memoli Buffa, P. and Del Corro, L. Entropy sentinel: Continuous LLM accuracy monitoring from decoding entropy traces in STEM. *arXiv:2601.09001*, 2026.

Pezeshkpour, P. and Hruschka, E. Large language models sensitivity to the order of options in multiple-choice questions. In *NAACL Findings*, 2024.

Rein, D., Hou, B. L., Stickland, A. C., et al. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv:2311.12022*, 2023.

Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Sharma, A. and Chopra, P. Think just enough: Sequence-level entropy as a confidence signal for LLM reasoning. *arXiv:2510.08146*, 2025.

Wang, Y., Ma, X., Zhang, G., et al. MMLU-Pro: A challenging multi-task language understanding benchmark. In *NeurIPS Datasets and Benchmarks*, 2024.

Xie, J., Chen, A. S., Lee, Y., Mitchell, E., and Finn, C. Calibrating language models with adaptive temperature scaling. In *EMNLP*, 2024.

Xu, B. and Lu, Y. TECP: Token-entropy conformal prediction for LLMs. *Mathematics*, 2025.

Zheng, C., Zhou, H., et al. Large language models are not robust multiple choice selectors. In *ICLR Spotlight*, 2024. arXiv:2309.03882.

A. Metric-Robustness Audit

The truncation null result in Section 4 is reported primarily in ECE_{10}, raising the concern that the null might be an artifact of that bin choice. We recompute each of the three on-disk vLLM conditions at k=5 vs the extended ∼99K-token reference across ECE with four bin counts ($m \in \{5, 10, 15, 20\}$), equal-mass ECE with 10 bins, Brier score, NLL, and AUROC. ECE and Brier are stable under truncation ($|\Delta| \leq 1.6 \times 10^{-5}$). When all answer tokens are visible, AUROC is identical because per-item confidence scores are identical. In incomplete-visibility cases, AUROC can in principle change; in our runs, observed changes were zero or negligible. NLL is the one metric that is *not* stable: on items where the ground-truth letter sits outside the top-5, the truncated model assigns $p_{\mathrm{true}} \approx 0$ and the per-item log-loss is inflated. This is consistent with the visibility audit and with the renormalization argument: the renormalized-confidence pipeline is invariant in argmax and relative-probability, while NLL depends pointwise on the true letter's probability and inherits the visibility gap.

B. Information-Theoretic Framework

We formalize the relationship between the truncated and full output distributions. Let $\mathbf{p} = (p_1, \ldots, p_{|V|})$ be the full output distribution at a token position, sorted so that $p_1 \geq p_2 \geq \cdots$.

Definition B.1 (Observed and missing mass)**.** The observed mass is $s = \sum_{i=1}^{k} p_i$ and the missing mass is $m = 1 - s$.

Definition B.2 (Partial and normalized entropy)**.** The partial entropy is $H_{\mathrm{partial}} = -\sum_{i=1}^{k} p_i \log_2 p_i$ and the normalized entropy is $H_{\mathrm{norm}} = -\sum_{i=1}^{k} \frac{p_i}{s} \log_2 \frac{p_i}{s}$.

Proposition B.3 (Lower bound on full entropy)**.** *$H_{\mathrm{full}} \geq H_{\mathrm{partial}} - m \log_2 m$ when $m > 0$, and $H_{\mathrm{full}} \geq H_{\mathrm{partial}}$ when $m = 0$.*

Proof. Decompose $H_{\mathrm{full}} = H_{\mathrm{partial}} + H_{\mathrm{tail}}$ where $H_{\mathrm{tail}} \geq -m \log_2 m$ by concavity. □

Empirical validation. We validated Theorem B.3 against real model distributions using vLLM extended logprobs (334 questions). Across 2,338 bound checks, we observed zero violations. At k=20, the mean gap between H_{full} and the lower bound is <0.001 bits.

Table 4. Power analysis for TOST equivalence detection. At n=167 and Δ=0.02, the study has 7% power; equivalence is confirmed because the observed δ is zero.

n	Δ=0.01	Δ=0.02	Δ=0.03	Δ=0.05	Δ=0.10
167	0.00	0.07	0.63	0.99	1.00
197	0.00	0.20	0.75	1.00	1.00
500	0.00	0.82	1.00	1.00	1.00

C. Power Analysis Tables

D. Cross-Provider Calibration on CalProbe-167

Table 5. Cross-provider calibration on CalProbe-167. ECE_{10}, renormalized confidence, bootstrap 95% CIs (n_{boot} = 10,000). †Groq CoT mode (n = 145); not directly comparable to direct-answer rows.

Model	Provider	Top-k	Acc.	ECE_{10}	Brier
GPT-4.1	OpenAI	20	0.850	0.125	0.118
GPT-4.1-mini	OpenAI	20	0.856	0.106	0.109
GPT-4.1-nano	OpenAI	20	0.737	0.197	0.208
GPT-4o	OpenAI	20	0.832	0.119	0.126
GPT-4o-mini	OpenAI	20	0.790	0.206	0.201
Qwen3-32B†	Groq	5	0.986	0.014	0.014
DeepSeek-V3.1	Together	20	0.892	0.280	0.300
Qwen3-235B	Together	20	0.946	0.066	0.065
Gemma-3n-E4B	Together	20	0.558	0.564	0.553
LFM2-24B	Together	20	0.772	0.508	0.493
Gemma-3n-E4B	vLLM	ext.	0.548	0.432	0.421
Qwen3-32B	vLLM	ext.	0.898	0.083	0.083

E. Token Variant Aggregation

API logprobs may represent the same answer letter as multiple tokens: "` A`" (with leading space), "`A`", "`a`", or "`A)`". We compared aggregated confidence against exact matching across all datasets (2,004 questions, 12 models). Mean ΔECE is $+0.02$: aggregation slightly *worsens* calibration. Exact single-letter matching is the appropriate extraction method.

F. MMLU-Pro–derived Per-Condition Results

Table 6 reports per-model calibration on the 500-question MMLU-Pro–derived set for the five OpenAI conditions sampled on this benchmark; the open-weight and Groq conditions discussed elsewhere in the paper were not run on this set. Accuracy ranges from 0.618 (GPT-4o-mini) to 0.744 (GPT-4o); ECE_{10} from 0.131 (GPT-4o) to 0.294 (GPT-4o-mini).

Table 6. Calibration on MMLU-Pro–derived questions (500 questions, 14 domains, reduced from 10 to 4 choices). Results are not directly comparable to standard MMLU-Pro benchmarks.

Model	Accuracy	ECE_{10}	Brier
GPT-4o	0.744	0.131	0.170
GPT-4.1-mini	0.706	0.205	0.221
GPT-4.1	0.738	0.219	0.221
GPT-4.1-nano	0.623	0.271	0.286
GPT-4o-mini	0.618	0.294	0.297

G. Temperature Effects on Calibration

Across five OpenAI models on CalProbe-167, ECE varies by less than 0.02 between $T = 0$ and $T = 1$ (Table 7).

Table 7. Temperature sweep effects on ECE for five OpenAI models, CalProbe-167.

Model	ECE(T=0)	ECE(T=1)	$\lvert\Delta\rvert$
GPT-4.1	0.125	0.131	0.006
GPT-4.1-mini	0.106	0.119	0.013
GPT-4.1-nano	0.197	0.211	0.014
GPT-4o	0.119	0.123	0.004
GPT-4o-mini	0.206	0.213	0.007

H. Open-Ended Generation: Where Truncation Does Matter

To verify that the MCQA conclusion does not extend to open-ended generation, we ran Qwen3-32B via vLLM on 140 open-text completion prompts (factual, scientific, creative) with 64 generated tokens each, requesting 1,000 logprobs per position. At k=20 (the OpenAI API limit), mean missing mass is 1.6% per token, with 20.4% of token positions losing more than 1% of probability mass and 4.9% losing more than 10%. At k=5, mean missing mass rises to 5.4%, with 39.2% of tokens affected. This contrasts with 4-choice MCQA, where missing mass at k=5 is effectively zero.

I. Probability Mass Captured by Top-k

Figure 3 shows captured probability mass against truncation level k for three models on the boundary multi-token experiment. Single-letter answers need $k \geq 5$ to capture all four admissible tokens; full-text answers are sharply peaked at $k = 1$ because the model copies text from the prompt.

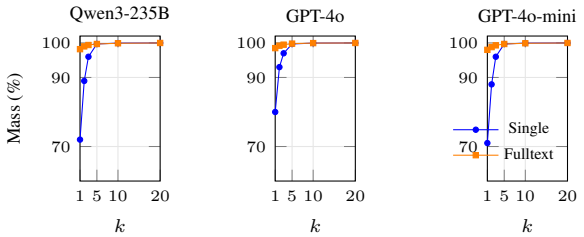

Figure 3. **Probability mass captured by top-k truncation.** Single-letter answers (blue) need $k \geq 5$ to capture all four options; full-text answers (orange) are sharply peaked at $k = 1$ because the model copies text from the prompt. Despite full-text retaining $>99\%$ mass at $k = 1$, calibration is much worse—driven by aggregator choice (Section 7), not truncation.

J. Cross-Benchmark Calibration-Repair Transfer Detail

Table 8 reports the per-model breakdown of the cross-benchmark calibration-repair transfer summarized in Section 6. The matched corpus comprises nine (provider, model, quantization) combinations present in both CalProbe-167 and GPQA Diamond. Across these conditions, temperature scaling worsens target ECE on 1/9 models, Platt on 2/9, and isotonic on 1/9; residual GPQA ECE remains above 0.10 on 8/9 models for every method.

Table 8. Cross-benchmark calibration-repair transfer, all matched models. Post-hoc calibrators (temperature / Platt / isotonic) are fit on CalProbe-167 and applied without refitting to GPQA Diamond. Δ is raw $-$ transferred (positive = improvement, negative = repair made it worse).

Model	Raw	T-ECE	T-Δ	Platt	P-Δ	Iso	I-Δ
DeepSeek-V3.1	0.189	0.011	$-.011$	0.470	$-.291$	0.389	$-.211$
GPT-4o	0.315	0.223	$+.092$	0.103	$+.212$	0.173	$+.142$
GPT-4.1-nano	0.328	0.207	$+.121$	0.140	$+.188$	0.166	$+.162$
Qwen3-235B	0.370	0.256	$+.114$	0.449	$-.078$	0.313	$+.058$
GPT-4.1-mini	0.397	0.218	$+.179$	0.100	$+.297$	0.082	$+.315$
LFM2-24B	0.429	0.153	$+.276$	0.401	$+.029$	0.380	$+.049$
GPT-4.1	0.469	0.273	$+.196$	0.169	$+.300$	0.225	$+.244$
GPT-4o-mini	0.478	0.290	$+.188$	0.190	$+.287$	0.234	$+.243$
Gemma-3n-E4B	0.582	0.224	$+.357$	0.264	$+.318$	0.290	$+.292$

K. Aggregator Pathology Detail

Table 9 reports per-model ECE and Brier scores for three confidence aggregators (geometric mean, min-token, sequence probability) on the constrained full-text answer-generation experiment of Section 7. Sequence probability achieves the lowest ECE on every model; the geometric-mean aggregator inflates ECE by 0.15–0.29 relative to that floor without changing model outputs.

Table 9. Calibration metrics by confidence aggregation strategy for constrained full-text answer generation on GPQA Diamond.

Model	Aggregator	ECE	Brier
GPT-4o	Geometric mean	0.562	0.542
GPT-4o	Min-token	0.297	0.310
GPT-4o	Sequence prob.	0.281	0.301
GPT-4o-mini	Geometric mean	0.616	0.604
GPT-4o-mini	Min-token	0.448	0.443
GPT-4o-mini	Sequence prob.	0.447	0.442
Qwen3-235B	Geometric mean	0.519	0.514
Qwen3-235B	Min-token	0.374	0.381
Qwen3-235B	Sequence prob.	0.374	0.377