Hallucination as Commitment Failure: Larger LLMs Misfire Despite Knowing the Answer

Jewon Yeom [1] **Jaewon Sok** [2] **Heejun Kim** [3] **Seonghyeon Park** [4] **Jeongjae Park** [1] **Taesup Kim** [1]

Abstract

Hallucination is often viewed as a direct consequence of missing knowledge: a model answers incorrectly when the correct answer is absent from its generation-time distribution, and correctly when it is present. We test this assumption by introducing a semantic notion of answer availability that aggregates token-level variants expressing the same answer concept, and asks whether the correct concept is already available at the moment the model commits to an answer. Across Qwen and Llama models from 0.8B to 72B in both Instruct and Base variants, 16–47% of Instruct hallucinations occur with substantial probability mass already on the correct concept, and the rate rises monotonically with scale. Comparing such failures against correct generations with matched semantic support, the distinguishing factor is not whether the correct concept is represented, but how its probability is distributed: correct generations concentrate mass on a single surface form, hallucinations disperse it across alternatives. The same sharpening asymmetry extends across multi-token generation and is detectable in pre-generation hidden states. Together, these results identify a single mechanism: instruction tuning sharpens answer commitment with scale, making helpfulness and confident hallucination two consequences of the same underlying disposition.

1. Introduction

Large language models (LLMs) frequently produce fluent but factually incorrect outputs—hallucinations—that undermine their reliability in safety-critical applications (Ji et al., 2023). A natural first question is *where in the generation trajectory* a hallucination is decided. If hallucination emerges everywhere, post-hoc analysis of full sequences is necessary; if it localizes at specific steps, both detection and intervention should target those steps.

Recent work in the reasoning literature suggests sharp localization. Inspecting token-level entropy $H(y_t \mid Q, y_{<t})$ during greedy decoding reveals that entropy is highly non-uniform across the sequence: at most steps it is near zero (the next token is essentially deterministic) but at a small number of steps it spikes sharply. Wang et al. (2025) report that $\sim$20% of tokens in chain-of-thought traces carry high entropy and act as "forking tokens" that determine reasoning paths; Vassoyan et al. (2025) identify "critical tokens" as decision points where models are most error-prone. Figure 1 shows the same phenomenon in a QA setting: an early spike fixes the domain of the answer (`Britain`), a later spike selects the answer entity (`Nicola`), and the steps in between are syntactic continuations following automatically.

A natural follow-up is whether entropy at these spikes is itself a hallucination signal. Existing work has established that it is not, in a stronger form than we will need: Simhi et al. (2025) document hallucinations produced with high certainty even when the model demonstrably has the correct knowledge, Xu et al. (2025) formalize "high-belief hallucinations" as a distinct phenomenon from confidence-based detection, and the original semantic entropy work of Farquhar et al. (2024) explicitly notes that confidently wrong outputs are a separate phenomenon from the confabulation regime that semantic entropy targets. The literature's response has been to develop more sophisticated estimators (Kuhn et al., 2023; Ma et al., 2025) or perturbation-based diagnostics (Simhi et al., 2025) that better separate hallucinated from non-hallucinated outputs. We pursue an orthogonal direction: rather than designing a better detector, we ask what the model's distribution is doing at the commitment step when the final answer is hallucinated, regardless of whether the wrong answer is emitted with high confidence

[1]Graduate School of Data Science, Seoul National University [2]Department of Rural Systems Engineering, Seoul National University [3]Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology [4]Department of Aerospace Engineering, Seoul National University. [†]Correspondence to: Taesup Kim <taesup.kim@snu.ac.kr>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.

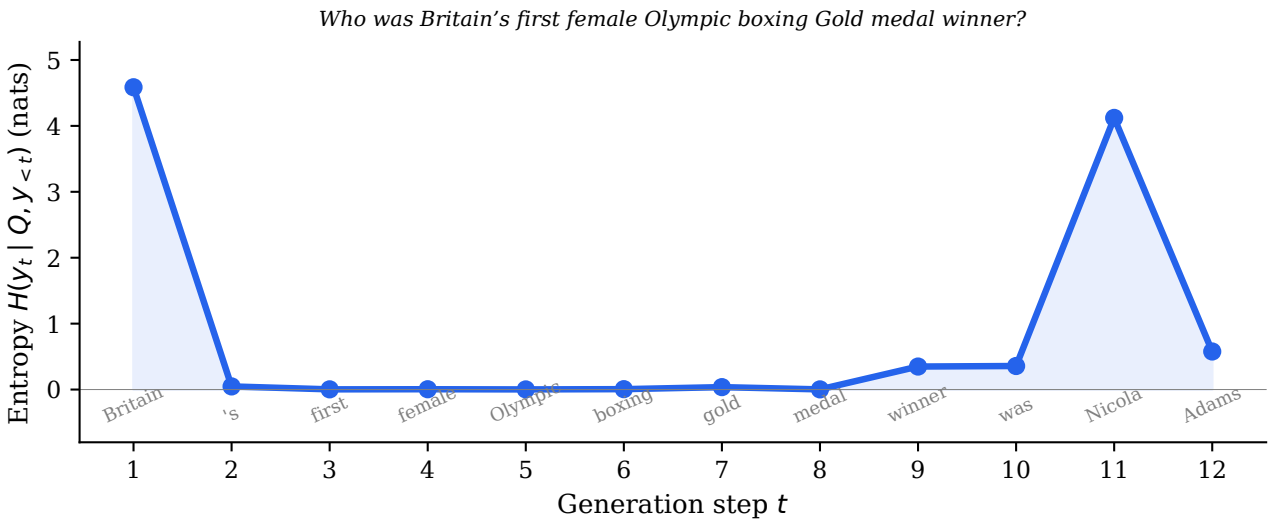

Figure 1. Token entropy $H(y_t \mid Q, y_{<t})$ across a representative generation trajectory (Qwen3.5-9B Instruct). Entropy is near zero at most steps but spikes sharply at a small number of *commitment steps*.

or not.

We refer to the answer-emission step as the *commitment step* t_c. In short-form QA with instruction-tuned models, $t_c = 1$ (§4.1), so the prompt format fixes the commitment step at $t = 1$ and lets us inspect the distribution at a single, known step. What it reveals is that "high entropy" has two structurally different sources: mass spread over genuinely different answers, and mass spread over different surface forms of the same answer (`Paris`, ` Paris`, `paris`, or `St`, `Saint`, `C` for `St. Basil's Cathedral`, Figure 2). To distinguish them, we define a *concept* as the equivalence class of token completions denoting the same answer and introduce the *per-step semantic probability mass* $P_{\text{mass}}(t; c) = \sum_{v \in S_c} P_\theta(v \mid Q, y_{<t})$, where S_c collects the first-token IDs of the concept's surface forms. With S_c built from ground-truth aliases, $P_{\text{mass}}(t; c^*)$ is an analytical probe—requiring the answer concept as input—rather than a deployable detector, but precisely this property lets us ask whether the model put substantial mass on the right answer at the moment of commitment.

The headline finding is that across nine instruction-tuned Qwen and Llama models from 0.8B to 72B, 16% to 47% of hallucinated outputs have $P_{\text{mass}}(t_c; c^*) \geq 0.2$: the model assigned non-trivial mass to the correct concept yet produced a wrong final answer. We call these *commitment failures*, and the rate rises monotonically with scale across both Qwen and Llama families. Commitment failures decompose into two cases: in ∼20%, the greedy first token does not match any surface form of c^* at all (*first-token selection failures*); in the remaining ∼80%, the greedy first token does land on a surface form of c^* but the continuation diverges (*multi-token divergences*). The first sub-population isolates a particularly clean question: when the model put substantial mass on c^* at the commitment step yet selected a token outside S_{c^*}, what does its distribution look like? We compare against *matched correct samples*—correct outputs whose $P_{\text{mass}}(t_c; c^*) \geq 0.2$, drawn from the same range of correct-concept mass—and find that selection failures have a three-fold lower maximum probability on any single surface-form token of c^* (0.26 vs. 0.78). The model has the same *amount* of mass on the correct concept; it just has it spread across alias forms (`Saint`, `St`, `C`) rather than concentrated on one, so a competing concept's single dominant token wins the argmax.

The empirical driver of the scale trend is instruction tuning, not scale itself. The probability assigned to the (wrong) greedy token in first-token selection failures rises monotonically across Instruct models—Qwen: 0.31 (0.8B) to 0.57 (72B); Llama: 0.33 (1B) to 0.49 (70B)—but stays flat at ∼0.30 across Base models of the same sizes. The same pattern extends to multi-token answers: in 70B+ Instruct, $H_{t=2}$ is ≈ 0.05 when the bigram (y_1, y_2) stays on a valid alias prefix of c^* and substantially higher when it diverges (Cohen's $d = 1.29$ across 18 models)—instruction tuning sharpens commitment specifically along c^*-aligned phrases. Instruction tuning sharpens commitments at multiple levels, and the same sharpening produces confident correctness when the committed phrase is right and confident misselection when it is wrong—making confident hallucination one face of the broader "alignment tax" that has been documented as accuracy loss (Ouyang et al., 2022), calibration loss (OpenAI, 2023; Hu et al., 2025), and mode collapse. The analytical probe P_{mass} (§3), the commitment-failure phenomenon and its scale dependence (§4.2), the within-population charac-

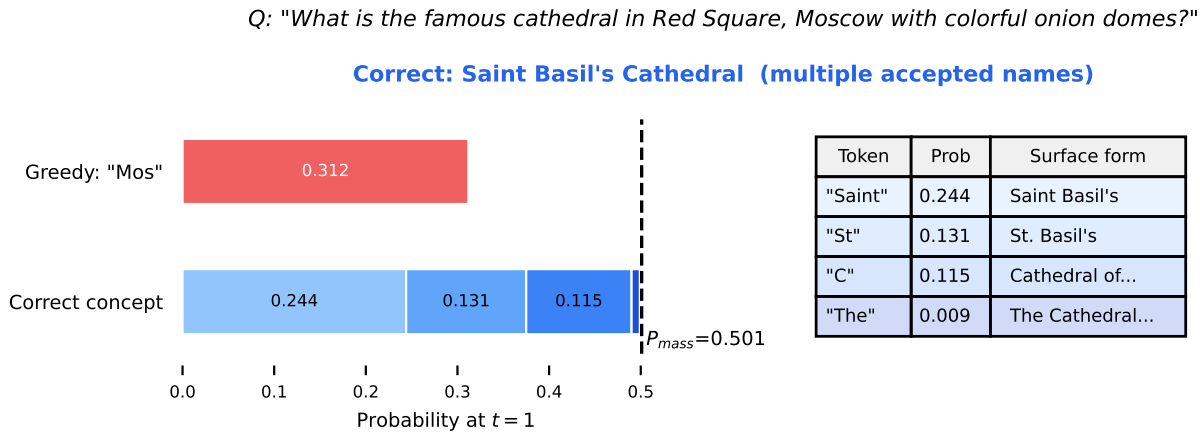

Figure 2. Vocabulary fragmentation at the commitment step: the correct concept's mass (0.501 total) is split across `Saint` (0.244), `St` (0.115), and `C` (0.131); the greedy token is the competing `Mos` (0.312). Greedy decoding hides what the distribution says about the correct concept. This pattern is most pronounced in small and Base models; in large Instruct models, fragmentation collapses (§4.3) and commitment failures arise instead from a wrong concept's token being even sharper than the correct concept's collapsed mass.

terization (§4.3), and the representation-level evidence for instruction-induced sharpening (§4.4) together constitute the contribution of this paper.

2. Related Work

Confident hallucination and uncertainty-based detection. Token- and sequence-level uncertainty has been the dominant lens for hallucination detection: perplexity (Ren et al., 2023), length-normalized NLL (Malinin & Gales, 2021), predictive entropy (Kadavath et al., 2022), importance-weighted scoring (MARS; Bakman et al., 2024), semantic entropy and its variants (Kuhn et al., 2023; Farquhar et al., 2024; Ma et al., 2025), sampling consistency (SelfCheckGPT; Manakul et al., 2023), consistency-confidence aggregation (CoCoA; Vashurin et al., 2025), and confidence elicitation (Xiong et al., 2024). A growing line of work documents that confident hallucination is itself a phenomenon: Farquhar et al. (2024) note that semantic entropy does not address confidently wrong outputs, Simhi et al. (2025) formalize "CHOKE" (certain hallucinations overriding known evidence), and Xu et al. (2025) characterize "delusions" as high-belief hallucinations. Calderon et al. (2026) characterize a closely related distinction ("empty shelves" vs. "lost keys"), finding that recall—not encoding—is the dominant bottleneck even in frontier models, through behavioral fact-level profiling; we provide complementary distributional analysis at the commitment step. These works establish the phenomenon at the response level; we provide its structural account at the commitment step (§4.2, 4.3).

Calibration and the alignment tax. Kadavath et al. (2022) showed pretrained LLMs are well-calibrated under appropriate elicitation, while OpenAI (2023) reported that pre-training yields well-calibrated probabilities but RLHF post-training degrades calibration substantially—a finding extended in Xie et al. (2024) for the token-level case and Chhikara (2025) for the question-type case. More broadly, the "alignment tax" originally framed as a drop in task accuracy after RLHF (Ouyang et al., 2022) is increasingly understood to include calibration loss and mode collapse: Hu et al. (2025) document that alignment makes models overconfident with reduced output diversity, framing this as a calibration–alignment trade-off. We give this its mechanism at the moment of commitment (§4.3): the same sharpening drives both confident correctness and confident misselection.

Token-level decision points. Wang et al. (2025) show that high-entropy "forking tokens" in chain-of-thought reasoning carry a disproportionate share of the learning signal in RLVR. Vassoyan et al. (2025) identify "critical tokens" as decision points where models are most error-prone. We extend the same phenomenon to short-form QA (Figure 1) and show that the relevant signal at the step is concept-grouped mass, not individual-token entropy.

Internal representations and first-token signal. Truthfulness is linearly decodable from hidden states (Burns et al., 2023; Azaria & Mitchell, 2023; Marks & Tegmark, 2024); DoLa (Chuang et al., 2024) and ITI (Li et al., 2023) act on this for decoding-time intervention. SEP (Kossen et al., 2024) introduces token-before-generation (TBG) probing. Token-level analyses have converged on first-token importance (Snel & Oh, 2025; Zhao et al., 2024); HaMI (Niu et al., 2025) adaptively selects informative tokens. We refine this: what matters is not position but the answer-level commitment step (first token in instruction-tuned short-form

QA, but migrating in long-form generation; §4.1), and we provide the first systematic Instruct–Base comparison for the TBG setting (§4.4).

3. Setup

Concept and P_{mass}. Given a query Q, an autoregressive LLM P_θ generates tokens $y_1, y_2, \ldots$ A *concept* c is an equivalence class of token completions denoting the same answer; its first-token surface forms collect into a *concept token set* S_c. We study the *per-step semantic probability mass*

$$P_{\text{mass}}(t; c) = \sum_{v \in S_c} P_\theta(v \mid Q, y_{<t}), \tag{1}$$

the total mass at step t on any first-token surface form of c. To analyze hallucination we set $c = c^*$, the ground-truth concept, with S_{c^*} built deterministically from the dataset's aliases (Appendix I). Under a latent-concept generation model, $P_{\text{mass}}(t; c^*)$ approximates the (unobservable) *concept belief* $P_\theta(c^* \mid Q, y_{<t})$ when S_{c^*} is alias-complete and concepts are well-separated; we make this precise in Appendix A (Proposition 1).

Models and data. Our primary scale ablation uses Qwen3.5 (Yang et al., 2025) at four sizes (0.8B, 2B, 4B, 9B) in both Instruct and Base variants, with all 4-bit NF4 quantization, paired with Llama-3.2 (1B, 3B) and Llama-3.1 (8B) (Grattafiori et al., 2024) in both variants—fourteen models total at small to mid scale. We extend the scale ablation to four large models (Qwen2.5-72B (Yang et al., 2024) and Llama-3.1-70B in both Instruct and Base) for the wrong-token sharpening and commitment-failure rate analyses. We use TriviaQA (Joshi et al., 2017) and NQ-Open (Kwiatkowski et al., 2019) for short-form QA (3,000 samples per model) and MMLU (Hendrycks et al., 2021) together with ARC-Challenge (Clark et al., 2018) for multiple-choice QA (2,672 samples per model); the representation analyses in §4.4 use a 1,500-sample subset (1,000 MCQA + 500 Short-QA).

Hallucination. Throughout, a response is a *hallucination* if it fails substring matching against the ground-truth aliases (Short-QA) or selects a wrong option (MCQA); we use "hallucinated" and "incorrect" interchangeably, following the convention of SE (Farquhar et al., 2024) and SEP (Kossen et al., 2024).

Metrics. Detection performance is AUROC with hallucination as the positive class. Probes are 5-fold CV logistic regression on hidden states. Calibration uses ECE (Guo et al., 2017) and Brier score.

4. Results

4.1. Where in the trajectory does correctness signal live?

Before turning to the central analysis, we verify the entropy/P_{mass} picture in our own data using long-form generation, where the commitment step t_c is not at $t = 1$. We collected 500 long-form Qwen3.5-9B Instruct responses (`Answer in a complete sentence`) and centered each trajectory on its commitment step t_c.

Figure 3(a) plots entropy as a function of step relative to t_c. Entropy peaks sharply at t_c for both correct and hallucinated samples, and hallucinated trajectories carry uniformly higher entropy at every relative step, but the per-sample $H(t_c)$ distributions overlap substantially (Wilcoxon $p = 0.055$). The max-entropy step t_H exactly matches t_c in only 20% of samples, within one step in 32%—a noisy localization at best.

Figure 3(b) re-plots the same trajectories using $P_{\text{mass}}(t; c^*)$ instead. P_{mass} is essentially zero everywhere except t_c, where it spikes to 0.92 for correct samples and 0.77 for hallucinated ones; generated-token probability $P(y_t)$ is nearly flat across both groups (Appendix Figure 7). The relevant fact is not that P_{mass} separates the two classes (it has access to ground-truth aliases) but that the gap is small: at the commitment step, hallucinated samples place a substantial 0.77 average mass on the correct concept yet still emit a different one. This sub-population is what we analyze in the rest of the paper.

The commitment step is where the information lives. Figure 3(c) makes this concrete: per-step detection AUROC of $P_{\text{mass}}(t; c^*)$ is at chance two or more steps before t_c, climbs to its peak at t_c, and decays back to chance within a few steps after. Generated-token probability $P(y_t)$ over the same window is flat (0.55–0.62 throughout, no peak): token-level confidence barely moves around t_c, while concept-grouped mass rises and falls sharply. This is the structural justification for studying the model's distribution at t_c specifically rather than aggregating across the trajectory.

4.2. Does the model have the answer when it hallucinates?

We now turn to short-form QA, where instruction-tuned models answer immediately and the commitment step is fixed at $t = 1$ (Appendix Table 8). This lets us inspect the model's distribution at a single, known step, and ask the central question: among hallucinated outputs, what does $P_{\text{mass}}(t_c; c^*)$ look like?

P_{mass} recovers a coherent confidence signal. P_{mass} is well-calibrated (ECE 0.023–0.096 across 7 Instruct models; Appendix Figure 9), with accuracy increasing monotonically across P_{mass} bins, and it outperforms the generated

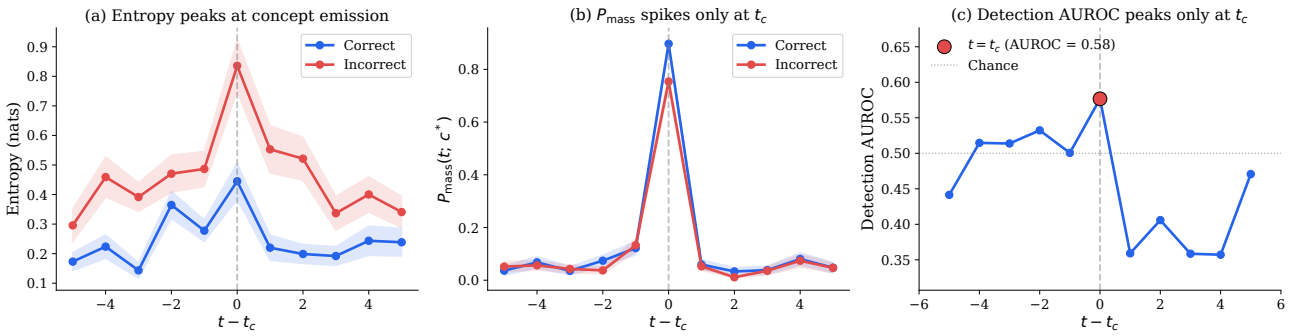

Figure 3. 500 long-form Qwen3.5-9B Instruct responses aligned to each trajectory's commitment step t_c. *(a) Entropy* peaks at t_c for both groups; hallucinated trajectories run uniformly higher, but per-sample distributions at t_c overlap. *(b)* $P_{\text{mass}}(t; c^*)$ is essentially zero except at t_c, where it concentrates to 0.92 (correct) and 0.77 (hallucinated). The relevant fact is not the separation between groups but the high P_{mass} on the hallucinated side: at the commitment step, the model often has substantial mass on the correct concept yet emits a competing one. *(c) Detection AUROC* of P_{mass} peaks sharply at t_c and decays to chance off-step, showing that the predictive information about correctness is localized to the commitment step itself.

greedy token's probability as a correctness signal in every one of the 18 models (detection AUROC $+0.07$ to $+0.57$; Appendix Table 10)—it captures the concept-level structure that token-level confidence misses, with the largest gains exactly where answers have many surface forms. The per-step result in Figure 3(c) confirms t_c is the right inspection point: aggregating P_{mass} across the trajectory underperforms the single-step quantity at t_c (Appendix M).

Commitment failures. A hallucinated sample is a *commitment failure* if $P_{\text{mass}}(t_c; c^*) \geq 0.2$: substantial mass on the correct concept yet a wrong final answer. The phenomenon arises because greedy emission depends on the maximum single-token probability, not on P_{mass}, so individual surface-form tokens of c^* may each be small enough that a competing concept's single dominant token wins, or the multi-token continuation may diverge after a correct first token. CF% rises monotonically from 16% at 0.8B to 47% at 70B Instruct (Table 1), holding across both Qwen and Llama families. Larger models do not produce uniformly fewer errors; they shift the error distribution toward commitment failures. The 0.2 threshold is conservative; the trend is robust to threshold choice in $[0.1, 0.4]$ (Appendix G).

Two structural sub-populations. Commitment failures decompose into two cases at the token level. In a *first-token selection failure*, the greedy emission y_1 does not match any surface form of c^* at all (e.g., the answer is `Saint Petersburg` and the model emits `Mos` to begin `Moscow`). In a *multi-token divergence*, y_1 does land on a surface form of c^* but the multi-token continuation diverges from any valid alias (e.g., the answer is `Adam Smith` but the model emits `Adam Levine`). The two are distributionally different and we treat them separately. Across the scale ablation, first-token selection failures account for roughly 20% of commitment failures, rising monotonically with scale: from 2.9% of all hallucinations at 0.8B to 6.0% at 72B in Qwen Instruct, and from 3.3% at 1B to 10.2% at 70B in Llama Instruct (Table 1). The remaining ∼80% are multi-token divergences—hallucinations where the first token is on track but the continuation is not. Multi-token divergences are analyzed in detail in §4.3 and Appendix F.

4.3. Why does the model commit to the wrong token?

We focus here on first-token selection failures, the strict sub-population where greedy emission $y_1 \notin S_{c^*}$ despite $P_{\text{mass}}(t_c; c^*) \geq 0.2$. These cases are directly inspectable at the token level: the model assigned substantial mass to the correct concept but a single token of a competing concept won the argmax. We ask two questions: (i) within a model, what distinguishes the distribution at a selection failure from a correct-but-comparable sample (one with a similar amount of mass on c^*)? (ii) How do these distributions change with scale?

Within-population: less concentrated correct mass. To isolate the token-level structure, we compare two groups in Qwen3.5-9B Instruct, both restricted to samples with $P_{\text{mass}} \geq 0.2$: first-token selection failures (those that hallucinated, $N = 128$) and *matched correct samples* (those that answered correctly, $N = 840$). Both groups have substantial mass on the correct concept; the only difference is the outcome. The maximum probability assigned to any single surface form of c^*, $\max_{v \in S_{c^*}} P_\theta(v)$, is dramatically smaller in the failure group: mean 0.26 vs. 0.78 (Welch $t = 47.6$, $p < 10^{-180}$, Cohen's $d = 2.98$; Appendix B). The same comparison across all 18 models gives $d < 0$ in 100% of cases (median $|d| = 1.93$, range $[1.01, 4.30]$); within Instruct models $|d|$ tends to grow with scale (Qwen Inst: 1.34→2.98→4.30 from 0.8B to 72B; Appendix Table 2). Vocabulary fragmentation across alias forms (Figure 2, e.g. `Saint`, `St`, `C`) is the most concrete realization in small and Base models; in mid-to-large Instruct models, the within-

Table 1. Commitment-failure rate (CF%) and decomposition into first-token selection failures (SF, greedy $\notin S_{c^*}$) and multi-token divergences (Div, greedy $\in S_{c^*}$ but final answer wrong) across the Instruct scale ablation. Acc: Short-QA accuracy. AUROC: $P_{\text{mass}}(t{=}1)$ AUROC. Halluc: hallucinated sample count. SF%: SF as fraction of all hallucinations. Full per-model details including Base models in Appendix Table 9.

Model	Acc	AUROC	Halluc	CF (CF%)	SF	Div	SF%
Qwen3.5-0.8B Inst	8.3%	.898	2,751	453 (16%)	81	372	2.9%
Qwen3.5-2B Inst	11.0%	.893	2,671	466 (17%)	97	369	3.6%
Qwen3.5-4B Inst	24.6%	.912	2,262	589 (26%)	128	461	5.7%
Qwen3.5-9B Inst	29.4%	.887	2,117	667 (32%)	128	539	6.0%
Qwen2.5-72B Inst	36.4%	.830	1,908	777 (41%)	114	663	6.0%
Llama-3.2-1B Inst	14.8%	.935	2,559	401 (16%)	84	317	3.3%
Llama-3.2-3B Inst	31.4%	.902	2,054	577 (28%)	157	420	7.6%
Llama-3.1-8B Inst	32.8%	.882	2,018	657 (33%)	178	479	8.8%
Llama-3.1-70B Inst	44.7%	.816	1,659	**780 (47%)**	170	610	10.2%

concept distribution has typically already collapsed onto a single alias and the SF–Corr gap arises from a different route, characterized below.

Across scale: monotonic sharpening, modulated by instruction tuning. Within first-token selection failures, the greedy emitted token is by definition outside S_{c^*}; we call its probability $P_\theta(y_1)$ the *wrong-token probability*. This rises monotonically with model size in instruction-tuned models: Qwen Instruct 0.31 (0.8B) $\rightarrow$ 0.36 $\rightarrow$ 0.40 $\rightarrow$ 0.44 $\rightarrow$ 0.57 (72B); Llama Instruct 0.33 (1B) $\rightarrow$ 0.43 $\rightarrow$ 0.46 $\rightarrow$ 0.49 (70B) (Figure 4, left). Base models behave differently: across the same size range, wrong-token probability stays near 0.30 (Llama Base: 0.26 at 1B to 0.33 at 70B; Qwen Base 0.8B/72B: 0.27/0.31). Scale alone does not produce sharpening; the combination of scale and instruction tuning does. The same sharpening that produces front-loaded correctness signal in §4.4 produces decisive misselection when the committed concept is wrong.

Sharpening extends to multi-token answers. The same sharpening operates beyond the first token, and the analysis instrument is the same in kind: the first-token analysis measures concentration as the maximum single-token probability within S_{c^*}; its multi-token analogue is the entropy of the next-token distribution conditional on the emitted prefix—$H_{t=2}$ measures phrase-level concentration exactly as $\max_v P(v)$ measures token-level concentration. We measure at $t = 2$ because the bigram is the earliest point at which alias alignment can break: entropy at a later forking token (e.g., the third token of `George Washington Carver`) reflects where a divergence is decided, whereas our claim concerns how early the model locks the phrase. Within multi-token divergences, $H_{t=2}$ (entropy of the next-token distribution after y_1) measures whether the model has committed to a specific multi-token phrase at $t = 1$ (low $H_{t=2}$) or is still deciding. We classify each divergence by whether the second token continues an alias of c^*. *Type A*: bigram (y_1, y_2) matches the start of a valid alias but the continuation diverges into a different entity (e.g., `George Washington Carver`—the agricultural scientist—when the answer is `George Washington` the first U.S. president; the bigram `George Washington` is shared with the alias prefix, but `Carver` fixes a different person). *Type B*: bigram diverges already at y_2 (e.g., `Adam Lambert` when the answer is `Adam Smith`: $y_1 =$ `Adam` is in S_{c^*} but `Lambert` breaks the alignment). Across 18 Instruct/Base models, Type A divergences have substantially lower $H_{t=2}$ than Type B (median Cohen's $d = 1.29$, $d > 1$ in 100% of models with sufficient N; Appendix F)—when the bigram is on track, the continuation is near-deterministic. The Type A fraction grows with both scale and instruction tuning, from 21–44% at 0.8B/1B to 77–83% at 70B+ Instruct. At 70B+ Instruct, $H_{t=2}$ within Type A divergences is 0.05–0.10—the model commits to wrong multi-token phrases with residual entropy comparable to deterministic continuations.

Two faces of commitment failure. The within-population analysis above measures top1 alias mass without normalizing by $P_{\text{mass}}(c^*)$. To separate *within-concept* structure (how the correct mass is distributed across alias forms) from *between-concept* structure (how the wrong greedy token compares to the correct concept), we measure two ratios on SF samples across all 18 models: $D_2 = \max_{v \in S_{c^*}} P(v)/P_{\text{mass}}(c^*)$ and $D_3 = P(\text{greedy})/P_{\text{mass}}(c^*)$. Within Llama Instruct, D_2 grows monotonically (1B 0.76 $\rightarrow$ 70B 0.99) while Llama Base plateaus (0.66 $\rightarrow$ 0.81); Qwen2.5-72B shows the same contrast (Inst 1.00 vs Base 0.76). D_3 follows the same pattern: Inst grows monotonically with scale (Llama 1.13$\rightarrow$1.45; Qwen 1.13$\rightarrow$1.65), Base stays flat ($\sim$1.0–1.13). The two ratios separate two failure modes: *fragmentation-driven failures* (low D_2, low D_3) in small or Base models where the correct mass is split across alias tokens (the regime Figure 2 illustrates), and *wrong-attractor failures* (high D_2, high D_3) dominating large Instruct models, where the correct concept has collapsed onto a single alias but a wrong concept's token is even sharper. Both arise from the same

instruction-induced sharpening: weak sharpening leaves correct mass spread; strong sharpening collapses fragmentation but strengthens wrong attractors at least as fast (Appendix E). The dichotomy is directly actionable: replacing greedy argmax with a cluster-argmax that aggregates normalized top-50 mass within concepts at t_c recovers 5.7–6.7% of selection failures in 70B+ Base models but only 1.8–2.4% in the corresponding Instruct models—fragmentation is recoverable at decoding time, wrong-attractor failures are not (Appendix E).

Instruction-induced sharpening therefore acts at three structural levels: at the first token (selection failures, sharper wrong-token probability), across multi-token answers (early commitment to specific phrase continuations), and within the correct concept's alias distribution (D_2 collapse).

4.4. When does the model "know" it is going to fail?

A separate but related observation is what makes $t_c = 1$ in instruction-tuned short-form QA. The prompt format alone is not enough: Base models, given the same short-form prompt, emit filler tokens first and delay t_c to later steps (Appendix Table 8). The Instruct–Base contrast lets us ask whether the front-loading is purely an output-formatting effect or reflects deeper changes in the representation.

Output-level detection. On MCQA, Instruct models reach near-perfect detection AUROC (0.974–0.999) using only $P(\text{correct option})$, while Base models span 0.558–0.748 (Figure 5, right; +0.29 average gap).

Attention to the question. At $t = 1$, Instruct models allocate a higher fraction of last-layer attention to the prompt segment containing the question and the answer-format template (+0.09 average; Figure 5, middle), consistent with attending to where the answer must be produced rather than first emitting filler. A token-level refinement shows the Instruct surplus concentrates on the template tokens immediately preceding the answer position rather than the question content itself—instruction tuning sharpens attention to the commitment position, which is the attention-level counterpart of the front-loading measured by the probes below.

Hidden-state probes (pre-generation). Logistic regression on the last-layer hidden state at $t = 1$, before any token is generated, yields Instruct $>$ Base in all four sizes (+0.08 average; Figure 5, left). The pattern holds across nearly all layers (Appendix J), peaking at mid-layers, ruling out a purely output-formatting explanation. This is the first systematic Instruct–Base comparison for the TBG setting of Kossen et al. (2024): pre-generation probe AUROC is 0.61–0.87 for Instruct versus 0.50–0.63 for Llama Base, and Base models gain substantially more from one generated token (avg pre$\rightarrow$post $\Delta_B = +0.030$ vs. $\Delta_I = +0.005$; Appendix Table 7). Correctness-predictive information is genuinely front-loaded in Instruct models. The pattern extends to 70B+: pre-generation probe AUROC reaches 0.830/0.844 for Qwen-72B/Llama-70B Instruct versus 0.785/0.786 for the corresponding Base models, with the same mid-to-late layer peak (Llama-70B Instruct: 0.857 at the mid layer) and the same front-loading asymmetry—post-generation probes lose AUROC relative to pre-generation in every 70B+ model, with Base losing far more (Llama-70B Base: -0.188) (Appendix Tables 6, 7).

The output-level gap (+0.29) is larger than the hidden-state gap (+0.08): instruction tuning's effect is partly representational (information is in the hidden states) and substantially in the output mapping (the projection amplifies it sharply).

5. Discussion and Limitations

Hallucination as commitment failure. Across 18 models from 0.8B to 72B, 16% to 47% of Instruct hallucinations leave non-trivial mass on the correct concept at the commitment step yet produce a wrong final answer, with the rate rising monotonically with scale. As models scale, more hallucinations come from commitment failures despite the population-level distribution including the correct answer, not from the answer being absent. The within-population finding sharpens this: at matched P_{mass}, failures consistently have lower top-token mass on c^* ($d < 0$ in 100% of 18 models, $|d|$ growing with scale within each Instruct family from 1.34 to 4.30; Appendix D). The structural difference between hallucination and correctness at comparable concept-level mass is whether any single surface form is concentrated enough to win. The empirical driver is instruction-induced sharpening: Instruct models sharpen first-token commitments with scale (0.31 to 0.57), Base models remain flat at $\sim$0.30. The same sharpening produces front-loaded correctness signal (§4.4) and decisive misselection when the committed concept is wrong.

Sharpening operates at multiple granularities. The first-token effect extends both inward and forward. Within multi-token divergences, the second-token entropy after an alias-prefix-aligned bigram falls steeply with scale and instruction tuning, reaching 0.05–0.10 at 70B+ Instruct (Appendix F); by the second token, the model has effectively committed to a specific multi-token continuation, and a wrong continuation is selected with the same residual entropy as a deterministic one. The same sharpening also operates within the correct concept's alias distribution: in 70B+ Instruct, mass on the correct concept has typically collapsed onto a single alias token (Appendix E), removing fragmentation as a recoverable failure mode. Confident hallucination is therefore not a momentary slip at $t = 1$ but the natural endpoint of a sharpening process operating at three structural levels—first-token selection, multi-token phrase commitment, and

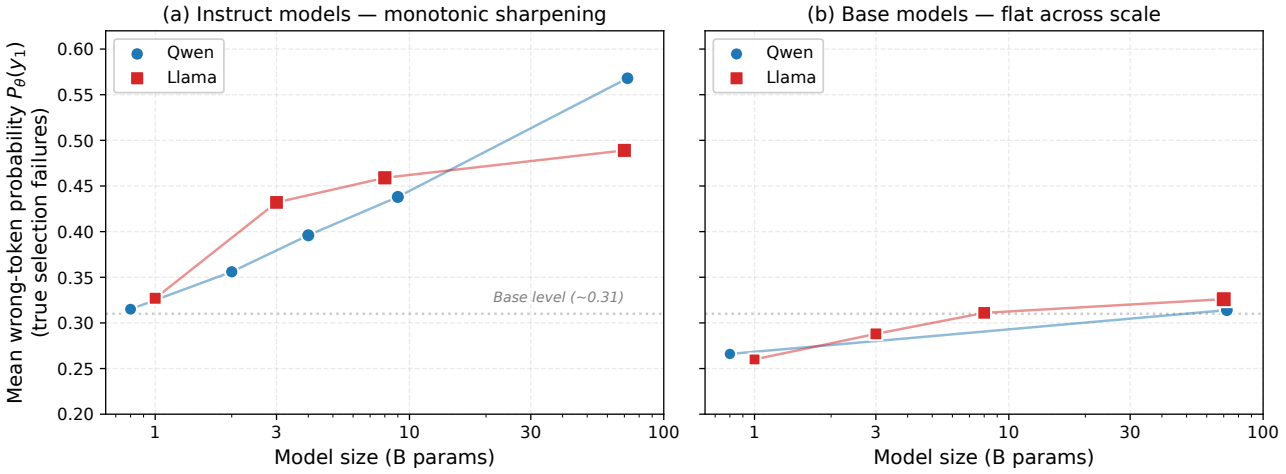

Figure 4. Within first-token selection failures, mean wrong-token probability $P_\theta(y_1)$ across models. *(a) Instruct models*: monotonic sharpening from ∼0.31 at 1B to ∼0.49–0.57 at 70B+, in both families. *(b) Base models*: flat at ∼0.26–0.33 across the same scale range. Marker size ∝ number of true selection failures; the dotted line at 0.31 marks the typical Base level. The contrast is family-independent: instruction tuning—not scale alone—is the driver.

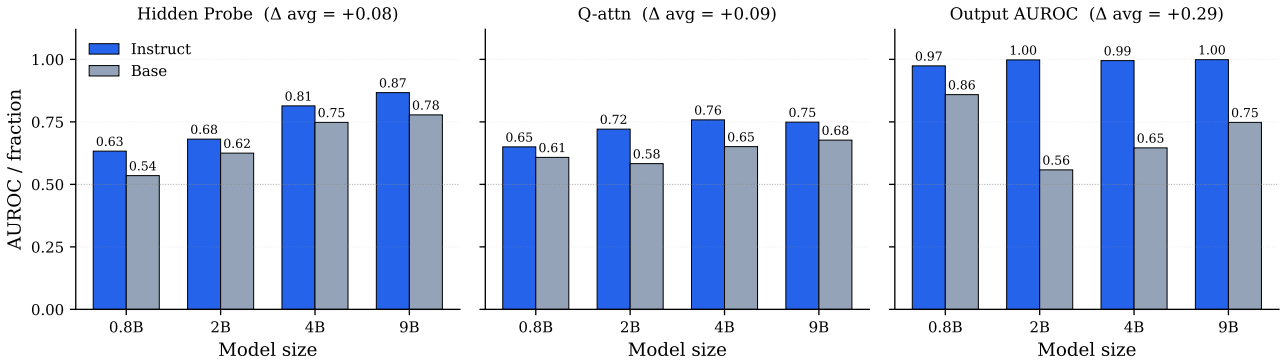

Figure 5. Three-level Instruct–Base comparison at $t = 1$. Hidden Probe: 5-fold CV AUROC on last-layer hidden states (MCQA, N=1,000). Q-attn: fraction of last-layer attention on question tokens (Short-QA, N=500). Output AUROC: P(correct option) on MCQA. Average Instruct–Base gaps: $+0.08$ (Hidden Probe), $+0.09$ (Q-attn), $+0.29$ (Output AUROC).

within-concept alias collapse—more decisive answers when right, more decisive misselections when wrong.

Implications for the picture of confident hallucination. A confident hallucination is one where the model places high probability on a wrong answer's tokens, and the standard framing treats this as evidence that the model has the wrong answer in its distribution and not the right one. Our results complicate this picture: in commitment failures the model has placed substantial mass on the *correct* concept ($P_{\text{mass}} \geq 0.2$) while still emitting a wrong token confidently. Token-level confidence is not coming from the absence of the right concept but from concentration on the wrong concept in spite of it. The same sharpness produces confident correctness when the committed concept is right. "Confidently wrong" and "confidently right" are two outcomes of one distributional disposition, not two different epistemic states—which may explain why uncertainty does not flag confident hallucinations (Simhi et al., 2025; Xu et al., 2025; Farquhar et al., 2024).

Limitations. P_{mass} is an analytical probe, not a deployable detector: it requires the ground-truth alias set S_{c^*} as input, so any practical use depends on inferring S_{c^*} from context (e.g., by clustering top-k tokens at t_c by semantic equivalence). Within multi-token divergences, $P_{\text{mass}}(t = 1; c^*) \geq 0.2$ does not distinguish concept-level belief from phrase-level commitment to specific multi-token continuations (Appendix F); the within-population analysis in §4.3 controls for this by restricting to first-token selection failures. Our experiments use 1–3 token answers and greedy decoding; concept-segmented P_{mass} for longer generation, and the behavior of commitment failures under temperature sampling or top-p truncation, are natural extensions. The analyses also span only Qwen and Llama families; we expect the same instruction-induced sharpening pattern in

other open-weight families, but cannot verify it for closed models without first-token distribution access.

Future work. The commitment-failure phenomenon directly suggests two directions. First, greedy decoding has a structural limitation when $P_{\text{mass}}(c^*)$ is high but split across surface forms: a *concept-aware* decoding rule that argmaxes over alias-clustered top-k tokens—rather than over individual vocabulary entries—would convert a meaningful fraction of selection failures into correct answers without retraining. Quantifying the upper bound and approximating it with semantic-similarity clustering of top-k is a concrete next step. Second, our setup fixes the commitment step at $t = 1$ via short-form prompting; long-form generation has multiple commitment events—domain commitment (`Britain`), answer commitment (`Nicola`), possibly others (rhetorical-frame, sub-claim)—that a richer typology should distinguish. Identifying these reliably is itself a problem: entropy is a noisy localizer (exact match in only 20% of long-form trajectories, §4.1), so non-entropy signals—hidden-state probes, attention concentration on Q, or P_{mass} rate of change $\Delta P_{\text{mass}}(t)$ at each candidate spike—are likely better candidates and worth systematic comparison.

6. Conclusion

We asked what is happening at the moment of hallucination, viewed through the model's distribution at the commitment step. Defining concepts as equivalence classes of token completions and introducing per-step semantic probability mass as an analytical probe, we found that a substantial fraction of Instruct hallucinations are commitment failures: the model puts non-trivial mass on the correct concept yet produces a wrong final answer, with the rate rising monotonically with scale. Larger models do not just know more; they also misfire more often on what they know. Within these failures, the structural difference from matched correct generations is not whether the correct concept is represented, but how its mass is distributed across surface forms. Across scale, the same sharpening pattern operates at the first token, across multi-token continuations, and within the correct concept's alias distribution—uniformly in Instruct models, absent in Base. This reframes confident hallucination as a structural consequence of how mass is shaped at the commitment step, and situates it as one face of the broader alignment tax—suggesting concept-aware decoding and finer commitment-step typologies as natural follow-ups.

References

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Bakman, Y. F., Yaldiz, D. N., Buyukates, B., Tao, C., Dimitriadis, D., and Avestimehr, S. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Calderon, N., Ben-David, E., Gekhman, Z., Ofek, E., and Yona, G. Empty shelves or lost keys? Recall is the bottleneck for parametric factuality. *arXiv preprint arXiv:2602.14080*, 2026.

Chhikara, P. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *Transactions on Machine Learning Research (TMLR)*, 2025. arXiv:2502.11028.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition, 1988.

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

Fisher, R. A. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Hu, T., Minixhofer, B., and Collier, N. Navigating the alignment-calibration trade-off: A Pareto-superior frontier via model merging. *arXiv preprint arXiv:2510.17426*, 2025.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.03341.

Ma, H., Pan, J., Liu, J., Chen, Y., Zhou, J. T., Wang, G., Hu, Q., Wu, H., Zhang, C., and Wang, H. Semantic energy: Detecting LLM hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*, 2025.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Manakul, P., Liusie, A., and Gales, M. J. F. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.

Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1): 50–60, 1947.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

Niu, M., Haddadi, H., and Pang, G. Robust hallucination detection in LLMs via adaptive token selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.07863.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., and Liu, P. J. Out-of-distribution detection and selective generation for conditional language models. *International Conference on Learning Representations (ICLR)*, 2023.

Simhi, A., Itzhak, I., Barez, F., Stanovsky, G., and Belinkov, Y. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14665–14688, 2025. arXiv:2502.12964.

Snel, J. and Oh, S. J. First hallucination tokens are different from conditional ones. *arXiv preprint arXiv:2507.20836*, 2025.

Vashurin, R., Goloburda, M., Ilina, A., Rubashevskii, A., Nakov, P., Shelmanov, A., and Panov, M. CoCoA: A minimum Bayes risk framework bridging confidence and consistency for uncertainty quantification in LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.04964.

Vassoyan, J., Beau, N., and Plaud, R. Ignore the KL penalty! boosting exploration on critical tokens to enhance RL fine-tuning. *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2025. arXiv:2502.06533.

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., Dang, K., Chen, X.-H., Yang, J., Zhang, Z., Liu, Y., Yang, A., Zhao, A., Yue, Y., Song, S., Yu, B., Huang, G., and Lin, J. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.01939.

Welch, B. L. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

Xie, J., Chen, A. S., Lee, Y., Mitchell, E., and Finn, C. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2409.19817.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Xu, H., Yang, Z., Zhu, Z., Lan, K., Wang, Z., Wu, M., Ji, Z., Chen, L., Fung, P., and Yu, K. Delusions of large language models. *arXiv preprint arXiv:2503.06709*, 2025.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhao, Q., Xu, M., Gupta, K., Asthana, A., Zheng, L., and Gould, S. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision (ECCV)*, 2024. arXiv:2403.09037.

A. Theoretical Analysis

We work under a latent-concept generation model: at each step the model implicitly considers candidate concepts $c \in \mathcal{C}$ before emitting tokens, so $P_\theta(y_t \mid Q, y_{<t}) = \sum_{c \in \mathcal{C}} P_\theta(y_t \mid c, Q, y_{<t}) \cdot P_\theta(c \mid Q, y_{<t})$. We refer to $P_\theta(c \mid Q, y_{<t})$ as the model's *concept belief* and to $P_\theta(y_t \mid c, Q, y_{<t})$ as the concept-conditioned emission distribution.

Proposition 1 (P_{mass} as Concept-Belief Proxy)**.** *Let $\gamma_c = P_\theta(y_t \in S_c \mid c, Q, y_{<t})$ (*completeness *of S_c under c) and $\epsilon = \max_{c' \neq c} P_\theta(y_t \in S_c \mid c', Q, y_{<t})$ (*leakage *from competing concepts). With $K = |\mathcal{C}|$,*

$$\left|P_{\text{mass}}(t; c) - P_\theta(c \mid Q, y_{<t})\right| \;\leq\; (1 - \gamma_c) + (K - 1)\epsilon.$$

Proof of Proposition 1. By the law of total probability,

$$P_{\text{mass}}(t) = \sum_{v \in S_c} \sum_{c \in \mathcal{C}} P_\theta(v \mid c, Q, y_{<t}) \cdot P_\theta(c \mid Q, y_{<t}) = \sum_{c \in \mathcal{C}} P_\theta(c \mid Q, y_{<t}) \cdot \alpha_c(t), \tag{2}$$

where $\alpha_c(t) = P_\theta(y_t \in S_c \mid c, Q, y_{<t}) \in [\gamma_c, 1]$ for the target concept and $\alpha_{c'}(t) \in [0, \epsilon]$ for $c' \neq c$. Then

$$P_{\text{mass}}(t) \leq P_\theta(c \mid Q, y_{<t}) + (1 - P_\theta(c \mid Q, y_{<t})) \cdot \epsilon, \tag{3}$$
$$P_{\text{mass}}(t) \geq P_\theta(c \mid Q, y_{<t}) \cdot \gamma_c. \tag{4}$$

Combining gives $|P_{\text{mass}}(t) - P_\theta(c \mid Q, y_{<t})| \leq (1 - \gamma_c) + (K - 1)\epsilon$. □

Proposition 2 (Posterior Concentration at Concept Emission, Auxiliary)**.** *Let t_c denote the first step at which a token from some concept's first-token set is emitted, $y_{t_c} \in S_{\hat{c}}$. Under the latent-concept model with completeness $\gamma_{\hat{c}}$ and leakage ϵ,*

$$P_\theta(\hat{c} \mid Q, y_{\leq t_c}) \;\geq\; \frac{\gamma_{\hat{c}} \cdot P_\theta(\hat{c} \mid Q, y_{<t_c})}{\gamma_{\hat{c}} \cdot P_\theta(\hat{c} \mid Q, y_{<t_c}) + (K - 1)\epsilon}. \tag{5}$$

With $\epsilon = 0$, the posterior concentrates to a delta function on $\hat{c}$, regardless of correctness. This formalizes the deterministic continuation between entropy spikes in Figure 1; it is auxiliary to the empirical analysis in §4.

Proof. Bayes' rule at step t_c gives $P_\theta(\hat{c} \mid Q, y_{\leq t_c}) = P_\theta(y_{t_c} \mid \hat{c}, Q, y_{<t_c}) \cdot P_\theta(\hat{c} \mid Q, y_{<t_c}) / P_\theta(y_{t_c} \mid Q, y_{<t_c})$. The numerator is at least $\gamma_{\hat{c}} \cdot P_\theta(\hat{c} \mid Q, y_{<t_c})$ by completeness. The denominator decomposes via total probability and is bounded above by $P_\theta(y_{t_c} \mid \hat{c}, Q, y_{<t_c}) \cdot P_\theta(\hat{c} \mid Q, y_{<t_c}) + (K - 1)\epsilon$, using $P_\theta(y_{t_c} \mid c', Q, y_{<t_c}) \leq \epsilon$ for $c' \neq \hat{c}$. With $P_\theta(y_{t_c} \mid \hat{c}, Q, y_{<t_c}) \leq 1$ in the denominator, the bound follows. □

B. Statistical Primer

We use a small number of standard statistics throughout this paper; this appendix is a brief reference for readers unfamiliar with them.

Cohen's d (effect size, Cohen, 1988). For two groups with means μ_1, μ_2 and pooled standard deviation σ, $d = (\mu_1 - \mu_2)/\sigma$. It expresses how far apart the two group means are in units of typical within-group spread. Conventional thresholds: $|d| < 0.2$ small, $|d| \approx 0.5$ medium, $|d| > 0.8$ large, $|d| > 1.5$ very large. Effect-size statistics like d complement p-values, which only tell you whether a difference is reliably nonzero, not how big it is.

Welch's t**-test** (Welch, 1947). A two-sample t-test that does not assume equal variances between the two groups; reports a t-statistic and a p-value for the null hypothesis "the two group means are equal."

Mann–Whitney U **test** (Mann & Whitney, 1947). A non-parametric two-sample test on whether one group's values systematically dominate the other's. It does not assume normality. We use it to confirm Welch's t-test results in cases where group distributions are skewed.

Pearson's r. Linear correlation coefficient between two scalar quantities, in $[-1, 1]$.

Fisher's combined p (Fisher, 1925). A way to combine k independent p-values into a single one: $-2\sum_i \ln p_i$ is χ^2-distributed with $2k$ degrees of freedom under a global null. Used here for pooled meta-analysis across models.

Cohen's d **sign convention**. Throughout, when we report negative d between SF and Correct groups, the sign indicates SF $<$ Corr (i.e., the failure group has lower top-token mass on c^*); the magnitude is what matters for the conclusion.

C. Spread-Based Diagnostics

The within-population result reported in §4.3 uses the simple statistic $\max_{v \in S_{c^*}} P_\theta(v)$. We tested an alternative summary statistic, the inverse Simpson diversity index (also called effective number of types):

$$\mathrm{Spread}(c^*; i, t) \;=\; \frac{\big(P_{\mathrm{mass}}(t; c^*)\big)^2}{\sum_{v \in S_{c^*}} P_\theta(v \mid Q_i, y_{<t})^2}.$$

$\mathrm{Spread} = 1$ when all mass is on a single token; $\mathrm{Spread} = k$ when uniformly distributed over k tokens. The two summary statistics measure different things—$\max$ measures the dominant token's probability, Spread measures the evenness of the distribution—and within first-token selection failures, the two give different pictures.

Across-scale comparison (Qwen and Llama Instruct). The table below reports $\mathrm{Spread}(c^*)$ *conditioned on the sample being a first-token selection failure*—i.e., the average over the SF subset, not over all samples. This is the relevant statistic for asking how dispersed the correct concept's mass is when a selection failure occurs.

Model	Avg $\mathrm{Spread}(c^*) \mid \mathrm{SF}$	CF%
Qwen3.5 0.8B Inst	1.91	16.5%
Qwen3.5 2B Inst	1.43	17.4%
Qwen3.5 4B Inst	1.53	26.0%
Qwen3.5 9B Inst	1.39	31.5%
Qwen2.5 72B Inst	1.11	40.7%
Llama-3.2 1B Inst	1.83	14.4%
Llama-3.2 3B Inst	1.30	28.1%
Llama-3.1 8B Inst	1.34	32.7%
Llama-3.1 70B Inst	1.18	47.0%

The SF-conditional Spread decreases with scale in both families: Qwen3.5–Qwen2.5 from 1.91 (0.8B) to 1.11 (72B), Llama-3.2–Llama-3.1 from 1.83 (1B) to 1.18 (70B). At first glance this might seem paradoxical: Spread measures how scattered the correct concept's mass is across alias tokens, and we showed in §4.3 that selection failures are precisely cases where the correct mass is spread out (within-population $d = 2.98$ on top-token mass), so one might expect models that hallucinate more to also have higher Spread. The resolution is that the across-scale decrease in SF-conditional Spread is a *consequence* of wrong-token sharpening, not a cause of it: as the wrong token's probability rises (0.31 to 0.57 across Instruct scale), the correct concept's remaining mass within selection-failure samples is squeezed and necessarily concentrated on fewer effective surface forms. The within-population fragmentation effect (which holds across all 18 models, $|d| \geq 1.0$, Appendix D) is a separate phenomenon from the across-scale Spread trend. Llama Instruct sharpens earlier (the largest drop is from 1B to 3B), Qwen more gradually, but both converge near 1.1–1.2 at the largest scales. Note that this SF-conditional Spread is distinct from the unconditional Spread averaged over all samples (which is more sensitive to the bulk of low-P_{mass} samples), and from the within-population test on $\max_{v \in S_{c^*}} P_\theta(v)$ in Appendix D (which controls for P_{mass}).

Within-population comparison using Spread (Qwen3.5-9B Instruct, $P_{\mathrm{mass}} \geq 0.2$). For completeness, we apply the within-population test of §4.3 but with Spread in place of $\max_{v \in S_{c^*}} P_\theta(v)$:

Group	N	Mean Spread	Median Spread
First-token selection failures	128	1.31	1.10
Correct samples (same P_{mass} range)	840	1.16	1.02

Welch $t = 2.94$, $p = 3.4 \times 10^{-3}$, Cohen's $d = 0.32$. The direction matches expectation (failures have slightly higher Spread than correct samples) and is statistically significant, but the effect size is much smaller than the corresponding test on $\max P_\theta(v)$ ($d = 2.98$ in §4.3). The reason: $\max P_\theta(v)$ is a sharper signal for selection at the argmax level than Spread is—Spread averages over the whole alias distribution and treats, say, "one alias token at 0.5 plus three at 0.05" similarly to "four alias tokens at 0.15 each" (similar Spread), even though only the first wins the argmax against a competing 0.4-probability token. The main-text analysis therefore uses $\max P_\theta(v)$ as the primary measure; we report Spread here for completeness.

Within-belief stratification by Spread **(commitment failures, all** P_{mass}**).** Among hallucinated samples with $P_{\mathrm{mass}} \geq 0.2$ in Qwen3.5-9B Instruct (i.e., commitment failures, including both first-token selection failures and multi-token divergences):

$P_{\mathrm{mass}}(t{=}1)$ band	$\mathrm{Spread} \in [1.0, 1.5]$	$\mathrm{Spread} \in (1.5, 2.5]$	$\mathrm{Spread} > 2.5$
$[0.2, 0.4)$	73.9% (N=199)	**83.3%** (N=36)	76.5% (N=17)
$[0.4, 0.6)$	58.6% (N=174)	62.5% (N=32)	50.0% (N=6)
$[0.6, 0.8)$	43.3% (N=178)	56.5% (N=23)	0% (N=2)
$[0.8, 1.0]$	31.2% (N=756)	29.5% (N=78)	50.0% (N=6)

A directional positive trend appears in three of four bands ($\sim$5–10pp), inconsistent in the highest bin where N is small.

D. Within-Population Effect Across Models

For each model, we replicate the within-population test of §4.3: among samples with $P_{\text{mass}} \geq 0.2$, compare $\max_{v \in S_{c^*}} P_\theta(v)$ between first-token selection failures (greedy $\notin S_{c^*}$, "SF") and correct samples in the same P_{mass} range. Table 2 reports the comparison across all 18 models. The within-population effect ($d < 0$) is uniform: SF samples have lower top-token mass on c^* than correct samples, in 100% of models. Effect sizes range from $|d| = 1.01$ (Qwen3.5-4B Base, $N_{\text{corr}} = 12$) to $|d| = 4.30$ (Qwen2.5-72B Inst), with median $|d| = 1.93$.

Table 2. Within-population effect at $P_{\text{mass}}(t_c; c^*) \geq 0.2$, comparing $\max_{v \in S_{c^*}} P_\theta(v)$ between first-token selection failures (greedy $\notin S_{c^*}$, "SF") and correct samples in the same P_{mass} range. d: Cohen's d on Top1, SF vs. Corr (negative means SF < Corr, the expected direction). p: Welch's t-test p-value. The within-population effect is consistent across all 18 models ($d < 0$ in 100%, $|d| \geq 1.0$, all $p < 10^{-2}$).

Model	N_{SF}	N_{Corr}	d	p
Qwen3.5-0.8B Inst	81	202	-1.34	5.8×10^{-28}
Qwen3.5-2B Inst	97	280	-1.88	3.8×10^{-64}
Qwen3.5-4B Inst	128	703	-2.79	3.8×10^{-199}
Qwen3.5-9B Inst	128	840	-2.98	2.2×10^{-186}
Qwen2.5-72B Inst	114	1,044	-4.30	3.6×10^{-133}
Llama-3.2-1B Inst	84	403	-2.18	1.2×10^{-89}
Llama-3.2-3B Inst	157	904	-2.73	1.9×10^{-233}
Llama-3.1-8B Inst	178	941	-2.80	1.7×10^{-229}
Llama-3.1-70B Inst	170	1,288	-2.97	5.4×10^{-263}
Qwen3.5-0.8B Base	62	34	-1.34	3.8×10^{-6}
Qwen3.5-2B Base	16	11	-1.28	1.6×10^{-2}
Qwen3.5-4B Base	34	12	-1.01	3.3×10^{-2}
Qwen3.5-9B Base	9	12	-1.45	5.2×10^{-3}
Qwen2.5-72B Base	119	988	-1.84	7.2×10^{-175}
Llama-3.2-1B Base	44	305	-1.64	7.2×10^{-38}
Llama-3.2-3B Base	78	728	-2.01	2.7×10^{-86}
Llama-3.1-8B Base	112	885	-1.99	2.1×10^{-126}
Llama-3.1-70B Base	210	1,049	-1.47	8.7×10^{-173}

Effect-size patterns. Two robust patterns emerge: (1) within Instruct models, $|d|$ grows monotonically with scale (Qwen Inst: 1.34→1.88→2.79→2.98→4.30 from 0.8B to 72B; Llama Inst: 2.18→2.73→2.80→2.97 from 1B to 70B); (2) Instruct models show larger $|d|$ than size-matched Base models in nearly every case (e.g., 9B: 2.98 Inst vs. 1.45 Base; 70B+: 2.97–4.30 Inst vs. 1.47–1.84 Base; 0.8B is the one exception, with both at $|d| = 1.34$). Statistical significance is overwhelming throughout: all 18 models give $p < 10^{-2}$, with $p < 10^{-130}$ for the 13 models with $N_{\text{SF}}, N_{\text{Corr}} \geq 100$. The Instruct–Base contrast in correct-sample top-token mass shows that the difference is one of distributional sharpness throughout, not specific to selection failures.

E. Within-Concept and Between-Concept Mass Decomposition (D_2, D_3)

This appendix supports the "two faces of commitment failure" analysis in §4.3. We measure two ratios on first-token selection failure (SF) samples:

$$D_2 = \frac{\max_{v \in S_{c^*}} P_\theta(v \mid Q)}{P_{\text{mass}}(t_c; c^*)} \quad \text{(within-concept top-1 share)} \tag{6}$$

$$D_3 = \frac{P_\theta(\text{greedy})}{P_{\text{mass}}(t_c; c^*)} \quad \text{(wrong-token dominance)} \tag{7}$$

D_2 measures how concentrated the correct concept's mass is on its single most-probable alias token: $D_2 \to 1$ means the mass has effectively collapsed onto one alias; lower D_2 indicates fragmentation across multiple alias forms. D_3 measures the wrong greedy token's mass relative to the correct concept's total: $D_3 > 1$ means the wrong token alone exceeds the entire correct concept.

Two failure modes. Together, D_2 and D_3 separate two structurally different mechanisms of commitment failure:

- *Fragmentation-driven* (low D_2, low D_3): correct mass is spread across alias surface forms (e.g., Figure 2); no single correct alias is large enough to win, but neither is the wrong token strongly dominant.
- *Wrong-attractor-driven* (high D_2, high D_3): correct mass has collapsed onto a single alias, but a wrong concept's token is even sharper.

Per-model statistics. Table 3 reports median D_2, median D_3, and the fraction of SF samples with $D_2 \geq 0.95$ (correct mass effectively collapsed) across all 18 models.

Table 3. Within- and between-concept mass ratios for SF samples across 18 models. D_2 measures how much of the correct concept's mass is in its top-1 alias token. D_3 measures the wrong greedy token's mass relative to the entire correct concept. Within Instruct models, D_2 grows monotonically with scale and the high-D_2 fraction climbs from 28% (1B) to 82% (72B); D_3 similarly grows from 1.13 to 1.65. Within Base models, both quantities stay flat. Sharpening collapses fragmentation in Instruct but strengthens wrong attractors in parallel.

Family	Variant	Size	N_{SF}	median D_2	median D_3	frac $D_2 \geq 0.95$
Qwen	Inst	0.8B	82	0.843	1.13	35.4%
		2B	104	0.941	1.21	45.2%
		4B	136	0.934	1.25	46.3%
		9B	127	0.976	1.42	57.5%
		72B	114	**0.999**	**1.65**	**81.6%**
Qwen	Base	0.8B	61	0.736	1.08	29.5%
		2B	17	0.935	1.14	41.2%
		4B	42	0.948	1.77	47.6%
		9B	13	0.940	1.55	46.2%
		72B	119	0.756	1.11	29.4%
Llama	Inst	1B	98	0.759	1.13	28.6%
		3B	164	0.982	1.28	61.6%
		8B	178	0.966	1.41	55.1%
		70B	170	**0.990**	**1.45**	**66.5%**
Llama	Base	1B	44	0.659	0.99	11.4%
		3B	81	0.682	1.00	16.0%
		8B	124	0.766	1.05	26.6%
		70B	210	0.812	1.13	31.9%

Patterns. Within Llama, the contrast is clean: Instruct D_2 grows monotonically ($0.76 \to 0.98 \to 0.97 \to 0.99$) while Base plateaus ($0.66 \to 0.68 \to 0.77 \to 0.81$), widening the Inst–Base gap from $+0.10$ at 1B to $+0.18$ at 70B. The $D_2 \geq 0.95$ fraction climbs from 28.6% (Llama-1B Inst) to 66.5% (Llama-70B Inst), while Llama Base stays in 11.4%–31.9% across the entire range. Qwen2.5-72B shows the largest absolute Inst–Base gap ($+0.24$, Inst 1.00 vs Base 0.76) and confirms the same pattern at the largest scale; for Qwen3.5 Base in the 2B–9B range, N_{SF} is small (13–42), making medians noisy estimates. D_3 tracks the same trend: monotonic growth in Instruct (Llama 1.13→1.45; Qwen 1.13→1.65) and flat in Llama Base (0.99–1.13).

Connection to within-population $|d|$. The within-population effect ($d = 2.98$ on Qwen-9B Inst, §4.3) measures the absolute top-1 alias probability difference between SF and matched-correct samples (0.26 vs. 0.78). D_2 normalizes this top-1 by $P_{\text{mass}}(c^*)$ to isolate within-concept structure, revealing that at Qwen-9B Inst the SF samples' top-1 already accounts for $\sim$98% of the correct concept's mass: the SF-Corr top-1 gap arises mostly from P_{mass} differences (SF samples have $P_{\text{mass}} \approx 0.26$, correct samples ≈ 0.79), not from fragmentation within the correct concept. This refines the §4.3 narrative: in mid-to-large Instruct models, the structural difference between SF and matched-correct is the absolute mass on c^* (and therefore on its top alias), with within-concept fragmentation playing a secondary role.

Connection to direct decoding intervention. Replacing greedy argmax with cluster-argmax over normalized top-50 tokens at $t = t_c$ recovers 5.7% of SF samples in Llama-70B Base and 6.7% in Qwen-72B Base, but only 2.4% and 1.8% in the corresponding Instruct models. The recovery rate is anti-correlated with D_2: when $D_2 \to 1$ there is no within-concept aggregation to do at decoding time. This confirms the mechanism dichotomy and supports the future-work direction (§5) of concept-aware decoding for non-Instruct or smaller models, where fragmentation is recoverable.

F. Phrase-Level Commitment: $H_{t=2}$ Analysis on Multi-Token Divergences

This appendix supports the phrase-level sharpening claim in §4.3. We restrict to multi-token divergences—commitment failures where greedy $y_1 \in S_{c^*}$ but the final answer is wrong—and use the entropy of the next-token distribution $H_{t=2}$, conditioned on the emitted y_1, to probe whether commitment is finalized at $t = 1$ (low $H_{t=2}$, deterministic continuation) or distributed across multiple tokens (high $H_{t=2}$, multiple candidate continuations).

Type A vs. Type B classification. For each multi-token divergence sample, we check whether the emitted bigram (y_1, y_2) matches the start of any ground-truth alias of c^* under the model's tokenizer. *Type A*: bigram matches an alias prefix—the model is on a trajectory consistent with a valid surface form of c^* at the first two tokens. *Type B*: bigram does not match any alias—y_1 lies in S_{c^*} but the second token already diverges from any valid c^* realization. Type A is the signature of phrase-level commitment to a c^*-aligned phrase at $t = 1$.

A subtlety: Type A samples are by definition still hallucinations (final substring match against ground-truth aliases failed), so they are cases where the model's bigram aligns with an alias prefix but the multi-token completion still diverges into a different entity than c^*. Common patterns include sharing a personal name with an unrelated figure (`George Washington Carver` when the answer is `George Washington`), sharing a place-name prefix with a different geographic entity (`Saint Petersburg Beach` when the answer is `Saint Petersburg`), or sharing an entity prefix with a related-but-distinct concept (`New York Times` when the answer is `New York`). In each case, the bigram is consistent with a valid alias of c^* but the continuation commits to a wrong concept. By contrast, `Adam Lambert` for `Adam Smith` is Type B—the bigram `Adam Lambert` matches no alias of `Adam Smith`.

Per-model results. Table 4 reports Type A fraction and $H_{t=2}$ statistics across 18 models. Two patterns are robust:

1. **Type A has substantially lower $H_{t=2}$ than Type B**, in 100% of models (median Cohen's $d = 1.29$, range $[0.71, 1.87]$). When the bigram aligns with an alias prefix, the continuation is near-deterministic.
2. **Type A fraction grows with both scale and instruction tuning.** Small Base: 21–53%. Small Instruct: 36–72%. Large Base (70B+): 59–74%. Large Instruct (70B+): **77–83%**.

Table 4. Multi-token divergence diagnostic across 18 models. N_{Multi}: number of multi-token divergences. Type A frac.: fraction where the bigram (y_1, y_2) matches a valid alias prefix. $H_{t=2}$ A / B: mean entropy at $t = 2$ for Type A / Type B divergences. d_{B-A}: Cohen's d comparing $H_{t=2}$ Type B vs. Type A. The d_{B-A} for Qwen3.5-9B Base is undefined because $N = 4$ Type B samples is too small.

Model	N_{Multi}	Type A frac.	$H_{t=2}$ A	$H_{t=2}$ B	d_{B-A}
Qwen3.5-0.8B Inst	372	44%	0.45	3.19	1.87
Qwen3.5-2B Inst	369	36%	0.57	2.63	1.69
Qwen3.5-4B Inst	461	67%	0.35	1.97	1.20
Qwen3.5-9B Inst	539	72%	0.20	1.72	1.29
Qwen2.5-72B Inst	663	77%	0.05	0.43	1.13
Llama-3.2-1B Inst	317	59%	0.61	2.36	1.26
Llama-3.2-3B Inst	419	72%	0.38	1.82	1.14
Llama-3.1-8B Inst	479	64%	0.26	1.32	1.03
Llama-3.1-70B Inst	610	**83%**	**0.10**	0.59	1.31
Qwen3.5-0.8B Base	374	21%	1.40	2.94	1.23
Qwen3.5-2B Base	21	38%	0.82	1.56	0.71
Qwen3.5-4B Base	25	36%	0.43	1.97	1.40
Qwen3.5-9B Base	4	75%	0.43	2.57	–
Qwen2.5-72B Base	706	74%	0.37	2.04	1.87
Llama-3.2-1B Base	435	24%	1.82	3.67	1.32
Llama-3.2-3B Base	508	40%	1.23	2.92	1.33
Llama-3.1-8B Base	435	53%	0.78	2.43	1.53
Llama-3.1-70B Base	531	59%	0.67	2.07	1.48

Pooled meta-analysis. Across all 18 models, the Type B vs. Type A $H_{t=2}$ comparison gives a median Cohen's $d = 1.29$, with $d > 0$ in 100% of models with sufficient N and a Fisher-combined $p < 10^{-200}$. Phrase-level commitment is real and pervasive: when the bigram aligns with an alias prefix (Type A), the continuation is near-deterministic; when it does not (Type B), the second token retains substantial entropy across multiple candidates.

Interpretation. The 70B+ Instruct numbers are striking: Llama-3.1-70B Instruct has $H_{t=2} = 0.10$ on Type A divergences

and Qwen2.5-72B Instruct has $H_{t=2} = 0.05$. These values approach the entropy of deterministic continuations—the model has committed to a specific multi-token phrase already at $t = 1$, with the second token essentially predetermined. This is the phrase-level analog of the first-token sharpening documented in §4.3: instruction-induced sharpening operates not just at the token level but across multi-token phrase commitments, and it grows monotonically with scale.

Caveat on P_{mass} interpretation. Within multi-token divergences, $P_{\text{mass}}(t = 1; c^*) \geq 0.2$ is consistent with two distributional pictures: (i) the model placed mass on c^*-aligned alias prefixes at $t = 1$ as part of a phrase-level commitment to a specific multi-token continuation (Type A); (ii) the model placed mass on alias first tokens that are also the start of competing concepts' phrases (Type B—e.g., generic `Sir`, `Saint` shared across many entities). P_{mass} does not distinguish these, but the within-population analysis in §4.3 (which restricts to first-token selection failures) does not depend on this distinction.

G. Robustness of CF% to the 0.2 Threshold

The 0.2 threshold defining commitment failures is conservative and somewhat arbitrary. Table 5 reports CF% at thresholds $\{0.1, 0.2, 0.3, 0.4\}$ across the scale ablation: the absolute level shifts but the monotonic increase with model size is preserved at all thresholds.

Table 5. CF% across thresholds $\theta \in \{0.1, 0.2, 0.3, 0.4\}$ defining commitment failures as hallucinated samples with $P_{\text{mass}}(t_c; c^*) \geq \theta$. The absolute level shifts with threshold, but the monotonic increase with model size is preserved at every column.

Family	Model	$\theta = 0.1$	$\theta = 0.2$	$\theta = 0.3$	$\theta = 0.4$
Qwen3.5	0.8B Inst	28.2%	16.5%	10.1%	6.5%
	2B Inst	28.9%	17.4%	12.3%	9.4%
	4B Inst	37.2%	26.0%	20.1%	16.7%
	9B Inst	41.7%	31.5%	26.0%	22.5%
Llama 3.2/3.1	1B Inst	26.8%	15.7%	10.4%	6.9%
	3B Inst	37.9%	28.2%	22.4%	17.8%
	8B Inst	41.2%	32.6%	27.1%	22.9%
Qwen2.5	72B Inst	44.3%	40.7%	38.1%	35.6%
	72B Base	54.9%	41.8%	34.4%	27.4%
Llama 3.1	70B Inst	53.6%	47.0%	42.0%	37.4%
	70B Base	55.7%	39.5%	27.0%	20.2%

H. Long-Form Generation: P_{mass} Trajectories

This appendix supports the discussion in §4.1 on how P_{mass} behaves when the commitment step is not at $t = 1$. We use Qwen3.5-9B Instruct on TriviaQA + NQ-Open and re-prompt each question with a long-form instruction (`Answer the following question in a complete sentence.`). The model now produces an answer like `The capital of France is Paris.` rather than `Paris`. The commitment step t_c is identified manually as the position of the answer entity within the generated sentence (e.g., `Paris` in the example above), and we align trajectories around t_c. We pre-screen samples to those whose long-form output contains a unique unambiguous reference to a single concept (correct or wrong) to make t_c well-defined.

Figure 6 shows the resulting $P_{\text{mass}}(t; c^*)$ trajectories. Correct samples have a sharp peak in P_{mass} at t_c (typically 0.6–0.9) and near-zero mass before and after, consistent with P_{mass} measuring the model's commitment to the correct concept at the moment of emission. Hallucinated samples sit near zero at all aligned steps—the model never put substantial mass on c^* in the trajectory. This is the long-form analog of "not-CF" hallucinations: cases where the model truly does not know the answer, distinct from CF/SF samples in short-form QA. Long-form P_{mass}-trajectories under the right prompting can therefore separate "model never had it" from "model had it but emitted something else," but the cleaner signal in our paper comes from instruction-tuned short-form QA where $t_c = 1$ is fixed.

For comparison, Figure 7 shows the analogous trajectory of the generated-token probability $P(y_t)$. Both correct and hallucinated trajectories are at ∼0.85–0.95 throughout, with only a small dip at t_c. Token-level confidence carries little of the correct/hallucinated signal that P_{mass} reveals, which is the central methodological point of the paper transferred to the long-form setting: the relevant signal lives at the level of concept-grouped mass, not individual-token entropy.

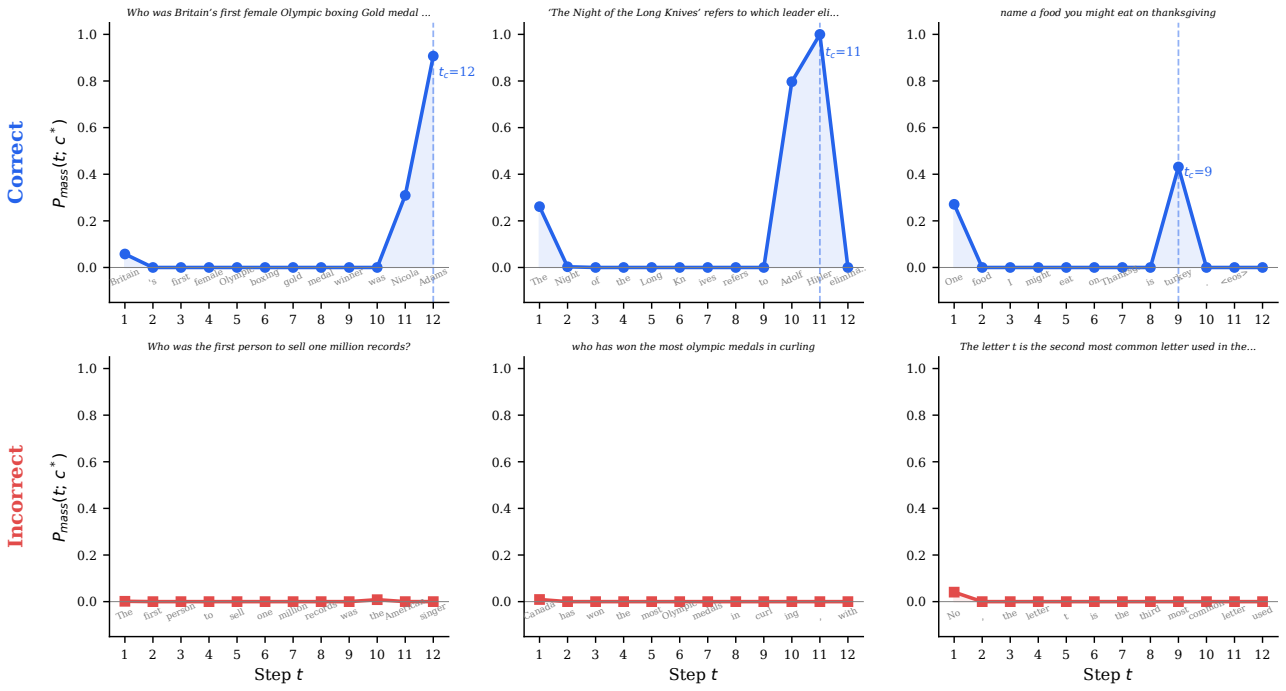

Figure 6. $P_{\text{mass}}(t; c^*)$ trajectories under long-form prompting (`Answer in a complete sentence`), Qwen3.5-9B Instruct. Top: correct samples; bottom: hallucinated. Dashed lines mark t_c (the position of the answer entity within the generated sentence). Correct samples: P_{mass} is near-zero before t_c, peaks at t_c, collapses afterward. Hallucinated samples: essentially no mass on c^* at any step.

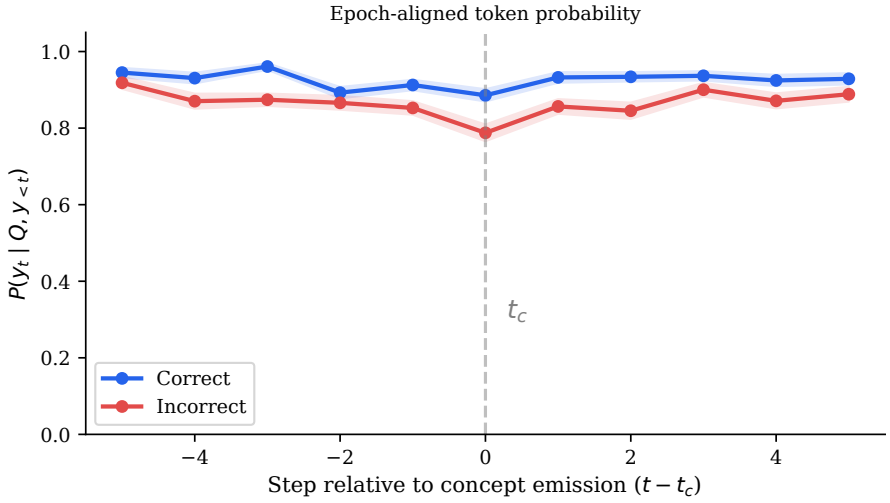

Figure 7. Generated-token probability $P(y_t)$ aligned to t_c in long-form generation. Both correct and hallucinated trajectories sit at ∼0.85–0.95 throughout, with only a small dip at t_c (∼0.78 vs. 0.89). Token-level confidence carries little of the correct/hallucinated signal that P_{mass} reveals (cf. Figure 3b).

I. S_c Construction

S_c is constructed deterministically with no LLM involvement:

- **Alias collection.** TriviaQA: `answer.value`, `answer.aliases`, `answer.normalized_aliases`. NQ-Open: all entries in `answer`. Typically 5–20 aliases per question.
- **Lexical variants.** For each alias a, six variants: original, lowercase, capitalized, and the same three with a leading space. We exclude `upper()` variants (single capital letters appear in many unrelated S_c) and newline-prefixed variants (the `\n` token appears in all S_c).
- **First-token extraction.** Each variant is tokenized with `add_special_tokens=False`; the first token ID is added to S_c.
- **Deduplication.** Final $|S_c|$ is typically 12–20.

J. Multi-Layer Probing

Table 6. Multi-layer probe AUROC (MCQA, logistic regression, 5-fold CV). Sub-10B rows are Qwen3.5; 70B+ rows extend the sweep to Qwen2.5-72B and Llama-3.1-70B, where the mid-to-late layer peak persists.

Model	First	Mid	Last−1	Last
0.8B Inst	0.565	0.624	**0.697**	0.637
0.8B Base	0.521	0.503	0.532	0.602
2B Inst	0.601	**0.718**	0.709	0.716
2B Base	0.513	0.553	**0.651**	0.691
4B Inst	0.690	**0.838**	0.827	0.798
4B Base	0.653	0.742	**0.774**	0.750
9B Inst	0.736	0.848	**0.867**	0.835
9B Base	0.726	0.747	0.776	**0.844**
72B Inst (Qwen2.5)	0.652	0.688	0.805	**0.830**
72B Base (Qwen2.5)	0.702	0.670	**0.793**	0.785
70B Inst (Llama-3.1)	0.613	**0.857**	0.850	0.844
70B Base (Llama-3.1)	0.532	**0.814**	0.803	0.786

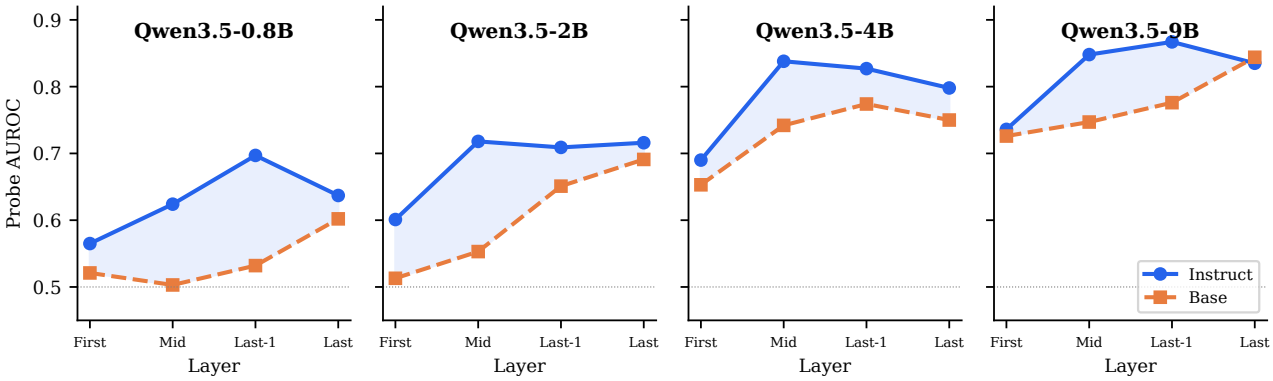

Figure 8. Layer-wise probe AUROC (Qwen3.5). Instruct (blue) > Base (orange) at nearly every layer; gap peaks at mid-layers.

The first-layer Instruct > Base difference (+0.01 to +0.04) rules out a purely output-formatting explanation. Mid-layer peaks ($\approx +0.06$ to +0.12 gap) indicate that instruction tuning amplifies correctness encoding most strongly in intermediate representations.

K. Pre- vs. Post-Generation Probes

Table 7. Pre-gen vs. post-gen probe AUROC (MCQA, last layer, N=1,000). At 70B+, post-generation probes lose AUROC in every model, with Base losing far more—the front-loading asymmetry strengthens with scale.

	Instruct			Base		
Size	Pre	Post	Δ_I	Pre	Post	Δ_B
0.8B	0.637	0.626	−0.011	0.602	0.610	+0.008
2B	0.716	0.730	+0.015	0.691	**0.771**	**+0.080**
4B	0.798	0.819	+0.021	0.750	0.784	+0.034
9B	0.835	0.830	−0.005	0.844	0.843	−0.001
72B (Qwen2.5)	0.830	0.828	−0.001	0.785	0.752	−0.033
70B (Llama-3.1)	0.844	0.764	−0.080	0.786	0.597	−0.188
Avg (sub-10B)			+0.005			+0.030

L. Extended Tables

Table 8. First-token behavior at $t = 1$ (Short-QA). Instruct: $t_c = 1$; Base: $t_c \geq 2$ due to a leading filler.

Model	Top-1 token y_1	$P_\theta(y_1 \mid Q)$	t_c
0.8B Inst	answer / `The`	30–83%	1
2B Inst	answer / `The`	39–60%	1
4B Inst	answer	57–83%	1
9B Inst	answer	52–82%	1
0.8B Base	`\n\n`	24–84%	$\geq$1
2B Base	`\n\n`	93–98%	$\geq$2
4B Base	`\n\n`	96–99%	$\geq$2
9B Base	`\n\n`	45–66%	$\geq$1

Table 9. Full scale ablation (Short-QA, N=3,000). TokP / P_{mass}: AUROC for generated-token probability and $P_{\text{mass}}(t{=}1)$. Spread: average $\text{Spread}(c^*)$. CF%: commitment-failure rate.

Family	Model	Acc	TokP	P_{mass}	Spread	CF%
Qwen3.5	0.8B Inst	8.3%	.759	.898	1.91	16%
	2B Inst	11.0%	.744	.893	1.43	17%
	4B Inst	24.6%	.832	.912	1.53	26%
	9B Inst	29.4%	.806	.887	1.39	32%
	0.8B Base	1.8%	.524	.902	1.71	15%
	2B Base	1.0%	.307	.881	1.46	1%
	4B Base	2.0%	.296	.833	1.40	2%
	9B Base	3.0%	.318	.795	1.21	0.4%
Llama 3.2/3.1 Inst	1B Inst	14.8%	.735	.935	1.83	16%
	3B Inst	31.4%	.776	.902	1.30	28%
	8B Inst	32.8%	.740	.882	1.34	33%
Llama 3.2/3.1 Base	1B Base	13.1%	.755	.887	2.09	18%
	3B Base	26.8%	.781	.880	2.05	27%
	8B Base	31.7%	.812	.899	1.73	27%
Qwen2.5	72B Inst	36.4%	.705	.830	1.11	41%
	72B Base	34.3%	.776	.848	1.78	42%
Llama 3.1	70B Inst	44.7%	.719	.816	1.18	**47%**
	70B Base	37.5%	.745	.845	1.76	40%

All 18 models (14 small + 4 large) now have full metric coverage; CF% values are verified against the data dump in §4.2.

Table 10. $P_{\text{mass}}(t = 1)$ vs. generated-token probability (AUROC) with calibration, across all 18 models. P_{mass} recovers a coherent confidence signal in every model; this is the prerequisite (not the headline) for the commitment-failure analysis. At 70B+ the Instruct–Base calibration gap widens to a 2.5–3.2× ECE ratio while detection AUROC slightly favors Base—the calibration-side signature of instruction-induced sharpening.

Family	Model	Acc	TokProb	P_{mass}	Δ	ECE	Brier
Llama Inst	1B	14.8%	.735	**.935**	+.200	.023	.070
	3B	31.4%	.776	**.902**	+.126	.065	.124
	8B	32.8%	.740	**.882**	+.142	.080	.144
Qwen Inst	0.8B	8.3%	.759	**.898**	+.139	.048	.064
	2B	11.0%	.744	**.893**	+.149	.055	.084
	4B	24.6%	.832	**.912**	+.080	.079	.116
	9B	29.4%	.806	**.887**	+.081	.096	.144
Llama Base	1B	13.1%	.755	**.887**	+.132	.024	.085
	3B	26.8%	.781	**.880**	+.099	.037	.125
	8B	31.7%	.812	**.899**	+.087	.046	.122
Qwen Base	0.8B	1.8%	.524	**.902**	+.378	.045	.026
	2B	1.0%	.307	**.881**	+.574	.006	.009
	4B	2.0%	.296	**.833**	+.536	.011	.018
	9B	3.0%	.318	**.795**	+.476	.020	.025
70B+ Inst	Qwen2.5-72B	36.4%	.705	**.830**	+.126	.199	.216
	Llama-3.1-70B	44.7%	.719	**.816**	+.097	.157	.200
70B+ Base	Qwen2.5-72B	34.3%	.776	**.847**	+.071	.063	.157
	Llama-3.1-70B	37.5%	.745	**.845**	+.099	.064	.160

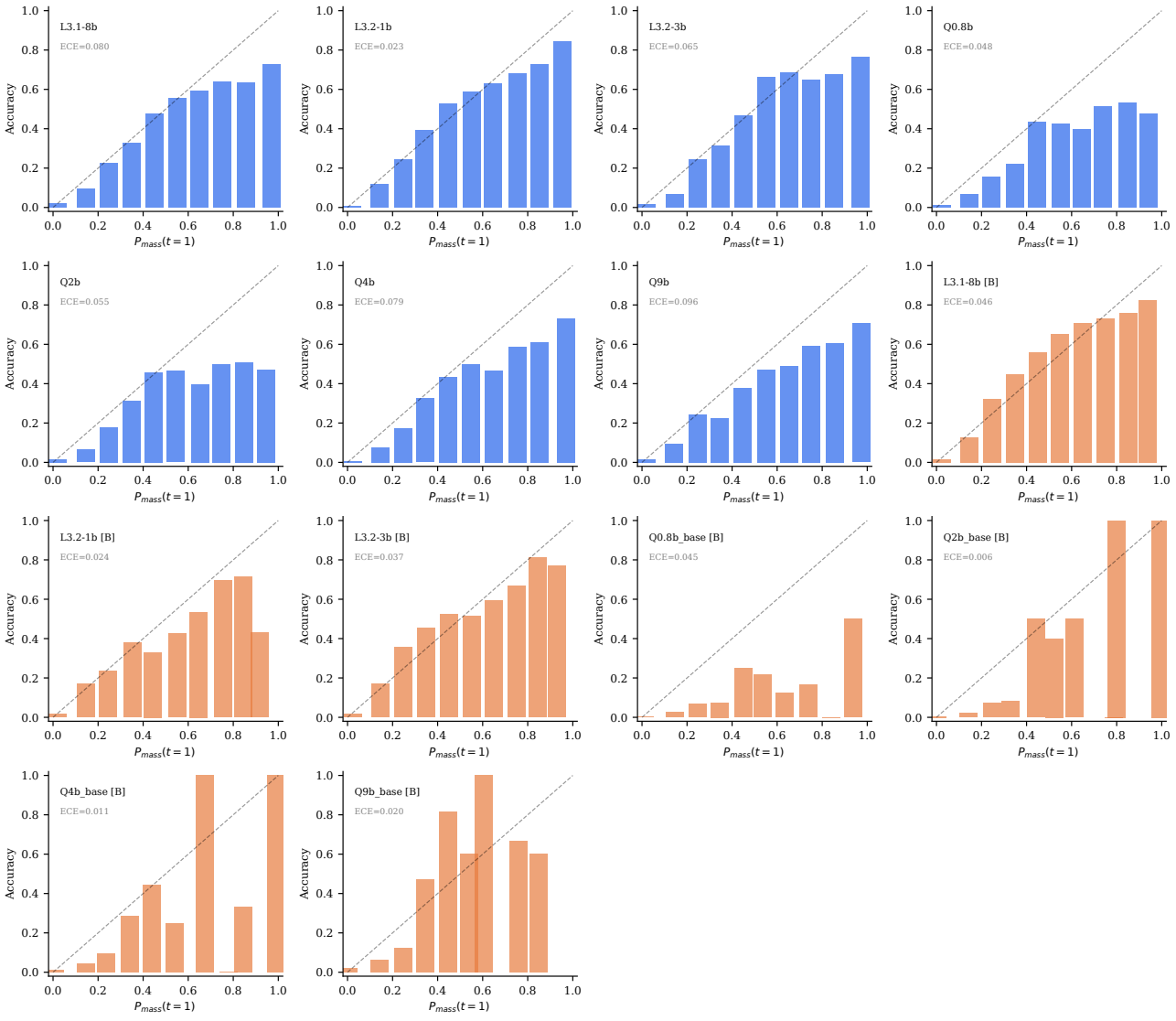

Figure 9. $P_{\text{mass}}(t = 1)$ calibration for all 14 models. Accuracy increases monotonically across P_{mass} bins (Instruct ECE 0.023–0.096).

M. Aggregation: Probability Space vs. Log Space

A natural question is how P_{mass} relates to standard sequence-level uncertainty estimators that aggregate per-step quantities across the full trajectory, typically in log space (mean log-probability, length-normalized NLL). For multi-token concepts these correspond to two different ways of forming a per-sequence score:

- *Probability-space first-step:* $P_{\text{mass}}(t = 1; c^*)$, our default.
- *Log-space full sequence:* $\frac{1}{T}\sum_{t=1}^{T} \log P_\theta(y_t \mid Q, y_{<t})$, the standard length-normalized log-likelihood (LN-NLL).

These differ in two ways: probability vs. log space, and single-step vs. trajectory-averaged. The two differences combine to give a sharp empirical contrast. On Qwen3.5-9B Instruct Short-QA (N=3,000), $P_{\text{mass}}(t = 1; c^*)$ achieves AUROC 0.887 while LN-NLL achieves 0.806 (the same gap as P_{mass} vs. TokProb in Table 10). Decomposing the gap by changing one factor at a time:

Estimator	AUROC
$P_{\text{mass}}(t = 1; c^*)$ (prob-space, single-step)	0.887
Average of per-step P_{mass} across full sequence (prob-space, averaged)	0.738
$\log P_\theta(y_1 \mid Q)$ (log-space, single-step)	0.812
LN-NLL (log-space, full sequence)	0.806

Two observations. First, the single-step $\rightarrow$ full-sequence drop is large in probability space (0.887 $\rightarrow$ 0.738) but small in log space (0.812 $\rightarrow$ 0.806): aggregating downstream tokens dilutes the probability-space signal because most steps are deterministic continuations whose P_{mass} is essentially zero. Second, at the commitment step itself, probability space (P_{mass}) outperforms log space ($\log P$) by 0.075 AUROC, because the relevant quantity at the commitment step is the total mass on the correct concept's surface forms—which is additive in probability, not in log-probability. The same picture holds for the relationship between greedy token probability and its log form: the probability is what is being compared by the argmax, so it is the natural quantity to inspect at the moment of commitment. We use probability space throughout the paper for this reason.

N. Compute Resources

All experiments were run on NVIDIA GPUs. We did not perform any model fine-tuning—all compute consists of forward passes only. Per-model wall-clock cost scales with model size:

- **Sub-10B models** (Qwen3.5-0.8B/2B/4B/9B, Llama-3.2-1B/3B, Llama-3.1-8B): single NVIDIA A100 (80GB), fp16, batch size 8; 0.5–2 hours per (model, dataset) combination.
- **72B/70B models** (Qwen2.5-72B, Llama-3.1-70B): single NVIDIA B200 (180GB), fp16, batch size 8; 4–8 hours per (model, dataset) combination. The B200's larger memory accommodates the 70B+ models without tensor parallelism.

The full 18-model evaluation across TriviaQA + NQ-Open (3,000 samples per dataset, top-50 token probabilities saved at the commitment step) totals approximately 53,000 forward passes. The Phase 2 within-concept and between-concept ratio analysis (D_2, D_3 in Appendix E) is a pure offline post-processing of the saved top-50 probabilities and required no additional GPU compute.

The probing experiments (Appendix J) use logistic regression on saved hidden states; each probe trains in under one minute on CPU.

O. Broader Impacts

This paper provides an analytical characterization of LLM hallucination structure: when a model places non-trivial mass on the correct concept yet emits a wrong final answer. The work introduces no new models, datasets, or deployable methods; the contribution is a probe and an empirical characterization. Potential positive societal impact: a clearer mechanistic account of confident hallucination may inform safer deployment and calibration practices, particularly in high-stakes settings where users treat fluency as evidence of reliability. We do not foresee direct negative societal impact from the analysis itself, beyond the general concern that deeper understanding of model failures could in principle inform adversarial exploitation; however, the analytical probe itself is not a generation or control method and does not enable any new attack surface.