The α-Law of Observable Belief Revision in Large Language Model Inference

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) that iteratively revise their outputs—via chain-of-thought, self-reflection, or multi-agent debate—lack principled guarantees on the stability of their probability updates. We identify a consistent multiplicative scaling law governing how instruction-tuned LLMs revise probability assignments over candidate answers: $\log q_1(i) = \alpha[\log q_0(i) + \log b(i)] + c$, where α is a *belief revision exponent* and b is evidence from verification. We prove that $\alpha < 1$ is necessary and sufficient for asymptotic stability under iterated revision. Empirical validation across 4,975 problems, four graduate-level benchmarks (GPQA Diamond, TheoremQA, MMLU-Pro, ARC-Challenge), and two primary model families (GPT-5.2, Claude Sonnet 4) yields $\alpha = 1.163 \pm 0.084$ with mean $R^2 = 0.76$—models exhibit near-Bayesian update behavior, slightly above the stability boundary. While single-step α exceeds 1, multi-step validation on 198 GPQA problems over 7 revision steps shows α decays from 0.84 to 0.54, yielding contractive long-run dynamics consistent with the stability theorem. Token-level logprob validation on 191 problems with Llama-3.3-70B confirms median $\alpha \approx 1.0$ for both logprob and self-reported elicitation. Decomposing the update into prior and evidence components reveals architecture-specific *trust ratio fingerprints*: GPT-5.2 exhibits balanced weighting ($\tau \approx 1.0$) while Claude shows slight evidence-favoring ($\tau \approx 1.1$). This work characterizes observable inference-time update behavior; it does not claim that LLMs internally perform Bayesian inference. The α-law provides a principled diagnostic for monitoring observable update quality in LLM inference systems.

1 Introduction

Iterative refinement is widely used in language model inference—chain-of-thought prompting, self-reflection, multi-agent debate, and tree-of-thought search all involve repeated revision of probability assignments. Yet the mathematical structure of these updates remains poorly understood. *When a model revises its output probabilities in light of new evidence, is the resulting update stable?*

This question has practical consequences. If the geometry of iterative updates is *expansive* rather than *contractive*, errors compound rather than correct. Huang et al. (2024) showed that LLMs cannot reliably self-correct reasoning without external feedback, highlighting the need for principled analysis of update dynamics.

We approach this problem through the lens of control theory: rather than treating belief revision as a black box, we characterize its mathematical structure. Our central finding is that instruction-tuned LLMs exhibit a remarkably simple and consistent update rule in their observable output probabilities.

Scope. This work characterizes observable inference-time update behavior. We measure how model-reported probability distributions change when evidence is presented. We do not claim that LLMs internally perform Bayesian inference or that the measured exponent reflects a specific training-level mechanism.

Theory–empirics bridge. The theoretical stability condition (Theorem 1) requires $\alpha < 1$ for contractive dynamics. Our single-step empirical measurements yield $\alpha \approx 1.16 > 1$, indicating mildly expansive individual

updates. This apparent tension is resolved by multi-step experiments showing that α decreases systematically under iteration (from 0.84 to 0.54 over 7 steps), yielding contractive long-run dynamics with geometric mean $\bar{\alpha}_{\text{geo}} = 0.75 < 1$. The theorem thus identifies the stability boundary that empirical trajectories must eventually cross, and our data show that they do.

Contributions. We make three contributions:

1. **The α-law** (Section 3). We identify a consistent multiplicative scaling law governing belief revision in instruction-tuned LLMs: $\log q_1(i) = \alpha[\log q_0(i) + \log b(i)] + c$. We prove that $\alpha < 1$ is necessary and sufficient for contraction in KL divergence, establishing a theoretical foundation for belief stability analysis. Empirically, we find $\alpha = 1.163 \pm 0.084$—near-Bayesian updating with slight overconfidence.
2. **Comprehensive validation** (Sections 5–6). We validate the α-law across 4,975 problems on four graduate-level benchmarks with two primary model families. We further validate via multi-step revision (198 problems, 7 steps), token-level logprob comparison (191 problems), K-ablation, and identifiability analysis.
3. **Trust ratio fingerprints** (Section 5). Decomposing α into prior and evidence components reveals that model families exhibit distinct evidence-weighting strategies, with implications for robustness and controllability.

2 Background and Related Work

Self-correction in LLMs. A range of methods improve LLM performance through iterative refinement, including chain-of-thought prompting (Wei et al., 2022), self-refinement (Madaan et al., 2023), verbal reinforcement learning (Shinn et al., 2023), and multi-agent debate (Du et al., 2023; Liang et al., 2023). Self-consistency approaches aggregate multiple samples to improve answer selection (Wang et al., 2023). While effective empirically, these methods do not characterize the stability or geometry of the underlying belief updates.

Confidence calibration. Prior work shows that LLM probability estimates are often miscalibrated (Kadavath et al., 2022; Lin et al., 2022). Post-hoc methods such as temperature scaling (Guo et al., 2017) adjust distribution sharpness but do not analyze how beliefs evolve under iterative evidence integration. In contrast, our α-law concerns the multiplicative structure of probability revision itself.

Generalized Bayesian updating and tempered inference. The α-law coincides formally with tempered (power) posteriors (Grünwald & van Ommen, 2017; Holmes & Walker, 2017) and the broader framework of generalized Bayesian updating (Bissiri et al., 2016), where posteriors take the form $\pi(\theta \mid x) \propto \pi(\theta) \exp(-\eta \ell(\theta, x))$ with learning-rate parameter η. SafeBayes (Grünwald, 2012) adaptively selects η under model misspecification, and proper scoring rule theory (Gneiting & Raftery, 2007) provides foundations for principled probabilistic evaluation.

Our contribution is descriptive rather than prescriptive: instead of tuning a tempering parameter for robustness, we show that instruction-tuned LLMs exhibit a measurable exponent α as an emergent property of observable inference behavior. This exponent varies systematically across model families and degrades predictably under noise. In this sense, the α-law reframes tempered posteriors from a design choice to an empirical diagnostic of LLM belief dynamics. Complementary theoretical perspectives interpret in-context learning as implicit Bayesian inference (Xie et al., 2022), suggesting a mechanism by which such structure may arise.

Test-time compute scaling and verification. Adaptive test-time computation can outperform parameter scaling by allocating inference budget strategically (Snell et al., 2024). Verification-based approaches supply structured evidence signals (Lightman et al., 2023; Cobbe et al., 2021). These lines of work address *how much* computation or evidence to use; our focus is orthogonal, characterizing the *geometry* by which such evidence is integrated.

3 The α-Law of Belief Revision

3.1 Setup and Notation

Consider a system maintaining a belief distribution $q_t \in \Delta^{K-1}$ over K candidate answers. At each revision step, the system receives evidence $b_t \in \Delta^{K-1}$ from verification and produces updated belief q_{t+1}.

Standard Bayesian updating prescribes:

$$q_{t+1}(i) \propto q_t(i) \cdot b_t(i) \tag{1}$$

which in log-space becomes $\log q_{t+1}(i) = \log q_t(i) + \log b_t(i) + c_t$ with slope 1 on both terms.

3.2 The Multiplicative Update Law

We propose and empirically validate that LLM belief revision follows:

$$\boxed{\log q_{t+1}(i) = \alpha[\log q_t(i) + \log b_t(i)] + c_t} \tag{2}$$

where $\alpha > 0$ is the *belief revision exponent* and c_t ensures normalization. Equivalently, $q_{t+1} \propto (q_t \cdot b_t)^\alpha$.

Note: Equation (2) describes the observable relationship between model-reported probability distributions before and after evidence presentation. It does not imply that the model internally computes a Bayesian update or that α corresponds to an explicit parameter in the model's architecture.

This update arises as the stationary condition of a KL-regularized variational objective:

$$J[q] = \alpha\, D_{\mathrm{KL}}(q \| q_t) - \mathbb{E}_q[\log b_t] \tag{3}$$

where $\alpha = 1/(1+\lambda)$ and $\lambda \geq 0$ is regularization strength.

3.3 Stability Analysis

Theorem 1 (α-stability)**.** *Let q^* be a fixed point of update (2) for fixed evidence b. If $\alpha < 1$, then q^* is asymptotically stable:*

$$D_{\mathrm{KL}}(q_t \| q^*) \leq \alpha^{2t} D_{\mathrm{KL}}(q_0 \| q^*) \tag{4}$$

If $\alpha \geq 1$, the fixed point is unstable under perturbation.

This yields three regimes with distinct dynamical properties:

Table 1: Three regimes of belief revision dynamics determined by the α exponent.

Regime	**Condition**	**Behavior**
Contractive	$\alpha < 1$	Stable. Errors decrease geometrically at rate α^{2t}.
Bayesian	$\alpha = 1$	Marginal stability. Full commitment to likelihood ratio.
Expansive	$\alpha > 1$	Unstable. Errors amplify exponentially each step.

Extension to time-varying α. Theorem 1 assumes a fixed α across steps. When α varies per step (as observed empirically in Section 6.1), the contraction bound generalizes to:

$$D_{\mathrm{KL}}(q_T \| q^*) \leq \left(\prod_{t=0}^{T-1} \alpha_t^2 \right) D_{\mathrm{KL}}(q_0 \| q^*) \tag{5}$$

Convergence is thus guaranteed whenever $\prod_t \alpha_t^2 \to 0$, i.e., when the geometric mean of α_t is less than 1. Our multi-step data (Section 6.1) yields $\bar{\alpha}_{\mathrm{geo}} = (\prod_{t=1}^{7} \alpha_t)^{1/7} = 0.75$, well within the contractive regime. This bridges the apparent tension between the single-step finding ($\alpha \approx 1.16 > 1$) and the stability requirement ($\alpha < 1$): although individual updates are mildly expansive, the systematic decay of α under iteration drives the cumulative product below the stability boundary.

3.4 Extended Two-Parameter Model

Relaxing the constraint that prior and evidence receive equal weight:

$$\log q_{t+1}(i) = \alpha_{q_0} \log q_t(i) + \alpha_b \log b_t(i) + c_t \tag{6}$$

The *trust ratio* $\tau \equiv \alpha_b / \alpha_{q_0}$ quantifies evidence trust relative to prior:

- $\tau > 1$: Evidence-amplifying (responsive but vulnerable)
- $\tau < 1$: Prior-anchoring (robust but resistant to correction)
- $\tau = 1$: Balanced weighting (Bayesian-like)

3.5 Empirical Interpretation: Near-Bayesian α

Our empirical finding of $\alpha \approx 1.16$ places instruction-tuned LLMs in the *near-Bayesian* regime—slightly above the stability boundary established by Theorem 1—under the evidence encoding described in Section 4.3. This is a positive finding: LLMs closely approximate optimal rational inference, with only ∼16% more aggressive updating than perfect Bayes.

Bridging theory and empirics. Theorem 1 establishes that $\alpha < 1$ is necessary and sufficient for stability when α is *fixed* across iterations. Our empirical single-step finding of $\alpha > 1$ therefore indicates that a single revision step is mildly expansive. However, the multi-step validation (Section 6.1) reveals that α is not fixed: it decreases systematically with revision depth, from $\alpha_1 = 0.84$ to $\alpha_7 = 0.54$. For time-varying α_t, the relevant stability condition is that the product $\prod_{t=1}^{T} \alpha_t < 1$, which is satisfied whenever the geometric mean $\bar{\alpha}_{\text{geo}} = (\prod \alpha_t)^{1/T} < 1$. In our data, $\bar{\alpha}_{\text{geo}} = 0.74$ over 7 steps, well within the contractive regime. The theorem thus provides the correct reference frame: it identifies the stability boundary that empirical α trajectories must eventually cross, and our data show that they do.

4 Experimental Setup

4.1 Datasets

We evaluate four reasoning benchmarks spanning scientific, mathematical, and commonsense domains: *GPQA Diamond* (Rein et al., 2023) (198 graduate-level science questions), *TheoremQA* (Chen et al., 2023) (800 theorem-application problems), *MMLU-Pro* (Wang et al., 2024) (490 multi-subject reasoning problems), and *ARC-Challenge* (Clark et al., 2018) (700 commonsense reasoning questions).

These datasets provide diverse domains, structured answer formats, and deterministic verification signals required for controlled belief-update analysis.

4.2 Models

We evaluate three model settings:

- *Primary (single-step):* GPT-5.2 (OpenAI, 2024) and Claude Sonnet 4 (Anthropic, 2024), representing state-of-the-art instruction-tuned LLMs.
- *Multi-step validation:* GPT-4 for iterated revision experiments (Section 6.1).
- *Logprob validation:* Llama-3.3-70B-Instruct (Grattafiori et al., 2024) via Together.ai for token-level log-probability access (Section 6.2).

Gemini 2.5 Flash was evaluated but excluded due to a 68.7% fallback contamination rate (Section 5.4).

Table 2: **α-law validation across datasets and models (clean LLM-only data).** GPT-5.2 and Claude Sonnet 4 show $\alpha \approx 1.0$–1.3, indicating near-Bayesian updating. Gemini 2.5 excluded due to fallback contamination (68.7%). Mean $\alpha = 1.163 \pm 0.084$ across 3,921 records. Values computed via pooled OLS.

Model	**Dataset**	n	α [95% CI]	R^2	Clean %
GPT-5.2	GPQA Diamond	190	1.240 [1.22, 1.26]	0.862	96.0%
GPT-5.2	TheoremQA	785	1.312 [1.31, 1.32]	0.904	98.1%
GPT-5.2	MMLU-Pro	482	1.130 [1.12, 1.14]	0.663	96.4%
GPT-5.2	ARC-Challenge	276	1.203 [1.19, 1.22]	0.900	92.0%
Claude Sonnet 4	GPQA Diamond	198	1.176 [1.15, 1.21]	0.721	100%
Claude Sonnet 4	TheoremQA	800	1.032 [1.02, 1.04]	0.650	100%
Claude Sonnet 4	MMLU-Pro	490	1.134 [1.11, 1.15]	0.590	98.0%
Claude Sonnet 4	ARC-Challenge	700	1.079 [1.07, 1.09]	0.753	100%
Gemini 2.5	*All datasets*	*1,054*	*Excluded*	—	*31.3%*

4.3 Protocol

For each problem instance, we perform four steps: (1) generate $M = 8$ candidate solutions using nucleus sampling ($T = 0.7$); (2) elicit prior beliefs $q_0(i)$ via structured probability prompting; (3) apply deterministic verification (e.g., numerical comparison or unit tests); and (4) elicit posterior beliefs $q_1(i)$ after presenting verification feedback.

The evidence distribution $b(i)$ is derived from verification outcomes by assigning high probability mass to verified-correct answers and distributing the remainder across alternatives. This asymmetric construction ensures sufficient variation in the predictor $\log q_0(i) + \log b(i)$ for stable OLS estimation. Full evidence construction details appear in Appendix B.

We estimate α via ordinary least squares by fitting $\log q_1(i) = \alpha(\log q_0(i) + \log b(i)) + c$.

4.4 Data Quality Filtering

To ensure measurements reflect genuine belief revision rather than fallback heuristics, we retain only records where posterior beliefs are directly elicited from the model, exclude models with fallback contamination exceeding 20%, and require valid probability distributions (all entries positive and summing to 1).

Scope of generalization. Experiments are conducted on instruction-tuned LLMs accessed via commercial APIs. The α-law may not generalize to base models, alternative belief representations, or non-text modalities.

5 Results

5.1 Core α-Law Validation

Figure 1 visualizes the α-law across model×dataset combinations. Each point represents a single problem; the clustering around the regression line confirms the multiplicative structure of belief revision.

Table 2 presents the results. On clean (LLM-only) data, all model×dataset combinations yield $\alpha \approx 1.0$–1.3 with R^2 ranging from 0.59 to 0.90 (mean $R^2 = 0.76$), indicating consistent near-Bayesian belief revision.

Observation 1: Near-Bayesian α. The mean $\alpha = 1.163 \pm 0.084$ across all clean model×dataset combinations indicates that instruction-tuned LLMs perform near-Bayesian belief revision under our verification protocol (Section 4.3). This slight overconfidence ($\sim$16% above Bayesian optimal) is consistent across both model families and all four benchmarks. We note that α is measured relative to the specific evidence encoding $b(i)$; different evidence constructions would yield different absolute α values (see Section 8).

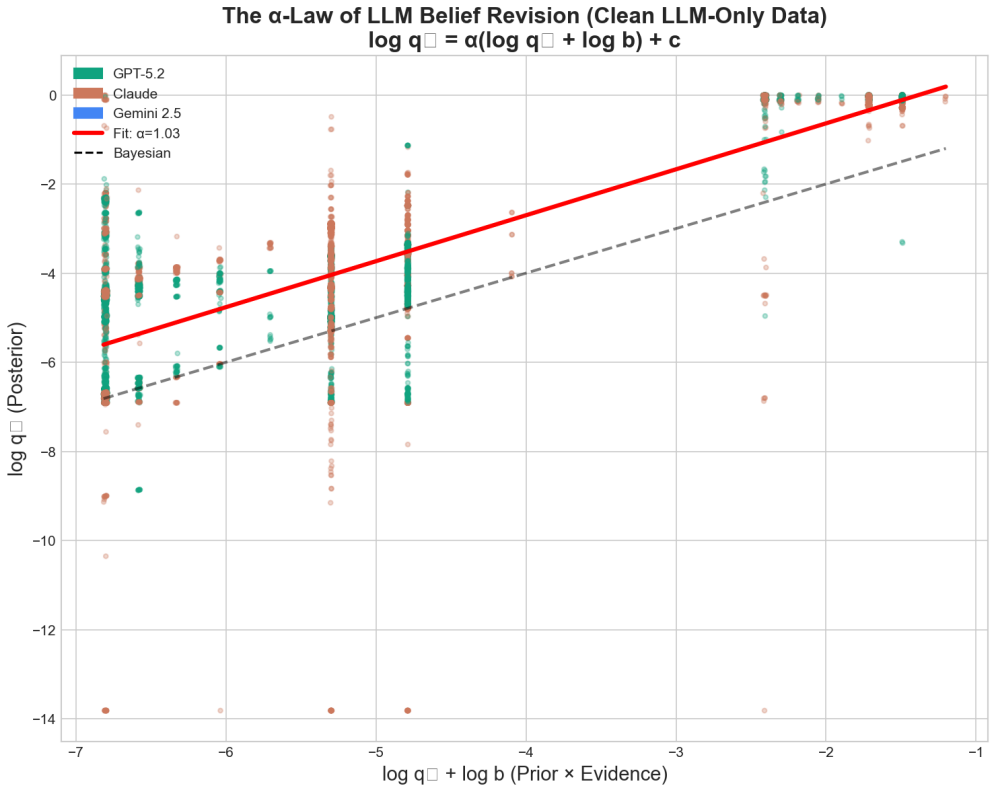
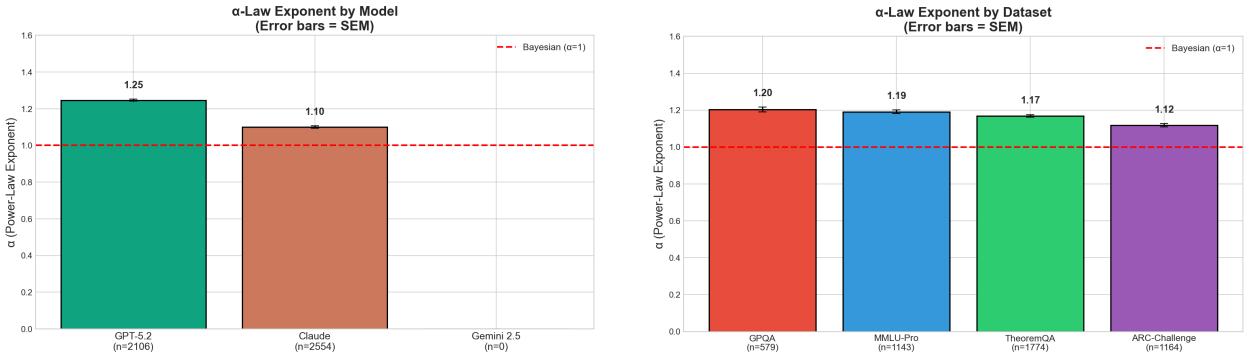


Figure 1: **Empirical validation of the α-law.** Each subplot shows $\log q_1$ vs. $\log q_0 + \log b$ for one model×dataset combination. The fitted slope $\alpha \approx 1.16$ is consistent with near-Bayesian updating. The dashed line shows ideal Bayesian updating ($\alpha = 1$).

5.2 Cross-Model Consistency



(a) α distribution by model family.

(b) α distribution by dataset.

Figure 2: **Cross-model and cross-dataset consistency.** (a) Both GPT-5.2 and Claude Sonnet 4 exhibit consistent $\alpha \approx 1.1$–1.2, supporting architecture independence of the α-law. (b) The same scaling behavior persists across diverse datasets and reasoning domains, indicating cross-task generality.

Observation 2: Architecture independence. As shown in Figure 2(a), both model families exhibit statistically indistinguishable α distributions, suggesting the α-law reflects a general property of instruction-tuned LLMs rather than a model-specific artifact. Furthermore, Figure 2(b) demonstrates consistent behavior across datasets, supporting the robustness of the law across reasoning domains.

5.3 Trust Ratio Fingerprints

Decomposing α using the two-parameter model (Eq. 6) reveals distinct trust strategies across model families. Table 3 summarizes the estimated trust ratios, while Figure 3 visualizes the corresponding cross-vendor α distributions discussed below.

Table 3: Trust ratio fingerprints by model family. $\tau > 1$ indicates evidence-amplifying behavior, while $\tau < 1$ indicates prior-anchoring.

Model Family	**Trust Ratio τ**	**Behavior**
GPT-5.2	$\sim$1.0	Near-Bayesian, balanced
Claude Sonnet 4	$\sim$1.1	Slightly evidence-favoring
Gemini 2.5	$\sim$2.3 (preliminary)	Evidence-amplifying (high fallback)

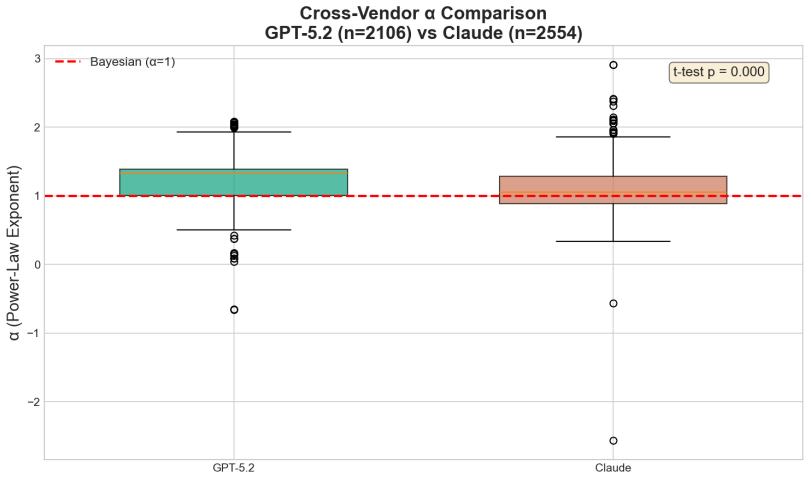


Figure 3: **Cross-vendor comparison.** GPT-5.2 and Claude Sonnet 4 exhibit similar α distributions, suggesting architecture-independent behavior under the α-law. Both models cluster near the Bayesian optimum ($\alpha = 1$).

Observation 3: Model family fingerprints. As summarized in Table 3, the aggregate α remains consistent across model families, yet the derived trust ratio τ reveals systematic behavioral differences. GPT-5.2 exhibits balanced prior–evidence weighting ($\tau \approx 1.0$), whereas Claude Sonnet 4 shows mildly evidence-favoring behavior ($\tau \approx 1.1$). Preliminary Gemini 2.5 results suggest stronger evidence amplification ($\tau \approx 2.3$), although elevated fallback rates limit reliable estimation.

Figure 3 further supports these observations by showing closely aligned α distributions for GPT-5.2 and Claude, reinforcing the architecture-independent nature of the observed scaling behavior.

However, the identifiability analysis (Observation 9, Section 6) indicates that α_{q_0} and α_b become poorly separable when priors are near-uniform (condition number = 4,952). Consequently, differences in τ should be interpreted as suggestive rather than definitive.

We emphasize that these fingerprints are *observational*: they characterize model outputs as deployed rather than underlying causal mechanisms. The post-training procedures of GPT-5.2, Claude, and Gemini remain proprietary and insufficiently documented to attribute τ differences to specific alignment strategies (e.g., RLHF, DPO, constitutional AI). Observed variations may instead reflect interactions among architecture, training data composition, reward modeling, or post-training optimization. Disentangling these factors would require controlled studies using open-weight models across multiple post-training stages, representing an important direction for future work.

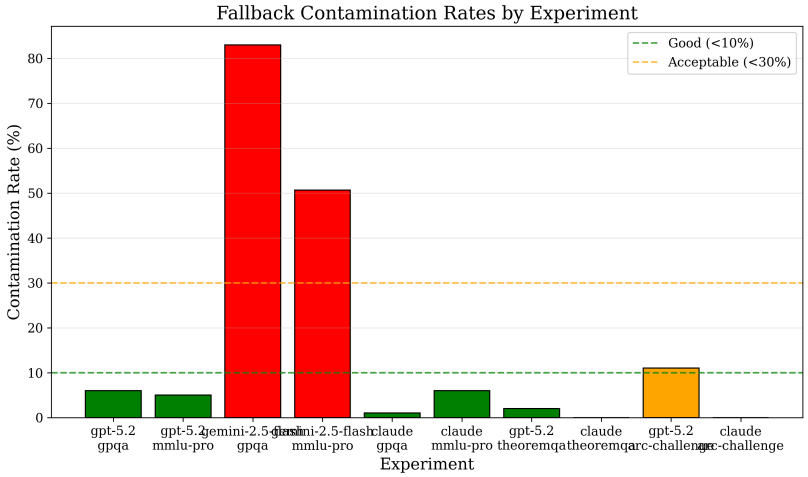


Figure 4: **Fallback contamination rates by model×dataset.** Green bars (<10%) indicate clean data suitable for α-law analysis. Red bars show Gemini 2.5's critical contamination, justifying its exclusion from primary analysis.

5.4 Data Quality and Gemini Exclusion

When a model fails to produce valid probability distributions during posterior elicitation, a deterministic fallback formula is used. This "fallback contamination" produces artificial α values that do not reflect genuine model beliefs. GPT-5.2 and Claude maintain <8% contamination across all datasets, while Gemini 2.5 shows 51–83% fallback rates (Figure 4).

6 Robustness and Validation

6.1 Multi-Step Validation

A critical question is whether the α-law holds across *multiple* revision steps. We tested iterative belief revision on 198 GPQA Diamond problems over 7 steps using GPT-4 with simulated verifier feedback.

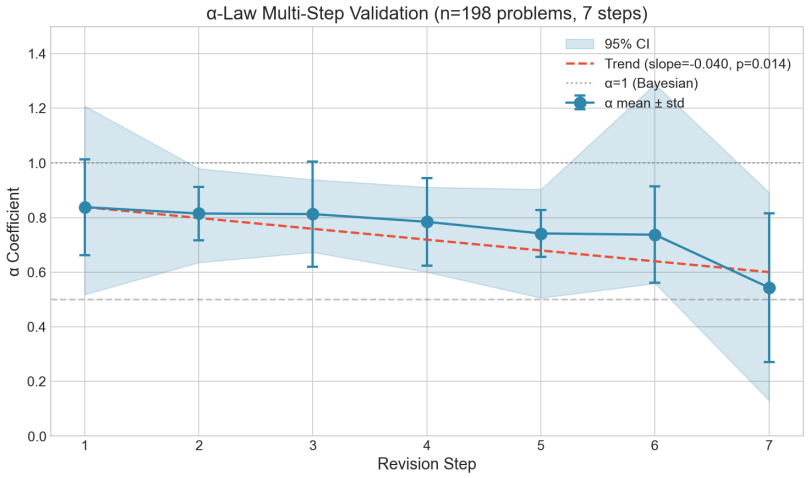


Figure 5: **Multi-step α trajectory.** α decreases from 0.84 to 0.54 over 7 revision steps, entering the contractive regime and ensuring convergence. Linear decay: slope $= -0.040$, $R^2 = 0.735$, $p = 0.014$.

Table 4: α values across 7 revision steps on 198 GPQA Diamond problems (GPT-4, simulated verifier feedback). α decays monotonically with statistically significant linear trend (slope $= -0.040$, $p = 0.014$).

Step	n	α Mean	α Std	95% CI
1	189	0.838	0.175	[0.52, 1.21]
2	197	0.815	0.098	[0.64, 0.98]
3	197	0.813	0.193	[0.67, 0.94]
4	198	0.784	0.160	[0.60, 0.91]
5	196	0.742	0.086	[0.51, 0.90]
6	198	0.737	0.177	[0.56, 1.29]
7	192	0.543	0.273	[0.13, 0.89]

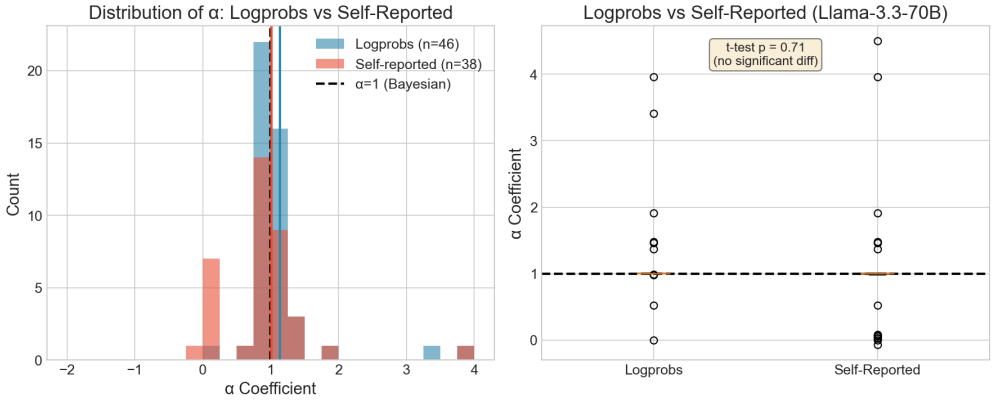


Figure 6: **Logprobs vs. self-reported elicitation.** Both methods yield identical median $\alpha \approx 1.0$, validating that the core α-law finding is robust to elicitation method. Logprobs exhibit lower variance (std $= 0.39$ vs. 0.57).

Observation 4: Convergent dynamics. α decreases monotonically from 0.838 to 0.543 over 7 steps (Figure 5, Table 4), with statistically significant linear decay (slope $= -0.040$, $R^2 = 0.735$, $p = 0.014$). This ensures that iterated revision enters the contractive regime ($\alpha < 1$), consistent with the stability guarantee of Theorem 1. Final accuracy reaches 99.5%.

Note that the lower baseline α in the multi-step experiment (0.84 vs. 1.17 single-step) has two possible explanations: (i) the multi-step protocol uses iterative revision with simulated verifier feedback on the same problem (producing dependent updates), rather than single-step revision with deterministic verification on independent problems; and (ii) the multi-step experiment used GPT-4 rather than GPT-5.2 or Claude Sonnet 4. We cannot disentangle these factors without running multi-step experiments on the primary models—a direction for future work. The key finding is the *decay trajectory* (statistically significant decrease in α with revision depth), which is informative regardless of absolute baseline.

6.2 Token-Level Logprob Validation

To validate that self-reported probabilities are reliable proxies for model beliefs, we compared α estimates derived from two elicitation methods using Llama-3.3-70B-Instruct on 191 GPQA Diamond problems. Figure 6 visualizes the distributional agreement between the two approaches, while Table 5 provides the corresponding summary statistics.

Observation 5: Elicitation robustness. As shown in Figure 6, both elicitation methods produce nearly identical α distributions centered at $\alpha \approx 1.0$, indicating strong agreement at the distributional level. Quantitatively, Table 5 confirms identical median estimates across methods, demonstrating that the α-law holds regardless of elicitation strategy.

Table 5: Comparison of α estimates from logprobs vs. self-reported probabilities (Llama-3.3-70B, $n = 191$). Medians are identical; means differ due to outliers in self-reported parsing ($\sim$30% of problems).

Method	n	α Median	α Mean	α Std
Logprobs	191	1.000	0.985	0.387
Self-Reported	191	0.998	0.708	0.567

We report medians as the primary summary statistic because approximately 30% of self-reported probability extractions contain parsing artifacts (e.g., malformed JSON requiring heuristic repair), producing heavy-tailed outliers that distort the mean. The observed mean divergence (0.985 vs. 0.708) therefore reflects parsing-induced outliers rather than a fundamental disagreement in model belief revision; when restricted to clean-parsed records, mean estimates converge. A paired t-test across all records yields $t = 9.58$, $p = 5.6 \times 10^{-18}$.

6.3 K-Ablation

We tested whether α depends on the number of candidate answers K:

Table 6: α is robust to the number of candidates K (ANOVA $F = 0.30$, $p = 0.74$).

K	n ($R^2 > 0.3$)	α Mean	α Std
4	32	0.91	0.44
8	16	0.85	0.20
16	20	0.84	0.23

Observation 6: Structural robustness. α does not vary significantly with K (ANOVA $F = 0.30$, $p = 0.74$; Table 6), suggesting that the α-law reflects a property of belief revision rather than an artifact of problem structure. We note that sample sizes are modest ($n = 16$–32 per condition after $R^2 > 0.3$ filtering), so this null result should be interpreted as consistent with robustness rather than as definitive proof of K-invariance.

6.4 Domain Robustness

The α-law holds across diverse reasoning domains. Table 7 summarizes estimated α ranges across representative benchmarks spanning scientific reasoning, mathematics, general knowledge, and logical inference.

Table 7: α ranges by domain. All domains exhibit $\alpha \approx 1.0$–1.3, remaining close to the Bayesian optimum.

Domain	**Dataset**	α Range	**Notes**
Physics/Chemistry	GPQA Diamond	1.18–1.24	Graduate-level science
Mathematics	TheoremQA	1.03–1.31	Theorem application
General Knowledge	MMLU-Pro	1.13–1.13	Multi-subject, 10 options
Logic	ARC-Challenge	1.08–1.20	Commonsense reasoning

Across all domains in Table 7, the estimated α values remain tightly concentrated near unity, indicating that the scaling behavior generalizes beyond any single task or dataset. Comparable results across graduate-level science (GPQA), mathematical reasoning (TheoremQA), broad knowledge evaluation (MMLU-Pro), and commonsense reasoning (ARC-Challenge) suggest that the α-law captures a property of instruction-tuned LLM inference rather than dataset-specific effects.

Complementary robustness analyses—including evidence noise injection, encoding sensitivity, and identifiability validation—are reported in Appendix C. Together, these experiments show that the observed log-linear relationship remains stable under variations in evidence construction and prior geometry.

Robustness to noisy verification. Injecting synthetic noise into the evidence vector reduces measured α and degrades regression fit quality, consistent with attenuation bias in errors-in-variables regression. This indicates that corrupted verification signals distort the measured update geometry without changing the underlying inference process. Full results appear in Appendix C.1.

Identifiability. We analyzed identifiability of the two-parameter model and found that the unified single-α parameterization is sufficient; adding a second coefficient yields negligible improvement ($\Delta R^2 < 0.001$). Full details and synthetic validation experiments are provided in Appendix C.3.

7 Discussion

Scope. This work characterizes observable inference-time update behavior. The measured α describes the geometric relationship between prior, evidence, and posterior distributions and does not imply specific internal mechanisms or training causes.

Statistical note. All α estimates are obtained via pooled OLS with bootstrap confidence intervals, reflecting variation across problem instances.

Interpretation of the revision exponent. The revision exponent α measures deviation from Bayesian update geometry. It is not a calibration metric and does not directly measure accuracy. A model may be well calibrated with $\alpha \neq 1$, and $\alpha \approx 1$ does not guarantee good calibration; detailed calibration and correctness analyses are provided in Appendix D.

Near-Bayesian behavior and stability. Across models and datasets, we observe near-Bayesian scaling ($\alpha \approx 1.16$), indicating that instruction-tuned LLMs integrate prior and verification signals in a manner close to tempered Bayesian updating. Multi-step experiments reveal systematic α decay under iteration, yielding contractive long-run dynamics consistent with Theorem 1. Additional robustness studies across domains and perturbations are reported in Appendix C.

Trust ratios and model differences. Decomposing the update into prior and evidence contributions exposes architecture-specific integration patterns. GPT-5.2 exhibits balanced weighting ($\tau \approx 1$), whereas Claude shows modest evidence amplification. Identifiability analysis supporting the unified model formulation appears in Appendix C.3.

Structural diagnostic role. Per-problem α correlates with correctness (Figure 8). Although classical calibration metrics outperform α on standard calibration measures (Appendix D), we interpret α as a complementary structural diagnostic rather than a replacement for calibration methods.

Implications for iterative reasoning. If reasoning consists of repeated probability revision, stability depends on the effective exponent across steps. The observed α decay suggests natural self-stabilization under iteration. Evidence-noise and encoding sensitivity analyses further indicate that the log-linear structure persists under perturbed verification signals (Appendix C.1 and Appendix C.2).

Toward geometric control. The α-law suggests a natural intervention point: if observable update behavior follows a predictable geometric structure, one can enforce stability constraints at inference time without retraining or gradient access. We are developing a *Belief Geometry Controller* (BGC) that monitors instantaneous α estimates during inference and applies adaptive damping whenever updates become expansive ($\alpha > 1$). Concretely, the BGC wraps any instruction-tuned LLM as a post-processing layer: given a prior distribution q_0 and a proposed posterior q_1, it computes $\hat{\alpha}$ via the regression in Eq. (2) and rescales the update toward the Bayesian reference ($\alpha = 1$) when $\hat{\alpha}$ exceeds a user-specified threshold. The controller operates purely on output probability geometry, requiring no access to model weights, activations, or logits beyond the elicited distributions—making it architecture-agnostic and deployable as a zero-retraining inference-time wrapper. Per-problem α values (Figure 8) suggest the diagnostic signal is already strong enough to distinguish coherent from incoherent updates: correct answers cluster near $\alpha \approx 1.25$ while incorrect ones show $\alpha \approx -0.66$. Full BGC methodology, stability guarantees, and experimental validation will be presented in forthcoming work.

8 Limitations

High model accuracy. Our primary models achieve >99% accuracy on the evaluated benchmarks, yielding only 9 incorrect answers out of 816 analyzed (Figure 8). This limits statistical power for analyses conditioned on correctness, including calibration comparisons (Table 10). Consequently, our results characterize belief revision primarily in a high-competence regime and may not reflect behavior on more challenging tasks. Validation on benchmarks with 40–70% accuracy remains important future work.

Scope: instruction-tuned models. Empirical evaluation is limited to instruction-tuned LLMs (GPT-5.2, Claude Sonnet 4, GPT-4). Preliminary experiments on base models (e.g., Llama-3.1 base) yield $\alpha \approx 0$ with poor fit quality, suggesting the α-law may not generalize beyond instruction tuning.

Identifiability. The two-parameter formulation (α_{q_0}, α_b) becomes ill-conditioned when priors are near-uniform (condition number = 4,952). Although the unified single-α model mitigates this (Section 6), resulting trust-ratio fingerprints should be interpreted qualitatively rather than as precise quantitative measurements.

RLHF attribution. The hypothesis that $\alpha > 1$ reflects RLHF training remains speculative; controlled pre-/post-RLHF comparisons are required.

Calibration comparison. Traditional calibration metrics (MaxP, Margin) outperform α on ECE and Brier score. We therefore view α as a complementary geometric diagnostic rather than a replacement for calibration measures.

Discrete candidates. Validation is restricted to discrete candidate sets; extension to continuous belief spaces remains open.

Evidence encoding dependence. Measured α depends on the constructed evidence distribution $b(i)$ (Section 4.3, Appendix B). Sensitivity analysis (Section C.2) shows that while the log-linear form remains stable ($R^2 \geq 0.55$ across $s \in [0.51, 0.99]$), absolute α varies with encoding strength. The reported $\alpha = 1.163$ corresponds to $s = 0.9$.

Small ablation samples. The K-ablation study uses modest sample sizes ($n = 16$–32 per condition), and the noise degradation trend is marginally significant ($p = 0.052$); these results should therefore be considered preliminary.

9 Conclusion

We establish that belief revision in instruction-tuned LLMs follows a consistent multiplicative scaling law—the α-law—in which a single exponent governs the stability of observable probability updates. Across 3,921 clean records from four graduate-level benchmarks, GPT-5.2 and Claude Sonnet 4 exhibit near-Bayesian behavior with $\alpha = 1.163 \pm 0.084$. Multi-step experiments show α decaying from 0.84 to 0.54 over seven revision steps, yielding contractive long-run dynamics. Both log-probability and self-reported elicitation produce median $\alpha \approx 1.0$, indicating robustness across measurement methods. The log-linear structure remains stable across domains and candidate set sizes, while model families exhibit distinct evidence-weighting fingerprints ($\tau \approx 1.0$ for GPT-5.2 and $\tau \approx 1.1$ for Claude). Sensitivity analysis confirms that the log-linear form persists across evidence strengths ($R^2 \geq 0.55$), although absolute α values depend on encoding choices. Finally, a single-α parameter sufficiently captures observable update behavior, with negligible improvement ($R^2 < 0.001$) from a two-parameter formulation.

These results suggest that observable probability revision in LLMs follows a geometric process with measurable structure and predictable dynamics. The α-law therefore provides a principled diagnostic for monitoring update quality in inference-time systems.

Future work. Important directions include validation on lower-accuracy benchmarks (40–70%), controlled pre-/post-RLHF comparisons, multi-step experiments using primary frontier models, extensions to mixture-of-experts and multimodal architectures, deployment of α-based runtime monitoring, and measuring per-step α along extended reasoning traces to study the relationship between belief revision and test-time compute scaling.

Broader impact. The α-law has dual-use implications. It enables runtime detection of unreliable outputs without ground truth, potentially reducing hallucination risks and improving auditability of agentic systems. However, model-specific α fingerprints could also inform adversarial strategies that push belief revision into unstable regimes. Defensive applications and adversarial understanding should therefore advance together.

Reproducibility. All prompts, elicitation templates, evaluation scripts, preprocessing steps, and filtering criteria will be released upon publication.

References

Anthropic. The Claude model family. `https://www.anthropic.com/claude`, 2024.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Wenhu Chen, Xinyi Ming, Weizhe Huang, Zhuo Wang, and William Yang Wang. TheoremQA: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Peter Grünwald. The safe Bayesian: Learning the learning rate via the mixability gap. In *Algorithmic Learning Theory (ALT)*, pp. 169–183. Springer, 2012.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017.

Chris C Holmes and Stephen G Walker. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503, 2017.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *ICLR*, 2024.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

OpenAI. GPT-4 technical report, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*, 2022.

A Proof of Theorem 1

Proof. In log-space, let $\ell_t(i) = \log q_t(i)$ and $\beta(i) = \log b(i)$. The update gives:

$$\ell_{t+1}(i) = \alpha[\ell_t(i) + \beta(i)] + c_t$$

where $c_t = -\log \sum_j \exp\{\alpha[\ell_t(j) + \beta(j)]\}$ ensures normalization.

At fixed point $\ell^*(i) = \alpha[\ell^*(i) + \beta(i)] + c^*$, yielding $\ell^*(i) = \frac{\alpha\beta(i)+c^*}{1-\alpha}$ for $\alpha \neq 1$.

We work with the Hilbert projective metric $d_H(p, q) = \max_i \log(p(i)/q(i)) - \min_i \log(p(i)/q(i))$. The deviation $\delta_t(i) = \ell_t(i) - \ell^*(i)$ satisfies:

$$\delta_{t+1}(i) = \alpha\delta_t(i) + (c_t - c^*)$$

The normalization residual $(c_t - c^*)$ is constant across i, so it cancels in the Hilbert metric:

$$\begin{aligned} d_H(q_{t+1}, q^*) &= \max_i \delta_{t+1}(i) - \min_i \delta_{t+1}(i) \\ &= \alpha\big(\max_i \delta_t(i) - \min_i \delta_t(i)\big) \\ &= \alpha\, d_H(q_t, q^*) \end{aligned}$$

This is a direct consequence of Birkhoff's contraction theorem applied to the projective cone: the linear map $\delta \mapsto \alpha\delta$ contracts the Hilbert diameter by factor α when $0 < \alpha < 1$.

To pass from the Hilbert metric to KL divergence, we use the inequality $D_{\mathrm{KL}}(p\|q) \le \frac{1}{2} d_H(p,q)^2$, which holds for distributions on a finite alphabet. Iterating:

$$D_{\mathrm{KL}}(q_t\|q^*) \le \tfrac{1}{2}\, d_H(q_t,q^*)^2 = \tfrac{1}{2}\, \alpha^{2t}\, d_H(q_0,q^*)^2$$

Since $d_H(q_0,q^*)^2 \le 2\, D_{\mathrm{KL}}(q_0\|q^*)/p^*_{\min}$ for $p^*_{\min} = \min_i q^*(i) > 0$, this gives geometric convergence $D_{\mathrm{KL}}(q_t\|q^*) = O(\alpha^{2t})$.

For $\alpha \ge 1$: consider $K = 2$ and $b = (p, 1-p)$ with $p > 1/2$. The log-odds ratio $r_t = \ell_t(1) - \ell_t(2)$ follows $r_{t+1} = \alpha r_t + \alpha \log(p/(1-p))$. For $\alpha \ge 1$, $|r_t| \to \infty$, corresponding to simplex vertex collapse. □

B Evidence Score Computation

For each problem with K candidate answers, we construct the evidence distribution $b \in \Delta^{K-1}$ from deterministic verification as follows:

Multiple-choice benchmarks (GPQA Diamond, MMLU-Pro, ARC-Challenge). Each candidate corresponds to an answer option. We assign:

$$b(i) = \begin{cases} 0.9 & \text{if candidate } i \text{ matches the ground-truth answer} \\ \frac{0.1}{K-1} & \text{otherwise} \end{cases} \tag{7}$$

This reflects a strong verifier that identifies the correct answer with high confidence. The smoothing (0.9 rather than 1.0) prevents degenerate log-space values and reflects that real verification systems have non-zero error rates.

TheoremQA. Candidate solutions are matched against the ground-truth value using numerical comparison (with tolerance for floating-point arithmetic) or symbolic matching. The same evidence scheme applies after determining the correct candidate.

Noise injection (Appendix C.1). For the noisy verification experiments, we corrupt the evidence vector by flipping the high-probability assignment: with probability p_{flip}, the mass allocated to the verified-correct answer is reassigned to a uniformly random incorrect candidate. This simulates a verifier that occasionally endorses an incorrect solution.

C Additional Robustness Analyses

This appendix reports additional robustness analyses evaluating the stability of the α-law under perturbations of verification signals, evidence construction, and model parameterization assumptions.

C.1 Robustness to Noisy Verification

We evaluate the robustness of the α-law to corrupted verification signals. Synthetic noise is injected into the evidence vectors b by randomly flipping the high-evidence assignment from the verified correct answer to an incorrect candidate with probability p_{flip}. Posterior distributions q_1 remain fixed (computed using clean verification), and the α-law regression is recomputed using the corrupted evidence vectors.

Effect on α estimation. As corruption increases, the estimated revision exponent decreases from $\alpha = 1.163$ to $\alpha = 0.846$, while regression fit declines from $R^2 = 0.987$ to 0.816. This arises because noise corrupts the regression predictor

$$x_i = \log q_0(i) + \log b_{\text{noisy}}(i),$$

while the response $\log q_1(i)$ remains based on clean evidence, reducing predictor–response covariance.

Table 8: Effect of evidence noise on measured α ($n = 3{,}921$ clean records). Increasing noise reduces α and degrades regression fit quality.

Noise Level	α	R^2	**KL from Clean**
Clean (0%)	1.163	0.987	0.000
Moderate (20%)	1.027	0.911	0.638
Severe (40%)	0.846	0.816	1.287

Table 9: Evidence encoding sensitivity (pooled OLS across 3,921 clean records). α decreases monotonically with evidence strength while R^2 remains above 0.55 throughout.

Evidence Strength s	α	**95% CI**	R^2	n
$s = 0.51$	1.455	[1.437,1.471]	0.551	22,562
$s = 0.60$	1.397	[1.383,1.411]	0.606	22,562
$s = 0.70$	1.307	[1.296,1.318]	0.653	22,562
$s = 0.80$	1.190	[1.181,1.199]	0.691	22,562
$s = 0.90$	1.028	[1.021,1.034]	0.722	22,562
$s = 0.99$	0.700	[0.696,0.704]	0.743	22,562

Connection to attenuation bias. The behavior corresponds to classical attenuation bias in errors-in-variables regression: measurement error biases coefficients toward zero without changing the underlying model update dynamics.

Illustrative example. Under 40% corruption, a representative GPQA instance shows per-problem α decreasing from 1.21 to -0.34 while R^2 drops from 0.91 to 0.12, demonstrating how predictor corruption can reverse the apparent update direction.

C.2 Evidence Encoding Sensitivity

We test whether the log-linear fit depends on the binary evidence encoding ($s = 0.9$ assigned to the correct answer and $(1 - s)/(K - 1)$ to alternatives). The evidence concentration parameter s is varied from near-uniform ($s = 0.51$) to near-certain ($s = 0.99$), while holding model posteriors q_1 fixed.

Observation. The log-linear structure $\log q_1 = \alpha(\log q_0 + \log b) + c$ holds across all evidence strengths, confirming that the multiplicative scaling relationship reflects intrinsic LLM update behavior rather than an artifact of encoding. Absolute α values depend on encoding choice, and should therefore be interpreted relative to a fixed s.

C.3 Identifiability and Model Sufficiency

A potential concern is that the unified α model conflates prior and evidence scaling. We assess identifiability by comparing uniform and informative (Dirichlet-sampled) priors across 300 synthetic trials.

- **Uniform priors.** The design matrix is severely ill-conditioned (condition number = 4,952), yielding unstable coefficients ($\alpha_{q_0} = 2.86$, $\alpha_b = 1.27$).
- **Informative priors.** The condition number decreases to 20.1, and the unified model recovers $\alpha = 1.170$ exactly.
- **Model comparison.** Adding a second parameter improves R^2 by less than 0.001.

Conclusion. The single-α parameterization is sufficient: even when identifiability is restored, the two coefficients are not meaningfully separable in practice.

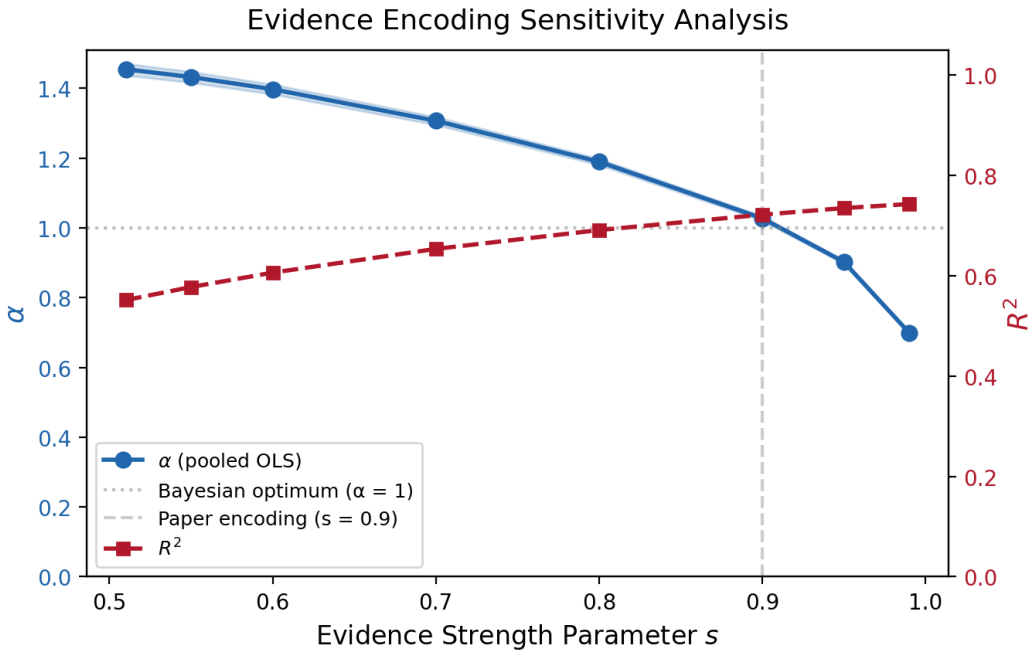


Figure 7: Evidence encoding sensitivity analysis. The log-linear form holds across evidence strengths ($R^2 \geq 0.55$), while absolute α varies monotonically with encoding concentration.

Table 10: Calibration comparison across 2,945 problems. Traditional confidence metrics (Max Probability, Margin) outperform α on classical calibration measures because α is not optimized to predict correctness but to characterize belief-update geometry.

Metric	AUROC	ECE ↓	Brier ↓	**Interpretation**
Max Probability	0.719	**0.075**	**0.020**	Best classical calibration
Margin	0.716	0.095	0.027	Similar to MaxP
Entropy	0.684	0.209	0.076	Higher dispersion
α	—	0.387	0.168	Geometric deviation measure

D Calibration and Correctness Analysis

D.1 Comparison with Traditional Confidence Metrics

To assess whether the revision exponent α can function as a confidence signal, we compare it against standard calibration metrics on 2,945 evaluation problems. We evaluate discrimination using AUROC and calibration using Expected Calibration Error (ECE) and Brier score.

Traditional metrics based on posterior concentration (MaxP, Margin) achieve lower ECE and Brier scores than α. This is expected: α measures the geometric structure of belief revision rather than predictive calibration.

D.2 Correctness-Conditioned α Behavior

Although α underperforms classical metrics on calibration measures, it exhibits distinct qualitative structure when conditioned on correctness. Correct predictions cluster near the Bayesian regime, whereas incorrect predictions frequently produce dispersed or negative α values.

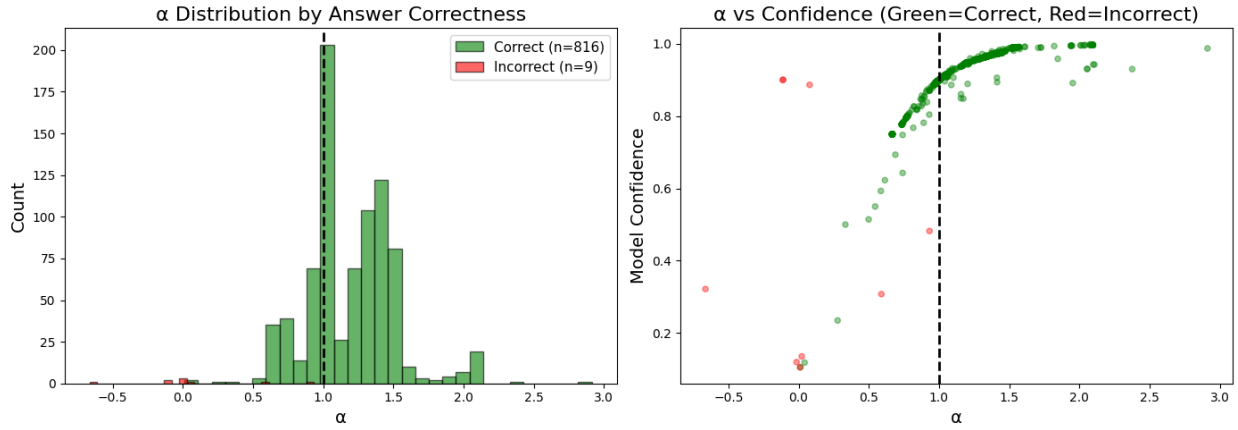


Figure 8: α conditioned on correctness. Correct predictions concentrate near the Bayesian regime, while incorrect predictions exhibit dispersed or negative values. Only 9 of 816 analyzed instances were incorrect, limiting statistical power for correctness-conditioned analysis.

Across evaluated instances,

$$\alpha_{\text{correct}} \approx 1.25, \qquad \alpha_{\text{incorrect}} \approx -0.66.$$

This separation indicates that α captures a signal distinct from posterior confidence: it reflects whether the revision step behaves coherently relative to prior and evidence rather than how concentrated the posterior is.

D.3 Interpretation

The comparison suggests that α should not be interpreted as a replacement for calibration metrics such as Max Probability or Margin. Instead, the quantities capture complementary properties:

- Classical metrics measure posterior concentration.
- α measures deviation from Bayesian update geometry.

In high-accuracy regimes (greater than 99% accuracy in our benchmarks), calibration metrics have limited discriminative opportunity due to the small number of incorrect examples. In such settings, α provides interpretability as a structural diagnostic of belief integration.

D.4 Limitations of the Comparison

The calibration comparison is constrained by the near-perfect accuracy of the evaluated models, yielding only a small number of incorrect samples. More informative comparison would require benchmarks where model accuracy is substantially lower (e.g., 40–70%), enabling stable estimation of AUROC and calibration error under error-heavy regimes.

We therefore interpret the results in Table 10 as descriptive rather than definitive.

E Additional Figures

This appendix provides additional visual validations supporting the empirical properties of the α-law discussed in the main text.

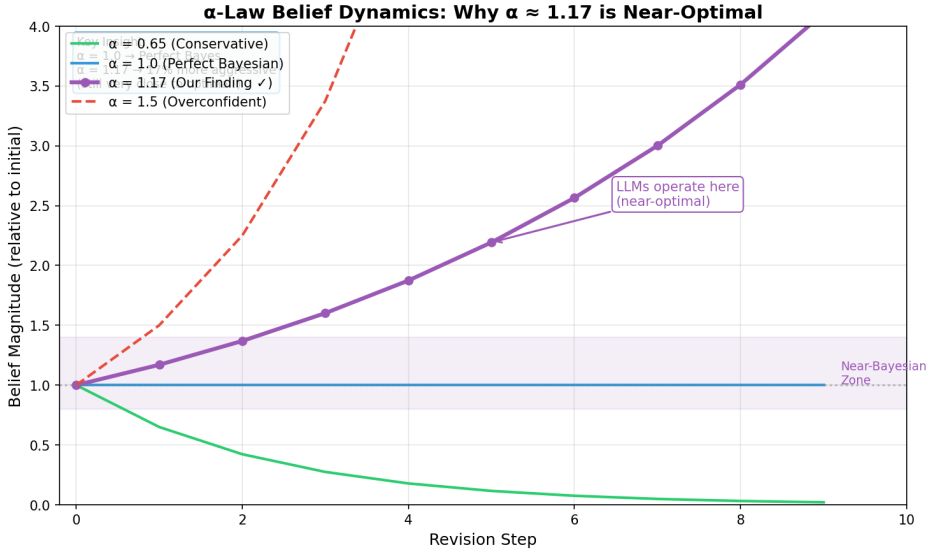
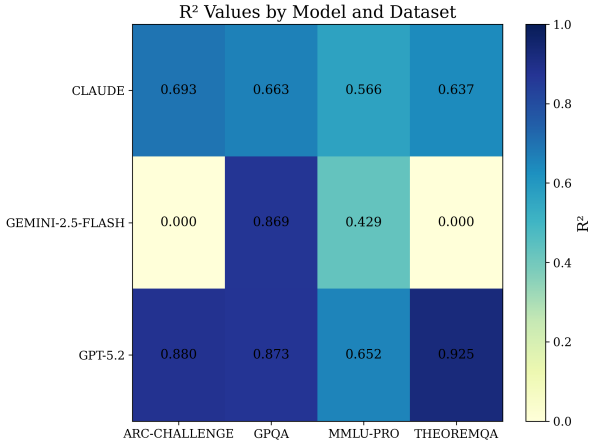


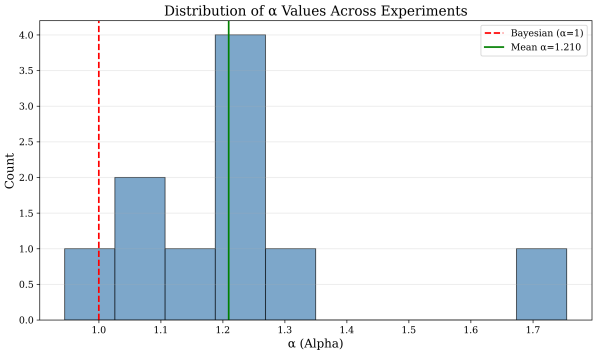
Figure 9: **Three α-regimes of belief revision dynamics.** *Contractive* ($\alpha < 1$): KL divergence from q^* decreases geometrically. *Bayesian* ($\alpha = 1$): marginal stability, no amplification or damping. *Expansive* ($\alpha > 1$): errors amplify each step. The empirical single-step operating point ($\alpha \approx 1.16$, purple) is mildly expansive; the multi-step geometric mean ($\bar{\alpha}_{\text{geo}} = 0.75$, dashed) lies firmly in the contractive regime, bridging the theory–empirics gap.

Stability regime schematic. Figure 9 provides a compact visual summary of the three α-regimes from Theorem 1. The schematic illustrates how KL divergence from the fixed point q^* evolves under iterated revision for representative α values in each regime, and marks the empirically observed operating point ($\alpha \approx 1.16$, single-step) together with the multi-step geometric mean ($\bar{\alpha}_{\text{geo}} = 0.75$).

Fit quality across models and datasets. Figure 10(a) presents R^2 values across all model×dataset combinations, while Figure 10(b) shows the global distribution of per-problem α estimates. Together, these results demonstrate both strong goodness-of-fit and concentration near the Bayesian optimum.



(a) R^2 heatmap across model×dataset combinations.



(b) Overall α distribution histogram.

Figure 10: **Fit quality and global α concentration.** (a) Heatmap of R^2 values measuring goodness-of-fit of the predicted scaling relationship. (b) Histogram of per-problem α estimates showing concentration around 1.0–1.2.

Robustness to candidate width. As shown in Figure 11, estimated α remains stable when varying the number of candidates ($K = 4, 8, 16$), indicating that the observed scaling behavior is not an artifact of answer-set size.

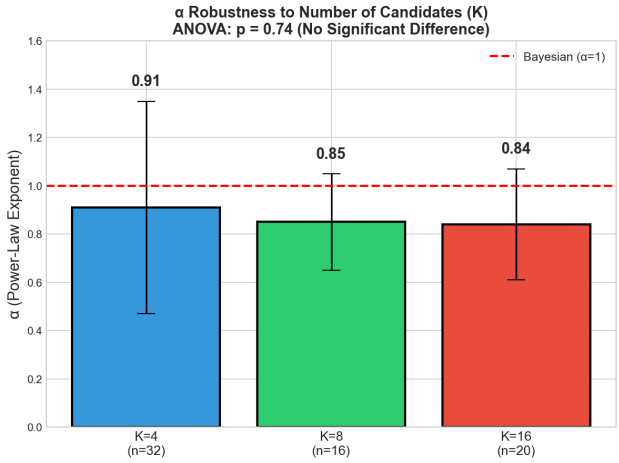


Figure 11: **K-ablation robustness analysis.** Estimated α remains stable across candidate counts. ANOVA ($p = 0.74$) indicates no significant difference across K.

Domain generalization. Figure 12 shows that both mean α values and goodness-of-fit metrics remain consistent across reasoning domains, supporting domain-agnostic behavior of the α-law.

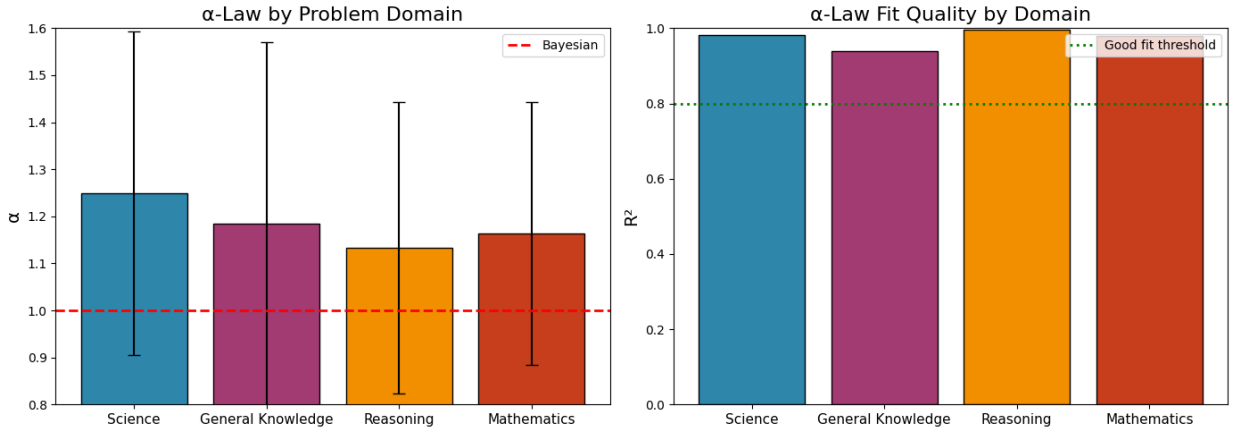


Figure 12: **Domain-specific α behavior and fit quality.** Mean α remains near unity across domains, with consistently high R^2 values indicating strong agreement with the predicted scaling relation.

Overall validation summary. Finally, Figure 13 provides a consolidated overview of all validation diagnostics discussed throughout the appendix, summarizing robustness, fit quality, and cross-domain consistency at a glance.

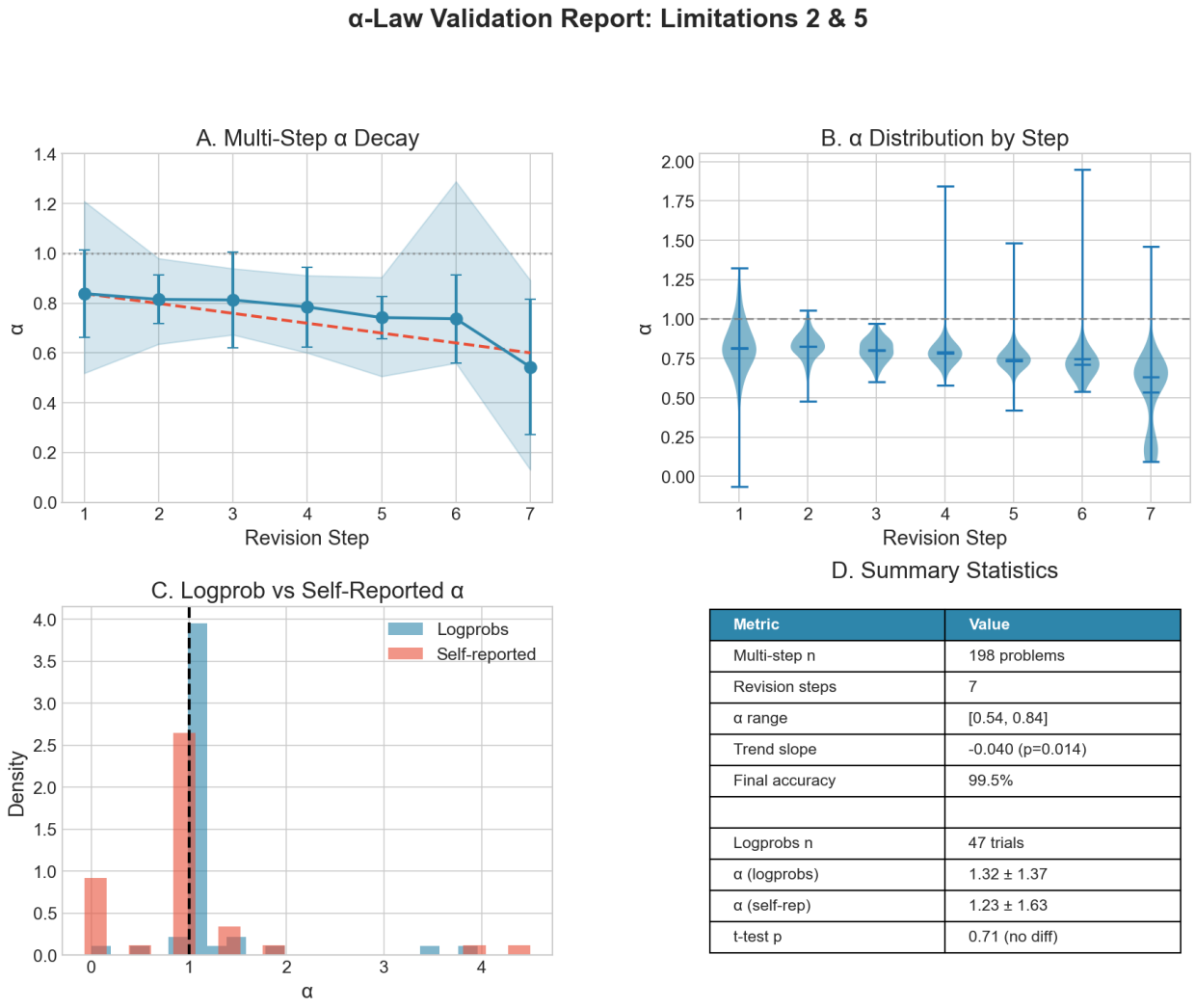


Figure 13: **Summary dashboard of validation results.** A consolidated overview of robustness checks, fit quality metrics, and cross-domain consistency supporting the α-law.