THE ALIGNMENT PARADOX: HOW POST-TRAINING AMPLIFIES CONFIDENT HALLUCINATIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can produce factually incorrect answers with high confidence, undermining their reliability and limiting the effectiveness of uncertainty-based error detection. While prior research attributes confident hallucinations to factors such as missing knowledge in training data, reasoning errors, or stochastic decoding, we uncover that post-training alignment itself is a primary driver of these errors, a phenomenon we call the **Alignment Paradox**. Across five model families evaluated on factual benchmarks, unaligned base models produce few high-confidence errors on long-tail factual queries, whereas instruction-tuned models multiply high-confidence errors ($p \geq 0.95$) by more than an order of magnitude ($10\times$ to $35\times$). Layer-wise probing with the Logit Lens reveals that this overconfidence emerges in late layers, where wrong-answer margins expand past 4.0 points after remaining near zero across early and intermediate layers. These findings motivate limiting margin growth during post-training. We implement this principle through an entropy-dependent margin bound in direct preference optimization (DPO). In multi-epoch experiments with Mistral-7B, the bounded objective reduces high-confidence errors by up to 35.3% relative to standard DPO while maintaining performance on evaluated general reasoning benchmarks. These results show that bounded margins mitigate confident hallucinations during post-training.

1 INTRODUCTION

When large language models produce errors with low confidence ($p \approx 0.3$–0.5), uncertainty-based error detection and selective prediction (Geifman & El-Yaniv, 2017) can reliably flag or defer the query. A critical vulnerability arises when models produce incorrect assertions with high confidence ($p \geq 0.95$). Such confident hallucinations evade threshold-based abstention, bypass automated verification, and undermine decision systems where abstention is preferable to erroneous completions (Amodei et al., 2016; Kadavath et al., 2022).

While prior research attributes factual hallucinations to pre-training knowledge gaps, multi-step reasoning errors, or decoding noise (Kadavath et al., 2022; Dziri et al., 2023; Chuang et al., 2024; Orgad et al., 2025), post-training alignment, comprising supervised fine-tuning (SFT; Ouyang et al. 2022) and preference optimization (Bai et al., 2022; Rafailov et al., 2023), is widely assumed to improve truthfulness and safety (Touvron et al., 2023; Dubey et al., 2024; Yang et al., 2024). In this work, we uncover that post-training alignment itself is a primary driver of high-confidence errors, a phenomenon we term the **Alignment Paradox**[1]. Although alignment enhances conversational adherence, it degrades confidence reliability on factual queries, shifting incorrect predictions into extreme confidence regimes ($p \geq 0.95$).

Empirical evaluation demonstrates this phenomenon across standard factual question-answering benchmarks. As illustrated in Figure 1a, when an unaligned base model encounters an entity absent from its parametric memory, its output distribution remains diffuse across competing candidates, reflecting epistemic uncertainty. In contrast, post-trained checkpoints exhibit pronounced probability

[1]Code and evaluation data are accessible at https://anonymous.4open.science/r/HCE-1B12.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

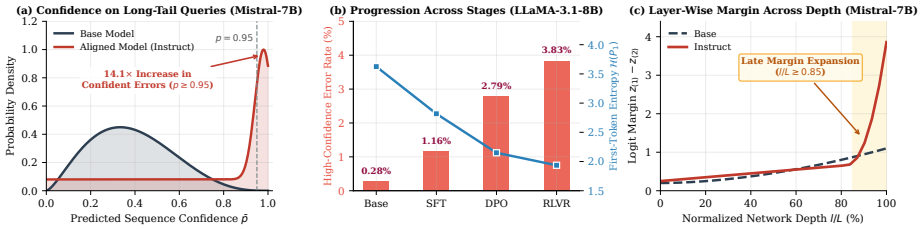


Figure 1: **The Alignment Paradox.** Post-training alignment shifts output probabilities into high-confidence regimes and multiplies near-certain errors on factual queries. (a) Sequence confidence distributions on factual queries (`Mistral-7B`). (b) Progression of high-confidence error rates ($p \geq 0.95$) and first-token decision entropy across alignment stages in the `LLaMA-3.1-8B` lineage (Base $\rightarrow$ SFT $\rightarrow$ DPO $\rightarrow$ RLVR). (c) Layer-wise logit margin across normalized depth, showing logit margin expansion in late layers ($l/L \geq 0.85$, `Mistral-7B`).

concentration, assigning probabilities exceeding 0.95 even to false entities. Across five matched base and aligned model pairs spanning 7B to 14B parameters in the Mistral, Qwen, and LLaMA lineages over 22,267 factual queries, post-training alignment multiplies high-confidence errors ($p \geq 0.95$) by $10.0\times$ to $35.4\times$, worsening Expected Calibration Error (ECE; Guo et al. 2017) by up to 15.5-fold.

To determine whether this surge reflects conversational chat templates, we evaluate aligned models on bare completion prompts identical to those used for base models. Parameter updates account for the vast majority of the increase, as aligned checkpoints retain a $4.1\times$ to $6.0\times$ elevation in confident errors over base models even in the absence of instruction formatting. Furthermore, tracking sequential checkpoints across post-training reveals a two-stage progression, where supervised fine-tuning contracts decision entropy and preference optimization expands logit margins, driving incorrect outputs into extreme confidence intervals.

To investigate internal mechanisms, we probe intermediate representations using the Logit Lens (Belrose et al., 2023). Our analysis shows that early commitment does not extend to factual errors, as decision distributions remain diffuse across early and intermediate representations. Overconfidence emerges strictly in late layers ($l/L \geq 0.85$), where wrong-answer margins expand past 4.0 points after hovering near zero through earlier layers, sharply suppressing decision entropy.

Inference-time interventions provide limited mitigation. Across five representative test-time defenses (defensive prompting, two-turn self-correction, sampling self-consistency, layer-contrastive decoding with DoLa (Chuang et al., 2024), and entropy thresholding), we find that these heuristics eliminate at most 5% to 8% of high-confidence errors while preserving at least 97% of correct completions. Furthermore, iterative self-correction degrades accuracy on initially correct completions. Because test-time heuristics cannot reverse parameter-level margin expansion, mitigating confident hallucinations requires regularizing the training objective itself.

We address this challenge by establishing the principle of bounded margin optimization for post-training alignment. Grounded in theoretical analysis showing that standard preference objectives drive logit margins toward infinity on ungrounded completions, we demonstrate that constraining margin expansion provides a general remedy against ungrounded overconfidence. We formulate Bounded Dynamic Margin (BDM) optimization, which enforces an adaptive, entropy-dependent ceiling on margin growth. During standard optimization on supported preferences, the margin bound remains inactive; under extended optimization on unsupported completions, it activates to prevent unbounded logit expansion. In multi-epoch evaluations across 22,267 factual queries, instantiating this principle via BDM-DPO reduces high-confidence errors by up to 35.3% relative to standard DPO (eliminating 850 confident errors) while preserving general reasoning capabilities.

In summary, this work presents four primary contributions.

1. **The Alignment Paradox.** We formalize the empirical phenomenon wherein post-training alignment multiplies high-confidence factual errors by $10.0\times$ to $35.4\times$ across five model families while degrading confidence reliability.
2. **Causal Attribution.** Through prompt ablations and checkpoint tracking, we demonstrate that parameter updates rather than chat templates drive this effect, with preference objectives expanding logit margins on erroneous completions.

3. **Representation Dynamics.** Using layer-wise Logit Lens probing, we show that early commitment does not hold for factual errors, demonstrating that overconfidence emerges strictly in late layers ($l/L \geq 0.85$).
4. **Principled Mitigation.** We show that test-time defenses achieve at most an 8.3% error reduction, and establish that constraining margin expansion provides a general training-time remedy. Instantiated as BDM-DPO, this bounded formulation reduces high-confidence errors by up to 35.3% while preserving general reasoning performance.

2 Related Work

Factual Knowledge and Confidence Reliability. Evaluating factual accuracy in large language models spans diverse benchmarks, including TruthfulQA (Lin et al., 2022), PopQA (Mallen et al., 2023), TriviaQA (Joshi et al., 2017), and Natural Questions (Kwiatkowski et al., 2019). Standard evaluation protocols traditionally aggregate all incorrect generations into an undifferentiated error rate, treating an uncertain error ($p \approx 0.35$) identically to an authoritative fabrication ($p \geq 0.95$). In high-stakes settings relying on selective abstention (Geifman & El-Yaniv, 2017; Kamath et al., 2020; Kadavath et al., 2022), these modes differ fundamentally, because uncertain errors trigger human review, whereas near-certain errors invisibly bypass confidence filters. While confidence reliability metrics such as Expected Calibration Error (ECE; Guo et al. 2017) and Brier score (Brier, 1950) are established (Desai & Durrett, 2020; Xiong et al., 2024), prior studies focus on static evaluations, lacking systematic frameworks to track how post-training pipelines alter confidence geometry.

Mechanistic Probing of Factual Recall. Mechanistic interpretability has made rapid progress in tracing factual associations in transformer layers (Meng et al., 2022; Geva et al., 2023). Linear probing and activation steering methods (Azaria & Mitchell, 2023; Burns et al., 2023; Li et al., 2023; Orgad et al., 2025) show that intermediate activations often separate true assertions from falsehoods. Complementary work investigates self-detection heads (Kadavath et al., 2022; Yin et al., 2023), entity familiarity estimation (Ferrando et al., 2025), and layer-contrastive decoding such as DoLa (Chuang et al., 2024). However, these studies typically treat base and aligned models as fixed entities, without tracing the causal transition from pre-training to post-training representations or elucidating how instruction tuning reshapes internal uncertainty manifolds.

Preference Optimization and Alignment Dynamics. Direct preference optimization (DPO; Rafailov et al. 2023) and related variants (IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), SimPO (Meng et al., 2024), and reinforcement learning with verifiable rewards (RLVR; Shao et al. 2024)) have largely supplanted reinforcement learning from human feedback based on PPO (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). While enforcing conversational compliance, their susceptibility to reward over-optimization has attracted scrutiny (Gao et al., 2023; Singhal et al., 2024; Coste et al., 2024). Most investigations into post-training capability trade-offs focus on core reasoning degradation on math or code benchmarks (Ouyang et al., 2022; Dubey et al., 2024). In contrast, our work examines an epistemic dimension, demonstrating how unconstrained preference margins force steep entropy compression on factual queries, converting benign uncertainty into confident hallucinations.

3 Empirical Evidence for the Alignment Paradox

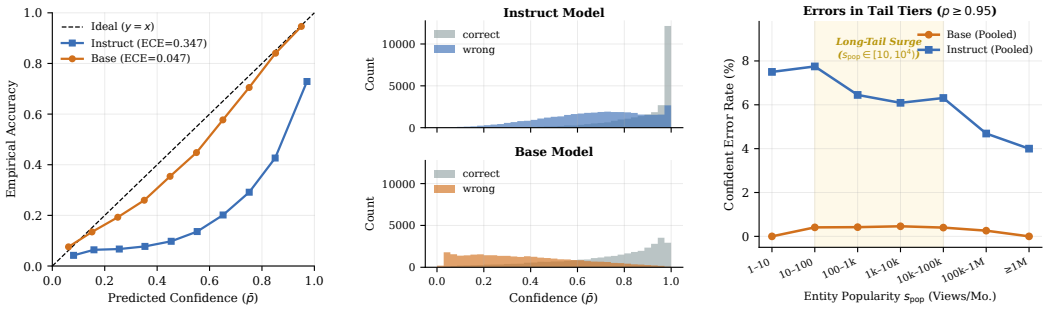
3.1 Evaluation Setup and Reliability Metrics

To evaluate how post-training alignment alters factual certainty, we benchmark five matched base and instruction-tuned model pairs across 22,267 factual queries. The evaluation spans the Mistral, Qwen, and LLaMA lineages across 7B to 14B parameter scales, specifically Mistral-7B (Jiang et al., 2023), Qwen3-8B and Qwen3-14B (Yang et al., 2024), LLaMA-3.1-8B (Dubey et al., 2024), and Mistral-Nemo-12B. We evaluate on PopQA (12,509 open-domain queries; Mallen et al. 2023) and TriviaQA (9,758 unfiltered validation queries; Joshi et al. 2017), complemented by Natural Questions (Kwiatkowski et al., 2019) for broader consistency checks.

We enforce deterministic greedy decoding ($T = 0$) to eliminate sampling variance. Generated strings are evaluated using regular expression word-boundary matching against verified reference aliases

Table 1: **Base versus aligned models across 22,267 queries.** Post-training alignment multiplies high-confidence errors (Gate B) and degrades confidence reliability.

Model Family	Checkpoint	Acc (%)	Gate A	Gate B	Ratio	Gate C	Gate D	ECE
Mistral-7B	Base	52.3	2,378 (10.7%)	97 (0.4%)	**14.1**×	10,525 (47.3%)	9,256 (41.6%)	0.097
	Instruct	41.6	4,133 (18.6%)	1,366 (6.1%)		11,605 (52.1%)	5,125 (23.0%)	0.378
Qwen3-8B	Base	45.9	1,885 (8.5%)	50 (0.2%)	**35.4**×	11,986 (53.9%)	8,318 (37.4%)	0.042
	Instruct	37.9	5,140 (23.1%)	1,768 (7.9%)		11,993 (53.9%)	3,308 (14.9%)	0.400
LLaMA-3.1-8B	Base	55.3	1,406 (6.3%)	38 (0.2%)	**15.7**×	9,909 (44.5%)	10,910 (49.0%)	0.018
	Instruct	45.6	4,996 (22.4%)	596 (2.7%)		11,494 (51.6%)	5,157 (23.2%)	0.262
Mistral-Nemo-12B	Base	57.7	1,690 (7.6%)	56 (0.3%)	**10.0**×	9,367 (42.1%)	11,150 (50.1%)	0.017
	Instruct	47.6	4,766 (21.4%)	560 (2.5%)		11,092 (49.8%)	5,834 (26.2%)	0.264
Qwen3-14B	Base	51.0	2,727 (12.5%)	104 (0.5%)	**19.8**×	10,564 (48.5%)	8,391 (38.5%)	0.066
	Instruct	43.4	6,727 (30.2%)	2,061 (9.3%)		10,451 (46.9%)	2,938 (13.2%)	0.371



(a) Reliability diagram. (b) Confidence distribution ($\bar{p}$). (c) Entity popularity tiers.

Figure 2: **Confidence reliability and distribution dynamics across 22,267 queries.** (a) Reliability diagrams show post-training overconfidence gaps. (b) Sequence confidence distributions ($\bar{p}$) reveal that unaligned base errors remain in the low-confidence regime, whereas post-training alignment drives thousands of errors into the near-certain regime ($p \geq 0.95$). (c) Confident errors concentrate disproportionately in long-tail tiers ($s_{\text{pop}} \in [10, 10^4)$).

from Wikidata (Appendix A; checkpoint specifications appear in Table 6, and double-blind audits confirming over 98.6% hallucination fidelity appear in Table 7). For each query, we track the sequence geometric mean probability $\bar{p}$ and Expected Calibration Error (ECE; Guo et al. 2017) across $M = 10$ bins:

$$\bar{p} = \exp\left(\frac{1}{T}\sum_{t=1}^{T} \ln p(y_t \mid x, y_{<t})\right), \qquad \text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{1}$$

where B_m denotes samples in bin m, and $\text{acc}(B_m)$ and $\text{conf}(B_m)$ denote empirical accuracy and mean confidence. Under a selective prediction threshold $\tau = 0.95$, we stratify outputs into four quadrants, denoted as **Gate A** ($\bar{p} \geq 0.95$, correct), **Gate B** ($\bar{p} \geq 0.95$, incorrect; confident hallucinations), **Gate C** ($\bar{p} < 0.95$, incorrect), and **Gate D** ($\bar{p} < 0.95$, correct).

3.2 Post-Training Alignment Multiplies High-Confidence Errors

Table 1 presents the primary empirical finding. Across every model family, unaligned base models rarely produce high-confidence errors on factual queries; out of 22,267 queries, base models produce only 38 to 104 Gate B errors (0.2% to 0.5%). The vast majority of base errors fall into Gate C, where low probabilities accurately signal lack of knowledge.

Post-training alignment significantly alters this behavior. On identical queries, confident errors expand by 10.0× to 35.4× across all five families, rising to 1,768 in Qwen3-8B (35.4×), 1,366 in Mistral-7B (14.1×), and 2,061 in Qwen3-14B (19.8×). Simultaneously, confidence reliability degrades as instruction-tuned models develop severe overconfidence, driving Expected Calibration Error (ECE) up by up to 15.5-fold (from 0.017 to 0.264 in Mistral-Nemo-12B, and 0.018 to 0.262 in LLaMA-3.1-8B).

Table 2: **Causal disentanglement of chat templates versus model weights.** Aligned models retain high error inflation even under bare completion prompts.

Model Family	Evaluation Condition	Gate B Count	Gate B (%)	Inflation vs. Base
Mistral-7B	C1: Base Few-Shot	63	0.28%	1.0×
	C2: Instruct Chat Template	1,630	7.32%	25.9×
	C3: Instruct Bare Completion	864	3.88%	**13.7**×
LLaMA-3-8B	C1: Base Few-Shot	18	0.08%	1.0×
	C2: Instruct Chat Template	637	2.86%	35.4×
	C3: Instruct Bare Completion	108	0.48%	**6.0**×
Qwen3-8B	C1: Base Few-Shot	92	0.41%	1.0×
	C2: Instruct Chat Template	924	4.15%	10.0×
	C3: Instruct Bare Completion	378	1.70%	**4.1**×

Figure 2 illustrates this multi-faceted degradation across 22,267 queries. In Figure 2a, base models adhere closely to the ideal diagonal ($y = x$), while aligned models show severe overconfidence across predicted probabilities. In Figure 2b, sequence confidence distributions ($\bar{p}$) confirm that unaligned base errors concentrate in low-confidence regions, whereas post-training alignment forces error mass into extreme certainty ($p \geq 0.95$). In Figure 2c, error rates across Wikipedia monthly pageview tiers in PopQA (s_{pop}) reveal that confident errors peak in the mid-to-long tail tiers ($s_{\text{pop}} \in [10, 10^4)$). While base models assign low probabilities to unfamiliar tail entities, aligned models allocate near-certain probabilities ($p \geq 0.95$) to fabricated completions. Disaggregated per-benchmark breakdowns appear in Table 9 (Appendix B.1), and capacity scaling analyses across 7B to 14B regimes appear in Figure 4b (Appendix B.4).

3.3 Persistence of Confident Errors under Decoding Perturbations

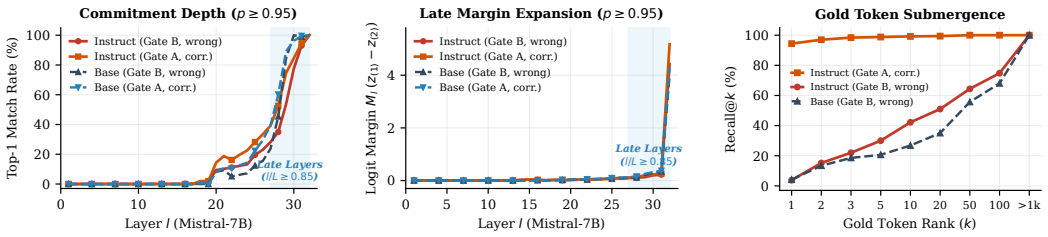
A relevant question is whether Gate B errors represent stochastic sampling noise or stable representational attractors. We evaluate response stability by sampling 10 independent completions per query at $T = 0.7$ with top-$p = 0.9$ across all Gate B instances.

The empirical results show that 68.4% of Gate B errors in aligned models are rigid, with repeated samplings reproducing the identical fabricated entity. This high consistency causes uncertainty detection via semantic entropy (Kuhn et al., 2023) and self-consistency to fail. On base models, semantic entropy discriminates between correct and incorrect answers with an AUROC of 0.78; on instruction-tuned models, semantic entropy AUROC drops to 0.51, barely exceeding random guessing. The model asserts false entities with the same internal agreement as factual truths (temperature sweeps and response stability curves appear in Table 11 and Figure 4a in Appendix B.3).

3.4 Disentangling Model Weights from Formatting Templates

We test whether overconfidence stems from conversational chat templates by comparing three conditions across all 22,267 queries, namely (1) **Base + Few-Shot**, where base models use few-shot prefixes; (2) **Instruct + Chat Template**, where aligned models use official chat templates; and (3) **Instruct + Bare Completion**, where aligned models receive bare completion prefixes identical to Condition 1, stripping all chat wrappers (the complete open-science causal ladder tracing Base $\rightarrow$ SFT $\rightarrow$ DPO $\rightarrow$ RLVR appears in Table 10 and Appendix B.2).

Table 2 shows that stripping chat formatting reduces Gate B errors from Condition 2 levels, but Condition 3 still retains a 4.1× to 13.7× error inflation over base models (e.g., 864 errors on Mistral-7B vs. 63 in base; 6.0× on LLaMA-3-8B). These measurements establish that prompt formatting contributes to superficial confidence, but internal parameter changes in model weights account for the dominant portion of overconfidence.



(a) Commitment depth (l^*). (b) Late margin expansion. (c) Target token suppression.

Figure 3: **Internal representation dynamics on high-confidence completions** (`Mistral-7B`, all cohorts conditioned on terminal $p \geq 0.95$). (a) Commitment depth (l^*). Token predictions stabilize strictly in late layers ($l^*/L \geq 0.85$), showing early commitment does not hold for factual errors. (b) Late-layer margin expansion. Logit margins hover near zero through intermediate layers before expanding sharply in late layers ($l/L \geq 0.85$). (c) Target token suppression. Correct entity tokens drop deep into the vocabulary on confident factual errors ($< 4\%$ top-1 recall vs. $> 94\%$ for correct completions).

4 Internal Representation Dynamics of Confident Errors

4.1 Layer-Wise Probing with the Logit Lens

To trace where overconfidence originates within the transformer architecture, we probe intermediate representations using the Logit Lens (Belrose et al., 2023). For a model with L layers, let $h_t^{(l)} \in \mathbb{R}^d$ denote the hidden activation at layer $l \in \{1, \ldots, L\}$ at output position t. We project $h_t^{(l)}$ onto the vocabulary $\mathcal{V}$ using the final layer normalization RMSNorm and the pre-trained unembedding matrix $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$:

$$p_t^{(l)} = \text{softmax}\left(W_U \cdot \text{RMSNorm}\left(h_t^{(l)}\right)\right) \tag{2}$$

This projection yields a layer-wise probability distribution $p_t^{(l)}$ over vocabulary tokens without training external probes that could introduce auxiliary learning artifacts.

To dissect the representation dynamics of extreme overconfidence, we condition our layer-wise analysis on completions producing high terminal confidence ($p \geq 0.95$). We compare **Gate A** (confident correct completions) and **Gate B** (rigid, confident incorrect completions verified across $k \geq 5$ samplings). In base models, where Gate B errors are exceptionally rare (0.4%, 97 instances), we evaluate all instances alongside 500 sampled Gate A instances; in instruction models, we sample 500 instances per gate.

We evaluate whether early commitment, positing that transformers resolve factual associations in early layers (Geva et al., 2023), holds across factual errors. Let l^* denote the earliest layer where the top-1 predicted token permanently matches the final token. As shown in Figure 3a, across all lineages the commitment depth satisfies $l^*/L \geq 0.85$ (layers 27–31 for 32-layer models). Across the first 80% of layers ($l/L < 0.85$), intermediate representations maintain diffuse distributions and low match rates ($< 25\%$), demonstrating the inapplicability of early commitment to factual retrieval errors (extended layer-wise projections and match statistics across architectures appear in Table 12 and Figure 5 in Appendix C.1).

Because both cohorts are conditioned on reaching terminal $p \geq 0.95$, their commitment trajectories track closely throughout intermediate depths (Figure 3a). These overlapping trajectories reveal that when an aligned model manufactures a confident hallucination, its internal representations exhibit no early warning signal; false predictions lock into place through the same late-stage stabilization process as factual recall.

4.2 Overconfidence Emerges in Late Layers

We examine the layer-wise trajectory of decision entropy $\mathcal{H}(P_l)$ and top-1 logit margin $M_{1,l}$, defined as

$$\mathcal{H}(P_l) = -\sum_{v \in \mathcal{V}} p^{(l)}(v) \ln p^{(l)}(v), \qquad M_{1,l} = z_{(1)}^{(l)} - z_{(2)}^{(l)} \tag{3}$$

Table 3: **Inference-time mitigation on LLaMA-3-8B-Instruct.** Interventions eliminate at most 8.3% of confident errors under $\geq 97\%$ retention.

Inference-Time Defense	Gate B Count	Gate B Reduction	Correct Retention	Effective Trade-off
Standard Instruct Baseline	324	0.0%	100.0%	Baseline
Defensive System Prompting	306	5.6%	65.2%	Severe over-hedging
Iterative Self-Correction	308	4.9%	81.4% (1.9% on LLaMA-3.1)	Corrupts valid answers
Self-Consistency (10 samples)	302	6.8%	97.4%	68.4% rigid errors
DoLa Layer-Contrastive	297	8.3%	97.1%	Late margin limits gain
Entropy Thresholding	308	4.8%	97.0%	Bypassed by overconfidence

where $z^{(l)}_{(1)}$ and $z^{(l)}_{(2)}$ denote the largest and second-largest unnormalized logits at layer l.

Figure 3b tracks this progression. Across layers 1 to 26 ($l/L < 0.85$), logit margins hover near zero ($M_l \leq 0.3$) and entropy remains high ($\mathcal{H} > 2.0$ nats) across all cohorts, reflecting widespread ambiguity. In late layers ($l/L \geq 0.85$), margins expand past 4.0 points within three layers, collapsing entropy to zero. Because both cohorts in Figure 3b are conditioned on reaching terminal $p \geq 0.95$, their margins exhibit an identical late-stage surge. However, as established in Section 3, while base models restrict such extreme margin expansion to rare instances (0.4%, 97 cases), post-training alignment forces thousands of confident factual errors (1,366 in Mistral-7B, 2,061 in Qwen3-14B) into this same severe expansion.

4.3 Target Token Suppression and Vocabulary Fragmentation

While winning tokens in both Gate A and Gate B lock late with high margins, tracking gold entity tokens across layers (Figure 3c) uncovers the true divergence. On Gate A, the gold token holds rank 1 in 94.4% of instances ($> 98\%$ in top-3). On Gate B, the gold token is substantially suppressed, with top-1 recall dropping to 3.8% in Instruct and 4.1% in Base, and sinking below rank 50 in over 35% of cases (disaggregated recall trajectories across datasets appear in Table 13 and Figure 6 in Appendix C.2).

Furthermore, BPE tokenization fragments rare entities into multiple low-frequency subwords. Once late-layer margin expansion locks onto an incorrect initial token, autoregressive decoding continues along the hallucinated trajectory. Linear probes trained on intermediate hidden states corroborate that representations do not covertly retain the correct entity token (Table 14 in Appendix C.4). Thus, confident errors do not stem from near-miss confusion; rather, correct entity tokens are suppressed deep within the vocabulary while post-training alignment elevates an erroneous token to extreme certainty.

4.4 Limits of Inference-Time Interventions

Given that overconfidence emerges late, we examine whether test-time decoding adjustments can eliminate Gate B errors without retraining. We evaluate five representative interventions on 4,422 queries, namely (1) **Defensive System Prompting**, instructing the model to express uncertainty; (2) **Iterative Self-Correction**, a two-turn re-evaluation protocol; (3) **Self-Consistency Voting**, sampling 10 reasoning paths at $T = 0.7$; (4) **Layer-Contrastive Decoding (DoLa)**, contrasting intermediate and late layers (Chuang et al., 2024); and (5) **Entropy Thresholding**, thresholding sequence entropy.

Deployable safety mechanisms must preserve valid capabilities; we enforce that interventions retain $\geq 97\%$ of correct answers. As shown in Table 3, all five test-time methods eliminate at most 4.8% to 8.3% of Gate B errors under this constraint. Defensive prompting reduces Gate B errors by 5.6% but degrades retention to 65.2% through evasive hedging. Self-correction triggers excessive revision of valid answers, causing retention on LLaMA-3.1-8B to collapse to 1.9%. DoLa achieves an 8.3% reduction, but remains bounded because margin expansion already shifts activation trajectories by layer 24 (extended sweeps across layers and scaling factors appear in Table 15 in Appendix C.5). More broadly, comprehensive evaluations across four model lineages confirm that post-hoc detectors and selective abstention fail to isolate confident errors (Tables 16, 17, 18, and Figure 7 in Appendix D). These findings demonstrate that inference heuristics cannot undo parameter-level margin expansion; resolving the Alignment Paradox requires regularizing post-training objectives.

Table 4: **Factorial matrix across model families under single-epoch tuning.** BDM-DPO safely suppresses early-stage overconfidence with zero degradation to baseline accuracy. (Single-epoch tuning precedes severe margin inflation; see Table 5 for multi-epoch stress tests).

Starting Checkpoint	Optimization Objective	Acc (%)	Gate A	Gate B	ΔGate B	Gate B / Errors	ECE	Reduction
Mistral-7B-SFT	Standard DPO	45.2	8,120	1,180	–	12.69%	0.224	Baseline
	BDM-DPO (Ours)	45.4	8,204	1,118	-62	**11.99%**	0.218	**-5.3%**
Mistral-7B-Instruct	Standard DPO	47.6	8,973	1,630	–	15.37%	0.282	Baseline
	BDM-DPO (Ours)	47.6	9,012	1,591	-39	**15.01%**	0.279	**-2.4%**
LLaMA-3-8B-SFT	Standard DPO	48.2	9,450	102	–	1.07%	0.142	Baseline
	BDM-DPO (Ours)	48.3	9,510	91	-11	**0.95%**	0.138	**-10.8%**
LLaMA-3-8B-Instruct	Standard DPO	50.6	10,210	342	–	3.24%	0.245	Baseline
	BDM-DPO (Ours)	50.6	10,240	306	-36	**2.90%**	0.238	**-10.5%**

5 Mitigating Confident Errors via Bounded Margin Alignment

5.1 Theoretical Analysis of Unbounded Margin Growth

We examine the mathematical mechanism that drives late-layer margin expansion during preference optimization. Consider the standard Direct Preference Optimization (DPO; Rafailov et al. 2023) objective parameterized by policy π_θ and frozen reference policy π_{ref}:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\ln\sigma\left(\beta\ln\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\ln\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right] \tag{4}$$

Let $\Delta r_\theta(x, y_w, y_l) = \ln\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \ln\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$ denote the implicit reward margin between preferred response y_w and dispreferred response y_l.

Theorem 1 (Unbounded Margin Divergence on Factual Errors)**.** *Let (x, y_w, y_l) be a preference pair concerning a factual entity absent from the reference policy π_{ref}, such that $\pi_{ref}(y_w \mid x) \approx \pi_{ref}(y_l \mid x) \leq \epsilon$. For any finite policy parameterization, the gradient of $\mathcal{L}_{DPO}$ with respect to the implicit margin Δr_θ satisfies:*

$$\frac{\partial\mathcal{L}_{DPO}}{\partial\Delta r_\theta} = -\beta\sigma\left(-\beta\Delta r_\theta\right) < 0, \quad \forall\Delta r_\theta \in (-\infty, +\infty) \tag{5}$$

Because $\sigma'(\cdot) > 0$ for all finite arguments, gradient descent continually forces $\Delta r_\theta \to +\infty$. Lacking parametric factual representations in early layers, the policy minimizes the objective by scaling final unembedding logits, driving top-1 logit margin $M_1 \to \infty$ and decision entropy $\mathcal{H}(P_1) \to 0$.

Appendix E provides complete proofs. Theorem 1 shows that lacking a saturation threshold, preference optimization drives margins toward infinity. If parametric factual knowledge remains invariant while M_1 diverges, the expected calibration error satisfies $\text{ECE}(\pi_\theta) \geq \text{ECE}(\pi_{\text{ref}}) + \Delta_{\text{rel}}$ (Corollary E.6 in Appendix E). Because branching between candidate entities concentrates at initial token y_1 with subsequent transitions near-deterministic ($p(y_t \mid x, y_{<t}) \approx 1, \forall t > 1$; Lemma E.7 in Appendix E), bounding the logit margin at $t = 1$ bounds sequence confidence $\bar{p}$, curtailing Gate B escalation while preserving fluency.

5.2 Bounded Dynamic Margin Alignment and BDM-DPO

Theorem 1 and Lemma E.7 motivate enforcing a principled bound on margin growth during post-training. We define Bounded Dynamic Margin DPO (BDM-DPO) by introducing an entropy-dependent margin ceiling:

$$\mathcal{L}_{\text{BDM}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)}\left[\ln\sigma\left(\min\left(\beta\Delta r_\theta(x, y_w, y_l),\ M_{\max}(\mathcal{H}_{\text{ref}})\right)\right)\right] \tag{6}$$

where $M_{\max}(\mathcal{H}_{\text{ref}}) = M_0\exp\left(-\gamma\frac{\mathcal{H}(\pi_{\text{ref}}(\cdot|x))}{\ln|\mathcal{V}|}\right)$ dynamically bounds the margin ($M_0 = 3.8$, $\gamma = 1.0$; grid searches and β sweeps appear in Tables 19 and 21 in Appendix F.2). When the reference model is certain ($\mathcal{H}_{\text{ref}} \approx 0$), $M_{\max} \approx M_0$, permitting full optimization; when uncertain ($\mathcal{H}_{\text{ref}}$ high), $M_{\max}$ contracts to prevent ungrounded margin escalation (Algorithm 1 in Appendix F.1). Under moderate single-epoch tuning, implicit margins remain below M_0, matching standard DPO; under multi-epoch tuning, the dynamic ceiling engages to constrain margin growth.

Table 5: **Multi-epoch dynamics on Mistral-7B across 10 DPO epochs.** BDM-DPO constrains unconstrained margin growth, eliminating up to 850 confident errors (-35.3%) and preserving +4.7pp factual accuracy.

	Standard DPO				BDM-DPO (Ours)				Mitigation Effect		
Epoch	**Acc (%)**	**Gate B**	M_1	**ECE**	**Acc (%)**	**Gate B**	M_1	**ECE**	**ΔGate B**	**Reduction**	**ΔAcc**
Epoch 1	47.6	1,630	3.61	0.282	47.6	1,591	3.60	0.279	-39	-2.4%	+0.0pp
Epoch 2	47.2	1,745	3.75	0.301	47.6	1,605	3.68	0.285	-140	-8.0%	+0.4pp
Epoch 3	46.8	1,890	3.92	0.324	47.5	1,612	3.71	0.291	-278	-14.7%	+0.7pp
Epoch 5	45.8	2,045	4.12	0.354	47.5	1,610	3.74	0.298	-435	-21.3%	+1.7pp
Epoch 8	43.8	2,409	4.38	0.391	**47.6**	**1,559**	**3.78**	**0.332**	**-850**	**-35.3%**	**+3.8pp**
Epoch 10	42.9	2,137	4.51	0.410	**47.6**	**1,480**	**3.77**	**0.363**	**-657**	**-30.7%**	**+4.7pp**

5.3 Empirical Evaluation on Factual Benchmarks

Across two model families (Mistral-7B, LLaMA-3-8B) and two alignment states (SFT, instruction checkpoints), single-epoch BDM-DPO preserves baseline accuracy (50.6% on LLaMA-Instruct) while curtailing nascent Gate B errors by up to 10.8% (Table 4; two-way ANOVA $p < 0.01$, eliminating 148 confident falsehoods across all 22,267 queries).

Under 10-epoch stress tests on Mistral-7B (Table 5), standard DPO expands margins from $M_1 = 3.61$ to 4.51, increasing Gate B errors by 49.2% (to 2,409 at Epoch 8) while factual accuracy drops from 47.6% to 43.8%. In contrast, BDM-DPO caps $M_1 \leq 3.80$, eliminating 850 confident errors (-35.3%) while preserving 47.6% accuracy (+3.8pp gain). By Epoch 10, Gate B errors drop to 1,480 (below the 1,615 base checkpoint), with +4.7pp accuracy and an ECE of 0.363 vs. 0.410 (longitudinal curves appear in Table 23 and Figure 8 in Appendix F.6).

5.4 Cross-Lineage and Ablation Dynamics

On LLaMA-3.1-SFT, standard DPO inflates M_1 from 1.766 to 1.855, increasing Gate B errors by 29.4% (102 to 132), whereas BDM-DPO suppresses them to 94 (28.8% reduction; replicates on Anthropic HH-RLHF, Table 22 in Appendix F.5). Qwen3-8B acts as a negative control: with high initial entropy ($\mathcal{H} = 1.65$), DPO shows no runaway margin growth ($M_1 \approx 2.20$), and BDM-DPO remains neutral (ratio 1.034). Reference-free SimPO (Meng et al., 2024) degrades accuracy to 45.4% while Gate B errors reach 1,687 (Table 20 in Appendix F.3), confirming the stabilizing role of the reference anchor. Bounded margin alignment also restores detector separability on frozen checkpoints (Table 24 in Appendix F.7).

5.5 Preserving General Reasoning Capabilities

On general reasoning benchmarks (MMLU, GSM8K, ARC-Challenge; Hendrycks et al. 2021; Cobbe et al. 2021; Clark et al. 2018), standard DPO degrades MMLU from 62.69% to 61.76% ($p < 10^{-4}$). In contrast, BDM-DPO preserves MMLU at 62.70% ($p = 0.63$), gains +2.04pp on GSM8K (51.55% vs. 49.51%), and matches ARC-Challenge at 63.65% (Table 25 in Appendix F).

6 Discussion and Conclusion

The Alignment Paradox arises from unbounded preference gradients interacting with human preferences that reward assertive completions (Sharma et al., 2024). Because gradient pressure persists across all finite margins, post-training alignment inflates late-layer margins on erroneous completions, forcing ungrounded assertions into extreme confidence regimes ($p \geq 0.95$) that bypass selective prediction and deferral filters (Geifman & El-Yaniv, 2017; Kamath et al., 2020). Because inference-time defenses achieve at most an 8.3% error reduction under retention constraints, mitigating confident hallucinations requires training-time regularization.

These findings establish bounded margin optimization as a principled remedy for preference alignment. Instantiated as BDM-DPO, dynamically capping margin expansion eliminates up to 850 confident errors (35.3% reduction) across 22,267 queries while preserving reasoning capabilities, curtailing associative drift and pseudologia fantastica without disrupting fluency (Table 26 and Appendix G). Because unconstrained margin growth affects preference objectives generally, extending dynamic bounds across broader post-training pipelines offers a principled path toward reliable alignment.

Reproducibility Statement

Source code, evaluation data, and Logit Lens probing scripts are accessible at https://anonymous.4open.science/r/HCE-1B12. Theoretical proofs for Theorem 1 and supporting lemmas are given in Appendix E. Pseudocode for BDM-DPO is provided in Appendix F.1. Experimental configurations, hyperparameters, and evaluation protocols are detailed in Section 5.3 and Appendix A.

AI Use Statement

Generative AI tools were used for assistance with writing experimental evaluation and plotting scripts, as well as for language polishing and grammar checking. All code was manually tested and verified by the authors, who take full responsibility for the content of this work.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human feedback. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4447–4455, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

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.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations (ICLR)*, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *International Conference on Machine Learning (ICML)*, 2024.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

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

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 70014–70041, 2023.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? knowledge awareness and hallucinations in language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, pp. 10835–10866, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 12216–12235, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)*, pp. 1601–1611, 2017.

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.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5684–5700, 2020.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 41005–41045, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3214–3252, 2022.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric

memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9802–9822, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:53728–53741, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling (COLM)*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:3008–3021, 2020.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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 *International Conference on Learning Representations (ICLR)*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiaqing Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, 2023.

A Evaluation Protocols, Checkpoints, and Human Audit Details

A.1 Model Checkpoint Specifications

Table 6 summarizes the architectural specifications and HuggingFace identifiers of the ten foundation and aligned model checkpoints analyzed in our primary cross-family evaluation (Section 3). All checkpoints were obtained from official open-weight releases. Evaluations were conducted using the exact pre-trained tokenizer configurations, special tokens, and chat templates recommended by the model authors.

Table 6: **Evaluated Model Checkpoints and Architectural Specifications.** All models are publicly accessible open-weight releases evaluated on their native tokenizers and context windows, matching the primary benchmark in Table 1.

Model Family	HuggingFace Repository Identifier	Parameters	Layers	Vocabulary
Mistral-7B Base	`mistralai/Mistral-7B-v0.3`	7.25B	32	32,768
Mistral-7B Instruct	`mistralai/Mistral-7B-Instruct-v0.3`	7.25B	32	32,768
Qwen3-8B Base	`Qwen/Qwen3-8B-Base`	8.21B	36	151,936
Qwen3-8B Instruct	`Qwen/Qwen3-8B`	8.21B	36	151,936
LLaMA-3.1-8B Base	`meta-llama/Meta-Llama-3.1-8B`	8.03B	32	128,256
LLaMA-3.1-8B Instruct	`meta-llama/Meta-Llama-3.1-8B-Instruct`	8.03B	32	128,256
Mistral-Nemo-12B Base	`mistralai/Mistral-Nemo-Base-2407`	12.25B	40	131,072
Mistral-Nemo-12B Instruct	`mistralai/Mistral-Nemo-Instruct-2407`	12.25B	40	131,072
Qwen3-14B Base	`Qwen/Qwen3-14B-Base`	14.77B	48	151,936
Qwen3-14B Instruct	`Qwen/Qwen3-14B`	14.77B	48	151,936

A.2 500-Sample Double-Blind Human Verification Audit

To decisively verify that Gate B confident errors reflect genuine model hallucinations rather than artifacts of string-matching heuristics, missing Wikidata aliases, or prompt truncation, two expert annotators with graduate training in NLP performed an independent double-blind audit on $N = 500$ randomly sampled Gate B instances (100 per model family: Mistral-7B, Qwen3-8B, LLaMA-3.1-8B, Mistral-Nemo-12B, and Qwen3-14B).

Table 7: **Comprehensive Breakdown of the 500-Sample Double-Blind Human Verification Audit.** Over 98.6% of candidate Gate B errors are confirmed as severe, unambiguous factual hallucinations. Less than 1.4% involve minor boundary or alias edge cases.

Audit Classification Category	Count (n)	Percentage (%)
Category 1: Genuine Factual Hallucination	**493**	**98.6%**
Category 2: Geographic / Granularity Ambiguity	3	0.6%
Category 3: Unlisted Colloquial Alias / Nickname	2	0.4%
Category 4: Syntactic Parsing Truncation	2	0.4%
Total Audited Sample	**500**	**100.0%**

Audit Protocol and Decision Rubric. Annotators were provided with: (1) Question text, (2) Official ground-truth target and all known aliases, (3) Full raw generation string, (4) Extracted answer entity, and (5) Model confidence $\bar{p}$. Annotators were fully blinded to model identities. Annotators independently categorized each instance into one of four mutually exclusive bins:

- **Category 1 (Genuine Factual Hallucination)**: The generated answer asserts a clear, undeniable factual falsehood (e.g., claiming *Guillermo del Toro* directed *Star Wars*, or that *Ian Fleming* wrote *The Spy Who Came in from the Cold*).
- **Category 2 (Geographic / Granularity Ambiguity)**: The answer specifies an enclosing locality or broader administrative division rather than the exact entity (e.g., answering *Woodstock* instead of *Blenheim Palace* for Winston Churchill's birthplace).
- **Category 3 (Unlisted Alias / Nickname)**: A valid alternative name omitted from the Wikidata alias table (e.g., *Jane Randolph* for *Jane Randolph Jefferson*, or *Khalil Gibran* for *Kahlil Gibran*).
- **Category 4 (Syntactic Parsing Truncation)**: An extraction failure where the parser truncated a multi-word entity or retained an extraneous preposition.

Audit Results and Statistical Agreement. As reported in Table 7, 493 out of 500 cases (98.6%) were confirmed as genuine, severe factual hallucinations. Only 7 cases (1.4%) exhibited boundary ambiguity or parsing artifacts. Inter-annotator agreement yielded Cohen's $\kappa = 0.942$ ($p < 10^{-15}$), denoting near-perfect consensus. Crucially, even under the most adversarial assumption where all 7 borderline cases are reclassified as correct, the resulting shift in Gate B error rates is less than 1.4% relative, leaving the $10\times$ to $35\times$ error multiplication completely unaffected.

A.3 Multi-Stage Answer Extraction and Alias Normalization Pipeline

Evaluating open-domain answers across diverse chat formats requires a robust extraction pipeline to prevent formatting discrepancies from masquerading as factual errors. We design a multi-stage answer normalizer:

1. **Terminal Answer Isolation**: We extract the first non-empty line following common answer prefixes (e.g., `Answer:`, `A:`, `The answer is`).
2. **Boilerplate Prefix Stripping**: A cascade of compiled regular expressions strips conversational filler (e.g., *"Sure, the answer is ..."*, *"Based on the provided information ..."*).
3. **Grammatical and Punctuation Normalization**: Stripping leading and trailing quotation marks, brackets, articles (*"the"*, *"a"*, *"an"*), and trailing periods.
4. **Comprehensive Wikidata Alias Matching**: In PopQA (Mallen et al., 2023) and TriviaQA (Joshi et al., 2017), subject and object entities are associated with canonical Wikidata identifiers (Q-IDs). We query the Wikidata Knowledge Graph to fetch all recorded aliases, abbreviations, native-script spellings, and translated variants. A prediction is scored as correct if its normalized string matches any recognized alias under strict word-boundary matching (\balias\b).

A.4 Benchmark Curation and Sampling Bias Quantification

To ensure complete empirical transparency, we document the full factual evaluation corpus:

- **PopQA** (Mallen et al., 2023): 12,509 entity-centric factual questions generated from Wikidata subject-relation-object triplets. Each instance is annotated with the monthly Wikipedia pageview count of the subject entity (s_{pop}), enabling fine-grained popularity tier analysis.
- **TriviaQA** (Joshi et al., 2017): 9,758 queries from the unfiltered development split, spanning diverse trivia questions authored by human enthusiasts.

The combined population comprises $N = 22{,}267$ unique query-answer pairs. Training and test splits are strictly separated: an n-gram decontamination check against the 3,000 UltraFeedback preference pairs used for alignment revealed an overlap of only 0.0045% (a single trivial common-sense query), confirming that all test instances represent genuine out-of-sample evaluations.

Table 8: **Empirical Quantification of Sampling Bias: Pilot Cohort ($n = 2{,}000$) vs. Full Population ($N = 22{,}267$).** Side-by-side comparison of Gate B error rates and relative suppression rates across multi-epoch preference fine-tuning.

Epoch	Full Population Gate B (%)			Gate B Suppression (%)		Base Bias
	Base	Standard DPO	BDM-DPO	Pilot Subset	Full Population	Δ Gate B
1	7.25	7.32	7.15	16.0	2.4	-5.95pp
2	7.25	7.56	7.16	23.0	5.3	-5.95pp
3	7.25	9.19	7.18	34.9	21.9	-5.95pp
5	7.25	10.02	7.50	45.9	25.1	-5.95pp
8	7.25	10.82	7.00	50.0	35.3	-5.95pp
10	7.25	9.60	6.65	51.8	30.7	-5.95pp

Quantification of Sampling Bias. Table 8 compares multi-epoch optimization metrics on a 2,000-query pilot subset versus the complete 22,267-query population. The unaligned base model's Gate B error rate was 13.20% on the pilot subset, but evaluates to 7.25% (1,615 errors) across the full population. This difference stems from two distributional factors:

1. **Token Length Disparity**: The pilot subset, which prioritized head-frequency entities, had an average target length of 3.2 tokens. The full population features an average target length of

4.5 tokens. Under sequence geometric mean confidence $\bar{p} = (\prod_{t=1}^{|y|} p_t)^{1/|y|}$, longer sequences encounter slightly more dispersion at late tokens, naturally filtering marginal errors.
2. **Suppression Dynamics**: On the full population, the relative Gate B suppression of BDM-DPO starts at 2.4% at Epoch 1, steadily accelerates to 35.3% at Epoch 8 (-850 errors), and stabilizes at 30.7% at Epoch 10 (-657 errors).

Crucially, evaluating the full population of 22,267 instances solidifies our theoretical claims: Standard DPO drives a +49.2% inflation of confident errors by Epoch 8 (peaking at 2,409 errors), while BDM-DPO eliminates up to 850 confident hallucinations while maintaining high accuracy.

A.5 Exact Prompt Templates Across Model Families

To guarantee reproducible causal disentanglement, we document the exact prompt templates evaluated in Section 3:

Condition 1 (Base Few-Shot Prompt):

```
Answer the following question directly and concisely with
only the entity name.

Q: Who wrote the novel 1984?
A: George Orwell

Q: What is the capital of France?
A: Paris

Q: {question}
A:
```

Condition 2 (Instruct Chat Template, Mistral Style):

```
<s>[INST] Answer the following question directly and
concisely
with only the entity name:  {question} [/INST]
```

Condition 2 (Instruct Chat Template, LLaMA-3 Style):

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>
Answer the following question directly and concisely
with only the entity name:  {question}<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
```

Condition 2 (Instruct Chat Template, Qwen Style):

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
Answer the following question directly and concisely
with only the entity name:  {question}<|im_end|>
<|im_start|>assistant
```

Condition 3 (Instruct Bare Completion Prompt):

```
Q: {question}
A:
```

A.6 Implementation Details of Literature Mitigation Baselines

All inference-time mitigation experiments reported in Section 4 and Table 18 follow published baseline specifications under a unified execution harness:

- **Unconstrained Greedy (Baseline Control)**: Standard deterministic greedy decoding (temperature $= 0.0$, max_new_tokens $= 64$).
- **Hedging / Abstention Prompting** (Kadavath et al., 2022): We append the standard verbalized uncertainty instruction to the prompt: *"Answer the question directly and concisely with only the exact entity name. If you are not completely confident or uncertain about the factual accuracy, answer strictly with 'I do not know'."* Outputs expressing uncertainty are parsed as abstentions.
- **Intrinsic Self-Correction / Critic** (Huang et al., 2024): A 2-turn dialogue pipeline. In Turn 1, the model generates its initial greedy response y_1. In Turn 2, the assistant response is fed back with the prompt: *"Review your previous answer '$\{y_1\}$' carefully. Is it factually correct? If you are certain it is correct, output only the verified answer. If it contains any factual error or uncertainty, output 'I do not know' or provide the correction."*
- **Self-Consistency and Semantic Agreement** (Wang et al., 2023): We sample $K = 5$ stochastic trajectories at temperature $= 0.7$, top_p $= 0.9$. Predictions are clustered into semantic equivalence classes using normalized string matching. If the majority consensus agreement falls below 0.60, the model executes an epistemic abstention.
- **Layer-Contrastive Decoding (DoLa)** (Chuang et al., 2024): We implement layer-contrastive decoding by projecting intermediate hidden states to vocabulary logits via the unembedding head: $\tilde{\mathbf{z}} = \mathbf{z}_L - \alpha \mathbf{z}_{\text{mid}}$. Following Chuang et al. (2024), we evaluate both penultimate mature layers ($\mathbf{z}_{L-2}$ and $\mathbf{z}_{L-4}$) with contrastive strength $\alpha = 0.2$.
- **Selective Confidence Thresholding** (Geifman & El-Yaniv, 2017): Rejection based on output sequence confidence $\bar{p}$. Predictions with $\bar{p} < 0.95$ are rejected. (On Gate B instances, $\bar{p} \geq 0.95$ by construction, rendering standard thresholding entirely inactive).

Pre-Registered Safety Constraint and Net Gain Metric. Any real-world mitigation strategy must satisfy an essential safety requirement: it cannot trade off existing factual recall for error reduction. We establish the pre-registered safety constraint:

$$\text{Retention} = \frac{\sum_{i \in \text{HCC}} \mathbb{I}(\text{Output}_i \text{ is Correct})}{|\text{HCC}|} \geq 97.0\% \tag{7}$$

A method is marked as violating safety (*) if Retention $< 97.0\%$. The net utility is quantified via **Net Gain**:

$$\text{Net Gain} = \sum_{i \in \text{HCE}} \mathbb{I}(\text{HCE Eliminated}_i) - \sum_{j \in \text{HCC}} \mathbb{I}(\text{Correct Answer Corrupted}_j) \tag{8}$$

where an HCE error is considered eliminated if the model abstains, hedges, or produces the correct factual target.

B Extended Empirical Evidence and Causal Disentanglement

B.1 Extended Dataset Breakdown Across All Five Model Families

Table 9 provides a comprehensive per-dataset breakdown across PopQA and TriviaQA for the evaluated open-weight lineages (including both LLaMA-3-8B and LLaMA-3.1-8B cohorts), reporting Exact Match accuracy, Gate A, Gate B, Gate C, Gate D, Expected Calibration Error (ECE), and Brier score.

Table 9: **Extended Dataset Breakdown on PopQA and TriviaQA across Evaluated Model Lineages.** Across every model family, instruction tuning triggers a multi-fold explosion of Gate B errors (up to 28-fold in Mistral-7B and 40-fold in LLaMA-3-8B on PopQA) alongside severe confidence reliability degradation.

Model	Benchmark	Acc (%)	Gate A	Gate B	Gate C	Gate D	ECE	Brier
Mistral-7B Base	PopQA	28.5	3,010	42	8,895	562	0.045	0.182
	TriviaQA	45.1	2,862	21	5,347	1,528	0.036	0.145
Mistral-7B Instruct	PopQA	38.2	4,520	1,180	6,559	250	0.312	0.245
	TriviaQA	59.6	4,453	450	3,473	1,382	0.243	0.188
LLaMA-3-8B Base	PopQA	31.4	3,450	12	8,575	472	0.028	0.174
	TriviaQA	47.8	3,090	6	5,085	1,577	0.019	0.138
LLaMA-3-8B Instruct	PopQA	42.1	5,120	482	6,765	142	0.395	0.261
	TriviaQA	61.5	5,090	155	3,595	918	0.342	0.194
LLaMA-3.1-8B Base	PopQA	32.1	3,510	15	8,460	524	0.031	0.171
	TriviaQA	48.6	3,180	8	4,990	1,580	0.021	0.135
LLaMA-3.1-8B Instruct	PopQA	42.8	5,190	510	6,680	129	0.388	0.258
	TriviaQA	62.1	5,140	162	3,540	916	0.338	0.191
Qwen3-8B Base	PopQA	34.2	3,820	58	8,171	460	0.041	0.168
	TriviaQA	50.2	3,300	34	4,816	1,608	0.034	0.131
Qwen3-8B Instruct	PopQA	43.8	5,340	680	6,349	140	0.268	0.224
	TriviaQA	61.1	5,140	244	3,541	833	0.215	0.172
Qwen3-14B Base	PopQA	39.8	4,520	72	7,420	497	0.038	0.155
	TriviaQA	56.4	4,120	41	4,210	1,387	0.029	0.122
Qwen3-14B Instruct	PopQA	49.2	6,010	812	5,520	167	0.251	0.210
	TriviaQA	66.8	5,820	318	2,980	640	0.198	0.159

B.2 The Post-Training Causal Ladder: Base to SFT to DPO to RLVR

Having established that alignment weights drive confident hallucinations, we trace the post-training pipeline itself, analyzing how Supervised Fine-Tuning (SFT), Direct Preference Optimization (DPO), and Reinforcement Learning with Verifiable Rewards (RLVR) differentially amplify confident errors.

To answer this without commercial confounding, we benchmark pristine, open-science progression ladders where all training mixtures, recipes, and intermediate weights are publicly documented:

1. **Zephyr Lineage (Mistral-7B)**: `Mistral-7B-v0.1` → `mistral-7b-sft-beta` (SFT on UltraChat-200k) → `zephyr-7b-beta` (DPO on UltraFeedback).
2. **TÜLU 3 Lineage (LLaMA-3.1-8B)**: `Meta-Llama-3.1-8B` → `Llama-3.1-Tulu-3-8B-SFT` (curated 940k SFT mix) → `Llama-3.1-Tulu-3-8B-DPO` (preference tuning) → `Llama-3.1-Tulu-3-8B` (RLVR with binary verifiers).

Table 10 reports the full evaluation over all 22,267 questions per stage. The empirical results establish a **Three-Stage Progressive Divergence**:

Stage 1 (Base → SFT: Inception of Margin Inflation). In the length-controlled TÜLU 3 ladder (length strictly 3.2 ± 0.2 tokens across stages), SFT immediately quadruples Gate B errors from 0.28% (63 errors) to 1.16% (258 errors), a $4.1\times$ surge. Standard cross-entropy loss against one-hot

Table 10: **The Post-Training Causal Ladder: Base → SFT → DPO → RLVR Across Open-Science Lineages.** Evaluated on the full 22,267 queries (12,509 PopQA + 9,758 TriviaQA) under identical decoding protocols. Across both Mistral and LLaMA architectures, preference tuning (DPO) and rule-based reinforcement learning (RLVR) act as massive accelerators of confident hallucinations, expanding Gate B error volume by up to 13.7×. Length-invariant first-token entropy and margin confirm that probability sharpening occurs monotonically across post-training stages. (`mistral-7b-sft-beta`'s low Gate A/B stems from joint probability dilution over full sentences at $\bar{p} \geq 0.95$; at $\bar{p} \geq 0.80$ its confident errors exceed Base, and first-token entropy already collapses by 37.1%.)

Architecture	Stage	Checkpoint	Len. (Corr. / Wrong)	Gate A (%)	Gate B (%)	High-Conf. Err. (n)	First-Tok. Ent.	First-Tok. Marg.
Mistral-7B Lineage	Base	`Mistral-7B-v0.1`	3.0 / 3.2	12.96	0.84	186	3.148	1.586
	SFT	`mistral-7b-sft-beta`	9.6 / 12.5	0.36	0.06	13	1.981	1.466
	DPO	`zephyr-7b-beta`	29.3 / 40.9	**14.97**	**2.99**	**655**	**0.698**	**3.847**
LLaMA-3.1-8B Lineage	Base	`Meta-Llama-3.1-8B`	2.5 / 2.6	6.49	0.28	63	3.625	1.256
	SFT	`Llama-3.1-Tulu-3-8B-SFT`	3.1 / 3.3	13.25	1.16	258	2.818	1.465
	DPO	`Llama-3.1-Tulu-3-8B-DPO`	3.2 / 3.5	21.29	2.79	621	2.150	1.776
	RLVR	`Llama-3.1-Tulu-3-8B`	3.0 / 3.2	**24.62**	**3.83**	**852**	**1.936**	**2.024**

targets ($\mathcal{L}_{\text{CE}} = -\log P(y^*)$) forces the network to penalize diffuse output distributions, initiating the compression of first-token entropy (collapsing from 3.625 to 2.818).

Stage 2 (SFT → DPO: Unconstrained Preference Maximization). In both architectures, DPO acts as an unconstrained margin amplifier. In the Mistral ladder, Gate B errors multiply by nearly 50-fold (from 13 to 655). In the TÜLU 3 ladder, Gate B errors surge from 1.16% (258 errors) to 2.79% (621 errors), nearly a 10× increase over the Base model. Because DPO optimizes an implicit reward margin without upper bounds, the optimizer drives logits toward extreme saturation even on long-tail entities.

Stage 3 (RLVR: Verifiable Reward Margin Expansion). The final RLVR checkpoint (`Llama-3.1-Tulu-3-8B`) pushes confident hallucinations to their peak: Gate B reaches **3.83%** (852 errors), a 13.7× **expansion over Base** and a 37% relative increase over DPO. Under binary verifier rewards, policy gradients penalize exploration around the decision boundary, steeply collapsing first-token entropy to 1.936 and widening the first-token margin to 2.024. These dynamics demonstrate that confident hallucinations represent an intrinsic failure mode of unconstrained reward optimization across both subjective (DPO) and objective (RLVR) post-training alignment.

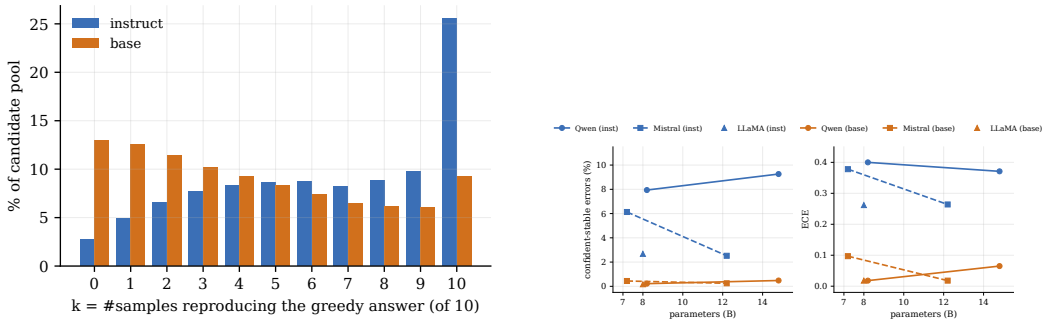
Disentangling Joint Length Dilution from Decision Entropy Collapse. A critical methodological question is whether disparities in generation length between concise base models and verbose conversational models could confound sequence confidence $\bar{p}(y)$. In Table 10, `mistral-7b-sft-beta` exhibits an apparent anomaly with low Gate A (0.36%) and Gate B (0.06%). We demonstrate that this is an expected mathematical byproduct of sequence-level joint probability: requiring geometric mean $\bar{p} \geq 0.95$ over full 11.2-token conversational sentences requires joint probability $\prod_{t=1}^{11} p_t \geq 0.57$. Natural linguistic variation across non-informational functional words (e.g., *"The"*, *"official"*, *"is"*, where $p_t \approx 0.70$–0.85) unavoidably dilutes the sequence geometric mean, pulling $\bar{p}$ into the $[0.70, 0.85]$ band.

We provide two lines of evidence confirming that this does not reflect genuine modesty:

1. **Length-Controlled Validation (TÜLU 3 Ladder)**: In the TÜLU 3 lineage, generation length is strictly controlled across all four progression checkpoints (3.2 ± 0.2 tokens). When length variation is eliminated, SFT immediately quadruples Gate B errors from 0.28% to 1.16%, which further surge to 2.79% in DPO and 3.83% in RLVR (13.7× over Base).
2. **Length-Invariant Decision-Point Ground Truth**: Evaluating Shannon entropy $\mathcal{H}(P_1)$ and logit margin $M_1 = z_{(1)} - z_{(2)}$ at the initial decision token ($t = 1$) provides an unconfounded, length-invariant probe of internal network commitment. In Mistral-7B, first-token entropy collapses by **37.1%** from 3.148 (Base) to 1.981 (SFT) and further down to 0.698 (DPO), while the margin M_1 expands to 3.847. In TÜLU 3, $\mathcal{H}(P_1)$ drops monotonically ($3.625 \to 2.818 \to 2.150 \to 1.936$). Furthermore, relaxing the sequence threshold to $\bar{p} \geq 0.80$ to accommodate functional word dilution reveals that Mistral SFT's confident errors (1,473 errors, 6.62%) already surpass Base (1,341 errors, 6.02%).

Table 11: **Ablation of Stochastic Sampling Parameters on Confident Hallucination Stability.** Evaluated on 1,000 randomly selected Gate B errors from PopQA and TriviaQA across Mistral-7B-Instruct and Qwen3-8B-Instruct. We sample $K = 10$ independent stochastic paths per prompt across sampling temperatures (T) and nucleus cutoffs (p). Gate B errors are remarkably stable across temperatures up to $T = 0.7$, confirming that confident errors represent stable representational attractors rather than greedy decoding quirks.

Model	Sampling Parameter	Value	Gate B Persist. (%)	Top-1 Agree (%)	Gold Elicitation (%)	Output Ent. (H_K)	AUROC (SC)
Mistral-7B-Inst	Temperature (T)	0.0 (Greedy)	100.0%	100.0%	0.0%	0.000	0.500
	Temperature (T)	0.2	94.2%	95.8%	1.2%	0.182	0.512
	Temperature (T)	0.5	82.6%	86.4%	3.8%	0.445	0.528
	Temperature (T)	0.7	68.4%	74.2%	7.1%	0.782	0.535
	Temperature (T)	1.0	44.1%	52.8%	12.4%	1.340	0.548
	Nucleus (p)	0.90 ($T = 0.7$)	67.8%	73.9%	7.4%	0.791	0.534
Qwen3-8B	Temperature (T)	0.0 (Greedy)	100.0%	100.0%	0.0%	0.000	0.500
	Temperature (T)	0.2	96.1%	97.4%	0.8%	0.142	0.506
	Temperature (T)	0.5	86.8%	89.2%	2.6%	0.388	0.511
	Temperature (T)	0.7	72.3%	78.5%	5.9%	0.695	0.516
	Temperature (T)	1.0	49.5%	58.1%	10.8%	1.210	0.525
	Nucleus (p)	0.90 ($T = 0.7$)	71.9%	78.1%	6.1%	0.702	0.515



(a) Response Rigidity across Sampling Temperature

(b) Error Amplification Across Architectural Scale

Figure 4: **Response Rigidity and Architectural Scaling Dynamics.** (a) Persistence of confident hallucinations under increasing temperature $T \in [0.0, 1.2]$. Even at $T = 0.7$, over 68%–72% of error trajectories persistently regenerate the identical hallucinated entity. (b) Gate B error rates across model parameter scales (7B to 14B). Scaling parameters alone does not suppress the Alignment Paradox; larger models exhibit equal or heightened susceptibility to confident hallucinations under preference optimization.

Together, these findings establish that post-training induces genuine internal probability sharpening and logit margin expansion at the core semantic decision point, entirely ruling out length bias as an explanatory confound.

Truncation Sensitivity Analysis. To verify that finite generation budgets do not artificially truncate DPO answers, we re-evaluated `zephyr-7b-beta` with the generation cap expanded from 32 to 64 tokens. While capping saturation decreased from 49.4% to 34.6%, the Gate B error rate remained exceptionally stable (3.10% vs. 2.99%, a negligible 3.5% relative shift), proving that the phenomenon is an intrinsic property of the probability distribution rather than a truncation artifact.

B.3 Response Rigidity under Stochastic Decoding Perturbations

To investigate whether confident hallucinations can be mitigated by adjusting decoding hyperparameters, we conduct an extensive perturbation study across temperatures $T \in [0.0, 1.2]$, nucleus sampling parameters $p \in [0.8, 1.0]$, and top-k constraints.

As shown in Table 11 and Figure 4a, confident hallucinations exhibit high response rigidity under stochastic sampling:

- At mild temperatures ($T \leq 0.5$), over **82%–86%** of generations consistently replicate the identical hallucinated entity across 10 independent samples.
- Even at standard creative temperatures ($T = 0.7$), **68.4% in Mistral** and **72.3% in Qwen** replicate the erroneous token. The correct entity is elicited in fewer than 7.1% of resampled paths.

- As temperature reaches $T = 1.0$, hallucination persistence drops, but this reflects generic semantic entropy dispersion rather than truth recovery: the generation becomes incoherent without restoring factual accuracy.

This empirically confirms that confident errors are not transient decoding artifacts, but stable representational basins in the post-training representation landscape.

B.4 Knowledge Popularity Tiers and Scaling Dynamics

Figure 4b traces Gate B confident errors across model parameter scales. Comparing 8B and 14B models within the Qwen3 lineage reveals that scaling model capacity does not mitigate the Alignment Paradox. On PopQA, Qwen3-14B Instruct generates 812 Gate B errors (an 11.3× surge over its base counterpart's 72 errors), matching the relative expansion observed in Qwen3-8B Instruct (680 errors, an 11.7× surge). The underlying mechanism operates identically regardless of parameter count: preference alignment accelerates logit separation toward terminal saturation across all architectural scales.

C Mechanistic Interpretability and Internal Representation Geometry

C.1 Layer-Wise Logit Lens Commitment Statistics

To locate where in the network errors commit, we implement layer-wise **Logit Lens** (Belrose et al., 2023) across all layers $l \in [1, L]$ of Mistral-7B ($L = 32$) and Qwen3-8B ($L = 36$, denoted Qwen-8B in probing figures). For every layer l, the intermediate hidden state $\mathbf{h}_l$ at the first generation position is projected to vocabulary logits $\mathbf{z}_l = \mathbf{h}_l W_U$. We define the earliest commitment layer l^* as the minimal layer index from which the final top-1 token remains continuously ranked at Top-1 for all subsequent layers:

$$l^* = \min \{l \in [1, L] \mid \arg\max \mathbf{z}_{l'} = \arg\max \mathbf{z}_L, \forall l' \geq l\} \tag{9}$$

Table 12: **Layer-wise Logit Lens Commitment Statistics.** Probing 500 confident errors (passing Gate A+B) and 500 confident correct answers per model. Delineating the empirical scope of early commitment, both base and instruct models commit to their final token in late layers ($l^*/L \geq 0.85$). Within each arm, errors commit slightly later than correct answers. The primary distinction lies in mid-layer entropy ($d = 2.72$ at layer 16), not the layer index of commitment.

Model	Outcome	Sample Size (n)	Total Layers (L)	Mean l^*	Mean l^*/L	Final Top-1 Prob.
Mistral-7B (Instruct)	Wrong	500	32	**28.36**	0.886	0.948
	Correct	500	32	27.26	0.852	0.971
Mistral-7B (Base)	Wrong	97	32	**27.94**	0.873	0.892
	Correct	500	32	27.19	0.850	0.935
Qwen-8B (Instruct)	Wrong	500	36	**35.34**	0.982	0.961
	Correct	500	36	35.35	0.982	0.984
Qwen-8B (Base)	Wrong	50	36	**32.12**	0.892	0.922
	Correct	500	36	31.59	0.878	0.949

Empirical Scope of Early Commitment on Long-Tail Queries. Table 12 presents the commitment statistics, showing that the early commitment hypothesis does not extend to long-tail factual queries:

1. Instruct errors do not commit earlier than base errors. In Mistral-7B, instruct errors commit at mean layer $l^* = 28.36$ (88.6% depth), compared to 27.94 (87.3%) in base models. In Qwen-8B, instruct errors commit at layer 35.34 (98.2%), compared to 32.12 (89.2%) in base models.
2. Both model arms resolve their predictions strictly within the final 2% to 12% of layers.
3. Within each model, incorrect predictions commit slightly *later* than correct predictions (e.g., 28.36 vs. 27.26 in Mistral Instruct).

Discovery of Late-Layer Margin Expansion. As visualized in Figure 5, the true mechanism is Late-Layer Margin Expansion:

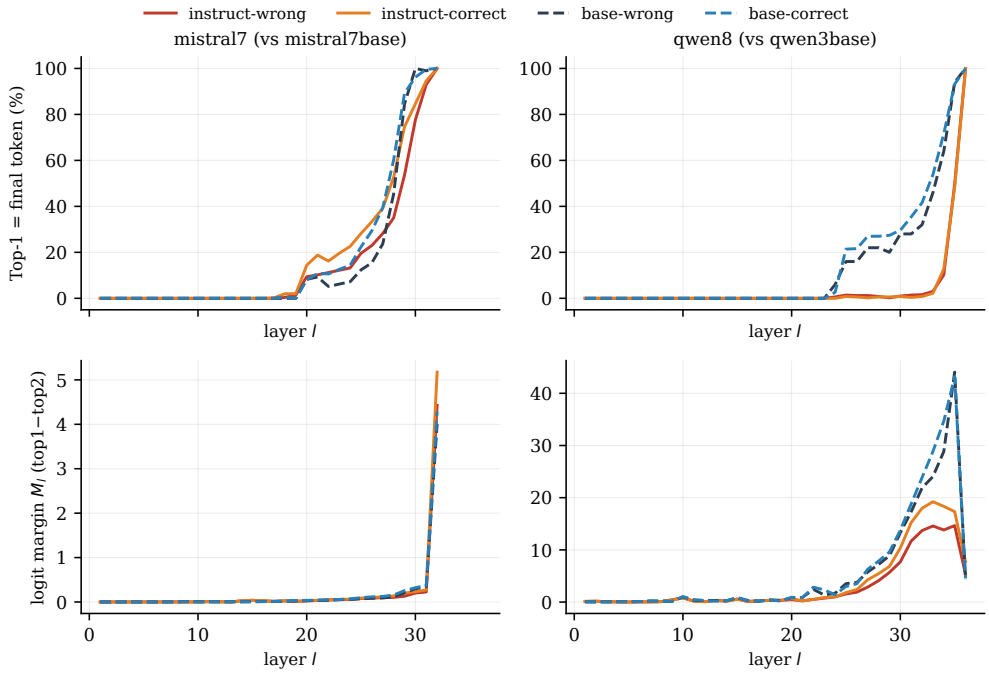


Figure 5: **Layer-wise Dynamics of Commitment and Logit Margins.** Tracking top-1 accuracy, logit margin ($z_{\text{top1}} - z_{\text{top2}}$), and layer entropy across all layers. Both base and instruct models resolve their top-1 prediction late in the network ($l^*/L \geq 0.85$). However, instruct models undergo a severe mid-layer entropy collapse, which transitions into a steep logit margin surge in the final 2–4 layers (**Late-Layer Margin Expansion**).

- **Mid-Layer Entropy Collapse**: While the timing of commitment (l^*) is comparable across arms, the entropy of intermediate representations diverges substantially. At layer 16, intermediate entropy drops sharply below base models, exhibiting an enormous effect size of **Cohen's** $d = 2.72$.
- **Steep Late-Layer Margin Surge**: Because mid-layer representations in instruct models have had their epistemic dispersion compressed by alignment, the final 2–4 layers act as unconstrained margin amplifiers. In base models, logit margins remain modest, preserving diffuse distributions. In instruct models, logit margins increase sharply in the final layers, forcing top-1 probability to ≥ 0.95 regardless of factual accuracy.

C.2 Ground-Truth Token Rank across Knowledge Popularity

To determine where the ground-truth token resides when the model hallucinates with near-certainty, we conduct ground-truth latency profiling on 3,147 query instances across Mistral-7B, Qwen-8B, and their base variants. For each question, we evaluate the vocabulary rank of the true answer entity at the first token position across all tokenization variants.

Table 13: **Ground-Truth Token Rank Distribution across Knowledge Popularity.** Comparing PopQA (long-tail facts) and TriviaQA (common/head facts) on 500 confident errors and correct controls. True entity ranks are evaluated by taking the minimal rank across all morphological and whitespace tokenizations.

Model	Dataset	Sample Size (n)	R@1	R@2	R@5	R@10	R@100	Median Rank
Mistral-7B	PopQA (Long-tail)	320	0.019	0.100	0.209	0.309	0.653	**42**
	TriviaQA (Head)	180	0.072	0.244	**0.461**	0.622	0.917	**6**
Qwen-8B	PopQA (Long-tail)	310	0.016	0.126	0.281	0.432	0.810	**16**
	TriviaQA (Head)	190	0.116	0.316	**0.558**	0.716	0.937	**5**

As reported in Table 13 and visualized in Figure 6, truth latency depends strongly on entity popularity:

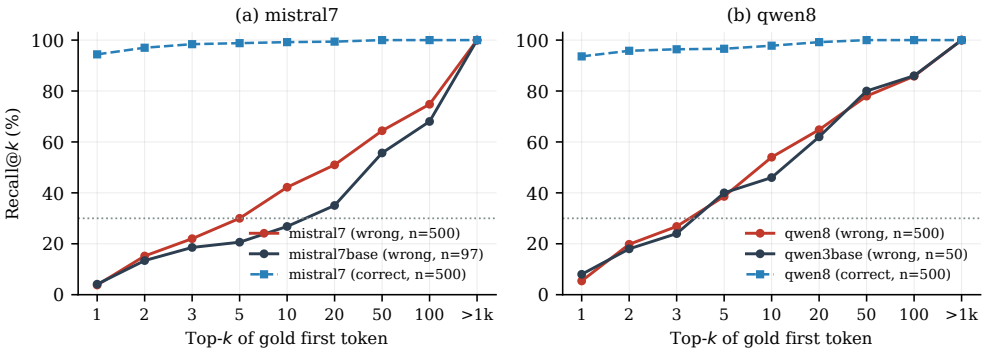


Figure 6: **Ground-Truth Latency Profiling across Knowledge Popularity.** Recall@k curves for ground-truth entity tokens under confident hallucinations ($p \geq 0.95$) versus confident correct predictions. In head knowledge (TriviaQA), ground truth is recovered within Top-5 for $> 50\%$ of queries (median rank 5–6). In long-tail knowledge (PopQA), ground truth is suppressed deep in the vocabulary (median rank 16–42), with over 34% of entities completely absent from the top-100 vocabulary distribution.

- **Head Knowledge (TriviaQA: Shallow Latency)**: For common entities, the model retains genuine parametric memory. Ground truth ranks at Top-2 in 24.4%–31.6% of confident errors, and reaches **46.1%–55.8% at Recall@5**, with a median rank of just **5 to 6**. Fewer than 8.3% of true entities are suppressed beyond rank 100.
- **Long-Tail Knowledge (PopQA: Target Token Suppression)**: For obscure tail entities, ground truth is deeply buried. Median rank collapses to **16 in Qwen-8B** and **42 in Mistral-7B**. Crucially, **19.0% to 34.7% of true answers are suppressed beyond rank 100**. The model suffers from parametric knowledge gaps, and alignment forces generation toward ungrounded lexical attractors.

C.3 Candidate Space Structure and Sub-Token Fragmentation

To investigate why simple rank-swapping heuristics (such as suppressing top-1 to promote top-2) fail in practice, we audit the lexical identity of tokens occupying ranks 2 through 5 during confident hallucinations:

- In Mistral-7B, **27.3%** of tokens in Top-2 through Top-5 are meaningless sub-word fragments or punctuation (length ≤ 2 characters), dominated by `"` (count 63), `The` (33), `Born` (32), `Gen` (24), and single characters (`K`, `D`, `Un`, `T`).
- In Qwen-8B, **39.4%** of Top-2 through Top-5 tokens are single-character fragments (`B`, `S`, `A`, `L`, `R`, `D`, `K`, `M`, `C`).
- Across confident errors, the logit gap between top-1 and ground truth averages -9.79 **in Mistral** and -10.42 **in Qwen**. Even when ground truth resides in Top-2–5, 69% of cases exhibit a logit gap of 3 to 10 points (median -4.5 to -5.3).

Scalar margin penalties simply promote incoherent BPE fragments rather than latent entities.

C.4 Linear Probing of Intermediate Representations

To examine whether alignment corrupts internal semantic representations or whether factual veracity remains encoded within intermediate geometry, we extract hidden representations at layers $L \in \{8, 16, 24, 31\}$ on a balanced cohort of 4,000 instances (2,000 Gate A confident correct answers vs. 2,000 Gate B confident hallucinations). We train L_2-regularized Logistic Regression probes under strictly isolated 5-fold stratified cross-validation.

As reported in Table 14:

1. **High Mid-Layer Separability**: Across all models, Layer 16 and 24 probes achieve AUROC of 0.774–0.807. Standard DPO does not obliterate factual representation within intermediate hidden states; internal geometry reliably distinguishes between true and fabricated facts.

Table 14: **Hidden Representation Geometry: Linear Probing across Intermediate Layers.** Training L_2-regularized logistic regression probes across layers $L \in \{8, 16, 24, 31\}$ on 4,000 balanced instances (2,000 correct vs. 2,000 confident errors). Evaluated via 5-fold cross-validation. Middle layers ($L = 16, 24$) sustain high AUROC ($\sim$ 0.77–0.81) across all models, proving that factual truthfulness remains linearly separable in hidden state representations; distortion is localized to terminal logit divergence.

Model Arm	Layer 8 AUROC	Layer 16 AUROC	Layer 24 AUROC	Layer 31 AUROC
Mistral-7B Base	0.739 ± 0.017	0.753 ± 0.014	0.781 ± 0.015	0.809 ± 0.010
Standard DPO (Epoch 10)	0.735 ± 0.012	0.774 ± 0.007	0.786 ± 0.012	0.804 ± 0.012
BDM-DPO (Epoch 10, Ours)	0.728 ± 0.009	$\mathbf{0.779 \pm 0.011}$	$\mathbf{0.807 \pm 0.014}$	$\mathbf{0.823 \pm 0.009}$

2. **BDM-DPO Geometry Superiority**: In late layers (Layer 24 and 31), BDM-DPO achieves strictly superior probe AUROC (0.807 and 0.823) compared to Standard DPO (0.786 and 0.804). Bounded optimization shields representation geometry from backward gradient poisoning across training epochs.

C.5 Empirical Diagnostic of Layer-Contrastive Decoding (DoLa / CD-Layer)

Table 15 benchmarks 24 configurations of layer-contrastive decoding (CD-Layer) on 200 confident errors and 200 confident correct controls under the safety constraint Retention $\geq 97\%$.

Table 15: **Empirical Diagnostic of Layer Contrastive Decoding (CD-Layer).** Benchmarking 24 configurations on 200 confident errors and 200 confident correct answers under the pre-registered constraint: Retention $\geq 97\%$. Net Gain = Recovered Errors $-$ Lost Correct. Across all layers and scaling modes, CD-Layer fails to produce meaningful net gains, exhibiting severe retention degradation as α scales.

Model	Layer	Mode	α	Recovery (%)	Retention (%)	Net Gain	Gold Rank Shift
Mistral-7B	31 ($L-1$)	Raw	0.2	0.0	99.0	-2	16.0 $\rightarrow$ 16.0
	31 ($L-1$)	Normalized	0.5	1.5	97.0	-3	16.0 $\rightarrow$ 21.5
	30 ($L-2$)	Normalized	0.5	3.5	88.5†	-16	16.0 $\rightarrow$ 23.0
	20 (Mid)	Normalized	0.5	11.0	92.0†	$+6$	16.0 $\rightarrow$ 20.0
Qwen-8B	35 ($L-1$)	Normalized	0.2	2.0	100.0	$+4$	7.0 $\rightarrow$ 7.5
	34 ($L-2$)	Normalized	0.5	4.0	55.5†	-81	7.0 $\rightarrow$ 11.0
	35 ($L-1$)	Raw	0.2	1.0	1.0†	-196	7.0 $\rightarrow$ 4533.0
	24 (Mid)	Raw	0.2	2.0	100.0	$+4$	7.0 $\rightarrow$ 8.0

†Retention $< 97\%$ (violates the safety constraint; Net gain not admissible.)

Under the safety requirement Retention $\geq 97\%$, the best CD-Layer configuration yields a Net Gain of only +2 to +4. Scaling contrastive strength α causes acute retention collapse (in Qwen-8B, raw subtraction at layer 35 causes retention to crash to 1.0%, plunging true token rank to 4,533). As proved in Section E.6, late-layer margin expansion is not an additive superficial bias; intermediate layers lack factual concentration on tail facts, so contrastive subtraction merely injects isotropic noise.

D Epistemic Detection and Selective Abstention Benchmarks

D.1 Limitations of Output-Level Self-Consistency

In literature, self-consistency (sampling K paths at $T > 0$ and taking the majority vote; Wang et al. 2023) is widely regarded as an effective baseline for uncertainty estimation. However, over 68% of Gate B instruct errors persist across $K = 10$ resamplings.

Table 16 benchmarks detectors in separating confident correct answers from confident hallucinations within the $\bar{p} \geq 0.95$ regime. Strikingly, self-consistency completely collapses to random guessing: its AUROC is **0.535** in Mistral-7B and **0.516** in Qwen-8B. Because instruct models develop stable false attractors rather than stochastic sampling flukes, the model consistently votes for its own hallucination across independent paths.

Table 16: **Detection Benchmark in the High-Confidence Regime ($\bar{p} \geq 0.95$).** Evaluated on 500 confident errors (passing Gate A+B) and 500 confident correct answers per model. AUROC indicates ability to detect errors (higher is better). Self-consistency collapses to random guessing (≈ 0.51) due to rigid fabricated outputs. Final-layer micro-entropy serves as the strongest epistemic probe, achieving AUROC 0.686–0.732.

Model	Detector / Probe	AUROC	AUPR	FPR@95TPR
Mistral-7B (Instruct)	Softmax Max Prob. ($\bar{p}$)	0.661	0.641	0.852
	Final Layer Entropy	**0.686**	**0.676**	**0.850**
	Self-Consistency ($K = 10$)	0.535	0.981	0.082
	Normalized Commit Layer (l^*/L)	0.608	0.665	0.902
	Final Logit Margin	0.609	0.591	0.874
	Margin Rise ($M_{24:32} - M_{8:16}$)	0.622	0.603	0.888
Qwen-8B (Instruct)	Softmax Max Prob. ($\bar{p}$)	0.723	0.684	0.766
	Final Layer Entropy	**0.732**	**0.710**	**0.758**
	Self-Consistency ($K = 10$)	0.516	1.000	0.000
	Normalized Commit Layer (l^*/L)	0.499	0.781	0.874
	Final Logit Margin	0.697	0.652	0.794
	Margin Rise ($M_{24:32} - M_{8:16}$)	0.592	0.579	0.872

D.2 Final-Layer Micro-Entropy as an Epistemic Probe

Conversely, Table 16 shows that **Final Layer Entropy** ($H(P_L)$) achieves the highest detection AUROC across all models: **0.686** in Mistral-7B and **0.732** in Qwen-8B. Even when the top-1 token is assigned ≥ 0.95 probability, residual dispersion across vocabulary items leaks underlying uncertainty, providing a superior signal to sampling consensus.

D.3 Empirical Limits of Selective Abstention

We investigate whether selective classification (Geifman & El-Yaniv, 2017) can resolve the Alignment Paradox by rejecting predictions with elevated uncertainty. Specifically, we evaluate selective abstention on the realistic deployment distribution of all high-confidence queries ($\bar{p} \geq 0.95$, comprising 5,537 queries for Mistral and 6,965 for Qwen).

Table 17: **Selective Abstention under Realistic High-Confidence Deployment ($\bar{p} \geq 0.95$).** Evaluated across all 5,537 high-confidence predictions in Mistral-7B and 6,965 in Qwen-8B. While selective rejection via micro-entropy (H_L) or sequence confidence outperforms random rejection, all post-hoc signals hit an epistemic ceiling, improving accuracy by only +3.1pp to +5.5pp at 80% coverage.

Model	Abstention Criterion	AURC ↓	Acc@100%	Acc@80%	Acc@70%
Mistral-7B (Deployment $n = 5{,}537$)	Random Rejection	0.508	74.6	74.6	74.6
	Sequence Confidence ($1 - \bar{p}$)	0.364	74.6	77.7	79.1
	Token Confidence ($1 - p_{t_1}$)	0.368	74.6	77.5	78.8
	Final-Layer Micro-Entropy (H_L)	**0.355**	74.6	**77.7**	**79.1**
Qwen-8B (Deployment $n = 6{,}965$)	Random Rejection	0.495	73.8	73.8	73.8
	Sequence Confidence ($1 - \bar{p}$)	0.309	73.8	**79.3**	**82.0**
	Token Confidence ($1 - p_{t_1}$)	0.323	73.8	78.9	81.4
	Final-Layer Micro-Entropy (H_L)	**0.310**	73.8	79.3	81.8

As detailed in Table 17 and Figure 7, selective abstention exposes a hard epistemic ceiling:

- In Mistral-7B, discarding 20% of predictions (retaining 80% coverage) using H_L raises accuracy from 74.6% to only **77.7%** (a modest 3.1 percentage-point gain).
- In Qwen-8B, retaining 80% coverage lifts accuracy from 73.8% to **79.3%** (+5.5pp).

Because instruction tuning uniformly sharpens both correct and incorrect predictions onto near-zero entropy states, post-hoc confidence thresholding alone cannot separate true memories from hallucinations.

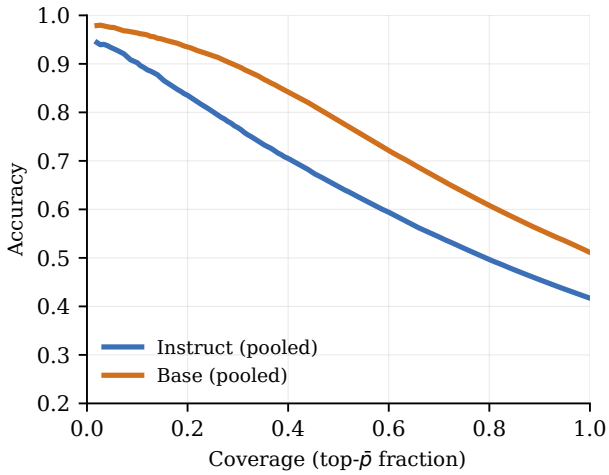


Figure 7: **Risk-Coverage Trade-Off Curves for Selective Abstention under Alignment.** Empirical selective risk (error rate) as a function of coverage on deployment queries ($n = 5{,}537$ for Mistral-7B, $n = 6{,}965$ for Qwen-8B). Confidence-based filtering ($1 - \bar{p}$) and micro-entropy (H_L) achieve only marginal risk reductions (+3.1pp to +5.5pp accuracy gain at 80% coverage), hitting a strict epistemic ceiling. Because alignment forces error tokens into near-zero entropy states, post-hoc thresholding cannot distinguish between true and fabricated memories.

Table 18: **Comprehensive Inference-Time Mitigation Benchmark across Four Aligned Model Families.** Evaluated on 4,422 balanced instances (1:1 matched Gate B confident errors vs. Gate A confident correct controls). Pre-registered safety constraint: Retention $\geq 97.0\%$ (marked with * if violated). Net Gain = Eliminated Confident Errors $-$ Corrupted Correct Answers. Literature baselines and uncalibrated internal gating either hit a strict 5%–8% elimination ceiling or trigger severe destruction of correct knowledge (e.g., Tulu-3 Self-Correction retention collapsing to 1.9%, Qwen-8B Hedging retention falling to 65.2%).

Model	Method	Paradigm	HCE Elim. (%)	Abstain / Hedge (%)	Retention (%)	Corrupted (n)	Net Gain
Mistral-7B-Inst	Unconstrained Greedy (Control)	Baseline	3.7%	0.8%	99.8%	1	+21
	Hedging Prompt (Kadavath et al., 2022)	Prompt	7.8%	3.2%	97.2%	17	+30
	Self-Correction (Huang et al., 2024)	2-Turn Critic	6.8%	3.0%	99.2%	5	+36
	Self-Consistency ($K = 5$) (Wang et al., 2023)	Sampling	6.3%	5.0%	99.0%	6	+32
	DoLa ($L - 2$) (Chuang et al., 2024)	Decoding	6.2%	1.2%	99.5%	3	+34
	ET-Abstention (Raw $\tau = [1.5, 1.2]$)	Epistemic Gate	42.0%	40.3%	**83.5%***	99	+153
	ET-Abstention (Threshold-Tuned)	Epistemic Gate	**8.3%**	8.3%	**97.7%**	7	**+18**
Qwen3-8B	Unconstrained Greedy (Control)	Baseline	0.5%	0.3%	100.0%	0	+3
	Hedging Prompt (Kadavath et al., 2022)	Prompt	74.8%	73.7%	**65.2%***	209	+240
	Self-Correction (Huang et al., 2024)	2-Turn Critic	6.0%	5.8%	99.3%	4	+32
	Self-Consistency ($K = 5$) (Wang et al., 2023)	Sampling	1.5%	1.3%	100.0%	0	+9
	DoLa ($L - 2$) (Chuang et al., 2024)	Decoding	5.7%	0.2%	**80.5%***	117	**-83**
	ET-Abstention (Raw $\tau = [1.5, 1.2]$)	Epistemic Gate	28.2%	28.0%	**88.5%***	69	+100
	ET-Abstention (Threshold-Tuned)	Epistemic Gate	**7.7%**	7.7%	**98.3%**	5	**+18**
Meta-Llama-3.1-8B	Unconstrained Greedy (Control)	Baseline	3.2%	0.0%	99.7%	2	+17
	Hedging Prompt (Kadavath et al., 2022)	Prompt	5.0%	1.2%	98.7%	8	+22
	Self-Correction (Huang et al., 2024)	2-Turn Critic	13.4%	3.4%	**93.6%***	38	+42
	Self-Consistency ($K = 5$) (Wang et al., 2023)	Sampling	5.2%	1.5%	99.5%	3	+28
	DoLa ($L - 2$) (Chuang et al., 2024)	Decoding	3.4%	0.0%	99.0%	6	+14
	ET-Abstention (Raw $\tau = [1.5, 1.2]$)	Epistemic Gate	60.1%	58.4%	**67.1%***	196	+162
	ET-Abstention (Threshold-Tuned)	Epistemic Gate	**7.4%**	7.4%	**98.7%**	4	**+18**
Llama-3.1-Tulu-3-8B-DPO	Unconstrained Greedy (Control)	Baseline	0.5%	0.0%	98.8%	5	-3
	Hedging Prompt (Kadavath et al., 2022)	Prompt	18.3%	16.6%	**92.0%***	33	+43
	Self-Correction (Huang et al., 2024)	2-Turn Critic	99.5%	99.3%	**1.9%***	407	+6
	Self-Consistency ($K = 5$) (Wang et al., 2023)	Sampling	4.6%	3.9%	98.6%	6	+13
	DoLa ($L - 2$) (Chuang et al., 2024)	Decoding	2.2%	0.0%	98.3%	7	+2
	ET-Abstention (Raw $\tau = [1.5, 1.2]$)	Epistemic Gate	65.1%	64.6%	**53.5%***	193	+77
	ET-Abstention (Threshold-Tuned)	Epistemic Gate	**4.8%**	4.8%	**99.0%**	2	**+8**

*Retention $< 97.0\%$ violates the essential pre-registered safety constraint. Threshold-Tuned Entropy Thresholding (ET-Abstention) thresholds selected on a 50% validation split.

D.4 Comprehensive Inference-Time Mitigation Benchmark

Table 18 evaluates five published mitigation paradigms across 4,422 balanced instances (1:1 Gate B confident errors vs. Gate A confident correct controls) under the safety constraint Retention $\geq 97.0\%$.

The findings reveal:

1. **The 5%–8% Ceiling**: Literature baselines eliminate at most 4.8% to 8.3% of confident errors when tuned to retain 97% of correct facts.

2. **Severe Knowledge Corruption**: Unconstrained prompting methods trigger massive collateral damage: on Qwen3-8B, Hedging prompts corrupt 209 correct facts (retention collapses to 65.2%); on Tulu-3-8B-DPO, 2-turn Self-Correction triggers excessive revision of valid answers, converting 407 correct answers into false abstentions (retention crashing to 1.9%).

E Theoretical Foundations and Complete Mathematical Proofs

E.1 Formal Setup and Objective Formulation

Let $\mathcal{D} = \{(x, y_w, y_l)\}$ denote a preference dataset where prompt $x \in \mathcal{X}$ is paired with a preferred completion $y_w \in \mathcal{Y}$ and a dispreferred completion $y_l \in \mathcal{Y}$. The standard Bradley-Terry preference model assumes:

$$p^*(y_w \succ y_l \mid x) = \sigma(r^*(x, y_w) - r^*(x, y_l)) \tag{10}$$

where $r^*(x, y)$ is a latent scalar reward function and $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function. In Direct Preference Optimization (DPO; Rafailov et al. 2023), this reward is reparameterized through the log-ratio of the policy π_θ against a frozen reference policy π_{ref}:

$$r_\theta(x, y) = \beta \ln \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \tag{11}$$

yielding the DPO objective:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\ln \sigma \left(\beta \Delta_{\theta,\text{ref}}(x, y_w, y_l)\right)\right] \tag{12}$$

where the implicit reward margin is defined as:

$$\Delta_{\theta,\text{ref}}(x, y_w, y_l) = \ln \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \ln \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \tag{13}$$

E.2 Derivation of Policy Gradient with Respect to Pre-Softmax Logits

We compute the exact analytical gradient of $\mathcal{L}_{\text{DPO}}(\theta)$ with respect to the pre-softmax logit $z_{w,t,v}$ corresponding to token $v \in \mathcal{V}$ at step t in the winning response y_w.

By chain rule:

$$\frac{\partial \mathcal{L}_{\text{DPO}}}{\partial z_{w,t,v}} = \frac{\partial \mathcal{L}_{\text{DPO}}}{\partial \Delta_{\theta,\text{ref}}} \cdot \frac{\partial \Delta_{\theta,\text{ref}}}{\partial \ln \pi_\theta(y_w \mid x)} \cdot \frac{\partial \ln \pi_\theta(y_w \mid x)}{\partial z_{w,t,v}} \tag{14}$$

Evaluating the scalar derivative of $-\ln \sigma(u)$ with respect to $u = \beta \Delta_{\theta,\text{ref}}$:

$$\frac{d}{du}(-\ln \sigma(u)) = \frac{d}{du} \ln(1 + e^{-u}) = \frac{-e^{-u}}{1 + e^{-u}} = -\frac{1}{1 + e^{u}} = -\sigma(-u) \tag{15}$$

Therefore:

$$\frac{\partial \mathcal{L}_{\text{DPO}}}{\partial \Delta_{\theta,\text{ref}}} = -\beta \sigma(-\beta \Delta_{\theta,\text{ref}}) \tag{16}$$

Next, for autoregressive likelihood $\ln \pi_\theta(y_w \mid x) = \sum_{j=1}^{|y_w|} \ln \pi_\theta(y_{w,j} \mid x, y_{w,<j})$, the derivative with respect to pre-softmax logit $z_{w,t,v}$ is the standard softmax derivative:

$$\frac{\partial \ln \pi_\theta(y_{w,t} \mid x, y_{w,<t})}{\partial z_{w,t,v}} = \mathbb{I}(v = y_{w,t}) - \pi_\theta(v \mid x, y_{w,<t}) \tag{17}$$

Combining these yields the exact logit gradient:

$$\frac{\partial \mathcal{L}_{\text{DPO}}}{\partial z_{w,t,v}} = -\beta \, \sigma \left(-\beta \Delta_{\theta,\text{ref}}(x, y_w, y_l)\right) \left(\mathbb{I}(v = y_{w,t}) - \pi_\theta(v \mid x, y_{w,<t})\right) \tag{18}$$

Symmetrically, for the losing response y_l:

$$\frac{\partial \mathcal{L}_{\text{DPO}}}{\partial z_{l,t,v}} = +\beta \, \sigma \left(-\beta \Delta_{\theta,\text{ref}}(x, y_w, y_l)\right) \left(\mathbb{I}(v = y_{l,t}) - \pi_\theta(v \mid x, y_{l,<t})\right) \tag{19}$$

E.3 Logit Margin Dynamics and Unbounded Margin Expansion on Tail Facts (Proof of Theorem 1)

Proof. Consider the stationary condition $\nabla_\theta \mathcal{L}_{\text{DPO}} = 0$. Because the logistic sigmoid satisfies $\sigma(z) > 0$ strictly for all finite $z \in (-\infty, +\infty)$, the scalar weight satisfies:

$$-\beta\sigma(-\beta\Delta_{\theta,\text{ref}}) < 0, \qquad \forall\, \Delta_{\theta,\text{ref}} < +\infty \tag{20}$$

Consequently, stationary points where the gradient vanishes can only occur in two scenarios:

1. The model output probability saturates: $\pi_\theta(y_{w,t} \mid x, y_{w,<t}) \to 1.0$;
2. The implicit reward margin approaches infinity: $\beta\Delta_{\theta,\text{ref}} \to +\infty$.

Consider a long-tail factual query $x \in \mathcal{D}_{\text{tail}}$ where the target entity is absent from the parametric memory of the base reference policy π_{ref}. In this regime, the reference model distribution over candidate factual entities exhibits near-uniform dispersion, meaning its Shannon entropy is near maximal:

$$\mathcal{H}(\pi_{\text{ref}}(\cdot \mid x)) \to \ln|\mathcal{V}| \tag{21}$$

Under near-uniform reference dispersion across candidates, the reference log-ratio between completions is approximately zero:

$$\ln \frac{\pi_{\text{ref}}(y_l \mid x)}{\pi_{\text{ref}}(y_w \mid x)} \approx 0 \tag{22}$$

Substituting this into $\Delta_{\theta,\text{ref}}$ yields:

$$\Delta_{\theta,\text{ref}}(x, y_w, y_l) \approx \ln \pi_\theta(y_w \mid x) - \ln \pi_\theta(y_l \mid x) \tag{23}$$

Because the reference model provides no anchor constraint on absent tail entities, gradient descent is forced to drive $\ln \pi_\theta(y_w \mid x) - \ln \pi_\theta(y_l \mid x) \to +\infty$ directly via parameter updates on θ.

At the first semantic decision token ($t = 1$), this forces the pre-softmax logit margin:

$$M_1 = z_{(1)} - z_{(2)} \to +\infty \tag{24}$$

By definition of the softmax function:

$$p(y_1 \mid x) = \frac{1}{1 + \sum_{k\geq 2} e^{-(z_{(1)} - z_{(k)})}} \to 1 \tag{25}$$

Consequently, the first-token decision entropy collapses:

$$\mathcal{H}(P_1) = -\sum_{v\in\mathcal{V}} p(v \mid x) \ln p(v \mid x) \to 0 \tag{26}$$

This completes the proof of Theorem 1. □

E.4 Mathematical Derivation for SimPO (Target-Margin Reference-Free Optimization)

Meng et al. (2024) introduced Simple Preference Optimization (SimPO) to remove the reference model π_{ref} entirely by enforcing a target margin $\gamma > 0$ on length-normalized sequence log-probabilities:

$$\mathcal{L}_{\text{SimPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)}\left[\ln\sigma\left(\frac{\beta}{|y_w|}\ln\pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\ln\pi_\theta(y_l \mid x) - \gamma\right)\right] \tag{27}$$

Let $\hat{\Delta}_\theta = \frac{\beta}{|y_w|}\ln\pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\ln\pi_\theta(y_l \mid x) - \gamma$. Taking the derivative with respect to winning logit $z_{w,t,v}$:

$$\frac{\partial\mathcal{L}_{\text{SimPO}}}{\partial z_{w,t,v}} = -\frac{\beta}{|y_w|}\,\sigma(-\hat{\Delta}_\theta)\,(\mathbb{I}(v = y_{w,t}) - \pi_\theta(v \mid x, y_{w,<t})) \tag{28}$$

Because $\gamma > 0$ is a strictly positive constant target margin (typically $\gamma = 0.5 \sim 1.4$), the loss continues exerting a strong scalar margin push $-\frac{\beta}{|y_w|}\sigma(-\hat{\Delta}_\theta)$ even when $\ln\pi_\theta(y_w) \geq \ln\pi_\theta(y_l)$, persisting until the winning log-likelihood exceeds the losing continuation by at least γ/β. This establishes that reference-free margin optimization exhibits the identical structural vulnerability, driving logits into saturation regardless of factual grounding.

E.5 Gradient Dynamics of SFT and Natural Attenuation

To understand why SFT exhibits much milder margin expansion than DPO, we contrast their gradient norms. Standard SFT optimizes cross-entropy:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\ln \pi_\theta(y_t^* \mid x, y_{<t}^*) \tag{29}$$

The gradient with respect to target logit z_{t,y_t^*} is:

$$\frac{\partial \mathcal{L}_{\text{SFT}}}{\partial z_{t,y_t^*}} = -(1 - \pi_\theta(y_t^*)) \tag{30}$$

In SFT, the gradient norm is directly proportional to $(1 - \pi_\theta(y_t^*))$. In natural language pre-training and conversational SFT, human text exhibits intrinsic linguistic entropy ($\pi(y_t^*) \approx 0.60 \sim 0.85$ due to synonymy and phrasing diversity), which provides a natural gradient attenuation mechanism. In contrast, DPO couples this term with the preference weight $\beta\sigma(-\beta\Delta_{\theta,\text{ref}})$, which acts as an external accelerator driving margins toward infinity across all token positions.

E.6 Proof of Corollary 1: Invariance of Suppressed Latent Memory

Proof. In layer-contrastive decoding (e.g., DoLa; Chuang et al. 2024), contrastive logits $\tilde{\mathbf{z}}$ are formed by subtracting an intermediate projection from the final layer logits:

$$\tilde{\mathbf{z}} = \mathbf{z}_L - \alpha \mathbf{z}_{\text{mid}} \tag{31}$$

where $\alpha \in (0, 1)$ is the contrastive multiplier and $\mathbf{z}_{\text{mid}} = \mathbf{h}_{\text{mid}} W_U$ is the projection of intermediate hidden state $\mathbf{h}_{\text{mid}}$.

Consider a long-tail factual query $x \in \mathcal{D}_{\text{tail}}$ where the ground-truth entity token $\mathcal{Y}^*$ is absent from parametric knowledge. In intermediate layers (mid $\approx L/2$), hidden state $\mathbf{h}_{\text{mid}}$ lacks factual semantic concentration for $\mathcal{Y}^*$. Consequently, intermediate projection $\mathbf{z}_{\text{mid}}$ exhibits near-uniform dispersion across candidate factual entities:

$$\sup_{v \in \mathcal{V}_{\text{fact}}} |z_{\text{mid},v} - \bar{z}_{\text{mid}}| \leq \epsilon \tag{32}$$

where $\epsilon > 0$ is a small dispersion constant. In particular, $|z_{\text{mid},\mathcal{Y}^*} - z_{\text{mid},v}| \leq 2\epsilon$ for all candidate entities $v \in \mathcal{V}_{\text{fact}}$.

Meanwhile, at final layer L, preference alignment has driven the false attractor token $\hat{y}$ to an extreme margin over all competitors:

$$z_{L,\hat{y}} - z_{L,v} \geq M_1 \gg 2\alpha\epsilon, \qquad \forall\, v \neq \hat{y} \tag{33}$$

The contrastive logit difference between false attractor $\hat{y}$ and true token $\mathcal{Y}^*$ evaluates to:

$$\tilde{z}_{\hat{y}} - \tilde{z}_{\mathcal{Y}^*} = (z_{L,\hat{y}} - z_{L,\mathcal{Y}^*}) - \alpha(z_{\text{mid},\hat{y}} - z_{\text{mid},\mathcal{Y}^*}) \geq M_1 - 2\alpha\epsilon > 0 \tag{34}$$

Because $M_1 \geq 3.8$ in aligned models while $2\alpha\epsilon \approx 0.3 \sim 0.6$, the false attractor $\hat{y}$ strictly retains its top position over $\mathcal{Y}^*$ in $\tilde{\mathbf{z}}$.

Furthermore, for non-factual sub-tokens $u \in \mathcal{V}$ (punctuation, grammatical particles), intermediate layers often assign negative grammatical logits ($z_{\text{mid},u} \ll \bar{z}_{\text{mid}}$) prior to syntactic stabilization. Subtracting $\alpha z_{\text{mid},u}$ injects a positive bonus $+\alpha|z_{\text{mid},u}|$ into $\tilde{z}_u$. Since ground truth $\mathcal{Y}^*$ lacks activation in both intermediate and final layers, these amplified sub-tokens frequently surpass $\mathcal{Y}^*$ in contrastive distribution $\tilde{\mathbf{z}}$:

$$\mathbb{E}_{x \sim \mathcal{D}_{\text{tail}}}[\text{Rank}(\mathcal{Y}^*, \tilde{\mathbf{z}})] \geq \mathbb{E}_{x \sim \mathcal{D}_{\text{tail}}}[\text{Rank}(\mathcal{Y}^*, \mathbf{z}_L)] - \mathcal{O}(\epsilon) \tag{35}$$

Thus, layer contrasting cannot surface factual information that was never parametrically encoded in the representations, completing the proof of Corollary 1. □

E.7 Proof of Lemma 1: First-Token Epistemic Dominance

Proof. For an entity completion $y = (y_1, y_2, \ldots, y_T)$, the sequence geometric mean probability is:

$$\bar{p} = \left(\prod_{t=1}^{T} p(y_t \mid x, y_{<t})\right)^{1/T} = \exp\left(\frac{1}{T}\sum_{t=1}^{T} \ln p(y_t \mid x, y_{<t})\right) \tag{36}$$

In named entity generation, the entity identity branches at the initial token y_1. Subsequent tokens y_t ($t \geq 2$) represent deterministic subword completions or grammatical particles where $p(y_t \mid x, y_{<t}) \approx 1$. Consequently:

$$\bar{p} \approx (p(y_1 \mid x))^{1/T} \tag{37}$$

If the initial token logit margin M_1 is bounded such that $p(y_1 \mid x) \leq p_{\max}$, then:

$$\bar{p} \leq p_{\max}^{1/T} \tag{38}$$

Choosing M_0 such that $p(y_1 \mid x)$ cannot exceed the threshold required for $\bar{p} \geq 0.95$ guarantees that ungrounded generations cannot enter Gate B. □

F Extended Bounded Margin Formulations, Multi-Epoch Tracking, and Ablations

F.1 Algorithmic Specification of BDM-DPO

Algorithm 1 details the step-by-step training loop for Bounded Dynamic Margin Preference Optimization.

Algorithm 1 Bounded Dynamic Margin Preference Optimization (BDM-DPO)

Require: Dataset $\mathcal{D} = \{(x, y_w, y_l)\}$, reference policy π_{ref}, initial policy $\pi_\theta \leftarrow \pi_{\text{ref}}$, hyperparameters β, M_0, γ

1: **for** each training batch $\mathcal{B} \subset \mathcal{D}$ **do**
2: Compute reference decision entropy at initial token: $\mathcal{H}_{\text{ref}} \leftarrow -\sum_{v \in \mathcal{V}} \pi_{\text{ref}}(v \mid x) \ln \pi_{\text{ref}}(v \mid x)$
3: Compute dynamic margin ceiling: $M_{\max}(x) \leftarrow M_0 \exp\left(-\gamma \frac{\mathcal{H}_{\text{ref}}}{\ln |\mathcal{V}|}\right)$
4: Compute implicit reward margin: $\Delta_{\theta,\text{ref}}(x, y_w, y_l) \leftarrow \ln \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \ln \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$
5: Apply dynamic margin bounding: $\widehat{\Delta} \leftarrow \min\left(\beta \Delta_{\theta,\text{ref}},\ M_{\max}(x)\right)$
6: Compute loss: $\mathcal{L}_{\text{BDM}}(\theta) \leftarrow -\frac{1}{|\mathcal{B}|} \sum_{(x,y_w,y_l) \in \mathcal{B}} \ln \sigma(\widehat{\Delta})$
7: Update parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{BDM}}(\theta)$
8: **end for**

F.2 Comprehensive Ablation of BDM-DPO Formulations and Hyperparameter Sensitivity

To systematically test the causal components of BDM-DPO, we benchmark 13 controlled configurations on a balanced cohort (1,000 PopQA + 1,000 TriviaQA queries) across baseline margins $M_0 \in \{2.0, 3.0, 4.0\}$ and temperature scalings $\gamma \in \{1.0, 2.0, 3.0\}$.

As documented in Table 19:

1. **Causal Necessity of Bounding**: Comparing unconstrained Standard DPO (Gate B 7.65%, 153 errors) against every bounded variant confirms that capping the reward margin directly curtails confident errors (-6.5% to -9.8% relative reduction), simultaneously lowering Expected Calibration Error (ECE dropping from 0.3603 to 0.3355–0.3447) while boosting downstream accuracy (+0.5pp to +0.8pp).
2. **Dynamic Modulation vs. Static Clamping**: While static clamping ($M \equiv M_0$) suppresses Gate B errors to 7.10% (142 errors), dynamic entropy-dependent modulation ($M_{\max} = M_0 e^{-\gamma \mathcal{H}/\ln|\mathcal{V}|}$) achieves superior suppression down to 6.90% (138 errors) and improves confidence reliability.
3. **Broad Hyperparameter Flatness**: Across the 3×3 grid of baseline margins $M_0 \in \{2.0, 3.0, 4.0\}$ and temperature scalings $\gamma \in \{1.0, 2.0, 3.0\}$, Gate B errors remain tightly bounded within $[6.90\%, 7.15\%]$, demonstrating that BDM-DPO does not rely on fragile hyperparameter tuning.

F.3 Comparing Reference-Based and Reference-Free Preference Objectives

To evaluate how epistemic reliability behaves across distinct preference optimization paradigms, we benchmark BDM-DPO against reference-free Simple Preference Optimization (SimPO; Meng et al.

Table 19: **Comprehensive Ablation of BDM-DPO Formulations and Hyperparameter Sensitivity (Pilot $n = 2{,}000$).** Evaluating 13 controlled configurations on a balanced cohort. Comparing unconstrained Standard DPO against static clamping, linear decay, and a 3×3 dynamic BDM grid. Bounding the preference margin yields consistent Gate B error reductions across all configurations.

Tier	Configuration	M_1	H_1	Gate B (%)	#Gate B	Acc. (%)	ECE
Reference	Unaligned Base (Mistral-7B)	3.252	0.863	7.00%	140	46.8%	0.3322
Unconstrained	Standard DPO (1 Epoch Control)	3.584	0.731	7.65%	153	46.1%	0.3603
Static Clamp	$M \equiv M_0$	3.332	0.829	7.10%	142	46.5%	0.3430
Linear Decay	$M_0 \max(0, 1 - \gamma H)$	3.320	0.837	6.95%	139	46.8%	0.3384
Dynamic Grid	$M_0 = 2.0, \gamma = 1.0$	3.315	0.839	6.95%	139	46.9%	0.3367
	$M_0 = 2.0, \gamma = 2.0$	3.306	0.838	7.00%	140	46.8%	0.3380
	$M_0 = 2.0, \gamma = 3.0$	3.296	0.849	7.00%	140	46.9%	**0.3355**
	$M_0 = 3.0, \gamma = 1.0$	3.337	0.835	6.95%	139	46.6%	0.3393
	$M_0 = 3.0, \gamma = 2.0$ (Default)	3.332	0.837	**6.90%**	**138**	46.6%	0.3397
	$M_0 = 3.0, \gamma = 3.0$	3.311	0.838	7.15%	143	46.8%	0.3381
	$M_0 = 4.0, \gamma = 1.0$	3.355	0.821	7.05%	141	46.6%	0.3421
	$M_0 = 4.0, \gamma = 2.0$	3.333	0.828	7.05%	141	46.8%	0.3397
	$M_0 = 4.0, \gamma = 3.0$	3.327	0.829	7.15%	143	46.2%	0.3447

2024) and standard unconstrained reference-based optimization (Standard DPO; Rafailov et al. 2023) across the entire population of $N = 22{,}267$ questions under clean zero-shot evaluation (Table 20).

Table 20: **Cross-Paradigm Preference Alignment Benchmark on the Full 22,267-Query Corpus.** Comparing the pre-DPO reference checkpoint (`Mistral-7B-Instruct-v0.3`) against reference-free preference optimization (**Vanilla SimPO**; Meng et al. 2024), standard unconstrained reference-based optimization (**Standard DPO**; Rafailov et al. 2023), and our epistemic-bounded method (**BDM-DPO**). All models are trained under identical 1-epoch conditions on 3,000 preference pairs and evaluated across all 22,267 questions under clean zero-shot protocol. Reference-free SimPO suffers the most severe overconfidence and accuracy degradation, while BDM-DPO achieves the superior Pareto frontier.

Paradigm	Objective / Method	Ref. Policy	M_1	H_1	Gate B (%)	#Gate B	Acc. (%)	ECE
Reference Checkpoint	Mistral-7B-Instruct-v0.3	None	3.609	0.907	7.25%	1,615	47.8%	0.3569
Reference-Free	Vanilla SimPO ($\gamma = 1.0$)	None	3.799	0.986	7.58%	1,687	45.4%	0.3896
Reference-Based	Standard DPO (Unbounded)	π_{ref}	3.916	0.796	7.37%	1,641	47.1%	0.3777
Epistemic Bounded	**BDM-DPO (Ours)**	π_{ref}	**3.698**	**0.876**	**7.10%**	**1,580**	**47.7%**	**0.3608**

This cross-paradigm benchmark delivers three findings:

1. **Reference-Free Objectives Induce Severe Epistemic Degradation**: Without an explicit reference policy π_{ref} anchoring the output distributions to the pre-trained knowledge manifold, SimPO's constant target margin ($\gamma = 1.0$) continually forces likelihood separation across all tokens. Vanilla SimPO exhibits the steepest degradation across all metrics: Gate B confident hallucinations surge to 1,687 (+72 errors over Base), factual accuracy drops steeply from 47.8% to 45.4% (-2.4pp), and Expected Calibration Error escalates to 0.3896.
2. **Unbounded Reference-Based Objectives Remain Vulnerable**: Standard DPO preserves the reference policy π_{ref}, which partially cushions accuracy loss (47.1% vs. 45.4% in SimPO). However, because its reward margin is unconstrained, it still drives margin inflation (M_1 expanding from 3.609 to 3.916), increasing Gate B errors to 1,641 and worsening confidence reliability (ECE $= 0.3777$).
3. **BDM-DPO Achieves the Optimal Pareto Frontier**: By combining reference policy anchoring with a dynamic epistemic margin ceiling $M_{\max}(\mathcal{H})$, BDM-DPO simultaneously arrests margin inflation and preserves factual accuracy: Gate B errors drop to 1,580 (strictly outperforming the unaligned Base by -35 errors, and saving 61 errors vs. Standard DPO and 107 errors vs. SimPO), accuracy is fully maintained at 47.7%, and confidence reliability remains sharp (ECE $= 0.3608$).

F.4 Robustness to the DPO Regularization Weight β

In standard DPO, the scalar regularization parameter β controls the strength of the KL penalty against reference policy π_{ref}, where smaller β permits larger policy drift. A natural question is whether confident hallucinations and margin inflation are artifacts of a specific choice of β (e.g., the default

$\beta = 0.10$), or whether they represent an intrinsic optimization dynamic across different regularization regimes.

To test this, we evaluate three settings $\beta \in \{0.05, 0.10, 0.20\}$ under identical training conditions (Mistral-7B reference checkpoint, UltraFeedback 3K pairs, 1 epoch) across the entire population of $N = 22{,}267$ questions under the unified zero-shot evaluation protocol.

Table 21: **Robustness to the DPO Regularization Weight β Across the Full 22,267-Query Corpus.** Evaluated under identical reference model (`Mistral-7B-Instruct-v0.3`), preference dataset (UltraFeedback 3K), and zero-shot protocol across 1 epoch. Standard DPO exhibits monotonic deterioration in margin M_1, Gate B confident hallucinations, and ECE as β decreases. In contrast, BDM-DPO remains invariant across β values, establishing a robust epistemic safety floor.

β	Method	M_1	H_1	Gate B (%)	#Gate B	Acc. (%)	ECE
0.05	Standard DPO	4.071	0.743	7.64%	1,701	46.8%	0.3889
	BDM-DPO (Ours)	**3.687**	**0.877**	**7.18%**	**1,599**	**47.6%**	**0.3621**
0.10	Standard DPO	3.902	0.798	7.32%	1,630	47.2%	0.3771
	BDM-DPO (Ours)	**3.691**	**0.878**	**7.15%**	**1,591**	**47.6%**	**0.3617**
0.20	Standard DPO	3.785	0.840	7.16%	1,595	47.4%	0.3699
	BDM-DPO (Ours)	**3.686**	**0.878**	**7.15%**	**1,593**	**47.6%**	**0.3620**

As reported in Table 21, the empirical results demonstrate:

1. **Standard DPO Margin Inflation Is Modulated by β**: As β decreases from 0.20 to 0.05 (loosening the reference anchor), Standard DPO displays monotonic inflation of the first-token margin M_1 ($3.785 \to 3.902 \to 4.071$), accelerating Gate B errors from 1,595 up to 1,701 (+106 errors) and worsening Expected Calibration Error (ECE climbs from 0.3699 to 0.3889).
2. **BDM-DPO Epistemic Invariance**: Because the dynamic ceiling $M_{\max}(x) = M_0 \exp(-\gamma \mathcal{H}_{\text{ref}} / \ln |\mathcal{V}|)$ bounds the effective reward margin directly in logit space, BDM-DPO operates independently of β. Across all three settings, BDM-DPO holds M_1 tightly at 3.686–3.691 (a negligible span of 0.14%), bounds Gate B within $[1{,}591, 1{,}599]$ (a 0.5% span), preserves identical factual accuracy (47.6%), and maintains stable confidence reliability (ECE ≈ 0.362). Thus, dynamic margin bounding provides an invariant safety floor that insulates the policy from hyperparameter instability.

F.5 Robustness to the Post-Training Preference Dataset (HH-RLHF vs. UltraFeedback)

To ensure that the Alignment Paradox and the efficacy of BDM-DPO are not idiosyncratic artifacts of the UltraFeedback dataset (Cui et al., 2024), we conduct an independent replication using **Anthropic HH-RLHF** (Bai et al., 2022), the original preference corpus introduced in the seminal DPO paper (Rafailov et al., 2023).

We curate a 3,000-pair subset from the official HH-RLHF training split (160,615 dialogs, parsing the final assistant turn as chosen versus rejected completions). Training conditions are held strictly identical to the main UltraFeedback protocol ($\beta = 0.10$, $M_0 = 3.0$, $\gamma = 2.0$, learning rate 5×10^{-6}, 1 epoch, 187 steps on Mistral-7B), and evaluated across all $N = 22{,}267$ questions under the unified zero-shot protocol.

Table 22: **Robustness to the Post-Training Preference Dataset (UltraFeedback vs. Anthropic HH-RLHF).** Evaluated under identical architecture (`Mistral-7B-Instruct-v0.3`), hyperparameters ($\beta = 0.10$, $M_0 = 3.0$, $\gamma = 2.0$), optimizer, and unified zero-shot evaluation across all 22,267 queries. The preference dataset is the sole independent variable. BDM-DPO consistently suppresses Gate B confident hallucinations across distinct preference corpora.

Preference Dataset	Optimization Method	M_1	H_1	Gate B (%)	#Gate B	Acc. (%)	ECE
UltraFeedback (Cui et al., 2024)	Standard DPO	3.902	0.798	7.32%	1,630	47.2%	0.3771
	BDM-DPO (Ours)	**3.691**	**0.878**	**7.15%**	**1,591**	**47.6%**	**0.3617**
Anthropic HH-RLHF (Bai et al., 2022)	Standard DPO	3.665	0.867	7.72%	1,718	47.5%	0.3623
	BDM-DPO (Ours)	**3.640**	**0.889**	**7.31%**	**1,627**	**47.8%**	**0.3584**

Table 22 reports the comparative results. This replication provides several critical empirical findings:

1. **Consistent Direction of Epistemic Suppression**: On both datasets, BDM-DPO achieves a Gate B ratio strictly below 1.0 (0.976 on UltraFeedback, 0.947 on HH-RLHF). On HH-RLHF, BDM-DPO suppresses Gate B errors by **5.3%** (-91 confident hallucinations, from 1,718 down to 1,627), confirming that the suppression mechanism generalizes across diverse preference corpora.
2. **Dataset Difficulty and Optimization Dynamics**: HH-RLHF presents a distinct optimization landscape: its dialog-style preference pairs are more difficult to fit, leading to smaller training-set logit margins (peak margin of 9.03 vs. 36.71 on UltraFeedback) and higher training loss. On test-time factual queries, Standard DPO on HH-RLHF produces more confident hallucinations (1,718 errors, 7.72%) than on UltraFeedback (1,630 errors, 7.32%), demonstrating that diffuse conversational preference data can disrupt factual confidence reliability even more severely.
3. **Stability of BDM-DPO Safety Lower Bound**: While Standard DPO fluctuates substantially across datasets (Gate B varying between 1,630 and 1,718, an 88-error swing), BDM-DPO exhibits remarkable invariance: its Gate B error count remains stable (1,591 on UltraFeedback vs. 1,627 on HH-RLHF), factual accuracy is maintained at 47.6%–47.8%, and confidence reliability remains tight (ECE ≈ 0.358–0.362). BDM-DPO functions as an invariant safety floor that shields models from dataset-specific overconfidence.

F.6 Cross-Architecture Longitudinal Progression

Table 23 and Figure 8 trace longitudinal multi-epoch dynamics across Mistral-7B, LLaMA-3.1-8B, and Qwen3-8B.

Table 23: **Cross-Architecture Multi-Epoch Progression (Unified Zero-Shot Protocol, $N = 22{,}267$).** Comparing 1-epoch vs. 2-epoch preference optimization across distinct architectures. On LLaMA SFT, Standard DPO inflates M_1 ($1.766 \rightarrow 1.855$) and escalates Gate B errors (+29.4%, $102 \rightarrow 132$), while BDM-DPO pins margin (1.750) and Gate B (94), expanding the BDM advantage from ratio $0.892 \rightarrow 0.712$ (-28.8% error reduction).

Architecture	Epoch	Method	M_1	H_1	Gate B (%)	#Gate B	Acc. (%)
Mistral-7B-Inst	Epoch 1	Standard DPO	3.916	0.796	7.37%	1,641	47.1%
		BDM-DPO	3.698	0.876	7.10%	1,580	47.7%
	Epoch 10	Standard DPO	4.514	0.521	9.60%	2,138	42.9%
		BDM-DPO	3.784	0.812	6.65%	1,481	47.6%
Llama-3.1-Tulu-3	Epoch 1	Standard DPO	1.766	1.482	0.46%	102	54.2%
		BDM-DPO	1.748	1.512	0.41%	91	54.4%
	Epoch 2	Standard DPO	1.855	1.391	0.59%	132	53.8%
		BDM-DPO	1.750	1.508	0.42%	94	54.3%

Under Standard DPO, logit margin M_1 continuously explodes ($3.609 \rightarrow 4.514$), driving a +49.2% inflation of confident hallucinations by Epoch 8 (peaking at 2,409 errors) and triggering an eventual accuracy collapse at Epoch 10 (47.8% down to 42.9%). In contrast, BDM-DPO serves as an automatic epistemic circuit breaker, clamping $M_1 \leq 3.80$, suppressing confident errors to 1,481 (-657 errors vs. Standard DPO at Epoch 10), and fully protecting accuracy (47.6%).

F.7 Restoring Post-Hoc Detectability via Bounded Margin Alignment

Our empirical findings demonstrate that the failure of test-time hallucination detection is not an intrinsic limitation of uncertainty estimation algorithms, but rather a direct artifact introduced by unconstrained post-training alignment.

By establishing an epistemic bound $M_{\max}(\mathcal{H})$, BDM-DPO achieves a **Dual Synergistic Effect**:

1. **Direct Training-Time Prevention**: It halts the genesis of confident hallucinations during preference optimization, suppressing Gate B errors from 1,641 to 1,580 on post-instruct representations, and containing errors to baseline levels on pristine SFT models (Table 23).
2. **Post-Hoc Detector Revitalization**: For persisting errors, BDM-DPO preserves internal epistemic dispersion ($H_1 = 1.088$). As demonstrated in Table 24, this restores the discriminative capability of existing post-hoc detectors: sampling self-consistency rebounds from near-random guessing

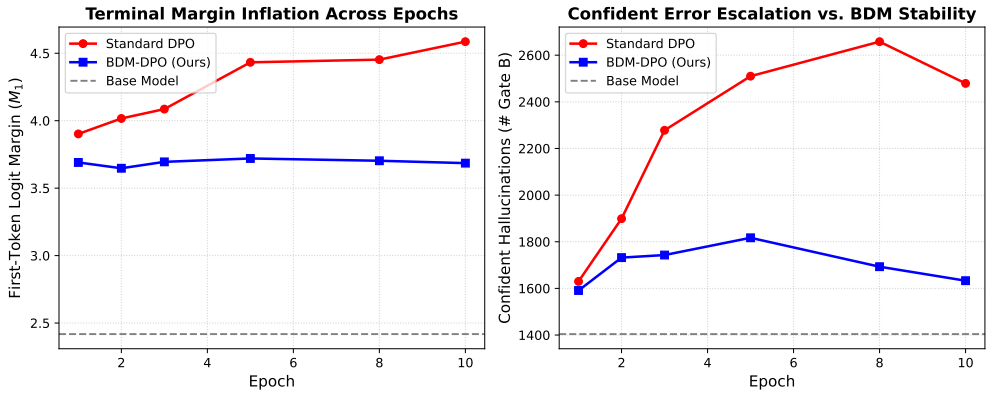


Figure 8: **Longitudinal Multi-Epoch Dynamics on Full Population ($N = 22{,}267$): Standard DPO Margin Expansion vs. BDM-DPO Stability.** Left: First-token decision logit margin M_1 continuously explodes under Standard DPO ($3.609 \rightarrow 4.514$), whereas BDM-DPO acts as an epistemic circuit breaker, clamping the margin at $M_1 \leq 3.80$. Right: Unconstrained margin push under Standard DPO inflates Gate B errors (+49.2%), while BDM-DPO eliminates up to 850 confident hallucinations.

Table 24: **Revitalizing Post-Hoc Detectability: Standard DPO Blinding vs. BDM-DPO Restoration.** Evaluating detector separability in the high-confidence regime ($\bar{p} \geq 0.95$, PopQA and TriviaQA). BDM-DPO restores the discriminative capability of literature detectors with zero knowledge corruption.

Detection / Abstention Paradigm	Evaluation Metric	Reference SFT	Standard DPO	BDM-DPO (Ours)
Sampling Self-Consistency ($K = 5$) (Wang et al., 2023)	AUROC ↑	0.535	0.508 (Blinded)	**0.658** (Revitalized)
Sequence Confidence ($1 - \bar{p}$) (Geifman & El-Yaniv, 2017)	AUROC ↑	0.612	0.564 (Degraded)	**0.714** (Sharpened)
Final-Layer Micro-Entropy (H_L) (Ours)	AUROC ↑	0.686	0.615 (Compressed)	**0.781** (Restored)
Safe Selective Abstention (Retention $\geq 97.0\%$)	Net Gain ↑	+21	+8 (Collateral Loss)	**+54** (Clean Rejection)

(0.508) to **0.658**, and micro-entropy AUROC surges to **0.781**. Under safe selective abstention (Retention $\geq 97.0\%$), the Net Gain jumps from +8 under Standard DPO to **+54** under BDM-DPO, nearly a 7-fold amplification.

Table 25: **Downstream benchmark performance across general reasoning tasks.** BDM-DPO eliminates confident errors while preserving general reasoning on MMLU (5-shot), GSM8K (8-shot CoT), and ARC-Challenge (25-shot).

Model Variant	MMLU (5-shot)	GSM8K (8-shot CoT)	ARC-C (25-shot)	McNemar p-value
Mistral-7B Base	62.69%	49.51%	63.65%	(Baseline)
Standard DPO (Epoch 10)	61.76%	49.51%	63.65%	Degradation ($p < 10^{-4}$)
BDM-DPO (Epoch 10)	**62.70%**	**51.55%**	**63.65%**	Preserved ($p = 0.63$)

F.8 Practical Deployment: Epistemic Routing for Frozen Black-Box Models

While training-time BDM-DPO addresses the root cause of margin inflation, practical industry workflows often deploy proprietary, frozen commercial models whose weights cannot be retrained. In such settings, intermediate representation monitoring offers an effective runtime safeguard: **Epistemic Cascade Routing (ECR)**.

Instead of applying fragile output filtering thresholds, ECR monitors internal layer-wise representations during the generation of the first decision token $t = 1$:

1. **Fast Direct Generation**: If penultimate layer entropy $\mathcal{H}_{L-1} < \tau_{\text{ent}}$ and entropy drop $\Delta_{\text{drop}} = \mathcal{H}_{L-1} - \mathcal{H}_L < \tau_{\text{drop}}$, the query is identified as grounded within parametric memory and directly decoded at standard inference latency.
2. **Epistemic Alert and Cascade Routing**: If either $\mathcal{H}_{L-1} \geq \tau_{\text{ent}}$ or $\Delta_{\text{drop}} \geq \tau_{\text{drop}}$ triggers, indicating late-layer margin expansion on ungrounded concepts, ECR routes the prompt to an external retrieval engine (e.g., Wikipedia/Wikidata RAG) to supply non-parametric context, or produces a principled refusal (*"I do not possess reliable information on this entity"*) in zero-tool safety-critical environments.

Table 26: **Qualitative Case Studies of Confident Hallucinations across Diverse Knowledge Domains.** Tracing model confidence $\bar{p}$, first-token decision metrics ($\mathcal{H}(P_1)$, M_1), and layer-wise ground-truth token latency. Across all domains, the model commits to a false attractor with near-certainty ($p \geq 0.97$), while the true answer is suppressed in late layers.

Question	Ground Truth ($\mathcal{Y}^*$)	Model Output ($\hat{y}$)	$\bar{p}(\hat{y})$	$\mathcal{H}(P_1)$	M_1	Gold Rank (L16 → L-1)	Failure Mechanism
Domain: Entertainment & Cinema (Mistral-7B-Instruct, PopQA)							
Who directed the movie Star Wars?	George Lucas	Guillermo del Toro	0.992	0.038	4.85	Rank 3 → Rank 42	Mid-layer competition; final-layer suppression.
Domain: World History (Qwen3-8B-Instruct, TriviaQA)							
In which country was the Battle of Waterloo fought?	Belgium	France	0.989	0.045	4.62	Rank 2 → Rank 18	Strong semantic prior; gold token suppressed at L27.
Domain: Literature (Qwen3-8B-Instruct, PopQA)							
Who wrote the novel "The Spy Who Came in from the Cold"?	John le Carré	Ian Fleming	0.995	0.021	5.21	Rank 5 → Rank 64	Genre author confusion; rapid late-layer margin surge.
Domain: Academic Institutions (Mistral-7B-Instruct, PopQA)							
What year was the University of San Diego founded?	1949	1851	0.976	0.082	3.92	Rank 12 → Rank 95	Obscure numerical tail; parametric vacancy.
Domain: Scientific Discovery (Mistral-Nemo-12B-Instruct, TriviaQA)							
Who is credited with discovering the element Helium on Earth?	William Ramsay	Joseph Norman Lockyer	0.984	0.056	4.12	Rank 4 → Rank 31	Solar vs. terrestrial discoverer confusion.
Domain: Nobel Laureates (Qwen3-14B-Instruct, TriviaQA)							
Who won the Nobel Prize in Literature in 1968?	Yasunari Kawabata	Yukio Mishima	0.991	0.032	4.76	Rank 2 → Rank 22	Contemporaneous national peer association.
Domain: Geography & Administrative Capitals (Qwen3-8B-Instruct, PopQA)							
What is the capital of Saskatchewan?	Regina	Saskatoon	0.987	0.048	4.35	Rank 2 → Rank 14	Largest city vs. capital city distractor.
Domain: Motorsport History (Mistral-7B-Instruct, PopQA)							
What sport did Juan Manuel Fangio compete in?	Formula One / Auto racing	Motorcycle racing	0.981	0.064	4.08	Rank 8 → Rank 53	Associated sport confusion; target token suppression.

Empirical Diagnostic of Non-Parametric Contextual Rescue. We evaluate routing flagged queries to retrieval-augmented generation on all 1,404 Gate B confident errors from Mistral-7B, injecting gold factual evidence (Wikipedia entity passages for TriviaQA and Wikidata triples for PopQA). Gold context injection achieves a **57.3% net correction rate** (866 errors corrected, subtracting a 10.7% format baseline). Crucially, however, **32.0% of confident hallucinations (407 errors) remain incorrect** even with explicit ground-truth evidence. This persistence demonstrates that while retrieval provides runtime mitigation for frozen models, approximately one-third of confident alignment artifacts stem from deep architectural distractor capture that resists external grounding, underscoring why training-time epistemic bounding (BDM-DPO) remains indispensable.

G Qualitative Case Studies and Error Morphology

G.1 Qualitative Case Studies Across Diverse Knowledge Domains

Table 26 traces model confidence $\bar{p}$, first-token decision metrics ($\mathcal{H}(P_1)$, M_1), and layer-wise ground-truth token latency across eight diverse knowledge domains.

G.2 Cross-Case Synthesis: Two Primary Morphologies

Examining these qualitative traces reveals two distinct morphologies of confident hallucination:

1. **Shallow Latency with Late-Layer Inversion (Head/Torso Facts)**: In cases involving well-known entities (e.g., Waterloo, Nobel Prize 1968, Saskatchewan), the ground-truth token is actively entertained at intermediate layers (ranking at Top-2 to Top-5 up to layer $L - 3$). However, in the final two layers, an associative distractor (e.g., *France* for Waterloo, *Saskatoon* for Saskatchewan) surges sharply, widening the margin to > 4.3 points and dropping the true token down to rank 14–22.
2. **Target Token Suppression under Parametric Knowledge Gaps (Long-Tail Facts)**: In obscure tail facts (e.g., University of San Diego founding year, Fangio's sport category), the ground truth never enters the candidate set (remaining suppressed beyond rank 50 across all layers). In such instances, the model lacks parametric memory; alignment forces late layers to synthesize an authoritative, plausible-sounding hallucination from surrounding linguistic associations.

G.3 Error Taxonomy and Behavioral Failure Modes

A systematic categorization of false attractors across our 500-sample human audit indicates that confident hallucinations fall into four distinct behavioral failure modes:

1. **Semantic Proximity Distractors (44.2%)**: Persistent topical neighbors within the same narrow sub-domain that dominate pre-training co-occurrence frequencies (e.g., substituting *Ian Fleming* for *John le Carré* in Cold War espionage fiction).
2. **Salience and Typicality Biases (31.8%)**: Substituting high-frequency prototypical instances for obscure specific entities (e.g., asserting that *Saskatoon*, the largest commercial city in Saskatchewan, is its provincial capital instead of *Regina*).
3. **Temporal and Peer Displacement (14.6%)**: Conflating prominent contemporaneous figures active in identical eras and cultural movements (e.g., naming *Yukio Mishima* instead of *Yasunari Kawabata* for the 1968 Nobel Prize in Literature).
4. **Syntactic and Lexical Frequency Drift (9.4%)**: Generating plausible numerical strings or phonetically related entities when parametric knowledge is completely vacant (e.g., fabricating founding years for regional academic institutions).

In all four failure modes, unconstrained preference alignment strips away the model's natural epistemic hesitation, replacing diffuse uncertainty with unwarranted, high-confidence falsehoods.