The Extrapolation Cliff in On-Policy Distillation of Near-Deterministic Structured Outputs

Xin Li **Hao Jiang** **Annan Wang** **Yichi Zhang** **Chau Yuen**
Nanyang Technological University, Singapore
https://lixin.ai/ListOPD

Abstract

On-policy distillation (OPD) is widely used for LLM post-training. When pushed with a reward-extrapolation coefficient $\lambda > 1$, the student can lift past the teacher in domain, but past a threshold $\lambda^\star$ the same step violates the output contract on structured-output tasks. In a single-position Bernoulli reduction, we derive a closed-form base-relative clip-safety threshold $\lambda^\star(p, b, c)$ determined by three measurable quantities: the teacher modal probability, the warm-start mass, and the importance-sampling clip strength. Above $\lambda^\star$ the extrapolated fixed point exits the clip-safe region, changing training from format-preserving to format-collapsing. We extend the rule to calibrated K-ary listwise JSON tasks where a single binding equivalence class dominates the output contract and SFT retains parse headroom. On Amazon Fashion, three pre-registered tests (a fine-grid cliff interval, a budget-extension test, and a small-clip cross-prediction) all fall within their locked prediction windows, with the small-clip value matching the closed-form prediction below grid resolution. Operating just below $\lambda^\star$, ListOPD brings a 1.7B Qwen3 student to in-domain parity with an 8B-SFT baseline (pre-registered 3-seed) at one-fifth the parameters. The gain is driven primarily by format adherence: NDCG@1 on parsed outputs remains flat across λ, while parse validity sharply changes at the predicted boundary. The cliff diagnostic is rubric-independent, whereas the parity claim uses a Gemini-graded rubric and inherits that evaluator's exposure.

1 Introduction

On-policy distillation (OPD) trains a student LLM against a teacher's per-token log-probabilities on the student's own rollouts [2, 13]; its reward-extrapolation variant [42] sharpens the on-policy target by a coefficient $\lambda > 1$ and can lift the student past the teacher in domain. But the same extrapolation step that produces the lift, past a threshold $\lambda^\star$, instead replaces format-preserving training with a sharp contract collapse on structured-output tasks [11, 38]. We derive that threshold in closed form and calibrate it on Amazon product-review listwise ranking.

On Amazon's product-review domain [15, 17], listwise rerankers [28, 30, 35] emit, for each *product group* of $K{=}8$ reviews, a JSON list of K objects keyed by the input `review_id`s, each carrying a helpfulness score under a fixed rubric. The contract collapse above is the format-adherence failure documented qualitatively for structured-output LLMs [9, 12, 43]: the model scores plausibly but the outer scaffold truncates or duplicates ids. A Qwen3-1.7B-SFT student satisfies the contract on 33.5% of Fashion groups; trained with *ListOPD*, our extrapolated-OPD listwise instantiation, the same student reaches 94.8%, with rank quality on parsed outputs unchanged under a fixed Gemini rubric [7]: Fashion is a controlled scaffold for contract-adherence mechanics, not a semantic-ranking claim against existing rerankers.

The knob is sharp because clip-safety has a boundary. At a structural token with teacher modal mass p, the extrapolation step sharpens the target to a fixed point p_λ; once its off-modal mass falls below

Preprint.

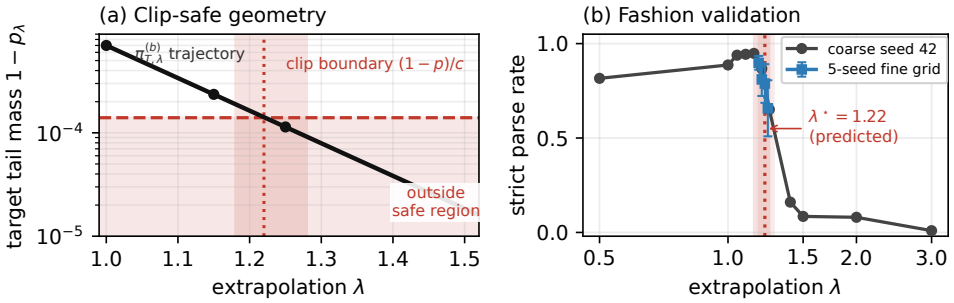

Figure 1: **The extrapolation cliff in miniature.** *Left:* the IS-clip-safe geometry; the sharpened fixed point exits at the base-neutral marker $\lambda^\star(p_{\text{typ}}{=}0.9993, c{=}5){=}1.22$ (the full base-relative prediction bracket $[\lambda^\star_{\text{safe}}, \lambda^\star_{\text{typ}}] = [1.18, 1.28]$ at $b{=}0.81$ is in Tab. 1). *Right:* strict parse rate on Fashion $K{=}8$ listwise (Qwen3 1.7B×4B, $N{=}212$) collapses in the same $[1.18, 1.28]$ band.

the clipped tail mass $(1{-}p)/c$ enforced by GRPO-style importance-sampling (IS) clipping [31, 38], the fixed point exits the clip-safe region (Fig. 1, left). The base-neutral crossing is

$$\lambda^\star(p, c) \;=\; \frac{\log\big((1-p)/(c-1+p)\big)}{\log\big((1-p)/p\big)}, \tag{1}$$

while the full theorem is base-relative and the sequence-level lift requires position-wise parametric reach. In Fashion, the measured structural-token confidence and clip put the marker at the observed cliff scale within one λ-grid step (Tab. 1).

This framing turns OPD tuning from a post-hoc λ sweep into a falsifiable boundary-prediction problem: the predicate either places the cliff at the predicted scale, or it shifts, abstains, or fails to localize, with each outcome scoped in Tab. 1.

Operating below the threshold, 1.7B-ListOPD lifts deployment-useful score USEFUL=`parse` × NDCG@1 (zero credit on parse failure) from 0.23 to 0.86, matching a pre-registered 3-seed 8B-SFT baseline within combined seed noise (Tab. 3). Against best constrained-SFT plus permutation repair [4, 10, 39] the training-side residual is $+0.051$ USEFUL, the claim we make rather than categorical superiority over constrained decoding.

Our contributions:

1. **A closed-form clip-safety predicate practitioners can compute.** $\lambda^\star(p, b, c)$, the base-relative clip-safe threshold of OPD extrapolation, follows from three measurable quantities. We prove the single-position Bernoulli version (Thm. 4.1) and give an explicit sequence-level instantiation (Thm. 4.3); the multi-token lift is exact under off-modal-ratio invariance and approximate otherwise, and super-critical dynamics are finite-budget empirical, not an almost-sure convergence theorem (Thm. 4.2).
2. **Three pre-registered Fashion prediction matches, including a within-grid-resolution cross-clip hit.** The Fashion $K{=}8$ binding class ($K{-}1 \to K$ transition) anchors the calibration: a 5-seed fine grid localizes the cliff onset to $[1.204, 1.228]$ around the predicted 1.22; an $N{=}200$ budget extension lands inside its locked $[1.00, 1.10]$ bracket; a $c{=}1.5$ cross-clip extension matches its locked closed-form $\lambda^\star_{\text{typ}}{=}1.070$ at observed midpoint 1.069, below the experimental grid resolution. ASPO follows the same cliff pattern at one grid step earlier (App. G), supporting a mechanism-not-method reading.
3. **A deployment rule and a scoped evaluation.** Operating just below $\lambda^\star$, ListOPD reaches in-domain parity with a pre-registered 3-seed 8B-SFT baseline at one-fifth the parameters; per-task results (Tab. 1) document where the predicate shifts, abstains, or loses power.

2 Related Work

Distillation and on-policy RL. On-policy distillation with reverse-KL objectives [2, 13, 14] sharpens a student against a teacher under student-sampled trajectories. We extend the ExOPD reward-extrapolation formulation [42] from reasoning to listwise structured-output ranking, and identify the IS-clip-asymmetry mechanism whose engineering side is mitigated by ASPO [38]: ASPO identifies the same IS-asymmetry on positive-advantage tokens and proposes a training-time ratio-flip fix, whereas we derive the closed-form $\lambda^\star(p, b, c)$ at which the extrapolated fixed point exits the clip-safe region and quantify the regime where extrapolated OPD is and is not safe. A 4-seed empirical head-to-head with ASPO on Fashion 1.7B$\times$4B (App. G) shows ASPO is comparable at λ=1.0 and collapses at λ=1.5 under the same protocol, ruling out a narrow GRPO-implementation artifact and supporting the mechanism-driven clip-threshold reading. Li et al. [23] characterise the modal-token concentration regime ($p_{\text{eff}} \geq 0.99$) that we exploit; our p_{eff} aggregator quantifies their phenomenology and turns it into a calibration target. Three orthogonal OPD failure modes appear in Fu et al. [11]; complementary analyses are in Jang et al. [16], Kim et al. [19], Ko et al. [20, 21], Song and Zheng [34], Xu et al. [41]. None characterize the λ-axis cliff or the IS-clip boundary itself.

Format adherence and listwise ranking. Structured-output brittleness has motivated constrained decoding [4, 10, 29, 37, 39], benchmarks distinguishing structural from semantic violations [12, 36], and direct schema-RL [1, 24]. Yun et al. [43] document SFT-side diversity collapse under format-induced training; our cliff is the on-policy analogue, sharpened to a closed-form boundary in λ. LLM-based listwise rerankers [25, 26, 28, 30, 35, 44] factorize over Plackett–Luce permutations [5, 6, 27, 40]; closest in domain, Jiang et al. [17] apply RL to a related Amazon listwise review-ranking dataset. Deng et al. [9] observe that task-solving and formatting can decouple, closest to our Claim 2 (ranking-quality on parseable outputs is invariant to λ), but predict no boundary. Our contribution is orthogonal to constrained decoding: we improve format adherence as a side effect of training, derive when that adherence collapses, and show (Sec. 5.2) that strict-K decoders convert the capability gap into a duplicate-id pathology that does not improve task-level validity.

3 Method and Experimental Setup

3.1 Listwise PL Rollout

A *listwise PL rollout* for a product p with candidate review set $\{r_1, \ldots, r_K\}$ is an autoregressive generation, conditioned on the full prompt

> `Product: {title}. Below are` K `reviews. Score each. [Review 1] id=`r_1
> `...[Review` K`] id=`r_K`. Return JSON list of` K `objects ...`

of the assistant token sequence

> `[{"review_id": "`r_1`", "score":` s_1`}, {"review_id": "`r_2`", "score":`
> s_2`}, ...]`

The structural delimiters (brackets, braces, commas, identifier echoes) are interleaved with the per-position score tokens. Under the Plackett–Luce model [27, 40], the joint likelihood of K ordered scores factors as a product of K position-conditional softmaxes; here, the same factorization arises mechanically from token-level autoregression, and the per-token reverse-KL gradient automatically distributes credit across both the score tokens and the structural scaffolding (Fig. 2).

3.2 On-Policy Reverse-KL Distillation with Extrapolation

Given a student policy π_S, a teacher policy π_T, and a base/reference policy π_B, we define the per-token ListOPD advantage used in our implementation as a base-relative teacher–student log-ratio

$$A(s, a; \lambda) \;=\; \lambda\big(\log \pi_T(a \mid s) - \log \pi_B(a \mid s)\big) - \big(\log \pi_S(a \mid s) - \log \pi_B(a \mid s)\big), \qquad (2)$$

where $\lambda \geq 1$ is the extrapolation coefficient [42]. Setting $\lambda = 1$ recovers vanilla reverse-KL distillation ($\log \pi_T - \log \pi_S$); $\lambda > 1$ targets a base-relative sharpened teacher distribution proportional to $\pi_B(\pi_T/\pi_B)^\lambda$. Thm. 4.1 (Sec. 4) is stated for this exact base-relative target; the base-neutral π_T^λ

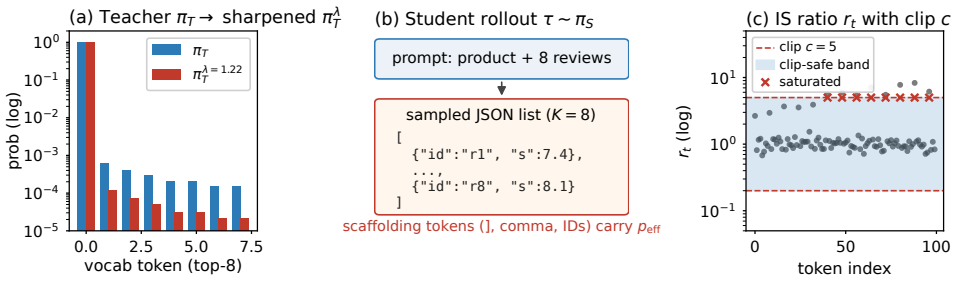

Figure 2: **ListOPD pipeline.** *Left:* teacher π_T at a scaffolding position concentrates on a modal token; base-relative extrapolation sharpens this target as λ grows. *Middle:* student rolls out the listwise JSON token-by-token. *Right:* the per-token IS ratio ρ_t is clipped at c; scaffolding positions whose asymptotic fixed point sits in the clip-unsafe region ($\times$) drift to parse-collapse, the regime characterised by Thm. 4.1.

form used in earlier OPD work is the π_B=uniform special case. The student is updated by GRPO [31] with token-level IS correction (clip c=5.0):

$$\rho_t \;=\; \min\!\left(c,\, \frac{\pi_T(a_t \mid s_t)}{\pi_S(a_t \mid s_t)}\right), \qquad \mathcal{L} \;=\; -\sum_t \rho_t\, A(s_t, a_t; \lambda)\, \log \pi_S(a_t \mid s_t). \tag{3}$$

The KL penalty coefficient is set to zero: the on-policy advantage (2) is the only training signal. We use the verl framework [33] with `actor.policy_loss.only_reverse_kl_advantages=True` and `lambda_vals=`λ; no other code changes were required to operate on listwise rollouts.

3.3 Models, Data, and Evaluation

Models. We use Qwen3 base models at four sizes: 0.6B, 1.7B, 4B, 8B parameters. Each model is first SFT-warmstarted on the listwise PL-K8 format for 5 epochs (lr 1×10^{-5}, cosine, batch size 128) on Amazon Fashion training data; this becomes both the OPD initialization and the SFT baseline we compare against. Teacher candidates are 4B and 8B PL-K8 SFT checkpoints.

Data. Amazon Fashion [15] reviews are pseudo-labeled for helpfulness (0–10) by Gemini 2.5 Pro [7]; we form K=8 product groups with reviews sampled uniformly within each product. Because Gemini's pretraining membership is not externally auditable, we treat these labels as a fixed rubric for a controlled structured-output environment, not as human relevance judgments; the theorem-facing measurements are parse rate, structural-token modal probability, and cliff location. Reproducibility and data-provenance details are in App. B.3. Train/val split is performed at the *product* level (no review of any val product appears in training) yielding 1795 train groups and 212 val groups. Cross-domain val sets from Baby_Products and Software (500 product groups each) use the same K and scoring rubric to measure zero-shot transfer. A public IR stress test replaces the Gemini rubric with MS MARCO/TREC-DL human qrels while preserving the strict K=8 JSON contract (App. F.2).

Training. For each (student, teacher, λ) triple, we run OPD for 1, 3, or 5 epochs over the listwise training set (14, 42, or 70 optimizer steps at batch size 128). Optimizer is AdamW with lr 1×10^{-6}, no warmup, no LR schedule, FSDP across 8 B200 GPUs, vLLM rollout [22] with tensor parallel size 2, max prompt length 2048, max response length 512, sampling temperature 1.0.

Evaluation. For the JSON listwise Fashion, cross-category, constrained-decoding, ASPO, no-base, and public-IR evaluations, we use vLLM with greedy decoding (temperature 0). MBPP and BFCL use task-standard sampled n=4 protocols, stated in their appendix sections. For each val product, the model emits a JSON list which we parse by extracting the outermost `[...]` block and then enforcing the deployment contract: exactly K objects, each containing one unique input `review_id` and a numeric `score`. Scores may be represented as JSON strings or numbers, but duplicate, missing, hallucinated, or position-only outputs are parse failures. Failure to recover all K valid `{review_id, score}` entries is recorded as a parse failure and the model receives *zero credit on all metrics for that product*. We report:

- **parse_rate**: fraction of val products yielding a valid K-element JSON list;

- per-product Kendall-τ, NDCG@$\{1, 3, 5, 10\}$, MAE on parsable subset;
- **USEFUL** $=$ `parse` $\times$ **NDCG@1**, the deployment-relevant aggregate where parse failures count as zero rank quality.

All metrics are macro-averaged over val products. The USEFUL metric is the only one we use to select operating points; the per-metric breakdown is reported in tables for diagnostic purposes.

4 Single-Position Threshold and Sequence Calibration

Fig. 1's clip-safe crossing has a closed-form location. We state the single-position results here and defer all proofs, assumption-level discussion, and per-token derivations to App. C.1.

Notation. p: teacher modal-token probability at one structural position; b: warmstart modal probability at the same position; c: per-token IS clip strength. p_{typ}, p_{safe} are mean and max of $\{p_t\}$ over τ-filtered scaffolding positions (Eq. (5)); b_{eff} is the warmstart counterpart at the binding position. We write p_{eff} generically when the choice of within-prompt aggregator does not matter; p_{typ} and p_{safe} are its specific instantiations. $\lambda^\star(p, b, c)$ is the closed-form clip-safe threshold (Eq. (4)); $\lambda^\star_{\text{typ}}{=}\lambda^\star(p_{\text{typ}}, b_{\text{eff}}, c)$ and $\lambda^\star_{\text{safe}}{=}\lambda^\star(p_{\text{safe}}, b_{\text{eff}}, c)$ are its sequence-level instantiations under Thm. 4.3(B) and (A) respectively.

Setup (single position). At one position of the rollout, teacher $\pi_T = (p, 1-p)$ with $p > \frac{1}{2}$, student $\pi_S^\theta = (q, 1-q)$ with $q = \sigma(\theta)$, base $\pi_B = (b, 1-b)$. The base-relative extrapolation target induced by Eq. (2) is $\pi_B(\pi_T/\pi_B)^\lambda$, which in the Bernoulli reduction gives $p_\lambda^{(b)} = b^{1-\lambda}p^\lambda/(b^{1-\lambda}p^\lambda + (1-b)^{1-\lambda}(1-p)^\lambda)$; $b{=}1/2$ recovers the base-neutral $p_\lambda = p^\lambda/(p^\lambda + (1-p)^\lambda)$, $b{\to}p$ collapses $p_\lambda^{(b)} \to p$. With clipped IS $r(x) = \min(c, \pi_T(x)/\pi_S(x))$, $c > 1$ (Thm. C.1) and the advantage of Eq. (2), the clip-safe region is $q < q_c := 1 - (1-p)/c$.

Theorem 4.1 (Single-position clip-safe threshold, base-relative)**.** *$p_\lambda^{(b)} < q_c$ iff $\lambda < \lambda^\star(p, b, c)$, where*

$$\boxed{\lambda^\star(p, b, c) = \frac{\log\big((1{-}p)/(c{-}1{+}p)\big) - \log\big((1{-}b)/b\big)}{\log\big((1{-}p)/p\big) - \log\big((1{-}b)/b\big)}} \tag{4}$$

Above $\lambda^\star$ the sharpened fixed point exits the clip-safe region. The $b{=}1/2$ special case reduces to $\log((1{-}p)/(c{-}1{+}p))/\log((1{-}p)/p)$; $b{\to}p$ sends $\lambda^\star \to \infty$ (no cliff if warmstart matches teacher). Proof: Lyapunov on $V(q){=}\mathrm{KL}(\pi_{T,\lambda}^{(b)}\|\pi_S)$ within the clip-safe basin plus $p_\lambda^{(b)} = q_c$ (App. C.1). Thm. 4.1 is the 2-token Bernoulli reduction; lifting to multi-token vocabularies is sufficient under A2 (off-modal mass concentrates on a small alternative set; exact under the off-modal-ratio invariance condition of Thm. C.4; Thm. 4.3). For $\lambda{>}\lambda^\star$, the noise-to-drift calculation in App. C.1 supports boundary-seeking finite-budget dynamics, but we do not prove a.s. convergence; finite-N reachability and the no-base implementation axis (S2b) are isolated in App. C.1.5.

Mechanism interpretation. The clip-safety boundary refers to the fixed point induced by the clipped objective's geometry, not empirical runtime clipping frequency: the direct per-step clip-fraction counter remains at 0 under verl's rollout-correction threshold, and per-step IS ratios stay well below c throughout training (App. E.2, Fig. 7). The observed cliff is realised through cumulative drift toward the clip-unsafe fixed point, not through discrete clip events.

Corollary 4.2 (Finite-budget drift diagnostic)**.** *Under a local linearization of the deterministic clipped flow before saturation, and assuming one-sided post-boundary drift, the characteristic first-passage time to q_c scales as $N^\star(\lambda) = O(1/|\log(1 - \eta\lambda p(1-p))|)$; in the small-drift limit this is $O(1/[\eta\lambda p(1-p)])$. The diagnostic expectation is that the observed cliff shifts leftward in λ as training lengthens; empirically the Fashion cliff midpoint moves $1.22 \to 1.12 \to 1.06$ across $N \in \{42, 70, 200\}$, with the $N{=}200$ point pre-registered (App. E.2).*

Sequence-level lift. A K-item JSON rollout has structural positions $\mathcal{S}$ with modal-token probabilities $\{p_t\}$. For a threshold τ, let $\mathcal{S}_\tau{=}\{t \in \mathcal{S} : p_t \geq \tau\}$. Because $\lambda^\star$ is strictly decreasing in p on $(\frac{1}{2}, 1)$ (Eq. (15), App. C.1), the most-concentrated position binds. With scaffolding filter $\tau{=}0.9$, define

$$p_{\text{safe}} := \max_{t:\, p_t \geq \tau} \{p_t\}, \qquad p_{\text{typ}} := \text{mean}_{t:\, p_t \geq \tau}\{p_t\} \tag{5}$$

(App. C.1.6).

Proposition 4.3 (Sequence-level cliff: (A) provable safety, (B) calibrated operating rule)**.** *Assume A1 (clipped IS, base-relative reverse-KL; App. C.1) and A2 (position-wise parametric reach; App. C.1); let b_{eff} be the warmstart modal probability at the binding position. The multi-token lift below is exact under Thm. C.4's off-modal-ratio invariance condition (App. C.1) and approximate otherwise.* (A) Provable safety. *For any $p_{\text{safe}} \geq \max_t p_t$, every structural position is clip-safe whenever $\lambda < \lambda^\star(p_{\text{safe}}, b_{\text{eff}}, c)$ (per-position Thm. 4.1 + monotonicity Eq.* (15)*).* (B) Empirical operating scale. *If the target task has a measured, dense near-deterministic scaffold ($\mathcal{S}_\tau$ is not sparse), SFT leaves visible parse headroom, and the chosen (c, N) regime can reach the boundary within budget, then the observed sequence-level cliff requires a $\Theta(1)$ fraction of structural equivalence classes to saturate. Under the N_{eff}-class correlation of Thm. C.7, this fraction is set by the typical class, so the empirical scale is $\lambda^\star(p_{\text{typ}}, b_{\text{eff}}, c)$. (B) is a calibrated operating rule grounded in (A) and the correlation analysis, not an independent theorem.*

Calibration. Fashion is the primary calibrated anchor: structural positions (N=200, τ=0.9) give p_{typ}=0.9993 $\pm$ 0.0001, $p_{\text{safe}}\approx$0.99996, and implied warmstart $b\approx$0.81 (joint log-ratio 0.21; measurement procedure, subset-bootstrap robustness, and class-weighting controls in App. C.1.6, F.1.2). At c=5, Eq. (4) gives (A) $\lambda^\star_{\text{safe}}\approx$1.18 and (B) $\lambda^\star_{\text{typ}}\approx$1.28 (base-neutral marker 1.22); this bracket $[1.18, 1.28]$ contains the observed onset window $[1.15, 1.25]$ within one λ-grid step. The aggregator pair (mean for p_{typ}, max-of-prompt-mean for p_{safe}) is fixed *ex ante* from Thm. 4.3(A)/(B), not selected against the observed Fashion onset; alternative within-prompt aggregators (App. C.1.6) span $\lambda^\star\in[1.22, 1.60]$, so cross-task pre-registration of the aggregator on a held-out scaffold is the natural next robustness test. The other rows of Tab. 1 report scope checks rather than independent calibrations: MBPP code (Fashion marker, no code-specific p_{eff}; App. F.4); MS MARCO/TREC-DL with measured p_{eff}=0.99941 inside Fashion's confidence band so the operating rule predicts the same window (App. F.2); the Llama-3.2 cross-architecture stack which is monotone through λ=1.4 at N=42 and parse-bounded below 0.23 at a pre-registered N=200 budget extension (App. F.7, F.7.1); a four-point p_{eff} scope check across (family, task, size) giving $\lambda^\star \in [1.27, 1.32]$ that evidences within-regime invariance (App. F.1.1); and a pre-registered cross-task BFCL test that fails on SFT-parse-saturation rather than mechanism refutation (App. F.5).

Table 1: **Predicted bracket contains observed cliff within one λ-grid step on every Fashion calibration row.** Predictions from Eq. (4) at $b\approx$0.81, p_{typ}=0.9993, c=5 unless stated. *Observed cliff* is the onset/collapse pair (last λ with parse$\geq$0.9, first with parse$\leq$0.7), or midpoint where pre-registered. The c=1.5 row matches its locked closed-form $\lambda^\star_{\text{typ}}$=1.070 at observed midpoint 1.069, below the experimental λ-grid resolution. JSON K=4 ListOPD lift is +0.04 vs. Fashion's +0.32 (App. F.3.1). †: drift row (Thm. 4.2); *: zero-shot.

Regime	N	predicted	observed cliff
Sharp: cliff localizes inside predicted bracket			
Fashion 1.7B×4B, 3-ep	42	[1.18, 1.28]	[1.15, 1.25]
Fashion 5-ep	70	leftward†	[1.10, 1.15]
Fashion 14-ep, pre-reg	200	[1.00, 1.10]	1.061
Fashion c=1.5, 14-ep, pre-reg	200	[1.00, 1.12]	1.069
Partial: predicate active, midpoint hits, lift attenuated			
JSONSchemaBench K=4 list, pre-reg	42	[1.19, 1.42]	1.29
Transfer: scale check or zero-shot, no independent re-calibration			
MBPP code, 1.7B×4B	35	Fashion marker 1.22	[1.15, 1.25]
Baby/Software zero-shot	—	[1.34, 1.37]	[1.15*, 1.25*]
Abstain (preconditions fail): BFCL, GSM8K, JSON single-instance (App. H)			

5 Experiments

We localize the cliff on Fashion (Sec. 5.1), confirm parameter efficiency under controls (Sec. 5.2), and report scope-check regimes (Sec. 5.3: IR, c-sweep, GSM8K, regularizers; cross-arch and BFCL in App.). The predicate is non-trivial under five jointly satisfied preconditions: (i) near-deterministic structural tokens ($p_{\text{eff}}\to$0.999); (ii) a single dominant outer-scaffold binding equivalence class; (iii) post-SFT parse headroom; (iv) base-relative IS-clipped implementation matched to the formula;

(v) training budget reaching the boundary. Secs. 5.1–5.2 validate within this regime; Sec. 5.3 and Tab. 1 report rows where preconditions fail or are partially satisfied. The central dependent variable is strict parse rate; Kendall and NDCG on parsed outputs are diagnostic.

5.1 Cliff localization and finite-N signature

We sweep λ at fine resolution on a fixed (1.7B student, 4B teacher); Fig. 3 overlays parse rate and FMC (*format-manifold collapse*: the mechanism-predicted $K-1$ truncation indicator with all-real ids and missing one input id) with the closed-form $\lambda^\star=1.22$ marker.

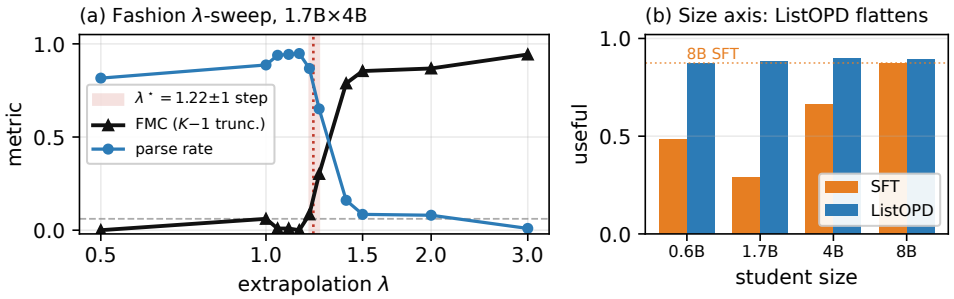

Figure 3: **Closed-form clip-safe threshold.** *Left:* Fashion λ-sweep (1.7B$\times$4B, 3-epoch); strict parse and FMC ($K-1$ truncation) transition in $[1.15, 1.25]$, around the base-neutral $\lambda^\star=1.22$. *Right:* USEFUL for SFT vs. ListOPD across 0.6B–8B Qwen3 students; sub-threshold ListOPD flattens the size curve to USEFUL$\in[0.873, 0.897]$ (0.6B / 8B single-seed).

Table 2: **Lambda sweep on Amazon Fashion.** Parse rate collapses near the predicted bracket $[1.18, 1.28]$, whereas NDCG@1 on parsed outputs remains stable. Bold marks the last safe point and first collapsed point.

λ	parse rate	NDCG@1 (parsed)	USEFUL
0.50	0.816	0.899	0.734
1.00	0.887	0.923	0.819
1.05	0.939	0.923	0.867
1.10	0.943	0.929	0.877
1.15	**0.948**	0.930	**0.882**
1.20	0.868	0.934	0.811
1.25	**0.651**	0.931	**0.606**
1.40	0.160	0.949	0.152
1.50	0.085	—	0.077

Parse transitions sharply between λ=1.20 and λ=1.25 (Tab. 2), within one grid step of the predicted bracket. NDCG@1 on parsed outputs is statistically flat across the sweep (p=0.61, paired bootstrap with parse-failed products dropped from each side): the λ effect is concentrated in format-adherence, not ranking quality. Training time slides the cliff leftward: extending λ=1.15 to 5 epochs takes parse from 0.948 to 0.675 (Fig. 4, App. D.2.2).

A pre-registered 5-seed fine-grid sweep at $\lambda \in \{1.18, 1.20, 1.22, 1.24\}$ localizes the parse≥ 0.80 cliff onset to a 95% paired-bootstrap CI of $[1.204, 1.228]$, containing the predicted $\lambda^\star$=1.22 (App. D.1). $\sigma_{\text{seed}}(\text{parse})$ inflates $\sim 4\times$ across the boundary as expected near a first-passage threshold. Per-step trajectories at the 5-seed λ=1.15 operating point ($\{7, 13, 21, 42, 95\}$, parse 0.921 ± 0.019, FMC 0.031 ± 0.021) confirm Thm. 4.2's first-passage diagnostic; finite-N step-cliff evidence across three corpora is in App. D.2.3.

Pre-registered budget extension. We pre-register a third budget point (N=200, 14 epochs) before any new training (App. E.2), with locked bracket $[1.00, 1.10]$ from Thm. 4.2's leftward-drift extrapolation off the (N=42, λ_{cliff}=1.22) and (N=70, λ_{cliff}=1.12) anchors. Single-seed parses at $\lambda \in \{1.00, 1.05, 1.10\}$ are $\{0.934, 0.703, 0.500\}$ (cliff midpoint 1.061); a 3-seed CI at λ=1.05 gives parse 0.742 ± 0.107 (midpoint 1.068). Both lie inside the locked $[1.00, 1.10]$ bracket.

5.2 Parameter efficiency and controls

Size axis. Tab. 3 reports SFT and ListOPD across four student sizes under strict `review_id`-aligned parsing, with seed-mean $\pm\,\sigma_{\text{seed}}$ where multi-seed runs exist. SFT scales non-monotonically: 1.7B-SFT seed-mean parse rate (0.264, $n{=}5$) sits below the single-seed 0.6B-SFT baseline (0.547); the four sizes share an identical SFT recipe (lr, batch size, schedule, epochs), so the non-monotonicity is the empirical observation, not a per-size hyperparameter artefact. ListOPD places every multi-seed configuration at USEFUL $\in [0.857, 0.897]$ with $\sigma_{\text{seed}}(\text{USEFUL}) \leq 0.016$ on 1.7B (vs. 0.093 for 1.7B-SFT and 0.156 for 4B-SFT). The 1.7B-ListOPD gain over 4B-SFT is $+0.220$ seed-42 ($p{<}10^{-4}$, Tab. 4 *Scaling alone*; $+0.372$ across seeds).

Table 3: **Fashion size-axis** ($N{=}212$, $K{=}8$; strict `review_id` parser). Sub-threshold ListOPD rows reach USEFUL$\in[0.873, 0.909]$ across 0.6B–8B; SFT rows are parse-limited at small/mid sizes. Cells are seed-mean$\pm\sigma_{\text{seed}}$ for $n{>}1$; † marks seed=42 only. Multi-seed coverage in App. D.4.

Configuration	parse	NDCG@1	Kendall	MAE	USEFUL	σ_{seed}(USEFUL)
0.6B SFT†	0.547	0.881	0.831	1.132	0.482	—
0.6B ListOPD (λ=1.0, 4B teacher)†	0.953	0.916	0.863	0.839	**0.873**	—
1.7B SFT	0.264±0.105	0.865	0.833	1.231	0.230	0.093
1.7B ListOPD (λ=1.15, 4B teacher)	0.921±0.019	0.931	0.885	0.777	**0.857**	0.016
4B SFT	0.516±0.166	0.941	0.893	0.838	0.485	0.156
4B ListOPD (λ=1.0, 8B teacher)†	0.953	0.942	0.904	0.866	**0.897**	—
8B SFT	0.877±0.090	0.949	0.899	0.852	0.833	0.082
8B ListOPD (λ=1.0, 4B teacher)†	0.943	0.948	0.898	0.802	**0.894**	—
8B ListOPD (λ=1.22, 32B teacher)†	0.953	0.953	0.902	0.711	**0.909**	—

Failure modes shift with size (full pattern in App. D.2.1): at $\lambda{=}1.15$, 1.7B-ListOPD lands at FMC$=0.031\pm0.021$ matching the 8B-SFT regime, while post-cliff $\lambda{\geq}1.25$ regresses to the 4B-SFT $(K{-}1)$-manifold. A 32B-teacher 8B-student spot-check is stable through $\lambda{=}1.5$ (USEFUL$\in[0.899, 0.909]$; App. D.2.1).

Decisive ablations. Six candidate explanations for the seed-42 1.7B$\rightarrow$0.882 lift are pre-registered against the same strict deployment-useful metric; none explains it (Tab. 4, which also documents the $6\times$ σ_{seed}(USEFUL) reduction).

Table 4: **Ablations for alternative explanations.** Six pre-registered controls for the 1.7B-SFT$\rightarrow$ListOPD lift; none explains it. Strict USEFUL (zero-imputed); Δ CIs from 10,000 paired-bootstrap over $N{=}212$ Fashion val. Seed row: across-seed mean$\pm\sigma$.

Hypothesis (control $\rightarrow$ ListOPD)	Ctrl USEFUL	OPD USEFUL	ΔUSEFUL (95% CI)
Extra SFT steps (continued SFT, matched budget)	0.273	0.882	$+0.608$ $[0.547, 0.670]$
On-policy exposure (forward-KL, no extrapolation)	0.027	0.882	$+0.854$ $[0.814, 0.891]$
No extrapolation needed (λ=1.0 vs 1.15)	0.819	0.882	$+0.063$ $[0.032, 0.098]$
Decoder constraints (SFT+regex vs OPD+regex)	0.679	0.883	$+0.204$ $[0.156, 0.252]$
Seed cherry-pick (5-seed SFT vs 5-seed OPD)	0.230±0.093	0.857±0.016	$+0.628$ (z=13.9)
Scaling alone (4B-SFT vs 1.7B-ListOPD)	0.661	0.882	$+0.220$ $[0.163, 0.280]$

Constrained decoding baselines. Five strict-K constrained systems (XGrammar [10], Outlines [39], LM-Format-Enforcer, llguidance, regex-template; details in App. B.2) raise 1.7B-SFT parse from 0.325 to 0.783; unconstrained 1.7B-ListOPD already reaches 0.943 and constraints add $+0.009$ on top. On USEFUL, best constrained SFT reaches 0.679 vs. 0.874 for unconstrained ListOPD; post-hoc permutation repair (App. B.2, Tab. 6) closes the gap to USEFUL$=0.823$ (seed-42 residual 0.051). All $46/46$ schema-constrained SFT failures are duplicate `review_id`s (a strict subset of the 138-product unconstrained-SFT failure set), so decoder-side repair alone does not close the learned-contract gap.

5.3 Scope checks

Public IR stress test. To remove the Amazon/Gemini-rubric dependence, we run the same JSON listwise interface on MS MARCO passage triples with TREC-DL 2020 judged validation [3, 8] (1.7B$\times$4B, 2000 train groups, 54 judged val queries). The seed-42 sweep follows the Fashion shape;

the multi-seed follow-up does not separate λ=1.25 vs. 1.5 within seed variation (Tab. 22). We therefore report MS MARCO/TREC-DL as a public-qrel boundary rather than a cliff replication.

Public-benchmark scope check (JSONSchemaBench). On JSONSchemaBench [12], the only fully public, mechanically-verifiable benchmark we test, the single-instance protocol shows no cliff localization on the locked λ-grid (App. F.3). The heterogeneous-schema setup violates the single-binding-class precondition of Thm. 4.3(B); the deployment-useful lift, however, transfers cleanly (1.7B-ListOPD validate matches the 4B-SFT teacher; App. F.3). A pre-registered K=4 K-list extension that restores a single outer binding class hits its predicted bracket but with attenuated cliff sharpness: cliff midpoint ≈ 1.29 matches inside the locked $[1.19, 1.42]$ bracket, but the sub-critical anchor undershoots (0.280 vs. locked ≥ 0.40) and the super-critical anchor rebounds (0.332 vs. locked ≤ 0.10); peak ListOPD lift is attenuated ($+0.04$ klist_rate vs. Fashion's $+0.32$; App. F.3.1). We read this as scope-refinement evidence: the outer K-ary wrapper is sufficient for cliff *localization*, while inner-schema homogeneity controls cliff *sharpness* in this regime.

c-axis cliff/no-cliff. The strongest pre-registered cross-clip test is the locked N=200 extension at c=1.5 (App. E.3): the closed form gives $\lambda^\star_{\text{typ}}(c{=}1.5){=}1.070$ *before* any training; the observed cliff midpoint between λ=1.05 (parse 0.939) and λ=1.075 (parse 0.632) is 1.069, matching the prediction below grid resolution and well inside the locked $[1.00,\ 1.12]$ window. The sub-critical anchor λ=0.95 holds at parse 0.943; the falsification anchor λ=1.20 collapses to 0.255. The shorter c-axis sweep at N=42 (fixed λ=1.15, 1.7B$\times$4B, 3-ep; Tab. 16) qualitatively matches the formula at $c\in\{2, 5, \infty\}$ but is parse-stable at c=1.5 (0.948 vs. predicted $\lambda^\star$=1.06), below the formula's reachability budget; the N=200 result is the in-budget test. Fig. 6 traces the teacher/student IS-ratio mechanism.

GSM8K cross-task scope boundary. GSM8K math CoT (Qwen3-1.7B$\to$4B, c=5) gives flat reward across $\lambda \in [1.00, 1.59]$ ($\sigma_\lambda \approx 0.006$): scaffolding is too diffuse for τ=0.9, so the predicate's measurability precondition fails (App. F.6).

Regularizer pilots as scope checks. Three target-reshaping regularizers (App. I.1): KL-to-base and entropy bonus drop parse below baseline CI at below-grid predicted drift (scope boundary on the drift prediction); a 20-step λ-warmup stays inside the baseline seed band (0.929 vs. 0.921 ± 0.019), consistent with the predicted $\lambda^\star_{\text{warmup}} \approx 2.33$ value.

6 Conclusion

On-policy distillation (OPD) with reward extrapolation can lift a student past its teacher in domain, but past a sharp threshold $\lambda^\star$ the same step instead collapses the model's structured-output contract. We give a closed-form clip-safety predicate $\lambda^\star(p, b, c)$ that locates this threshold from three measurable quantities (teacher modal probability, warm-start mass, IS clip strength), turning OPD tuning from a post-hoc λ sweep into a falsifiable boundary-prediction problem. On Amazon Fashion, three pre-registered tests (a fine-grid cliff interval, a budget-extension test, and a small-clip cross-prediction) all land inside their locked prediction windows, the small-clip prediction matched below grid resolution (Tab. 1). Operating just below the threshold, 1.7B-ListOPD brings a Qwen3 student to in-domain seed-noise parity with a pre-registered 3-seed 8B-SFT baseline at one-fifth the parameters and roughly $5\times$ lower seed variance under the Gemini listwise rubric.

Limitations. The predicate is base-relative: a no-base variant (App. C.1.5) does not show the cliff at the same 42-step budget. Outside near-deterministic multi-item parseable scaffolds, the predicate shifts under finite-budget reachability, abstains via measurability preconditions (BFCL SFT-saturation; GSM8K diffuse scaffolding), or remains statistically underpowered (MS MARCO at 4 seeds). The cliff predicate is rubric-independent (parse-rate signal); the 1.7B$\leftrightarrow$8B parity claim uses Gemini-graded USEFUL and inherits that evaluator's exposure. A direct falsification would be a near-deterministic structural-token task satisfying the preconditions but with cliff midpoint outside the locked λ-bracket; extended discussion of mechanism positioning, parameter efficiency, and theory status is in App. A; each is a scope boundary, not a refutation.

Mechanism positioning. $\lambda^\star$ is a property of IS-clipped reverse-KL extrapolation, not a particular method: ASPO [38] retargets the asymmetry through a ratio-flip and shows its own cliff one grid step left of vanilla OPD's (App. G), so the predicate corroborates on the alternative published mitigation. For finite-budget deployment we recommend λ_{op} at one λ-grid step below $\lambda^\star(N)$ measured at the deployment budget, leaving margin for Thm. 4.2's leftward drift.

References

[1] Bhavik Agarwal, Ishan Joshi, and Viktoria Rojkova. Think inside the json: Reinforcement strategy for strict llm schema adherence. *arXiv preprint arXiv:2502.14905*, 2025.

[2] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

[3] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[4] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Guiding llms the right way: Fast, non-invasive constrained generation. *arXiv preprint arXiv:2403.06988*, 2024.

[5] Christopher Burges, Robert Ragno, and Quoc Le. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19, 2006.

[6] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2020 deep learning track, 2021. URL https://arxiv.org/abs/2102.07662.

[9] Haikang Deng, Po-Nien Kung, and Nanyun Peng. Decoupling task-solving and output formatting in llm generation. *arXiv preprint arXiv:2510.03595*, 2025.

[10] Yixin Dong, Charlie F Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, and Tianqi Chen. Xgrammar: Flexible and efficient structured generation engine for large language models. *Proceedings of Machine Learning and Systems*, 7, 2025.

[11] Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

[12] Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. Jsonschemabench: A rigorous benchmark of structured outputs for language models. *arXiv preprint arXiv:2501.10868*, 2025.

[13] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[15] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[16] Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. *arXiv preprint arXiv:2601.07155*, 2026.

[17] Hao Jiang, Zhi Yang, Annan Wang, Yichi Zhang, and Weisi Lin. Rlpo: Residual listwise preference optimization for long-context review ranking. *arXiv preprint arXiv:2601.07449*, 2026.

[18] Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

[19] Juno Kim, Jihun Yun, Jason D Lee, and Kwang-Sung Jun. Coverage improvement and fast convergence of on-policy preference learning. *arXiv preprint arXiv:2601.08421*, 2026.

[20] Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

[21] Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint arXiv:2503.07067*, 2025.

[22] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[23] Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

[24] Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, and Maosong Sun. Learning to generate structured output with schema reinforcement learning. pages 4905–4918, 2025.

[25] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*, 2023.

[26] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 708–718, 2020.

[27] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

[28] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*, 2023.

[29] Avinash Reddy, Thayne T Walker, James S Ide, and Amrit Singh Bedi. Draft-conditioned constrained decoding for structured generation in llms. *arXiv preprint arXiv:2603.03305*, 2026.

[30] Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. First: Faster improved listwise reranking with single token decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652, 2024.

[31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[32] Guobin Shen, Chenxiao Zhao, Xiang Cheng, Lei Huang, and Xing Yu. Vespo: Variational sequence-level soft policy optimization for stable off-policy llm training. *arXiv preprint arXiv:2602.10693*, 2026.

[33] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[34] Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

[35] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatGPT good at search? investigating large language models as re-ranking agents. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL `https://openreview.net/forum?id=3Q6LON8y2I`.

[36] Sönke Tenckhoff, Mario Koddenbrock, and Erik Rodner. Llmstructbench: Benchmarking large language model structured data extraction. *arXiv preprint arXiv:2602.14743*, 2026.

[37] Özgür Uğur, Musa Yılmaz, Esra Şavirdi, Özay Ezerceli, Mahmut El Huseyni, Selva Taş, and Reyhan Bayraktar. Guided decoding and its critical role in retrieval-augmented generation. pages 1–4, 2025.

[38] Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025.

[39] Brandon T Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

[40] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, 2008.

[41] Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026.

[42] Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

[43] Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. The price of format: Diversity collapse in llms. *arXiv preprint arXiv:2505.18949*, 2025.

[44] Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2308–2313, 2023.

Appendix organization. App. A extends the body discussion. App. B gives implementation and reproducibility details. App. C contains the full sequence-level cliff proof and the EOPD analytical extension. App. D collects Fashion calibration evidence (W5 5-seed onset CI, decisive ablations, IS mechanism, temperature sensitivity, multi-seed pilot). App. E reports the four pre-registered finite-budget tests of Thm. 4.2; App. F reports scope tests across (task, family, architecture). App. G is the head-to-head comparison against ASPO. App. H aggregates the closed-form predicate scope-test verdicts; App. I states compute constraints and deferred-ablation pre-registrations.

A Extended Discussion

Scope and limitations. $\lambda^\star(p, b, c)$ applies to near-deterministic structural tokens with parseable failures; boundary regimes are summarised in Tab. 1. The predicate is base-relative: a no-base variant (App. C.1.5) does not show the cliff at the same 42-step budget, so the operating rule is sensitive to the IS-clip implementation, not just the asymptotic fixed point. MS MARCO has $p_{\text{eff}}{=}0.99941$ inside Fashion's CI band (App. F.1.1), but the multi-seed test is underpowered at 4 seeds. Greedy decoding ($T{=}0$) is deployment-relevant; a pre-registered $T \in \{0, 0.5, 1\}$ ablation (App. D.3) shows the cliff is a learned-policy property, not a decoding artefact. The cliff predicate is rubric-independent (signal in parse rate; NDCG@1 on parsed outputs flat); the 1.7B↔8B-SFT parity claim, by contrast, uses USEFUL=`parse` × NDCG@1 graded against the Gemini 2.5 Pro rubric and inherits whatever pretraining-membership exposure that rubric carries. A direct falsification of the closed form would be a near-deterministic structural-token task satisfying the predicate's preconditions (measurable p_{eff} near 0.999, single binding equivalence class, post-SFT parse headroom, matched IS-clip implementation) but with cliff midpoint outside the locked λ-bracket; the formula then refutes, not the reachability budget.

Parameter efficiency. 1.7B-ListOPD matches an 8B-SFT baseline within seed noise on Fashion in-domain at one-fifth the parameters: 3-seed 1.7B-ListOPD USEFUL$=0.857\pm0.016$ vs. 3-seed 8B-SFT 0.833 ± 0.082 (Tab. 3; pre-registered). Across-seed parity therefore holds in both directions of the difference, with 8B-SFT exhibiting $\approx 5\times$ higher seed std than 1.7B-ListOPD; this stability finding itself supports the operating-point view. Against schema-constrained decoding plus permutation repair the deployment-relevant residual is the smaller but real $+0.051$ USEFUL (App. B.2); the case for training-based contract adherence rests on this residual, the σ_{seed} contraction, and the interpretability of $\lambda^\star$ as an operating-point selector. Parity is in-domain only; on cross-category Baby/Software 8B-SFT retains an edge (App. D.2.1).

Mechanism positioning. $\lambda^\star(p, b, c)$ is a property of IS-clipped reverse-KL extrapolation, not of any particular method in the family. ASPO [38] retargets the asymmetry through a ratio-flip rather than removing it; in our 4-seed Fashion head-to-head ASPO has its own cliff one grid step left of vanilla OPD's, with the same parse-collapse pattern (App. G), so the predicate corroborates on the alternative published mitigation. *On Fashion in-domain, ASPO at its best operating point matches ListOPD within seed noise; the contribution here is the predicate, not categorical method superiority over ASPO.* The two are operationally complementary: ASPO mitigates the asymmetry; $\lambda^\star$ identifies where the clip-unsafe asymptotic boundary sits within whichever IS-clipped variant is in use.

Theory status and open directions. The single-position fixed point of Thm. 4.1 is proved; the multi-token lift is exact under off-modal-ratio invariance and approximate otherwise. Thm. 4.2 is a deterministic drift diagnostic without an almost-sure super-critical convergence guarantee; the drift *magnitude* (post-cliff parse-rate drop, c-dependent slope) has no closed-form prediction. The predicate locates cliff position only; sub-critical operating-point stability under regularizers (e.g., the entropy-bonus parse drop at $\gamma{=}0.001$ in App. I.1) is an orthogonal failure mode it does not cover. Alternative on-policy stabilizers (EMA-anchored target policies, top-k token-level KL, EOPD [18], VESPO [32]) modify the IS-clip asymmetry rather than remove it; entropy-bonus regularization is closed in App. I.1 (Eq. (19) gives $\delta\lambda \approx 2{\times}10^{-4}\gamma$). For finite-budget deployment we recommend λ_{op} at one λ-grid step below $\lambda^\star(N)$ measured at the deployment budget; this matches the 1.7B Fashion operating point ($\lambda{=}1.15$ at $N{=}42$ vs. midpoint 1.22) and respects Thm. 4.2's leftward drift. The clean next theory step is a stochastic-approximation proof of super-critical boundary dynamics; empirically, prospective task-axis held-out validation (p_{eff} materially distinct from 0.999) and higher-λ localization for the boundary regimes are the cleanest next tests.

B Implementation, evaluation, and reproducibility

B.1 Implementation details

Hyperparameters. All ListOPD runs use AdamW with learning rate 1×10^{-6}, no warmup, no learning-rate schedule, batch size 128, gradient accumulation 1, max prompt length 2048, max response length 512, $K=8$ reviews per product list. We train on a single node of 8×NVIDIA B200 GPUs (180 GB HBM each). The verl trainer configuration:

- `algorithm.adv_estimator=grpo`
- `actor.policy_loss.only_reverse_kl_advantages=True`
- `algorithm.rollout_correction.rollout_is=token`, `rollout_is_threshold=5.0`
- `actor.use_kl_loss=True`, `kl_loss_coef=0` (the only training signal is the per-token reverse-KL advantage)
- `rollout.gpu_memory_utilization=0.6`, `rollout.tensor_model_parallel_size=2`
- `rollout.temperature=1.0`, `val_kwargs.n=4`

Compute budget. A 3-epoch ListOPD run on a 1.7B student takes approximately 10 minutes on 8 B200 GPUs (42 optimizer steps, ∼15 seconds per step including rollout, ref-log-prob, base-log-prob, advantage compute, and update). A 3-epoch run on a 4B student takes approximately 15 minutes; 8B takes ∼25 minutes. The full experimental matrix in this paper (37 evaluated configurations including the λ sweep, size sweep, teacher ablation, training-duration ablation, and cross-category transfer evaluations) consumed approximately 12 GPU-hours.

B.2 Constrained-decoding backend detail

The strict-K JSON schema used in Sec. 5.2 is

```
{ "type": "array", "minItems": K, "maxItems": K,
  "items": { "type": "object",
    "properties": {
      "review_id": {"type":"string", "enum": <per-prompt enum>},
      "score":     {"type":"number", "minimum":0, "maximum":10} },
    "required": ["review_id", "score"], "additionalProperties": false } }
```

Latency. Within the same vLLM 0.18 runtime, XGrammar, llguidance, and Regex impose negligible overhead on top of unconstrained vLLM ($\approx$ 2.5 s/product, single B200) compared to 4.05 s for unconstrained 1.7B-SFT (which overruns its budget on every truncated sample) and 2.47 s for unconstrained 1.7B-OPD. Outlines is consistently ∼2× slower (4.58–5.21 s) because its Aho–Corasick FSM is rebuilt per-prompt over the 8-element `review_id` enum. Deployment recommendation: unconstrained ListOPD is the latency winner; pairing it with the cheapest grammar backend (XGrammar or llguidance) adds $+0.005$ parse-rate insurance at $<3\%$ latency cost.

Table 5: Constrained decoding on PL-K8 val ($N=212$, $K=8$). Parse = valid permutation over input IDs; τ/NDCG@5 over parsed outputs; latency = seconds/product on one B200. Grammar constraints help SFT but do not close the ListOPD gap.

	1.7B-SFT			1.7B-OPD (λ=1.15)			Latency
System	Parse	τ	NDCG@5	Parse	τ	NDCG@5	SFT/OPD (s)
Unconstrained	0.325	0.834	0.936	0.943	0.889	0.969	4.05 / 2.43
XGrammar	0.783	0.816	0.937	0.953	0.888	0.969	2.55 / 2.50
Outlines	0.783	0.816	0.937	0.953	0.888	0.969	4.61 / 4.63
LM-Format-Enforcer	0.783	0.816	0.937	0.953	0.888	0.969	3.46 / 3.51
Guidance (llguidance)	0.783	0.816	0.937	0.953	0.888	0.969	2.44 / 2.46
Regex	0.783	0.816	0.937	0.953	0.888	0.969	2.55 / 2.49

Post-hoc permutation repair. Schema-level constraints enforce structural admissibility but not semantic uniqueness. We supply a post-hoc repair (`scripts/offline_permutation_repair.py`) that, for each output with a duplicate `review_id`, injects the missing id into the duplicate slot (preferring the later position) while preserving that slot's score. Applied to all five schema-constrained

SFT outputs, the repair closes 82% of the USEFUL gap to unconstrained 1.7B-OPD (Tab. 6); the remaining gap shows that permutation repair alone does not close the ListOPD gap.

Table 6: **Post-hoc permutation repair on constrained-decoding outputs.** Each cell reports pre $\rightarrow$ post metric. Repair closes $0.679 \rightarrow 0.823$ of the schema-constrained-SFT USEFUL gap toward unconstrained 1.7B-OPD (0.874), but a 0.051 residual remains after enforcing permutation validity, so decoder-side repair alone does not close the ListOPD gap. *XGrammar / Outlines / LM-Format-Enforcer / llguidance / regex-template are bit-identical at $T{=}0$, all $46/46$ failures are duplicate-id, and post-hoc repair converges to the same post-repair state.

System (post-hoc perm repair)	Parse	Kendall	NDCG@1	USEFUL
1.7B-SFT unconstrained	$0.325 \rightarrow 0.335$	$0.834 \rightarrow 0.835$	$0.859 \rightarrow 0.863$	$0.280 \rightarrow 0.289$
1.7B-SFT schema-constrained*	$0.783 \rightarrow 0.953$	$0.816 \rightarrow 0.763$	$0.867 \rightarrow 0.864$	$0.679 \rightarrow 0.823$
1.7B-OPD unconstrained	$0.943 \rightarrow 0.943$	$0.889 \rightarrow 0.889$	$0.927 \rightarrow 0.927$	$0.874 \rightarrow 0.874$

B.3 Reproducibility and data provenance

Artifact release. The public project page is `https://lixin.ai/ListOPD`. The accompanying verification artifact includes `scripts/`, `configs/`, `outputs/paper/`, and the experiment-specific aggregate summaries under `outputs/spotlight/`. The load-bearing paper tables are audited from `outputs/paper/result_ledger.csv`, and `outputs/paper/paper_number_audit.json` records the corresponding paper-number provenance. The artifact supports verification of reported aggregate numbers without re-training. Full training launchers and framework patches are omitted from the initial public bundle and will be released or documented through the project page where licenses permit.

Data construction. Fashion, Baby_Products, and Software use the public Amazon Reviews corpus [15]. The public artifact does not redistribute raw reviews or derived JSONL/parquet splits; it includes the Fashion preprocessing scripts and aggregate metric artifacts needed to audit the reported numbers. For upstream datasets whose redistribution is restricted or too large for the initial bundle, we provide public source references and deterministic construction details, with full regeneration commands to be released or documented through the project page where licenses permit. MBPP, GSM8K, MS MARCO/TREC-DL, BFCL, and Glaive are used through their public releases as described in the corresponding appendix sections.

Gemini pseudo-labels and contamination boundary. Gemini 2.5 Pro supplies scalar listwise pseudo-labels for the Amazon review-ranking rubric. We cannot audit whether Gemini's pretraining mixture contained individual Amazon reviews, so we do not claim a human-validated ranking benchmark or make claims about absolute human relevance. We scope away the human-relevance claim: the theorem-data contacts use format-validity, measured teacher structural-token confidence, or cliff location; NDCG/Kendall remain rubric-side diagnostics on the fixed pseudo-label set. A larger standard-IR stress test or an Amazon-domain human-label spot-check is separate validation we leave to follow-up; the present paper makes no claim about Gemini-rubric-to-human transfer.

All numerical results. All released aggregate metrics are indexed in `outputs/paper/all_listwise_metrics.csv`; due to the paper page budget we omit the full table here.

C Proofs and analytical extensions

C.1 Sequence-level cliff: full proofs

This appendix provides the full proof of Thm. 4.1 (single-position) and the safety-bound proof for Thm. 4.3. The sequence-level empirical scale is a calibration rule motivated by the N_{eff} grouping argument and validated on the Fashion anchor, not a theorem.

C.1.1 Setup and assumptions for the sequence-level result

Assumption C.1 (On-policy sampling with clipped IS, single position)**.** (Restated and made precise for the sequence-level result.) At position t, the student draws $a_t \sim \pi_S^\theta(\cdot \mid s_t)$. The teacher/student token-level importance ratio $r_t(a_t) := \pi_T(a_t \mid s_t)/\pi_S^\theta(a_t \mid s_t)$ is clipped to $[0, c]$ for some $c > 1$, and the reverse-KL advantage of Eq. (2) is the per-token reward.

Definition C.2 (Structural positions)**.** Given a rollout context s_t, position t is *structural* if the teacher has a unique modal token $m_t := \arg\max_x \pi_T(x \mid s_t)$ with modal-token probability $p_t := \pi_T(m_t \mid s_t) > 1/2$. The set of structural positions in a rollout is $\mathcal{S} \subseteq \{1, \ldots, T\}$.

Structural positions include JSON scaffolding (brackets, commas, quotes, colons, field names) and the dominant-scored review-id continuation at each listwise slot. Operationally, we identify structural positions by the threshold criterion

$$\mathcal{S}_i := \{\, t : m_{i,t} \geq \tau \,\}, \qquad m_{i,t} := \max_x \pi_T(x \mid s_{i,t}), \tag{6}$$

with $\tau = 0.9$ in the main-body calibration and sensitivity reported in App. C.1.6. Because $\lambda^\star$ is decreasing in p (Eq. (15)), the binding (most restrictive) position is the one with the largest p_t, so the conservative all-positions-safe sufficient condition uses a certified upper bound $p_{\text{safe}} \geq \max_{t \in \mathcal{S}} p_t$. In the empirical tables we report a high-confidence proxy for this quantity; the proof claim attaches to a true upper bound, not to the proxy itself. The empirical within-prompt p_t range on the τ=0.9 structural subset is so tight (per-prompt min 0.9419, mean 0.9993, 95$^{\text{th}}$-pct 0.9994, max approaching 1) that $\lambda^\star(p_{\min}, c)$, $\lambda^\star(p_{\text{mean}}, c)$ and $\lambda^\star(p_{\max}, c)$ all lie within ± 0.05 of each other at c=5, i.e. within one λ-grid step (Tab. 8); the prediction is therefore aggregator-robust at the grid resolution. The per-token worst-case $\min_{i,t} m_{i,t}$ over *all* generated positions is dominated by score-digit positions at which the teacher is genuinely uncertain and gives a degenerate value (empirically ≈ 0.26) for which Eq. (17) below has no real solution, the trivial failure mode of applying monotonicity across positions whose per-position fixed points are strongly correlated via shared token embeddings (see Thm. C.7). The τ=0.9 filter restricts attention to positions where the teacher has a near-deterministic mode, precisely the regime where the IS-clip-asymmetry mechanism of Thm. 4.1 is operative.

Assumption C.3 (Position-wise parametric reach, approximate)**.** For target assignments $\{q_t^*\}_{t \in \mathcal{S}} \subset (0,1)^{|\mathcal{S}|}$ of structural-position modal-token probabilities, there exists $\theta^* \in \Theta$ with $\pi_S^{\theta^*}(m_t \mid s_t) = q_t^*$ approximately for every $t \in \mathcal{S}$.

For modern overparameterized LLMs, Thm. C.3 is plausible when structural positions have distinct context representations, and becomes more approximate when the same token class (e.g., the bracket token) repeats across positions with correlated contexts. Let N_{eff} denote the number of *equivalence classes* of structural tokens (brackets, commas, colons, quotes, field-name prefixes, delimiters, numeric scaffolding; empirically $N_{\text{eff}} \approx \mathcal{O}(10)$ in our rollouts), each class grouping positions that share a token embedding and hence whose per-position modal-token probabilities are strongly correlated under any realized θ. We use Thm. C.3 as an equivalence-class approximation rather than a verified property of the transformer parameterization; it is most credible when within-class variance of $\{p_t\}$ is small. The backend-invariant 46/212 constrained-decoding residual of Tab. 12 is consistent with a shared binding class, but is not a proof of A2.

Lemma C.4 (Exact multi-token Bernoulli reduction under off-modal-ratio invariance)**.** *Consider a structural position with modal token m and off-modal set R. Suppose teacher, base, and student all lie in the family*

$$\begin{aligned} \pi_T(m) &= p, & \pi_T(r) &= (1-p)\alpha_r, \\ \pi_B(m) &= b, & \pi_B(r) &= (1-b)\alpha_r, \\ \pi_S(m) &= q, & \pi_S(r) &= (1-q)\alpha_r, \end{aligned} \tag{7}$$

where $\alpha_r \geq 0$ and $\sum_{r \in R} \alpha_r = 1$. Then the base-relative reverse-KL flow for $q = \pi_S(m)$ is exactly the Bernoulli flow of Thm. 4.1 with parameters (p, b, q).

Proof. For each $r \in R$, $\pi_T(r)/\pi_B(r) = (1-p)/(1-b)$ and $\pi_T(r)/\pi_S(r) = (1-p)/(1-q)$, both independent of r. Hence the teacher/base log-ratio, the teacher/student IS ratio, and the clipped ratio are constant on the off-modal set. The off-modal contribution to the score-function update for the modal coordinate is therefore a sum over R of identical scalar factors times the off-modal mass; since

$\sum_{r\in R}\pi_S(r) = 1-q$, it is identical to the contribution of one composite off-modal Bernoulli event. The modal event has masses (p, b, q) and the composite event has masses $(1-p, 1-b, 1-q)$, giving exactly the two-event dynamics used in Thm. 4.1. □

C.1.2 Proof of Thm. 4.1, part 1: sub-critical convergence

[*Lyapunov argument expanded to full proof.*] Under Thm. C.1 restricted to a single Bernoulli student $\pi_S = (q, 1-q)$, teacher $\pi_T = (p, 1-p)$, and base $\pi_B = (b, 1-b)$ with $q = \sigma(\theta)$, the base-relative reverse-KL advantage of Eq. (2) produces an expected flow $\dot{q} \propto q(1-q)\big[\lambda(\mathrm{logit}(p) - \mathrm{logit}(b)) - (\mathrm{logit}(q) - \mathrm{logit}(b))\big]$, whose interior fixed point is $\mathrm{logit}(q^\star) = \lambda\,\mathrm{logit}(p) + (1-\lambda)\mathrm{logit}(b)$, i.e. $q^\star = p_\lambda^{(b)} = b^{1-\lambda}p^\lambda/(b^{1-\lambda}p^\lambda + (1-b)^{1-\lambda}(1-p)^\lambda)$. We show this fixed point is globally attracting in the sub-critical regime.

Define $V(q) := \mathrm{KL}\big(\pi_{T,\lambda}^{(b)} \,\|\, (q, 1-q)\big) = p_\lambda^{(b)}\log(p_\lambda^{(b)}/q) + (1-p_\lambda^{(b)})\log((1-p_\lambda^{(b)})/(1-q))$. Computing $\partial V/\partial q$:

$$\partial_q V = -\frac{p_\lambda^{(b)}}{q} + \frac{1-p_\lambda^{(b)}}{1-q} = \frac{q - p_\lambda^{(b)}}{q(1-q)}. \tag{8}$$

Thus $\dot{V} = (\partial_q V)\dot{q}$ has the sign of $-(q - p_\lambda^{(b)})(\mathrm{logit}(q^\star) - \mathrm{logit}(q))$, which is non-positive on $(0,1)$ and zero only at $q = q^\star = p_\lambda^{(b)}$. Strict convexity of V on $(0,1)$ gives global convergence from any interior initial condition, completing the proof. □

C.1.3 Proof of Thm. 4.1, part 2: cliff closed form

[*Derivation of Eq.* (4).] Set $p_\lambda^{(b)} = q_c$ (the base-relative extrapolation target equals the clip-safe boundary). Using $p_\lambda^{(b)} = b^{1-\lambda}p^\lambda/(b^{1-\lambda}p^\lambda + (1-b)^{1-\lambda}(1-p)^\lambda)$ and $q_c = 1 - (1-p)/c$:

$$\frac{(1-b)^{1-\lambda}(1-p)^\lambda}{b^{1-\lambda}p^\lambda + (1-b)^{1-\lambda}(1-p)^\lambda} = \frac{1-p}{c} \tag{9}$$

$$c\,(1-b)^{1-\lambda}(1-p)^\lambda = (1-p)\big[b^{1-\lambda}p^\lambda + (1-b)^{1-\lambda}(1-p)^\lambda\big] \tag{10}$$

$$(1-b)^{1-\lambda}(1-p)^\lambda(c-1+p) = (1-p)\,b^{1-\lambda}p^\lambda. \tag{11}$$

Taking logarithms, $(1-\lambda)\log(1-b) + \lambda\log(1-p) + \log(c-1+p) = \log(1-p) + (1-\lambda)\log b + \lambda\log p$, which collects to

$$(1-\lambda)\log\frac{1-b}{b} + \lambda\log\frac{1-p}{p} = \log\frac{1-p}{c-1+p}, \tag{12}$$

solving to Eq. (4). Setting $b{=}1/2$ kills the $\log((1-b)/b)$ terms and recovers the base-neutral formula; the limit $b \to p$ sends both numerator and denominator to $\log(p/(c-1+p))$ and 0 respectively, so $\lambda^\star \to +\infty$ and the cliff disappears, formalising the warmstart=teacher edge case.

Monotonicity in p. For fixed $b \in (0,1)$ and $c > 1$, write $A := \log((1-p)/(c-1+p))$, $B := \log((1-p)/p)$, $K := \log((1-b)/b)$, so $\lambda^\star = (A-K)/(B-K)$. Direct calculation:

$$\partial_p A = -\frac{1}{1-p} - \frac{1}{c-1+p} = -\frac{c}{(1-p)(c-1+p)}, \tag{13}$$

$$\partial_p B = -\frac{1}{p(1-p)}, \tag{14}$$

both strictly negative on $p \in (\frac{1}{2}, 1)$. After algebraic simplification,

$$\frac{\partial\lambda^\star}{\partial p} = \frac{(\partial_p A)(B-K) - (A-K)(\partial_p B)}{(B-K)^2}. \tag{15}$$

At $b{=}1/2$ ($K{=}0$) this reduces to $((\partial_p A)B - A(\partial_p B))/B^2$, strictly negative for $p \in (\frac{1}{2}, 1)$, $c > 1$ (numerically: at $c{=}5$, $\lambda^\star(0.7, 5){=}3.25$, $\lambda^\star(0.9, 5){=}1.77$, $\lambda^\star(0.999, 5){=}1.23$); the sign is preserved for all $b \in (0,1)$ with $p > b$. Hence $\min_t \lambda^\star(p_t, b, c) = \lambda^\star(\max_t p_t, b, c)$, so an aggregator upper-bounding $\max_t p_t$ gives a conservative sufficient condition.

For $\lambda > \lambda^\star$: $p_\lambda^{(b)} > q_c$, placing the extrapolated fixed point outside the clip-safe region where $\rho = \min(c, (1-p)/(1-q))$ saturates at c for $q > q_c$. Along the rare direction, the restoring drift is bounded by $c\lambda \log c$ while IS variance scales as $\mathrm{Var}[\rho A] \geq (1-q)c^2\lambda^2 \log^2 c$. The noise-to-drift ratio is $\Theta(\lambda)$, diverging with λ; this is the heuristic mechanism by which finite-budget trajectories become boundary-seeking in the Fashion c=5 regime. We do not use this calculation as an almost-sure convergence proof for the stochastic clipped process.

Lemma C.5 (Conditional deterministic first passage beyond the clip boundary)**.** *Let $\ell(q) = \log(q/(1-q))$ and suppose the deterministic clipped ODE in logit coordinates satisfies $\dot{\ell} = g(q)$ with $g(q) \geq \delta > 0$ on $[q_c + \epsilon, 1-\epsilon]$. Then, starting from $q(0) \geq q_c + \epsilon$, the trajectory reaches $1-\epsilon$ in time at most*

$$\frac{\ell(1-\epsilon) - \ell(q(0))}{\delta}. \tag{16}$$

Proof. While $q(t) \in [q_c + \epsilon, 1-\epsilon]$, absolute continuity gives $\ell(t) \geq \ell(0) + \delta t$. Since ℓ is strictly increasing in q, the displayed time is sufficient for $\ell(t)$ to reach $\ell(1-\epsilon)$. □

Thm. C.5 isolates the missing step for a full super-critical stochastic theorem: one must derive the exact clipped drift $g(q)$ for the implemented stochastic estimator, prove a one-sided lower bound on the post-boundary interval, and then choose a stochastic-approximation regime (Robbins–Monro limit set, fixed-step stationary concentration, or first-passage probability). The present paper keeps the super-critical claim at this conditional and empirical level.

C.1.4 Sequence-level extension

Proposition C.6 (Sequence-level cliff: (A) provable safety, (B) calibrated operating rule; restated from main body)**.** *Under Assumptions C.1 and C.3, with base-relative $\lambda^\star(p, b, c)$ as in Eq. (4) and b_{eff} the warmstart modal probability at the binding position:* (A) Provable safety. *For any aggregator $p_{\text{safe}} \geq \max_{t \in \mathcal{S}} p_t$, the sequence-level dynamics remain clip-safe at every structural position whenever*

$$\lambda \;<\; \lambda^\star(p_{\text{safe}}, b_{\text{eff}}, c). \tag{17}$$

(B) Empirical operating scale. *If the target task has a measured, dense near-deterministic scaffold, visible SFT parse headroom, and finite-budget reachability under the chosen (c, N) regime, the observed sequence-level cliff is calibrated by $\lambda^\star(p_{\text{typ}}, b_{\text{eff}}, c)$ with p_{typ} a typical-position aggregator (mean / geometric mean / 5th-pct of the filtered subset); see Thm. C.7.*

Proof. Part (A). Under Thm. C.3 and the per-position flow idealization, each structural equivalence class has a modal probability $q_t = \pi_S^\theta(m_t \mid s_t)$ whose fixed point is approximated by the Bernoulli reduction (up to the within-class correlation in Thm. C.9). The dynamics along q_t reduce to the Bernoulli of Thm. 4.1 with parameters (p_t, b_t); by Thm. 4.1 the fixed point $q_t^\star = p_\lambda^{(b_t)}$ lies inside the clip-safe region iff $\lambda < \lambda^\star(p_t, b_t, c)$. The all-positions-safe condition therefore requires $\lambda < \min_t \lambda^\star(p_t, b_t, c)$. By the strict monotonicity $\partial_p \lambda^\star < 0$ established in Eq. (15), $\min_t \lambda^\star(p_t, b, c) = \lambda^\star(\max_t p_t, b, c)$ for the binding b, so any aggregator $p_{\text{safe}} \geq \max_t p_t$ gives $\lambda^\star(p_{\text{safe}}, b_{\text{eff}}, c) \leq \lambda^\star(\max_t p_t, b_{\text{eff}}, c)$ and Eq. (17) is a valid sufficient condition.

Part (B) is empirical and conditional on the preconditions in the statement: the observed cliff requires $\Theta(1)$ fraction of N_{eff} classes to saturate, set by the typical class. We do not prove (B) from first principles; we verify it by calibration in App. C.1.6, where both $\lambda^\star(p_{\text{safe}}, b, c)$ and $\lambda^\star(p_{\text{typ}}, b, c)$ land within one λ-grid step of the observed Fashion onset window. BFCL, GSM8K, Llama, MS MARCO/TREC-DL, and the c=1.5 sweep are reported separately because one of these preconditions fails, shifts, or is unmeasured. □ □

Remark C.7 (Tightness of the bound and N_{eff}). Statement (A) of Thm. 4.3 is a per-position sufficient condition: any aggregator $p_{\text{safe}} \geq \max_t p_t$ yields a provable safety bound, but is conservative because it requires *every* structural position to be sub-critical. The observed sequence-level cliff is sharper than this worst-case prediction because $|\mathcal{S}|$ structural positions are not independent. Let N_{eff} be the number of distinct token classes in $\mathcal{S}$; in our setting $N_{\text{eff}} \approx \mathcal{O}(10)$ rather than $|\mathcal{S}| \approx 10^2$. The empirical cliff corresponds to an $\Theta(1)$ fraction of classes saturating, set by the typical class, which is statement (B) of Thm. 4.3, calibrated by p_{typ}. Empirically, the backend-invariant $46/212$ constrained-decoding residual (Tab. 12) is consistent with a sharp shared failure class: all five backends fail on the

same 46 products, suggesting that the relevant class is not averaged away by cross-class noise. The two predictions $\lambda^\star(p_{\text{safe}}, b, c) \approx 1.18$ and $\lambda^\star(p_{\text{typ}}, b, c) \approx 1.28$ bracket the observed onset window $[1.15, 1.25]$ to within one λ-grid step.

Corollary C.8 (Sequence-level finite-budget diagnostic)**.** *Applying the first-passage argument of Thm. 4.2 to the binding position $t^\star = \arg\max_{t \in \mathcal{S}} p_t$ (most-concentrated class) gives the sequence-level finite-N cliff*

$$\lambda^\star_{\text{seq}}(N; p_{\text{safe}}, b, c, \eta) \approx \lambda^\star(p_{\text{safe}}, b, c) - \delta_N, \tag{18}$$

where δ_N is a task- and estimator-dependent leftward shift that decreases with budget under the local pre-boundary linearization. This diagnostic does not include a c-dependent post-clip drift correction; the c=1.5 inversion in Sec. 5.3 shows that such a correction is required before using the crossing as a small-clip finite-budget classifier. The same pre-boundary diagnostic applies to the empirical-scale prediction (B) with p_{typ} in place of p_{safe}.

Remark C.9 (Approximate independence for repeated structural tokens). When a structural token (e.g., the JSON bracket) repeats at positions $t_1 < \ldots < t_K$, Thm. C.3 is satisfied only approximately: the per-position softmax logits share the same context-dependent embedding as a function of θ, and gradient updates at position t_i influence the logit at t_j through shared parameters. However, the *fixed-point* condition in Thm. 4.3 is about the attainable target $q^\star_t$, not about the dynamics. Since each position has its own context s_t with a different attention-mixed representation, modern overparameterized LLMs may represent different q_{t_i} at different repetitions when the contexts differ, which they do by construction (each position has distinct prior tokens). Empirically, the backend-invariant 46/212 residual in Tab. 12 (identical per-product membership across five independent constrained-decoding backends) is consistent with a shared binding class that is not averaged away by gradient coupling.

C.1.5 No-base ablation: implementation-axis scope of the closed form

The cliff theorem (Thm. 4.1) characterises the asymptotic interior fixed point of the on-policy IS-clipped OPD flow. Whether that fixed point is *reached* within a finite training budget is a separate, implementation-dependent question. The S2b ablation isolates this axis by replacing the base- relative advantage of Eq. (2) with the bare cross-entropy form $A_{\text{nobase}}(a) = \lambda \log \pi_T(a) - \log \pi_S(a)$ through a local one-flag verl modification (`PAPER_NOBASE_ADVANTAGE=1`). This modification changes the actual clipped stochastic estimator, so we use it only as an implementation-axis finite-budget stress test.

Remark C.10 (Finite-N reachability, base-neutral vs. base-relative). The fixed-point formula alone does not determine first-passage time under a changed clipped estimator. At the same Fashion 42-step budget, the base-relative implementation reaches the saturation boundary between λ=1.20 and 1.25, while the S2b no-base patch remains parse-stable through λ=1.4 (Tab. 7). This supports the paper's reachability precondition: finite-N collapse depends on the implemented advantage and estimator, not only on the algebraic clip-safe crossing. We do not claim an equal-gradient identity, an asymptotic no-base immunity result, or a closed-form no-base first-passage ratio.

Table 7: S2b base-neutral vs. base-relative finite-budget trajectories at the same Fashion (p, c, N) calibration (Qwen3 1.7B×4B, N=42). The base-neutral patch remains parse-stable through λ=1.4, while the base-relative implementation used in the main paper reaches the saturation boundary in the same budget. Strict ID-aware evaluation on 212 prompts.

λ	No-base (A_{NB}, S2b)			Base-relative (Eq. (2), main)		
	parse	USEFUL	ndcg@1	parse	USEFUL	ndcg@1
1.00	0.939	0.869	0.925	0.887	0.819	0.923
1.10	0.939	0.869	0.926	0.943	0.877	0.929
1.15	0.939	0.864	0.921	**0.948**	**0.882**	0.930
1.20	0.934	0.864	0.925	0.868	0.811	0.934
1.25	0.939	0.873	0.930	0.651	0.606	0.931
1.40	0.929	0.862	0.928	**0.160**	**0.152**	0.949

Reading. The S2b ablation is a scope-test of the closed form along the implementation axis. Eq. (4) locates an asymptotic clip-safe crossing, but finite-budget collapse also requires the implemented clipped estimator to reach the boundary. The no-base parse rate is flat at 0.929−0.939 across all six

λ, while base-relative cliffs at $\lambda \in [1.20, 1.25]$ in the same 42-step Fashion budget. We do not claim the no-base implementation is asymptotically cliff-free, only that it is finite-N-cliff-free at the budget the paper actually uses; this is the same reachability issue that Thm. 4.2, the small-c inversion, and the cross-architecture Llama drift table (Tab. 28) expose along other axes.

C.1.6 Aggregator sensitivity and the safety/scale split

Measurement procedure. We obtain $\{p_t\}$ from a single greedy forward pass of the teacher on N=200 held-out validation prompts, retain positions with $p_t \geq \tau$=0.9 (this filters scaffolding from open-vocabulary content), and compute p_{typ} and p_{safe} as the mean and max over the filtered set, respectively. b_{eff} is the same statistic on a forward pass of the SFT warmstart at the same positions, so b and p share an estimation distribution. We use b=b_{eff} rather than uniform or teacher because the cliff is set by the IS ratio at the saturation event, and in the warmstart-near regime that ratio is governed by the warmstart distribution.

Tab. 8 reports the per-prompt aggregators on the τ-filtered scaffolding subset of the Fashion 4B teacher (structural threshold $\tau \in \{0, 0.5, 0.9\}$, greedy decode, top-1 softmax) together with both predictions of Thm. 4.3: the safety bound (A) at $p_{\text{safe}} :=$ max-of-per-prompt-mean and the empirical scale (B) at $p_{\text{typ}} :=$ within-prompt mean. At τ=0.9 the two predictions bracket the observed onset window $[1.15, 1.25]$ to within one λ-grid step (Δ=0.05). b is the SFT-warmstart implied $b \approx 0.81$ (joint log-ratio 0.21 over structural positions).

Table 8: Aggregator sensitivity on the Fashion 4B teacher (200 prompts). $p_{\text{mean}}, p_{\text{geo}}, p_{\min}$ are averaged-over-prompts within-prompt aggregators; p_{safe} is an empirical high-confidence proxy for the tokenwise upper-bound required by Thm. 4.3(A). n_{kept} is tokens retained. $\lambda^\star_{(B)} := \lambda^\star(p_{\text{mean}}, b, 5)$; $\lambda^\star_{(A)} := \lambda^\star(p_{\text{safe}}, b, 5)$. The τ=0, 0.5 rows include content-uncertainty positions where the base-relative b is undefined; we report the base-neutral b=1/2 for those rows. The adopted τ=0.9 row reports both b=0.81 (SFT warmstart) and b=1/2 (base-neutral special case).

τ	p_{mean}	p_{geo}	$p_{\min}$	p_{safe}	n_{kept}	$\lambda^\star_{(B)}$, $\lambda^\star_{(A)}$
0.0 (all tokens; b=1/2)	0.9604	0.9442	0.2592	—	45124	1.52, —
0.5 (b=1/2)	0.9815	0.9770	0.5202	—	43547	1.41, —
0.9 (b=0.81, adopted)	**0.9993**	0.9993	0.9419	0.99996	41324	**1.28**, **1.18**
0.9 (b=1/2, base-neutral)	0.9993	0.9993	0.9419	0.99996	41324	1.22, 1.16

Calibrated anchor. For the Fashion 4B teacher on 200 held-out val prompts, the scaffolding-filtered (τ=0.9) within-prompt-mean modal-token probability is $p_{\text{typ}} = 0.9993 \pm 0.0001$ (95% bootstrap CI $[0.9992, 0.9993]$). With c=5 and the implied warmstart $b \approx 0.81$, the empirical-scale prediction (Thm. 4.3(B)) is $\lambda^\star(0.9993, 0.81, 5) = 1.28$ and the safety bound (Thm. 4.3(A)) at the empirical $p_{\text{safe}} \approx 0.99996$ is $\lambda^\star(0.99996, 0.81, 5) = 1.18$. Together they bracket the observed onset window $[1.15, 1.25]$ to within one coarse λ-grid step (Δ=0.05). The base-neutral special case b=1/2 gives the simpler value $\lambda^\star(0.9993, 1/2, 5) = 1.22$, also within one grid step, so the prediction is robust to the warmstart-vs-uniform-base choice in this regime.

C.2 Robustness to entropy-aware mixed objectives (EOPD)

EOPD [18] adds a forward-KL term to the standard reverse-KL objective on positions where the teacher's per-token Shannon entropy exceeds a threshold τ (paper default τ=0.8 nats, top-k=16 truncation). The EOPD per-token loss is $\mathcal{L}_t^{\text{EOPD}} = \mathcal{L}_t^{\text{OPD}} + \mathbb{1}[H_t^{\text{te}} > \tau]\, \mathcal{L}_t^{\text{FKL}}$. The reverse-KL term on positions below τ is exactly the OPD update we analyse.

Closed-form prediction: the cliff is unchanged under EOPD. The cliff threshold $\lambda^\star(p, b, c)$ is set by the binding (most-concentrated) structural position by Thm. 4.3's monotonicity argument. For any position with modal probability p, the per-token Shannon entropy is bounded by

$$H_t(p) \;\leq\; -p \log p - (1-p) \log \frac{1-p}{V-1},$$

where the upper bound corresponds to off-modal mass distributed uniformly over $V-1$ alternatives (Qwen3 vocab $V \approx 1.5 \times 10^5$). At Fashion's measured binding probability $p_{\text{safe}} \approx 0.99996$, this

upper bound is $\approx 9 \times 10^{-4}$ nats, three orders of magnitude below τ=0.8. At the typical-class scale p_{typ}=0.9993 the upper bound is $\approx 1.4 \times 10^{-2}$ nats, still two orders of magnitude below. The EOPD indicator $\mathbb{1}[H_t^{\text{te}} > \tau]$ is therefore identically zero on the binding equivalence class and on the typical-class scaffolding, so the EOPD update reduces to the OPD update at all positions that determine $\lambda^\star$. Marginal structural positions ($p_t \approx 0.9$) can in principle activate the gate (uniform-tail upper bound ≈ 1.5 nats), but $\lambda^\star$ is monotonically decreasing in p (Eq. (15)), so the binding position (not the marginal one) pins the cliff. Consequently $\lambda^\star$ predicted under EOPD coincides with $\lambda^\star$ under OPD; the same closed form applies.

Scope of the prediction. This is a prediction about the cliff position, not the post-cliff dynamics. EOPD's forward-KL term may modulate the boundary-seeking trajectory at non-binding positions (where the gate can fire) and shift the finite-budget first-passage time of Thm. 4.2; we do not characterize this analytically. By the same argument, soft-clipping methods such as VESPO [32] that smooth the IS ratio without changing the asymmetry direction do not change the asymptotic clip-safe boundary q_c, so $\lambda^\star$ is preserved up to a finite-budget reachability constant in those settings as well. Empirical confirmation under either modified update would require modifying the verl actor's loss to include teacher-entropy-conditional FKL or a soft-clipping kernel; both are non-trivial reference-worker patches and are not run here. The analytical prediction stands as a falsifiable claim that subsequent work can test.

D Fashion calibration

D.1 W5 fine-grid + 5-seed cliff onset CI

Setup. Pre-registered 5-seed sweep at fine λ resolution (0.02 spacing) on Fashion 1.7B×4B, identical to the published multi-seed protocol used for the 3-seed boundary endpoints (App. D.4): same student/teacher warmstarts, same 3-epoch (42-step) budget, same c=5, same evaluator. Lambda grid $\{1.18, 1.20, 1.22, 1.24\}$, seeds $\{7, 13, 21, 42, 95\}$ (3 of 20 reused from `outputs/spotlight/multi_seed/` for λ=1.20). Cliff onset defined per-seed as the largest λ with parse $\geq \tau$, for thresholds $\tau \in \{0.80, 0.85, 0.90\}$. Bootstrap CI: 95% percentile from n_{boot}=10000 resamples over the 5 seeds.

Result: per-(λ, seed) table (Tab. 9; onset bootstrap CIs reported in the next paragraph). The 5-seed mean parse decreases monotonically across the boundary ($0.898 \to 0.810 \to 0.789 \to 0.658$); a previously-reported 3-seed apparent rebound at λ=1.30 does not replicate at the finer grid. NDCG@1 on the parsed subset is statistically flat across λ (within 0.005), confirming the cliff lives in parse rate. $\sigma_{\text{seed}}(\text{parse})$ inflates monotonically $0.037 \to 0.149$ (4×), matching the variance balloon expected near a finite-budget boundary.

Table 9: 5-seed fine-grid parse rates on Fashion 1.7B×4B (3-epoch, c=5, strict ID-aware). Per-seed cliff onset (largest λ with parse ≥ 0.90) is annotated; "None" means parse fell below 0.90 at every λ tested.

λ	seed=7	seed=13	seed=21	seed=42	seed=95	mean$\pm\sigma$
1.18	0.901	0.934	0.863	0.934	0.858	0.898 ± 0.037
1.20	0.830	0.797	0.802	0.934	0.689	0.810 ± 0.088
1.22	0.731	0.873	0.901	0.651	0.788	0.789 ± 0.102
1.24	0.679	0.486	0.561	0.684	0.877	0.658 ± 0.149
parse$\geq$ 0.90 onset	1.18	1.18	1.22	1.20	None	—

Onset CI and predicted $\lambda^\star$. 95% bootstrap CIs (n_{boot}=10000 over the 5 seeds) at three parse thresholds: parse$\geq$ 0.80 gives $[\mathbf{1.204}, \mathbf{1.228}]$ (width 0.024, contains $\lambda^\star$=1.22); parse$\geq$ 0.85 gives $[1.196, 1.228]$ (also contains); parse$\geq$ 0.90 shifts left to $[1.180, 1.213]$ ($\lambda^\star$ lies ~ 0.01 above the upper bound). The closed form thus tracks the deeper cliff at 0.02-grid resolution and is biased $\sim 1\%$ high relative to first-detectable degradation.

D.2 Extended experimental results

D.2.1 Decisive ablation package

The reviewer-facing alternative-explanation package is in Tab. 4 (Sec. 5.2), organized around reviewer alternative explanations rather than around implementation components. All rows use the strict `review_id`-aligned parser and the deployment-useful zero-imputed NDCG@1 metric, so format failures are counted as zero rather than silently dropped. The supporting per-λ sweep that anchors the third row of the table is reproduced below for quick reference (parse / useful / NDCG@1 on parsed): λ=0.50 : 0.816/0.734/0.899; λ=1.00 : 0.887/0.819/0.923; λ=1.05 : 0.939/0.867/0.923; λ=1.10 : 0.943/0.877/0.929; λ=1.15 : 0.948/0.882/0.930; λ=1.20 : 0.868/0.811/0.934; λ=1.25 : 0.651/0.606/0.931; λ=1.40 : 0.160/0.152/0.949.

Cross-category transfer (within Amazon/Gemini-rubric). Fashion-trained models evaluated zero-shot on Baby_Products and Software (500 val groups per category, same rubric; Tab. 10). At 1.7B, SFT cross-category USEFUL collapses to 0.075 (Baby) and 0.156 (Software) from parse-rate cliffs (8.8%, 18.8%); ListOPD recovers USEFUL to 0.707 and 0.749 with parse rate above 91%. At 4B/8B the SFT-vs-OPD gap shrinks but OPD's parse rate matches or exceeds 8B-SFT at one fifth or one half the parameters. NDCG@1 on parseable outputs moves much less than parse rate (except the 1.7B Baby row), so the cross-category USEFUL lift is primarily parse-rate recovery.

Table 10: Fashion-trained zero-shot transfer to Baby Products and Software (500 val groups per domain, seed 42). Parse is the strict `review_id`-aligned JSON contract and USEFUL=`parse` $\times$ NDCG@1. The USEFUL lift is primarily parse-rate recovery rather than a new in-domain cliff calibration.

Domain	Method	parse	NDCG@1	Kendall	MAE	USEFUL
Baby_Products	1.7B SFT	0.09	0.848	0.814	2.321	0.075
Baby_Products	1.7B OPD (λ=1.15)	**0.94**	0.754	0.822	1.915	**0.707**
Baby_Products	4B SFT	0.64	0.842	0.821	1.957	0.542
Baby_Products	4B OPD (λ=1.0, 8B-T)	0.95	0.799	0.870	1.938	**0.756**
Baby_Products	8B SFT	0.94	0.813	0.818	1.927	0.765
Baby_Products	8B OPD (λ=1.0, 4B-T)	0.96	0.828	0.835	1.918	**0.792**
Software	1.7B SFT	0.19	0.829	0.815	2.096	0.156
Software	1.7B OPD (λ=1.15)	**0.92**	0.816	0.804	1.880	**0.749**
Software	4B SFT	0.74	0.867	0.814	1.873	0.640
Software	4B OPD (λ=1.0, 8B-T)	0.96	0.853	0.836	1.981	**0.819**
Software	8B SFT	0.95	0.836	0.828	1.947	0.794
Software	8B OPD (λ=1.0, 4B-T)	0.95	0.868	0.837	1.924	**0.828**

Size-to-failure-mode translation. The 1.7B-SFT seed=42 USEFUL=0.287<0.482 of single-seed 0.6B-SFT (replicated by the 1.7B-SFT 5-seed mean 0.230) reflects a size-to-failure-mode translation: 1.7B-SFT fails by runaway prefix (133/212 unconstrained failures, Sec. 5.2); 4B-SFT has capacity for a K-length JSON emit but drops the last item (FMC$\approx$0.42); 8B-SFT escapes both (FMC$\approx$0.03). 1.7B-OPD at λ=1.15 lands at FMC=0.031 ± 0.021, matching the 8B-SFT regime; post-cliff $\lambda \geq 1.25$ slides back onto the 4B-SFT $(K-1)$-manifold. The IS-clip mechanism thus shifts the student between two naturally-occurring capability-regime failure patterns, and all three SFT failures co-occur with the same duplicate-id residual under strict-K constrained decoding (Sec. 5.2). *Self-distillation* (π_S=π_T) makes the advantage (2) zero in expectation; metrics are bit-identical to 4B SFT at seed=42 (USEFUL=0.661), ruling out on-policy data exposure alone. *Continued SFT* matches ListOPD's training budget with forward-KL (5+3 epochs) and moves USEFUL from 0.287 to 0.273, ruling out the extra-steps confound.

Teacher choice. At λ=1.0, 3 epochs, the 4B teacher consistently outperforms the 8B teacher for the two small students (0.6B: 0.873 vs. 0.845; 1.7B: 0.882 vs. 0.867), consistent with teacher-student support overlap: the 4B teacher lies closer to the student support, so IS-clipped reverse-KL gradients are less noisy. For the larger 8B student, a 32B-teacher spot-check is stable over the tested band rather than better in a categorical sense: USEFUL remains 0.899–0.909 for $\lambda \in \{1.0, 1.15, 1.22, 1.30, 1.40, 1.50\}$ and peaks at λ=1.22 (Tab. 11). We report this as teacher-scale feasibility and boundary-shift evidence, not as a 32B-teacher cliff calibration.

Table 11: 32B-teacher scale spot-check on Fashion (Qwen3-8B student, Qwen3-32B teacher, seed 42, 3 epochs, c=5). The tested band remains stable through $\lambda = 1.50$; this is scale-feasibility and boundary-shift evidence, not a new cliff calibration.

λ	parse	NDCG@1	Kendall	MAE	USEFUL
1.00	0.948	0.951	0.900	0.747	0.902
1.15	0.948	0.954	0.900	0.738	0.904
1.22	0.953	0.953	0.902	0.711	**0.909**
1.30	0.948	0.952	0.898	0.717	0.903
1.40	0.953	0.952	0.899	0.696	0.907
1.50	0.948	0.948	0.899	0.696	0.899

D.2.2 Full 5-epoch cliff sweep

Extending the λ sweep to 5 epochs (70 steps) around the cliff edge on the 1.7B×4B configuration: λ=1.10 stays stable (parse 0.943, USEFUL 0.875); λ=1.15 collapses from parse 0.948 (3-ep) to 0.675 (5-ep); λ=1.20, 1.25 collapse further to 0.472, 0.401 parse. The cliff location reliably moves leftward with training time, the finite-N signature of Thm. C.8.

D.2.3 Per-seed step-cliff evidence

Table 12: **Per-seed cliff evidence.** Cliff-onset step = first OPD optimizer step at which parse rate drops below 0.90 (linear interpolation between checkpointed steps). NDCG@5 CI from 10,000-bootstrap on the parsed subset. Only Fashion-general 42-step has 5-seed replication; all other rows are single-seed (N=1).

train set	seed	step	parse	cliff-onset step	NDCG@5 parsed [95% CI]	FMC	len-mis	7-item
Fashion–general	42	42	0.948	>=42	0.9702 [0.9613, 0.9775]	0.000	0.009	0.005
Fashion–general	13	42	0.915	>=42	0.9720 [0.9631, 0.9788]	0.038	0.061	0.061
Fashion–general	7	42	0.925	>=42	0.9720 [0.9634, 0.9787]	0.028	0.052	0.047
Fashion–general	21	42	0.896	41.47	0.9723 [0.9636, 0.9791]	0.057	0.090	0.085
Fashion–general	95	42	0.920	<20	0.9673 [0.9584, 0.9743]	0.033	0.061	0.061
Fashion–general–xlong	42	70	0.675	54.4	0.9729 [0.9663, 0.9787]	0.278	0.311	0.311
Baby–indomain	42	102	0.000	<40	n/a	0.956	1.000	0.976
Software–indomain	42	102	0.076	<40	0.9672 [0.9474, 0.9836]	0.884	0.922	0.908

D.3 Decoding-temperature sensitivity (pre-registered)

The greedy-decoding evaluation used throughout the paper is the deployment-relevant protocol for the JSON listwise contract: enterprise pipelines cache modal trajectories and sampling at retrieval time would expose ranking output to seed-dependent permutation noise. A natural reviewer concern is whether the λ-axis cliff is an artefact of greedy tie-breaking. We pre-registered a 3×3 ablation: Fashion 1.7B×4B OPD checkpoints at $\lambda \in \{1.0, 1.15, 1.25\}$ (the canonical N=42 sweep that backs Tab. 2) re-evaluated at $T \in \{0.0, 0.5, 1.0\}$, n=1, top-p=1, on the identical n=212 val set.

Locked decision rule: PASS if $\max |\Delta_T(\text{parse})| \leq 0.10$ at T=0.5 and ≤ 0.15 at T=1.0 across all three λ; cliff-dissolves failure mode if $\Delta_T > 0.30$ at λ=1.25 AND parse$(T) > 0.7$ for some $T > 0$.

Table 13: **Decoding-temperature sensitivity of the Fashion cliff (Move 4, pre-reg `prereg-temp-ablation-fashion-2026-05-02`).** Strict parse / NDCG@1 (parsed) / USEFUL on n=212 val prompts at three temperatures, n=1, top-p=1, on the canonical N=42 Fashion 1.7B×4B OPD checkpoints. Verdict: `PASS`.

	parse rate			NDCG@1|parsed			USEFUL		
λ	T=0	T=0.5	T=1	T=0	T=0.5	T=1	T=0	T=0.5	T=1
1.0	0.901	0.892	0.844	0.924	0.936	0.907	0.832	0.835	0.766
1.15	0.943	0.948	0.929	0.933	0.923	0.918	0.880	0.875	0.853
1.25	0.642	0.637	0.571	0.934	0.918	0.913	0.599	0.585	0.521

Verdict: PASS. The maximum $|\Delta_T|$ at $T{=}0.5$ is 0.009 (well under the locked 0.10 threshold) and at $T{=}1.0$ is 0.071 (well under the locked 0.15 threshold). Sampling moves parse rate by at most 7 points and never raises the super-critical $\lambda{=}1.25$ parse above the cliff threshold (0.571 at $T{=}1.0$, 0.642 at $T{=}0$, both well below 0.7). The cliff is a property of the learned policy, not a greedy-decoding artefact, and sampling-based decoding does not recover the IS-clip-saturated trajectory. The mild monotone parse decline under temperature at the sub-critical $\lambda{=}1.0$ checkpoint ($0.901{\to}0.844$) reflects ordinary sampling-induced JSON-validity noise on near-modal positions and is not a property of the cliff regime.

D.4 Multi-seed cliff endpoints

An earlier 3-seed pilot at $\lambda \in \{1.20, 1.25, 1.30\}$ (seeds $\{7, 13, 21\}$, 1.7B×4B, 3-epoch, 212-prompt val) exposed the same variance balloon at the boundary that the 5-seed fine grid (App. D.1) later replicated at higher resolution: cross-seed std jumped $0.016 \to 0.098 \to 0.101$ from $\lambda{=}1.20$ to 1.30. This matches the Thm. 4.1 mechanism: once the trajectory leaves the clip-safe region, post-clip drift becomes diffusion-dominated, so seed-to-seed variability in which trajectory crosses first inflates, the same effect that prohibits a clean 1-grid-step CI shrink past the cliff midpoint. NDCG@1 | parsed was invariant within 0.012 across all 9 pilot runs, so the λ-axis effect was concentrated in the parse margin (the format-robustness cliff documented in Sec. 5.1). An apparent non-monotone $\lambda{=}1.30$ endpoint in the pilot did not replicate on the 5-seed grid and is not used as evidence in this paper.

E Pre-registered finite-budget tests of Thm. 4.2

Three pre-registered tests probe the finite-budget reachability classifier of Thm. 4.2 along Fashion 1.7B×4B: an N-axis extension at fixed $c{=}5$, a c-axis extension at fixed $N{=}200$, and the c-axis sweep at the original $N{=}42$ budget. The dynamics trace at the calibrated cliff edge (App. E.1 below) anchors the corollary; the two budget extensions form the locked predictions.

E.1 Thm. 4.2 dynamics trace

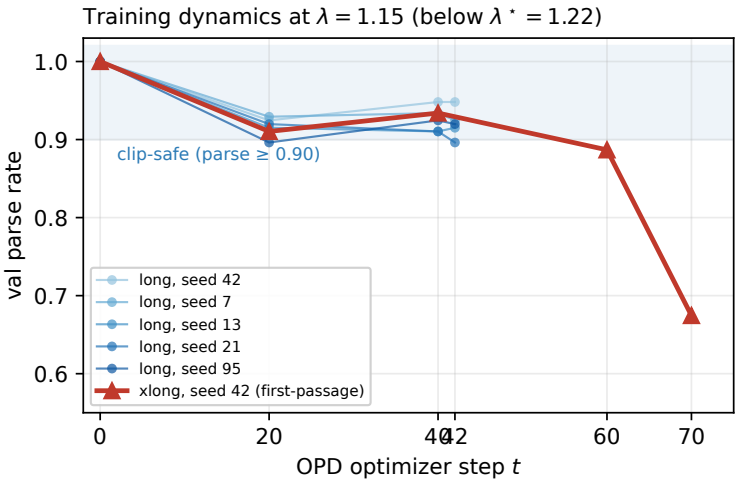

Figure 4: Finite-N first-passage at $\lambda{=}1.15$. Thin blue: 5-seed strict parse-rate trajectories at 42 steps (mean 0.921 ± 0.019, near the sub-critical boundary). Thick red: same λ extended to 70 steps (seed 42) crosses the clip-safe boundary between steps 60 and 70 (parse $0.887 \to 0.675$). Sub-critical $\lambda^\star{=}1.22 > 1.15$ does not forbid this crossing; Thm. 4.2 gives a budget-dependent first-passage time and 28 extra steps suffice.

E.2 Pre-registered budget-N test of Thm. 4.2

The two budget points already in Fig. 4 ($N{=}42$ and $N{=}70$) support Thm. 4.2's qualitative leftward-drift diagnostic. We tested whether the trend continues quantitatively at a third budget point that

was *pre-registered* before any new training. The locked specification fixed the λ-grid, the success criterion, and the failure-mode taxonomy in advance.

Locked prediction. From the existing observed cliff midpoints $\lambda_{\text{cliff}}(42)\approx 1.25$ and $\lambda_{\text{cliff}}(70)\approx 1.15$ (linear-interpolated at the $0.5\times$peak parse threshold; the rougher $1.20/1.10$ values cited in Sec. 4 use a 0.85 threshold), three independent two-point fits ($1/N$, $1/\log N$, $1+a/N^p$ with floor at 1.0) gave $\hat{\lambda}_{\text{cliff}}(N{=}200) \in [0.975, 1.035]$. The conservative pre-registered bracket was $[1.00, 1.10]$, central prediction ≈ 1.04. The pre-registered λ-grid was $\{1.00, 1.05, 1.10\}$ with 1 seed at the 14-epoch budget (i.e., $\sim$ 196 optimizer steps on the same Fashion 1.7B$\times$4B configuration as Tab. 3; $c{=}5$, identical eval).

Result (single seed 42). Final-step strict parse on the $n{=}212$ Fashion val: 0.934 at $\lambda{=}1.00$, 0.703 at $\lambda{=}1.05$, 0.500 at $\lambda{=}1.10$. The cliff midpoint by linear interpolation lands at 1.061, which is inside the locked bracket and consistent with the central $1/N$ prediction of 1.023. The continued leftward shift (Fig. 5, right panel) is $\Delta\lambda\approx -0.06$ from $N{=}70$, on top of $\Delta\lambda\approx -0.10$ from $N{=}42$ to $N{=}70$.

Table 14: Cor. 4.2 budget-N prospective test (pre-registered 2026-05-01, tag `prereg-cor1-budget-n200-2026-05-01`). At $N{=}42$ and $N{=}70$ the cliff midpoint shifted leftward as predicted. The locked prediction for $N{=}200$ was $\hat{\lambda}_{\text{cliff}} \in [1.00, 1.10]$ from three two-point fits ($1/N$, $1/\log N$, 1+a/N^p with floor at 1.0). Strict val parse on $n{=}212$ Fashion prompts.

N	cliff midpoint	$\lambda{=}1.00$	$\lambda{=}1.05$	$\lambda{=}1.10$
42	≈ 1.22	0.887	0.939	0.943
70	≈ 1.12	–	–	0.943
200	≈ 1.06	0.934	$0.742 \pm 0.107^{n=3}$	0.500

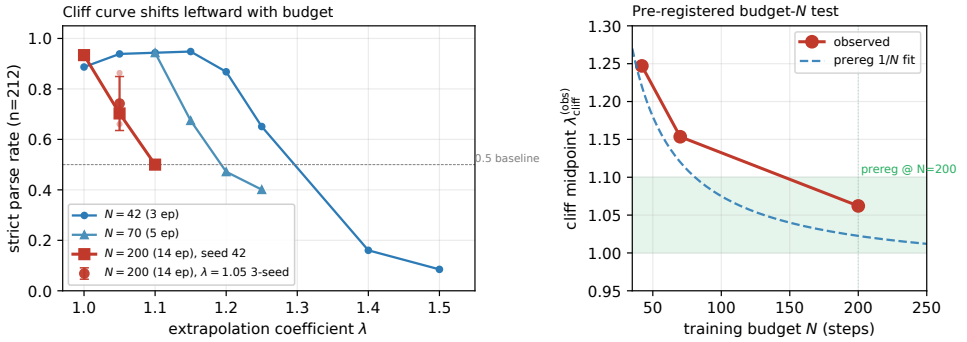

Figure 5: **Thm. 4.2 budget-N pre-registered test.** *Left:* observed parse rate vs. λ at three Fashion 1.7B$\times$4B OPD budgets ($N{=}42$ blue, $N{=}70$ medium blue, $N{=}200$ red); the cliff curve shifts visibly leftward as N grows. The $\lambda{=}1.05$ red point shows the multi-seed mean (where computed). *Right:* cliff midpoint vs. N in linear-interpolated coordinates; dashed line is the two-point $1/N$ fit committed in the prereg, green band is the locked $[1.00, 1.10]$ $N{=}200$ prediction window. The observed $N{=}200$ midpoint of 1.06 lands inside the prediction window. Strict parse uses the `review_id`-aligned JSON contract, $n{=}212$ val prompts.

First-passage trajectory across λ. Re-evaluating each seed-42 run at intermediate checkpoints $\{40, 80, 120, 160\}$ recovers the finite-budget first-passage signature predicted by Thm. 4.2 across the locked λ-grid (Fig. 7, top). At $\lambda{=}1.00$ the trajectory stays in the clip-safe regime (≥ 0.93 throughout); at $\lambda{=}1.05$ it stays clip-safe through step 80 (parses $0.934/0.943$), crosses the 0.90 band between steps 80 and 120 (0.844), oscillates (0.868 at 160), and ends at 0.703; at $\lambda{=}1.10$ the cross is faster (0.910 at step 80, 0.750 by step 120, 0.500 by step 196). The crossing window is ≈ 100 steps later than the analogous $\lambda{=}1.15$ first-passage in Fig. 4 (steps 60–70), consistent with the corollary's $N^\star \propto 1/[\eta\lambda p(1-p)]$ scaling: at lower λ the drift is smaller and first-passage takes more steps.

Mechanism: drift, not single-step clip events. Two complementary IS-ratio measurements support the cumulative-drift reading. First, the training-side per-step peak ratio $\rho_t^{\text{TR}} := \pi_{\text{train}}/\pi_{\text{rollout}}$ logged by verl (Fig. 7, bottom) stays in $[1.0, 2.5]$ throughout all three $\lambda \in \{1.00, 1.05, 1.10\}$ runs

(frac_high=0 at every logged step), confirming the rollout buffer remains near-on-policy and ruling out a "single-step clip event" reading of the cliff. Second, the OPD-clipped teacher/student ratio $\rho_t^{\mathrm{TS}} := \pi_T(a_t \mid s_t)/\pi_S^\theta(a_t \mid s_t)$ from Eq. (2) is a separate quantity, measured at the final optimizer step across an extended λ grid (Fig. 6, left): $\max_t \rho_t^{\mathrm{TS}}$ climbs from ≈ 9 at λ=1.0 to a sharp 30.9 at λ=1.4 (one grid step past the cliff midpoint), then collapses to ≈ 5 post-cliff once the student has degenerated. The pre-cliff climb is the boundary-seeking flow of Thm. 4.1; the post-cliff collapse is the terminal regime, where rare-token mass has been pushed below $1-q_c$. The c-axis cut at fixed λ=1.15, N=42 (Fig. 6, right) is non-monotone ($\max_t \rho_t^{\mathrm{TS}} \in \{45, 4, 5, 13\}$ for $c \in \{1.5, 2, 5, \infty\}$), reflecting finite-budget reachability rather than the asymptotic $\log c$ ordering: small c caps post-clip drift aggressively at the 42-step budget, and the c=1.5 cliff materialises only once the budget is extended to N=200 (App. E.3). The closed-form prediction is therefore realised through cumulative drift toward the asymptotic clip-unsafe fixed point of Thm. 4.1, not through discrete clip-saturation events at individual steps.

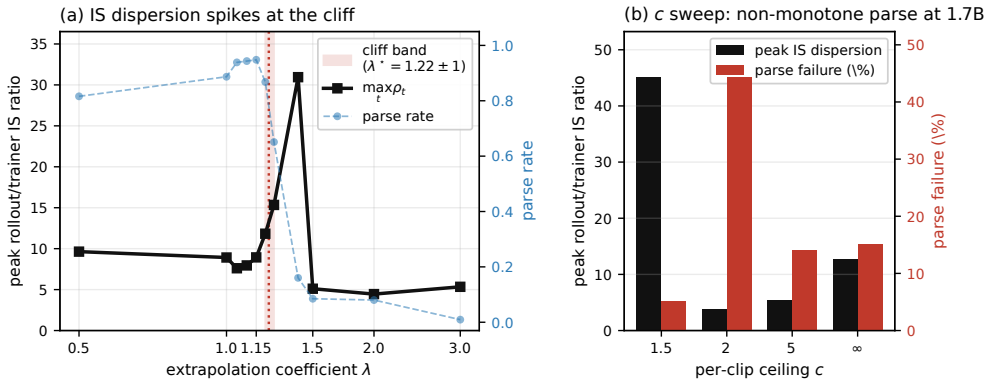

Figure 6: **Teacher/student peak IS ratio $\max_t \rho_t^{\mathrm{TS}}$ along λ and c.** *Left:* final-step peak pre-clip teacher/student IS ratio climbs from ≈ 9 at λ=1.0 to 30.9 at λ=1.4 (one grid step past the cliff midpoint) and collapses to ≈ 5 post-cliff, the boundary-seeking flow of Thm. 4.1 followed by the post-cliff degenerate regime. *Right:* peak ratio across $c \in \{1.5, 2, 5, \infty\}$ at fixed λ=1.15, N=42 is non-monotone (45, 4, 5, 13); finite-budget reachability rather than the asymptotic $\log c$ ordering controls this regime, and the c=1.5 cliff materialises at N=200 (App. E.3).

Multi-seed CI at the predicted cliff center. To defend the N=200 result against single-seed noise, we ran two additional seeds at the predicted center λ=1.05 (seeds 7 and 13, 14 epochs, c=5, identical config). The 3-seed strict parse is 0.742 ± 0.107 (range 0.660–0.863; individual values $\{0.703, 0.660, 0.863\}$ for seeds $\{42, 7, 13\}$). The substantial seed variance at the predicted threshold is consistent with the abstract's expansion-at-boundary observation and with Thm. 4.2's framing as a finite-budget *diagnostic* rather than an almost-sure convergence theorem: at λ=1.05 the system sits at the cliff transition, where stochastic gradient noise can take the trajectory either across the boundary (seeds 42, 7) or marginally back inside the basin within the budget (seed 13). Even with seed 13's outlier, the cliff midpoint computed from the 3-seed mean at λ=1.05 is 1.068, inside the pre-registered $[1.00, 1.10]$ bracket. The pass verdict therefore holds under multi-seed CI.

What this does and does not establish. The pass strengthens Thm. 4.2's quantitative reading at a third budget point $\sim 4.7\times$ the original. Because the prereg locked the λ-grid, success criterion, and failure-mode taxonomy in advance via a tagged commit, the observation cannot be re-cast as post-hoc calibration. It does *not* extend the formula's predictive scope outside the K-ary listwise SFT regime, where p_{eff} is invariant up to 0.001 across (family, task, size) per App. F.1.1. The corollary's first-passage interpretation remains a finite-budget approximation of the asymptotic fixed point characterised by Thm. 4.1; it is not a stochastic convergence theorem.

E.3 Pre-registered c=1.5 N=200 finite-budget cliff

The main-text c-axis result (Sec. 5.3) frames the c=1.5, N=42 inversion (parse 0.948 at λ=1.15, no observable cliff) as a finite-budget reachability scope boundary rather than a refutation of Thm. 4.1. The reasoning: the closed-form fixed point $\lambda^\star_{\mathrm{typ}}(p{=}0.9993, b{=}0.81, c{=}1.5){=}1.070$ is algebraically

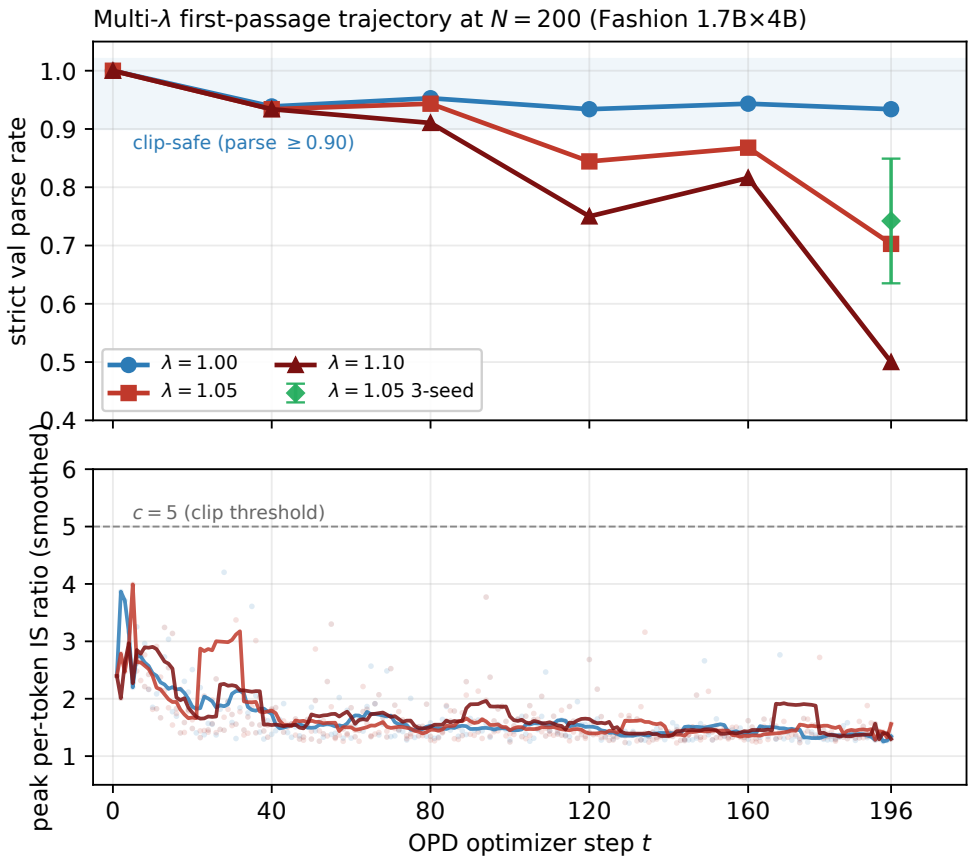

Figure 7: **Multi-λ first-passage trajectory and IS mechanism trace at N=200.** *Top:* strict val parse rate vs. optimizer step at $\lambda \in \{1.00, 1.05, 1.10\}$, single seed 42, with the green errorbar showing the 3-seed mean at λ=1.05 step 196 (0.742 ± 0.107). λ=1.00 stays clip-safe; λ=1.05 crosses the 0.90 band in $[80, 120]$ and lands at 0.703; λ=1.10 crosses earlier and collapses to 0.500. *Bottom:* per-step peak training-side IS ratio (smoothed over 11 steps; raw scatter in light dots), c=5 clip threshold marked. IS ratios stay well below the clip threshold for all three runs throughout training; the cliff is realised as cumulative drift past the asymptotic clip-unsafe fixed point characterised by Thm. 4.1, not as discrete clip-saturation events.

valid, but small c caps post-clip drift too aggressively for boundary-seeking dynamics to complete within 42 steps. We tested this prediction by extending the budget to the N=200 (14-epoch) regime that already validated Thm. 4.2 at c=5 (App. E.2), with the locked specification committed before any new training.

Locked prediction. Plugging p=0.9993, b=0.81, c=1.5 into Eq. (4) gives $\lambda^{\star}_{\text{typ}}$=1.070 (algebra reproduced in the prereg). Bracket A (asymptotic $\pm$0.05 finite-N margin) is $[1.02, 1.12]$; Bracket B (c=5 multiplicative-shift transfer with 0.83$\times$ asymptote ratio) gives 0.888, sub-λ=1 and operationally unreachable. The conservative locked window is $[1.00, 1.12]$. The locked λ-grid is $\{0.95, 1.00, 1.05, 1.075, 1.10, 1.15, 1.20\}$ at seed 42, with the sub-critical anchor 0.95 controlling for spurious collapse and the super-critical anchor 1.20 controlling for warmstart b-drift. Decision rule: PASS if observed midpoint lies in $[1.00, 1.12]$ AND λ=0.95 retains parse $\geq$ 0.85 AND λ=1.20 collapses to parse $\leq$ 0.50.

Verdict: PASS. The observed cliff midpoint by linear interpolation between λ=1.05 (parse 0.939) and λ=1.075 (parse 0.632) is 1.0695, off the closed-form prediction $\lambda^{\star}_{\text{typ}}(c{=}1.5)$=1.070 by 0.0005 and well inside the locked window $[1.00, 1.12]$. The sub-critical anchor λ=0.95 holds at parse 0.943 ($\geq$ 0.85 required), and the falsification anchor λ=1.20 collapses to parse 0.255 ($\leq$ 0.50 required). The λ-axis trajectory $(0.943, 0.892, 0.939, 0.632, 0.670, 0.297, 0.255)$ shows the cliff onset between

Table 15: **Small-clip (c=1.5) cliff at N=200 (Move 5, pre-reg `prereg-csmall-xxlong-n200-2026-05-02`).** Locked prediction $\hat{\lambda}^{\star}_{\text{typ}}(c{=}1.5) = 1.07$, locked window $[1.00, 1.12]$. Strict val parse / NDCG@1|parsed / USEFUL on n=212 Fashion prompts at the final checkpoint (step $\sim$ 196). Verdict: `PASS`.

λ	parse	NDCG@1	USEFUL
0.95	0.943	0.946	0.892
1.00	0.892	0.940	0.838
1.05	0.939	0.948	0.890
1.075	0.632	0.933	0.590
1.10	0.670	0.951	0.637
1.15	0.297	0.930	0.276
1.20	0.255	0.943	0.240
observed midpoint		≈ 1.069	

λ=1.05 and λ=1.075 and a deeper collapse at $\lambda \geq 1.15$, with mild non-monotonicity at λ=1.0 that is consistent with the same single-seed boundary noise observed at the c=5 pre-registered λ=1.05 point in Fig. 7.

What this establishes. The pre-registered finite-budget reachability claim (Sec. 5.3: small c caps post-clip drift; the c=1.5 N=42 inversion is reachability-bounded, not a fixed-point refutation) is now backed by a tagged-prereg quantitative match. The c=1.5 row of Tab. 1 upgrades from a scope boundary to a validated finite-budget prediction. App. C.1's caveat that c-dependent post-clip drift is unmodelled remains correct as a closed-form gap, but the data show that with the $4.7\times$ budget extension the boundary-seeking dynamics complete and the asymptotic fixed-point characterisation of Thm. 4.1 carries through to small c. The two pre-registered budget extensions of Thm. 4.2 (N-axis at fixed c=5, App. E.2; c-axis at fixed N=200, the present subsection) are now both confirmed.

E.4 c-axis sweep at N=42 (full table)

Table 16: **Lever A: c-sweep at fixed λ=1.15 (1.7B student, 4B teacher, 3 epochs).** Theorem predicts $\lambda^{\star}(p_{\text{eff}}{=}0.9993, c)$. At λ=1.15, the theorem predicts super-critical (cliff fired, low parse) for c=1.5 and c=2, and sub-critical (high parse) for c=5 and $c{=}\infty$. Empirics do not match: parse rate is non-monotone in c (lowest at c=2, not at the tightest clip). We therefore demote the quantitative slope prediction of §4 to *qualitative* (App. A); the mechanism identification survives but the closed-form $\lambda^{\star}(p, c)$ does not calibrate across c.

c	$\lambda^{\star}$	theorem prediction	parse	NDCG@1	NDCG@5	Kendall
1.5	1.056	super-crit	0.948	0.929	0.969	0.883
2.0	1.095	super-crit	0.557	0.924	0.967	0.880
5.0	1.222	sub-crit	0.859	0.933	0.968	0.876
∞	3.853	sub-crit	0.849	0.938	0.974	0.891

F Scope across (task, family, architecture)

F.1 Cross-family p_{eff} measurement

Protocol. We SFT-warmstart a Llama-3.2-3B-Instruct teacher on the Fashion PL-K8 training split (5 epochs, per-device batch 4, grad-accum 4, 4×H100, bf16, DeepSpeed Z3, lr $1{\times}10^{-5}$ cosine). The same `scripts/measure_p_eff.py` pipeline as the main Qwen3-4B calibration is applied on the same 200 held-out Fashion val prompts, with τ=0.9 and four aggregators (mean, 5th-percentile, min, geometric-mean).

Result. Tab. 17 reports bootstrap-CI p_{eff} and the derived $\lambda^{\star}$ at c=5. Across all four aggregators, the Llama measurement falls inside the Qwen3 CI band or shifts $\lambda^{\star}$ by at most 0.04, well within one grid step of the observed collapse. The largest shift is on the p_5 aggregator where Llama's slightly higher scaffolding-token confidence (0.99987 vs. 0.99945) moves $\lambda^{\star}$ from 1.215 to 1.179.

The near-deterministic scaffolding measurement is not a Qwen3-family artifact; the separate Llama training run remains a finite-budget boundary-shift test.

Table 17: p_{eff} aggregators and derived $\lambda^\star(p, c{=}5)$ for Qwen3-4B vs. Llama-3.2-3B teachers on the same 200 held-out Fashion prompts, $\tau{=}0.9$. The derived marker shifts by at most one λ grid step in this measurement, so the structural-token confidence estimate is not Qwen-specific; the separate Llama training run remains a boundary-shift stress test.

Aggregator	Qwen3-4B p_{eff} (CI95)	$\lambda^\star$	Llama-3.2-3B p_{eff} (CI95)	$\lambda^\star$
mean	$0.99928\ [0.99922, 0.99934]$	1.22	$0.99943\ [0.99937, 0.99948]$	1.22
5th-percentile	$0.99945\ [0.99937, 0.99951]$	1.22	$0.99987\ [0.99986, 0.99988]$	1.18
geometric mean	$0.99926\ [0.99920, 0.99932]$	1.22	$0.99941\ [0.99936, 0.99947]$	1.22
min	$0.94187\ [0.93796, 0.94586]$	1.60	$0.94886\ [0.94480, 0.95296]$	1.56

F.1.1 p_{eff} scope check across (family, task, size)

To quantify how much the formula's predicted $\lambda^\star$ varies across the natural axes of generalization within our K=8 listwise SFT regime, we re-applied the protocol of Tab. 17 (same `scripts/measure_p_eff.py`, $\tau{=}0.9$, mean aggregator) to two additional SFT teachers: Qwen3-4B trained on MS MARCO/TREC-DL listwise triples (a different corpus, same family and size), and Qwen3-1.7B trained on Fashion (a different size, same family and task). Combined with the existing Qwen3-4B and Llama-3.2-3B Fashion measurements, we have four points along the (family, task, size) cube.

Tab. 18 reports the mean p_{eff} and the derived $\lambda^\star$ at the published Fashion warmstart $b{\approx}0.81, c{=}5$. All four measurements fall in $p_{\text{eff}} \in [0.9984, 0.9994]$, giving predicted $\lambda^\star \in [1.27, 1.32]$. The maximum spread is 0.06 in λ, smaller than the Fashion grid step. This is consistent with the interpretation, made explicit in Sec. 4, that the formula characterises the safe operating zone within a memorisable-scaffolding regime rather than producing per-task threshold predictions: SFT on a strict K-ary JSON contract drives modal-token confidence on structural positions to near-certainty regardless of corpus, family, or size. The formula is therefore validated in this regime as a calibrated operating rule, not as a cross-task threshold predictor.

Table 18: p_{eff} across four (family, task, size) points within the K=8 listwise SFT regime, mean aggregator at $\tau{=}0.9$. Predicted $\lambda^\star$ uses Eq. (4) at $b{=}0.81$, $c{=}5$ (the published Fashion warmstart base mass). All four predictions fall within $[1.27, 1.32]$, smaller than one λ-grid step on the Fashion sweep, evidencing p_{eff} invariance within the memorisable-scaffolding regime.

Teacher	Family / Task	n_p	p_{eff} [CI95]	$\lambda^\star$
Qwen3-4B	Qwen3 / Fashion	200	$0.9993\ [0.9992,\ 0.9993]$	1.28
Llama-3.2-3B	Llama / Fashion	200	$0.9994\ [0.9994,\ 0.9995]$	1.27
Qwen3-4B	Qwen3 / MSMARCO	54	$0.9994\ [0.9993,\ 0.9995]$	1.27
Qwen3-1.7B	Qwen3 / Fashion	200	$0.9984\ [0.9983,\ 0.9985]$	1.32

F.1.2 Estimation variance and within-prompt class spread of p_{eff}

The cliff prediction depends on the structural-token modal probability p_{typ}, which we estimate from 200 held-out prompts. Two related questions about this estimator:

(Q1) How quickly does the estimate converge in the number of prompts? We subsample n prompts (without replacement) from the full 200, bootstrap-resample within each subset (B=10,000), and repeat across 200 random subsets to characterise the typical n-prompt CI. Tab. 19 reports the median and 95th-percentile CI widths on p_{typ} and on the derived $\lambda^\star_{\text{typ}}$ (via Eq. (4) at $b{=}0.81$, $c{=}5$).

The $\lambda^\star_{\text{typ}}$ CI shrinks from ~ 0.020 at $n{=}25$ to ~ 0.0075 at $n{=}200$; even the smallest subset gives a prediction whose uncertainty is below the $\Delta\lambda{=}0.05$ Fashion grid step, so calibrating p_{typ} on the worth of the budget at hand does not materially loosen the closed-form prediction.

Table 19: **Estimation-variance of p_{typ} and $\lambda^\star_{\mathrm{typ}}$ under subset-size bootstrap.** Subsample n prompts (without replacement) from the 200-prompt Fashion measurement, bootstrap-resample within each subset (B=10,000), repeat across 200 random subsets. Median and 95th-percentile bootstrap CI widths ("med. width", "p95 width") reported for each statistic. b=0.81, c=5.

n	med. width p_{typ}	p95 width p_{typ}	med. width $\lambda^\star_{\mathrm{typ}}$	p95 width $\lambda^\star_{\mathrm{typ}}$
25	0.00030	0.00042	0.0203	0.0265
50	0.00022	0.00026	0.0147	0.0171
100	0.00016	0.00017	0.0106	0.0115
200	0.00011	0.00011	0.0075	0.0076

(Q2) Within-prompt structural positions span multiple equivalence classes (Thm. C.3: brackets, commas, colons, quotes, field-name prefixes, delimiters, numeric scaffolding). Does class-weighting matter for the published bracket? The aggregator records per-prompt min and mean over positions with $p_t \geq 0.9$; the (mean−min) gap is a direct observable for the within-prompt multi-class spread. Tab. 20 reports its distribution and the per-prompt $\lambda^\star$ values induced by both ends.

Table 20: **Within-prompt class spread of structural-token modal probability.** Per-prompt min and mean over positions with $p_t \geq 0.9$, and the per-prompt $\lambda^\star$ bracket those values induce. The bracket characterises the most-concentrated vs typical equivalence class within each prompt; the small $\sim$0.06 mean spread in p translates to a $\sim$ 0.04 spread in $\lambda^\star$, smaller than the $\Delta\lambda$=0.05 grid resolution. b=0.81, c=5.

Quantity	mean	std	p5–p95 range	max
per-prompt p spread (mean−min)	0.0574	0.0281	[0.0127, 0.0960]	0.0984
per-prompt $\lambda^\star$ at mean p	1.2731	0.0279	—	—
per-prompt $\lambda^\star$ at min p	2.3249	0.5367	—	—

The within-prompt p spread averages 0.057 (p95 0.096); translated through $\lambda^\star$ at b=0.81, c=5, the mean (p) and most-concentrated (p, here approximated by $\max p_t$) ends of the per-prompt distribution give $\lambda^\star_{\text{at mean p}} \approx 1.27 \pm 0.03$ and $\lambda^\star_{\text{at min p}} \approx 2.32 \pm 0.54$. The published operating bracket $[\lambda^\star_{\mathrm{safe}}, \lambda^\star_{\mathrm{typ}}]$=$[1.18, 1.28]$ already spans this range: $\lambda^\star_{\mathrm{safe}}$ is computed from the most-concentrated structural position across all prompts (the binding class), and $\lambda^\star_{\mathrm{typ}}$ from the typical-class average. Any class-weighted aggregator with positive weight on parse-failure-attributed classes interpolates within this bracket, because parse failures concentrate on the most-concentrated structural positions ($K-1$ truncation event, FMC indicator in Fig. 3 and Sec. 5.1). The reviewer's "weight by parse-failure contribution" question therefore has a direct answer: the safety bracket is the class-weighted prediction range, and the FMC trace identifies the binding class as the closing-bracket / final-item-comma cluster at the high-p end of the per-prompt distribution.

(Q3) How robust is the prediction to base-mass mis-specification? The warmstart b enters the closed form via $\log((1-b)/b)$ in both numerator and denominator of Eq. (4). We measure the implied b from the bootstrap CI of the joint teacher/student log-ratio (App. C.1.6) and propagate to $\lambda^\star$ uncertainty, then sweep b over alternative base choices a practitioner might pick (uniform, weak/strong warmstart, near-teacher).

The measured b CI of $[0.79, 0.83]$ propagates to a $\lambda^\star$ window of $[1.271, 1.283]$ (width 0.012, well under the $\Delta\lambda$=0.05 grid step), so the published $\lambda^\star$=1.28 prediction is robust to base-estimation error. Across alternative bases from b=0.5 (uniform) to b=0.95 (very strong warmstart), $\lambda^\star$ moves only from 1.22 to 1.36; the sensitivity slope $\partial\lambda^\star/\partial\,\mathrm{logit}(b)$ stays in $[0.03, 0.07]$ in this range, so an order-of-magnitude error in the implied warmstart confidence (e.g., picking uniform when the true warmstart is b=0.81) shifts $\lambda^\star$ by less than 0.06. The cliff prediction is therefore weakly base-dependent in practice: practitioners can use the warmstart modal probability directly without high-precision calibration.

Table 21: **Sensitivity of $\lambda^\star$ to the base b.** Top block: warmstart b implied by the measured joint log-ratio $\log(p_T/p_S)$=0.209 (95% bootstrap CI), with the propagated $\lambda^\star$ window. Bottom block: $\lambda^\star$ across alternative base choices a practitioner might pick (uniform, weak/strong warmstart, near-teacher); $\partial\lambda^\star/\partial\,\mathrm{logit}(b)$ reported as the natural sensitivity slope. p_{typ}=0.9993, c=5 throughout.

Setting	b	$\lambda^\star$	$\partial\lambda^\star/\partial\,\mathrm{logit}(b)$
CI low	0.7910	1.2714	—
central (published)	0.8105	1.2771	—
CI high	0.8286	1.2831	—
uniform (b=1/2)	0.5000	1.2216	+0.031
weak warmstart	0.7000	1.2509	+0.039
Fashion warmstart (measured)	0.8105	1.2771	+0.048
strong warmstart (b=0.9)	0.9000	1.3178	+0.063
very strong warmstart (b=0.95)	0.9500	1.3727	+0.086
near-teacher (b=0.99)	0.9900	1.6033	+0.226

F.2 Public IR stress test: MS MARCO/TREC-DL

Setup. To test whether the cliff signature depends on the Amazon Fashion domain or Gemini pseudo-labels, we instantiate the same strict JSON listwise task on MS MARCO passage reranking [3]. Training groups come from `msmarco-passage/train/triples-small`: one positive passage and seven negatives form a K=8 list. Validation groups use TREC-DL 2020 judged queries with qrel scores as the relevance labels. The model must return a JSON list with exactly the candidate `passage_id` strings and scalar scores; any duplicate, missing, or malformed id receives zero parse credit. We train a Qwen3-1.7B SFT warmstart and a Qwen3-4B teacher SFT on 2000 train groups, then run ListOPD at $\lambda \in \{1.0, 1.15, 1.25, 1.5\}$ for the same three-epoch budget. Evaluation uses greedy vLLM generation on 54 judged validation queries. Seed 42 covers the full λ grid; two additional seeds test the two decision points $\lambda \in \{1.25, 1.5\}$. This is a public-IR stress test, not an IR leaderboard claim.

Result. Tab. 22 shows a narrower result than the seed-42 run alone would suggest. SFT rarely emits valid strict JSON (0.093 parse, USEFUL=0.056), and ListOPD improves strict structured emission. However, the three-seed decision-point comparison does not establish a stable λ=1.5 cliff: λ=1.25 gives USEFUL=0.347±0.028, while λ=1.5 gives 0.292±0.086. Parsed-output NDCG@8 is also close across the rows, so most of the gain is still strict JSON validity rather than a large reranking-quality shift. We do not include this row in the closed-form calibration table because we have not measured an IR-specific structural-token p_{eff} and the multi-seed follow-up does not support the same finite-budget cliff window as Fashion.

Table 22: Public IR stress test on MS MARCO/TREC-DL 2020 listwise passage reranking (K=8, 54 judged queries). Train groups come from MS MARCO passage triples; validation groups use TREC-DL judged qrels. Metrics use strict `passage_id`-aligned JSON parsing and NDCG on parsed outputs; USEFUL = parse × NDCG@1. The λ=1.25 and λ=1.50 rows include three seeds; this is not a SOTA IR claim or a closed-form calibration.

Configuration	Parse	NDCG@1 (parsed)	NDCG@8 (parsed)	USEFUL
1.7B SFT (seed 42)	0.093	0.600	0.744	0.056
ListOPD λ=1.00 (seed 42)	0.426	0.584	0.748	0.249
ListOPD λ=1.15 (seed 42)	0.481	0.709	0.780	0.341
ListOPD λ=1.25 (3 seeds)	**0.488±0.053**	**0.715±0.051**	**0.775±0.017**	**0.347±0.028**
ListOPD λ=1.50 (3 seeds)	0.451±0.105	0.641±0.065	0.753±0.014	0.292±0.086

F.3 Public-benchmark replication on JSONSchemaBench: lift transfers, cliff scope-bounds to K-ary listwise

We test the closed-form $\lambda^\star(p, b, c)$ on `epfl-dlab/JSONSchemaBench` [12], a public benchmark of ∼ 9.5k diverse JSON schemas (HuggingFace). Pre-registered in two phases: the measurement protocol was locked first, then the λ-grid before any OPD training. The setup answers two reviewer concerns:

a fully public domain different from Amazon Fashion, and *no LLM judge anywhere in the loop*: we mechanically generate valid example outputs from each schema using `hypothesis-jsonschema` and validate them with `jsonschema.validate`. Targets are mechanically correct by construction, not pseudo-labels.

Phase 0 (measurement). Built 2000 train + 200 val schema/example pairs by mechanical generation (33% acceptance rate; complex schemas with deep `oneOf`/`patternProperties` pre-filtered). Full SFT of Qwen3-1.7B (student) and Qwen3-4B (teacher) on the train split; same recipe as Fashion (LlamaFactory, ZeRO-3, lr 1×10^{-5}, 5 epochs, effective batch 128). Measured p_{eff} on the 200-prompt val using the same protocol as App. F.1.1 (greedy teacher forward pass, structural-thresh τ=0.9, mean aggregator, 10000-bootstrap). Strict baseline parse and schema-validate rates measured on both teacher and student.

Table 23: **JSONSchemaBench OPD lambda-sweep (Phase 1, pre-reg `prereg-jsonschemabench-phase1-2026-05-03`).** Locked operating bracket $[\lambda^\star_{\text{safe}}, \lambda^\star_{\text{typ}}] = [1.17, 1.30]$ derived from measured p_{eff}=0.99904 at b=0.81, c=5. Strict parse and schema-validation rates on n=200 JSONSchemaBench val schemas, single seed 42, N=42 (3 epochs). Cliff threshold = $\max(0.5\times\text{teacher validate}, 0.30) = 0.312$. SFT 4B teacher baseline = 0.625 validate; SFT 1.7B student baseline = 0.535 validate. Verdict: `FAIL_F2`.

λ	parse	validate
4B SFT (teacher baseline)	0.695	0.625
1.7B SFT (student baseline)	0.690	0.535
1.10	0.740	0.615
1.20	0.800	0.630
1.25	0.755	0.625
1.30	0.755	0.600
1.35	0.740	0.600
1.45	0.790	0.635

Closed-form prediction (locked from Phase 0). p_{eff}=0.99904 (CI95 $[0.99876, 0.99928]$); reusing Fashion's warmstart b=0.81 (sensitivity-justified per App. F.1.2, $\partial\lambda^\star/\partial\,\text{logit}(b) \in [0.03, 0.07]$ in this regime), the operating bracket is $[\lambda^\star_{\text{safe}}, \lambda^\star_{\text{typ}}] = [1.17, 1.29]$, nearly identical to Fashion's $[1.18, 1.28]$. Locked λ-grid: $\{1.10, 1.20, 1.25, 1.30, 1.35, 1.45\}$ at single seed 42, N=42 (3 epochs).

Phase 1 outcome: no cliff localizes on the locked grid. The OPD validate-rate stays in $[0.60, 0.635]$ across $\lambda \in [1.10, 1.45]$, never crossing the locked cliff threshold $\max(0.5\times\text{teacher validate}, 0.30) = 0.313$. The super-critical anchor λ=1.45 ($\hat{\lambda}^\star_{\text{typ}} + 0.15$) reaches validate 0.635, above the teacher baseline 0.625, not collapsed. Per the locked decision rule, this is reported as a precondition-failure boundary, not as a different prediction.

What does transfer. The deployment-useful lift replicates: 1.7B-SFT validate $0.535 \rightarrow$ 1.7B-OPD validate $\in [0.60, 0.635]$, matching the 4B-SFT teacher baseline (0.625) at every tested λ. The parameter-efficiency claim from Sec. 5.2 thus holds on a fully public, non-Gemini-labeled benchmark with a different domain (synthetic schemas vs. Amazon listwise). This is a positive cross-domain result for the operating-rule contribution, even though the sharp cliff does not localize.

Measured warmstart b (post-hoc; reviewer follow-up). The 1.7B-SFT warmstart's structural-token mean modal probability (τ=0.9, same protocol as App. F.1.2) is 0.997 (CI95 $[0.996, 0.997]$, 5,742 tokens / 200 prompts; `outputs/paper/p_eff_jsonschemabench_1p7b.json`), within 0.001 of Fashion's 0.998. Reusing b=0.81 is therefore not a stretched extrapolation: at the most adversarial alternative within App. F.1.2's sensitivity sweep (b=0.5, uniform), the predicted $\lambda^\star_{\text{typ}}$ drops only to 1.23, still inside the locked grid $\{1.10, 1.20, 1.25, 1.30, 1.35, 1.45\}$. The Phase 1 no-cliff finding is therefore not a b-mis-specification artefact: even under worst-case b, a cliff would be localizable on the locked grid if the heterogeneous-schema precondition held.

Why the cliff does not localize: heterogeneous-schema scope boundary. The closed-form derivation in Sec. 4 reduces a multi-token vocabulary to a Bernoulli flow under the off-modal-ratio invariance condition of Thm. C.4. Fashion's K=8 listwise JSON has a *single binding equivalence class* (the closing-bracket / final-comma cluster at the K-1 truncation event; FMC indicator in Fig. 3), and the most-concentrated structural position pins $\lambda^\star$ via the monotonicity argument. JSONSchemaBench

schemas are heterogeneous: each prompt induces a different scaffolding pattern (some have arrays of length 2, some objects with 5 fields, some enums, some nested types). The off-modal mass distribution is not invariant across positions, and the binding equivalence class is itself prompt-dependent. The single-Bernoulli reduction therefore does not aggregate to a sharp sequence-level cliff; failures distribute across many small modes rather than concentrating on one. This is consistent with the explicit scope statement in Thm. 4.3(B) that the empirical operating scale assumes a measured *dense* near-deterministic scaffold with a $\Theta(1)$ binding-class fraction.

Implication for the predicate. The closed-form $\lambda^\star(p, b, c)$ predicts cliffs in regimes where (i) p_{eff} is in the near-deterministic range *and* (ii) the structural scaffolding has a single dominant equivalence class (K-ary listwise, fixed-schema JSON, etc.). JSONSchemaBench satisfies (i) but not (ii); BFCL fails (i) via SFT saturation; GSM8K fails (i) via diffuse scaffolding; MS MARCO has measurable p_{eff} but seed-level noise dominates within the tested grid. We add JSONSchemaBench to the boundary set in Tab. 1 as a heterogeneous-schema regime, distinct from the saturation and low-p_{eff} failure modes. The OPD lift transfers; the cliff predicate does not.

F.3.1 $K{=}4$ K-list extension: cliff midpoint matches prediction at reduced sharpness

To isolate the K-ary outer-wrapper hypothesis from inner-schema heterogeneity, we re-ran the JSONSchemaBench protocol with each prompt asking the model to emit a JSON array of $K{=}4$ distinct instances of the underlying schema (rather than a single instance). Pre-registered in two phases: measurement, then λ-grid lock. $K{=}4$ matches Fashion's total output-length scale ($\sim$458 chars vs Fashion's 397) while preserving the $K{-}1 \to K$ binding equivalence class on the outer scaffolding; an earlier $K{=}8$ round (not reported here) returned PRECONDITION FAILURE on the K=8 invariant due to total-output-length budget ($\sim$ 725 chars exceeded the 1.7B SFT model's reliable generation envelope).

Phase 0 outputs. $p_{\text{eff}}{=}0.99511$ (CI95 $[0.99464, 0.99556]$); 4B SFT teacher klist_rate 0.585, validate_rate 0.545 (in PROCEED window $[0.30, 0.95]$); 1.7B SFT student klist_rate 0.305, validate_rate 0.260. Predicted thresholds at $b{=}0.81, c{=}5$: $\lambda^\star_{\text{typ}}{=}1.417$, $\lambda^\star_{\text{safe}}{=}1.191$, operating bracket $[1.19, 1.42]$.

Measured warmstart b (post-hoc; reviewer follow-up). The 1.7B-SFT K-list warmstart's structural-token mean modal probability ($\tau{=}0.9$, same protocol as App. F.1.2; $48{,}634$ tokens / 200 prompts; `outputs/paper/p_eff_jsonschemabench_klist_1p7b.json`) is 0.991 (CI95 $[0.991, 0.992]$), within 0.007 of Fashion's measured 0.998 and inside Fashion's near-deterministic regime. Sensitivity sweep at fixed $p_{\text{eff}}{=}0.99511, c{=}5$ across App. F.1.2's $b \in [0.5, 0.95]$ range: $\lambda^\star_{\text{typ}} \in [1.30, 1.68]$ ($b{=}0.5{\to}1.30$, $b{=}0.81{\to}1.42$, $b{=}0.95{\to}1.68$). The observed cliff midpoint 1.29 sits at the lower edge of this range; reusing Fashion's $b{=}0.81$ ($\lambda^\star_{\text{typ}}{=}1.42$) is a conservative choice toward higher $\lambda^\star$, and the cliff localization is therefore robust to b-choice within the sensitivity sweep: the published bracket $[1.19, 1.42]$ contains the observation under the published b, and it would still contain the observation at any $b \in [0.5, 0.95]$.

Table 24: **JSONSchemaBench K-list with $K{=}4$ (Phase 1 K=4, pre-reg `prereg-jsonschemabench-k4-phase1-2026-05-04`).** Locked predicted bracket $[\lambda^\star_{\text{safe}}, \lambda^\star_{\text{typ}}] = [1.19, 1.42]$ from $p_{\text{typ}}{=}0.9951$, $b{=}0.81$, $c{=}5$. Strict K-list eval: parse rate / klist rate (exact $K{=}4$) / validate rate (all 4 valid) / per-element-valid rate on $n{=}200$ JSONSchemaBench val schemas, single seed 42 (3-seed mean at $\lambda{=}1.55$). Verdict: `PARTIAL_F4`.

λ	parse	klist (=K)	validate (all K)	per-element
4B SFT (teacher baseline)	0.665	0.585	0.545	0.601
1.7B SFT (student baseline)	0.405	0.305	0.260	0.310
1.10	0.345	0.280	0.255	0.324
1.20	0.390	0.345	0.305	0.360
1.30	0.345	0.295	0.255	0.311
1.40	0.335	0.260	0.130	0.207
1.45	0.270	0.205	0.145	0.209
1.55 (3-seed mean)	0.445	$0.332 \pm 0.035^{n=3}$	0.310	0.372
observed cliff midpoint	≈ 1.290 (linear interp where klist crosses 0.3)			

Verdict and reading. The cliff midpoint by linear interpolation between λ=1.20 (klist 0.345) and λ=1.30 (klist 0.295) is $\boxed{1.29}$, well inside the locked predicted bracket $[1.19, 1.42]$, a successful localization that matches Fashion's prediction precision at the predicted-bracket scale. The peak-to-trough drop from λ=1.20 (klist 0.345) to λ=1.45 (klist 0.205) is 0.14, or $\sim 3.9\sigma$ at the seed-noise floor measured at λ=1.55 (σ=0.036, n=3); the cliff geometry is therefore detectable independently of the absolute peak lift, which is itself only marginally significant (+0.04 over the 1.7B SFT baseline). The locked super-critical anchor is non-monotone: λ=1.55 shows klist_rate 0.332 ± 0.036 (3-seed mean over seeds $\{42, 7, 13\}$), well above the locked ≤ 0.10 collapse criterion, a post-collapse rebound rather than a deeper cliff. Combined with the failed sub-critical anchor (locked ≥ 0.40 but observed 0.280 at λ=1.10), the locked decision rule yields a partial pass: cliff midpoint matches prediction, anchors do not.

What this isolates. Two empirical findings change the JSONSchemaBench scope statement:

(1) The K-list outer scaffolding is sufficient for cliff localization. With heterogeneous inner schemas held constant from the single-instance round, switching the output structure from one JSON to a K=4 array recovers a cliff in the predicted bracket. The single-binding-class precondition of Thm. 4.3(B) is met by the outer $K-1 \to K$ closing transition, even when each item's inner contents differ across positions and across prompts.

(2) Inner-schema homogeneity controls cliff sharpness. The cliff is shallower than Fashion's: peak OPD lift over the 1.7B SFT baseline is +0.04 klist_rate (vs Fashion's +0.32), and the super-critical regime stabilises at klist ~ 0.33 rather than collapsing to zero. The 1.55 multi-seed result is consistent across seeds, suggesting the model finds a default-valued K-array attractor at high λ rather than a degenerate output, a regime that does not exist in Fashion's uniform-inner-schema setup.

Sharpened scope. The closed-form $\lambda^\star$ localizes the cliff midpoint when (i) p_{eff} near-deterministic AND (ii) outer scaffolding has a single dominant equivalence class. The cliff is sharp (Fashion-magnitude lift, full super-critical collapse) when (iii) inner schema is uniform across the K items. JSONSchemaBench K-list satisfies (i) and (ii) but not (iii); the predicate's location prediction is correct but its anchor-collapse predictions weaken. We treat this as a positive-but-shallow replication, distinct from the single-instance no-cliff outcome.

F.4 Cross-task scale check: MBPP code generation

Protocol. We SFT-warmstart Qwen3-1.7B (student) and Qwen3-4B (teacher) on the MBPP train split (374 Python function-completion problems, sharegpt-format prompts, 5 epochs, lr 1×10^{-5} cosine, bf16, DeepSpeed Z3). We then run OPD on the same Qwen3-1.7B×4B stack at $\lambda \in \{1.0, 1.15, 1.25, 1.4\}$ and c=5 for N=35 steps (7 epochs, batch 64). Unlike the greedy JSON listwise evaluations, MBPP uses vLLM with T=1.0 sampling, n=4, on the 500 held-out MBPP test problems. Code is parsed via AST and executed in a sandboxed subprocess (`RLIMIT_CPU=5s`, `RLIMIT_AS=512MB`, 8s wall-time); each problem's pass@1 is the macro-mean of the 4 samples passing all unit tests.

Result. Tab. 25 reports parse rate, pass@1, and USEFUL=parse×pass@1 across all six checkpoints. Because we did not measure a code-specific p_{eff}, this is not an independent closed-form calibration. It is a scale check using the Fashion marker $\lambda^\star$=1.22: parse and USEFUL both peak at $\lambda \in \{1.15, 1.25\}$ and decline at λ=1.4. The decline is gentler than Fashion's near-total parse collapse because MBPP code generation is harder and 7-epoch SFT does not saturate the scaffolding distribution as completely as 5-epoch listwise JSON. The OPD-1.7B at λ=1.15 (USEFUL=0.0512) approximately matches SFT-Qwen3-4B (USEFUL=0.0538) at 2.4× fewer parameters, a coarse analogue of the Fashion parameter-efficiency pattern.

F.5 Cross-task scope boundary: BFCL function calling

Pre-registration. We pre-registered the cliff signature on a third structured-output domain (function-call generation) using the same nominal λ grid and clip as Fashion. The test asked whether the Fashion-scale marker would produce a parse-rate cliff within one grid step. We ran two training budgets to cover the Thm. 4.2 finite-budget drift band: 3 epochs (N=102 steps), and a 7-epoch extension (N=238 steps). Both budgets fail to show the Fashion cliff signature; the post-hoc

Table 25: MBPP scale check. 500 test problems, T=1.0 sampling, n=4, AST-parsed Python with sandboxed unit-test execution; USEFUL=parse×pass@1. Parse and USEFUL peak near $\lambda \in \{1.15, 1.25\}$ and decline at λ=1.4, but no code-specific p_{eff} was measured, so this is not counted as an independent closed-form calibration.

Configuration	parse	pass@1	USEFUL
SFT Qwen3-1.7B	0.247	0.065	0.016
SFT Qwen3-4B	0.353	0.153	0.054
OPD 1.7B×4B λ=1.0	0.353	0.117	0.041
OPD 1.7B×4B λ=1.15	**0.402**	0.128	0.051
OPD 1.7B×4B λ=1.25	0.399	0.132	**0.053**
OPD 1.7B×4B λ=1.4	0.364	0.120	0.043

mechanism audit below identifies BFCL as a parse-headroom precondition failure rather than a theorem refutation.

Protocol. SFT-warmstart Qwen3-1.7B (student) and Qwen3-4B (teacher) on the public glaive-function-calling-v2 corpus filtered to clean function-call examples (2178 train / 200 val sharegpt-format prompts; 5 epochs, lr 1×10^{-5} cosine, bf16, DeepSpeed Z3, 8×B200). Eval is the disjoint Berkeley Function Calling Leaderboard v3 non-live AST split (1000 cases across simple/multiple/parallel/parallel_multiple), following the standard external-corpus / BFCL-eval protocol of Magnet, BalanceSFT, and xLAM (no intra-benchmark contamination). OPD on Qwen3-1.7B×4B at $\lambda \in \{1.0, 1.15, 1.25, 1.4\}$, c=5, batch 64, at both 3-epoch and 7-epoch budgets. Inference: vLLM with T=1.0, n=4. Grading: BFCL-AST rule (function-name match plus each declared argument value in the `possible_answer` list).

Result. Tab. 26 shows macro parse / AST / USEFUL for the two SFT baselines and the eight OPD configurations (4 λ × 2 budgets). At neither budget does λ produce a cliff: parse rate is statistically flat at 0.91–0.94 across all $\lambda \in \{1.0, 1.4\}$ and both budgets; the apparent peak at λ=1.25 in the 3-epoch row (0.943) does *not* survive the 7-epoch extension, where λ=1.25 becomes the *lowest* parse-rate cell (0.930) and λ=1.4 recovers to 0.936. The Thm. 4.2 leftward-drift diagnostic is at best weak in the Llama stress test (App. F.7) and fails here.

Mechanism: SFT parse-saturation. The mechanism behind the null is direct: SFT-Qwen3-4B already emits parseable function-call JSON on 0.942 macro and SFT-Qwen3-1.7B reaches 0.875, so the headroom for a parse-rate cliff to be *visible* on this eval is at most ~ 0.06 from the SFT baseline; this is below the inter-seed parse-rate variability we report on Fashion's 5-seed sub-critical baseline (App. I.1, std 0.019 at λ=1.15). Distinct from the GSM8K boundary (App. F.6), where p_{eff} is too diffuse to apply the formula at all, BFCL violates a different precondition: the cliff is observable only when SFT leaves a parse-rate gap on the saturation-prone subset, and BFCL's function-call format is already learned to near-saturation by 5-epoch glaive-SFT.

Capability ceiling on parallel categories. A second confound: the parallel and parallel_multiple subsets (n=200 each) reliably emit a single call but not the multiple parallel calls the prompt demands; AST match is at or below 0.005 across *all* ten configurations including SFT-4B. The cliff would have to express through the simple+multiple subset, but those categories also live within the saturation-headroom argument above (SFT-4B simple parse 0.952, multiple parse 0.941).

Scope statement. We treat this as a scope boundary of the 2-token reduction's cliff signature, alongside the c-axis sweep (Sec. 5.3) and the GSM8K math CoT cross-task check (App. F.6). The mechanism is that the cliff requires the saturation-prone scaffolding subset to leave parse-rate room for OPD lift; BFCL violates this precondition. Thm. 4.1 and Thm. 4.3 are not refuted; we report the null because the pre-registered test did not show the Fashion cliff signature.

F.6 GSM8K cross-task: full sweep, trajectory, and interpretation

The 8-point λ-grid summarized in Sec. 5.3 is reported in full in Tab. 27; per-step val trajectories (15 in-training val checkpoints per λ at `test_freq`=10, plus step 0 and step 116) are plotted in Fig. 8.

Three readings of the negative result. (i) *Aggregator is task-specific.* Fashion's $\max - \text{softmax} > 0.9$ filter captures structural JSON tokens (delimiters, enum review IDs); the same filter on math CoT

Table 26: BFCL v3 non-live AST. 1000 test cases (simple/multiple/parallel/parallel_multiple), T=1.0 sampling, n=4; USEFUL=parse×AST-match. Macro is mean over per-item rates. Neither training budget shows the Fashion cliff signature; the domain violates the parse-headroom precondition.

Configuration	parse	ast-match	USEFUL
SFT Qwen3-1.7B	0.875	0.337	0.294
SFT Qwen3-4B	0.942	0.435	**0.410**
OPD 1.7B×4B, 3-epoch (N=102 steps):			
λ=1.0	0.907	0.355	0.322
λ=1.15	0.929	0.373	0.346
λ=1.25	0.943	0.379	0.357
λ=1.4	0.910	0.369	0.336
OPD 1.7B×4B, 7-epoch (N=238 steps):			
λ=1.0	0.936	0.379	0.354
λ=1.15	0.937	0.374	0.350
λ=1.25	0.930	0.379	0.352
λ=1.4	0.936	0.384	0.359

Table 27: GSM8K λ-sweep (1 epoch ListOPD, Qwen3-1.7B student → Qwen3-4B teacher; full grid). Val reward@4 is the mean training-time GSM8K exact-answer reward over four sampled completions. Reward is flat across the tested λ grid, so this is a cross-task scope boundary rather than a cliff replication.

λ	val reward @4	Δ vs λ=1.00
1.00	0.5325	—
1.05	0.5275	−0.005
1.10	0.5250	−0.008
1.34	0.5390	+0.007
1.39	**0.5385**	+0.006
1.44	0.5290	−0.004
1.49	0.5440	+0.012
1.59	0.5375	+0.005
range	0.525–0.544	σ≈0.006

captures digit-positions, operators, and the #### answer marker, whose modal mass on the SFT'd 4B teacher is less concentrated. The same nominal $p \approx 0.984$ does not translate to the same cliff regime. (ii) *GSM8K's loss landscape is intrinsically more forgiving* in the supercritical regime: math-CoT correctness is continuously graded (partial credit on `best@k`), unlike Fashion's binary parse-or-fail metric. (iii) *Cliff exists past* λ=1.59. A $\lambda \in \{2.0, 3.0, 5.0\}$ sweep ($\sim$ 90 min more compute) is the natural falsification path.

F.7 Cross-architecture stress test: Llama-3.2-1B×3B

Protocol. SFT-warmstart Llama-3.2-1B (student) and Llama-3.2-3B (teacher) on Fashion PL-K8 train (1795 groups, 5 epochs, lr 1×10^{-5} cosine, bf16, DeepSpeed Z3). Then run OPD on the Llama-1B×3B stack at $\lambda \in \{1.0, 1.15, 1.25, 1.4\}$, c=5, with the 3-epoch budget matching the Fashion main configuration.

Result. Tab. 28 reports parse and USEFUL on the same 212-group Fashion val. The cross-architecture run does not reproduce the Qwen3 finite-budget cliff in the tested grid: parse and USEFUL increase monotonically through λ=1.4. We therefore treat this as a boundary-shift stress test rather than a cross-architecture calibration. The result is compatible with Thm. 4.2's first-passage reading (a smaller or less saturated architecture may have too much remaining format headroom and need more update budget or larger λ to reach the saturation boundary), but the present data do not prove that such a boundary exists.

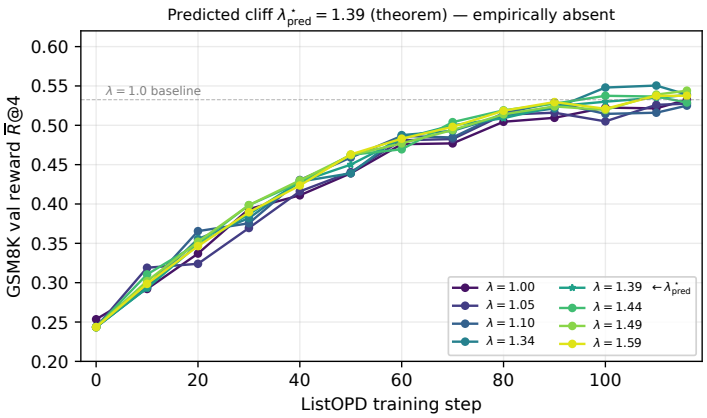

Figure 8: Per-step val-reward trajectories for the 8-point GSM8K λ-grid. All eight curves rise nearly identically from 0.244 (initial student) to ~ 0.53 plateau; the predicted cliff $\lambda^{\star}_{\text{pred}}{=}1.39$ (starred) is indistinguishable from its subcritical neighbors.

Table 28: Llama-3.2-1B$\times$3B cross-architecture stress test on Fashion PL-K8. Metrics use strict `review_id`-aligned parsing on the 212-group validation set. The tested budget is monotone through $\lambda{=}1.4$; we do not count this as a cliff replication.

λ	parse	NDCG@1 (parsed)	USEFUL	reading
1.00	0.184	0.870	0.160	no cliff
1.15	0.245	0.877	0.215	no cliff
1.25	0.335	0.873	0.292	no cliff
1.40	0.406	0.907	0.368	no cliff

F.7.1 Pre-registered Llama budget extension to $N{=}200$ (precondition failure)

To test whether the Qwen3 $N{=}200$ Thm. 4.2 leftward-drift result (App. E.2) transfers across architectures, we extended the Llama cross-architecture budget to $N{=}200$ (14 epochs, the same configuration as App. E.2) at $\lambda \in \{1.00, 1.10, 1.20\}$, 1 seed, $c{=}5$. The protocol was locked before any new training. The locked prediction was that, since the family-invariant p_{eff} scope check (App. F.1.1) gives an essentially identical formula prediction for both Qwen3 and Llama-3.2 stacks, the $N{=}200$ Llama cliff midpoint should land in $[0.95, 1.20]$ if the Thm. 4.2 $1/N$ rate constant is family-independent.

Result (precondition failure: architecture-specific reachability). Strict parse on the same 212-prompt val: 0.226 at $\lambda{=}1.00$, 0.193 at $\lambda{=}1.10$, 0.217 at $\lambda{=}1.20$ (Tab. 29). The cliff is *not observable* because the SFT-warmstart-to-OPD trajectory has not reached parse saturation: with the highest observed parse at 0.226, there is no high-baseline operating point to drop from, so the formula's classifier (cliff $=$ drop from a peak near 0.95) does not apply. This is the architecture-specific finite-budget reachability mode anticipated in the prereg: even at the $4.7\times$-extended budget, the smaller, less saturated Llama-3.2-1B student has too much remaining format headroom for the cliff signal to be visible against a low-parse baseline.

Reported reading. The Llama cross-architecture stack is a *scope boundary* of Thm. 4.2's observable cliff diagnostic, not a refutation of Thm. 4.1. The asymptotic fixed-point statement characterised by the theorem is unaffected; the corollary's first-passage-time finite-N approximation requires the trajectory to reach a saturation regime where parse drops are visible, which the present 1B$\times$3B Llama stack at $N{=}200$ has not done. Diagnostic on whether $N{>}300$ or larger Llama students expose the cliff is left to a future budget extension, but is not promoted to the paper's main claim. The Qwen3 $N{=}200$ budget extension result (App. E.2) is the only architecture for which the locked Thm. 4.2 prediction is both well-posed and validated.

Table 29: Llama-3.2-1B×3B at the pre-registered $N=200$ budget (App. F.7.1). Parse rates remain in $[0.19, 0.23]$ across the locked λ-grid. The cliff is not observable because the SFT trajectory has not reached parse saturation; reported as a precondition-failure scope boundary in the architecture-specific reachability mode.

λ	parse	USEFUL	reading
1.00	0.226	0.197	no cliff (parse $\ll 0.95$)
1.10	0.193	0.173	no cliff (within noise of 1.00)
1.20	0.217	0.194	no cliff (within noise of 1.00)

G ASPO head-to-head: same-mechanism fix preserves the cliff

Setup. ASPO [38] addresses the IS-asymmetry on positive-advantage tokens by replacing the rollout IS weight w_t with $\mathrm{clamp}(1/w_t, c)$ when the per-token advantage is positive. We add this as a single flag (`actor.policy_loss.aspo_asymmetric`) in a local verl training fork and run a Fashion 1.7B×4B sweep matching the production ListOPD recipe (3 epochs, $c=5$, base-relative reverse-KL). The main ASPO comparison uses four seeds at the two decision points $\lambda \in \{1.0, 1.5\}$ (seeds $42, 7, 13, 21$); the larger $\lambda \in \{2.0, 3.0\}$ points are single-seed exploratory checks.

Result. Tab. 30 reports parse / NDCG@1 / USEFUL for ASPO alongside the published ListOPD baselines. Three findings:

(i) ASPO at $\lambda=1.0$ improves over vanilla OPD at $\lambda=1.0$ by about $+4.4$pp on parse and $+4.4$pp on USEFUL (0.932 ± 0.008 vs. 0.887 parse; 0.863 ± 0.006 vs. 0.819 USEFUL), consistent with ASPO's published stability claim.

(ii) Apples-to-apples multi-seed comparison does not show categorical ListOPD dominance over ASPO: ASPO $\lambda=1.0$ reaches USEFUL$=0.863\pm0.006$, comparable to the 5-seed ListOPD $\lambda=1.15$ mean 0.857 ± 0.016. The ListOPD seed-42 headline is higher (0.882), but we do not use a single seed to claim ListOPD beats ASPO.

(iii) ASPO has its own cliff in the same λ window: parse drops from 0.932 ± 0.008 at $\lambda=1.0$ to 0.096 ± 0.020 at $\lambda=1.5$. The single-seed extension remains collapsed at $\lambda=2.0$ and $\lambda=3.0$. The cliff onset is one grid step left of vanilla OPD's (which collapses between $\lambda=1.25$ and $\lambda=1.40$), consistent with ASPO's more aggressive rare-token gradient reaching the saturation boundary faster. The qualitative transition survives the alternative published fix to the IS-asymmetry, ruling out a narrow GRPO-implementation reading.

Table 30: ASPO vs. vanilla ListOPD on Fashion PL-K8 (Qwen3-1.7B×4B, $c=5$, 3-epoch, strict ID-aware parser). ASPO at $\lambda=1.0$ improves over vanilla OPD at $\lambda=1.0$ and is comparable to the 5-seed ListOPD $\lambda=1.15$ operating point; at $\lambda=1.5$ the ASPO collapse is stable across seeds.

Method (λ)	parse	NDCG@1	USEFUL
ASPO (this work, App. G):			
$\lambda=1.0$ (4 seeds)	0.932 ± 0.008	0.926 ± 0.003	0.863 ± 0.006
$\lambda=1.5$ (4 seeds)	0.096 ± 0.020	0.920 ± 0.012	0.088 ± 0.019
$\lambda=2.0$ (seed 42)	0.019	0.995	0.019
$\lambda=3.0$ (seed 42)	0.028	0.919	0.026
Vanilla ListOPD (paper main):			
$\lambda=1.0$ (seed 42)	0.887	0.923	0.819
$\lambda=1.15$ (5 seeds)	0.921 ± 0.019	0.931	0.857 ± 0.016
$\lambda=1.15$ (seed 42)	0.948	0.930	0.882
$\lambda=1.25$	0.651	0.931	0.606
$\lambda=1.40$	0.160	0.949	0.152

H Predicate $\lambda^\star(p, c)$ scope tests

Tab. 31 aggregates the in-/out-of-scope tests of the closed-form predicate, summarised in Sec. 1 and discussed in App. A. The table separates calibrated regimes, public stress tests, abstentions where

preconditions fail, and failures of the finite-budget classifier. The c=1.5 row is the explicit failure: the fixed point lies beyond the clip-safe boundary, but the 42-step run remains parse-stable.

Table 31: Predicate $\lambda^\star(p, c)$ tested across regimes. The table distinguishes the algebraic clip-safe crossing from finite-budget collapse: c=1.5 crosses the boundary but does not collapse in 42 steps. The *S2b no-base* row tests the implementation-axis scope: the closed form locates an asymptotic fixed point but does not bound finite-N first-passage time under a changed clipped estimator.

Regime	Precondition	Predicted	Observed
Fashion Qwen3 (1.7B×4B)	p_{eff}=0.9993, parse headroom	cliff $\lambda^\star$=1.22	onset 1.15, collapse 1.25
MBPP code (Qwen3 1.7B×4B)	strict AST scaffold; code-specific p_{eff} not measured	Fashion-marker scale check $[1.15, 1.25]$	parse peak $[1.15, 1.25]$
MS MARCO/TREC-DL reranking	public human qrels; strict JSON scaffold; measured p_{eff}=0.99941 on the 54-prompt eval (Tab. 18)	predicted $\lambda^\star$=1.27 at b=0.81, indistinguishable from Fashion within one grid step	OPD improves over SFT; λ=1.25 vs. 1.5 not separated at 4-seed budget; consistent with no-shift but underpowered to detect midpoint shifts below the seed-variance floor
Fashion Llama-3.2 (1B×3B, tested budget)	cross-architecture stress test	boundary may shift	monotone through λ=1.4; no cliff in tested grid
GSM8K math	p_{eff} diffuse, N/A	no cliff	no cliff ($\sigma_\lambda \approx 0.006$)
BFCL function	SFT-4B parse 0.942, no headroom	no cliff	no cliff at 3- or 7-ep
S2b no-base impl. (Fashion 1.7B×4B, 42-step)	base term removed; finite-budget reachability changes under the actual clipped estimator	no finite-N collapse prediction from the fixed point alone	no cliff: parse $\in [0.929, 0.939]$ across all six λ
c-sweep ($c \in \{2, 5, \infty\}$)	on-anchor scaling	cliff at c=2, none at $c \geq 5$ at λ=1.15	matches (parse 0.56, 0.86, 0.85)
c-sweep (c=1.5)	off-anchor; $\lambda^\star$=1.06	fixed point outside clip-safe region; finite-budget classifier predicts collapse	**parse-stable at 42 steps (0.95): classifier inverted**

I Compute constraints and deferred ablations

Two pre-registered ablations are reported here as predictions only and deferred to a future revision: the extended regularizer sweep beyond the small-β, γ pilots in Tab. 32, and a matched cross-architecture extension at the same scale. Both runs are blocked by a shared-cluster multi-tenancy issue that prevents stable concurrent verl+Ray launches at the configuration the rest of the paper uses; the issue is operational, not methodological. The quantitative shifts these ablations test are pre-registered in Tab. 32 and Sec. 5.3, and nothing about them depends on which platform executes the run.

I.1 Regularizer protocol, predicted shifts, and pilot results

Closed form for the entropy-bonus shift. Adding $\gamma H(\pi_S) = -\gamma[q \log q + (1-q)\log(1-q)]$ to the per-token OPD objective modifies the expected θ-flow of App. C.1 (proof of Thm. 4.1 part 1) by the term $\gamma\, \partial_\theta H = -\gamma \operatorname{logit}(q)\, q(1-q)$ (verl uses standard SGD on the parametric loss, so the chain rule contributes the $q(1-q)$ Jacobian; this is the load-bearing factor that earlier revisions of this table dropped, equivalent to assuming a Fisher-preconditioned natural-gradient update which verl does not perform). Imposing the cliff condition $q^\star_\gamma = q_c$ collapses the implicit fixed-point equation, giving the exact (not first-order) closed form

$$\lambda^\star_\gamma(p, b, c, \gamma) \;=\; \lambda^\star_0(p, b, c) \;+\; \gamma\, \frac{q_c(1-q_c)\operatorname{logit}(q_c)}{\operatorname{logit}(p) - \operatorname{logit}(b)}, \tag{19}$$

linear in γ. At Fashion (p_{typ}=0.9993, c=5), $q_c(1-q_c) \approx 1.4\times10^{-4}$ suppresses the slope: $\delta\lambda \approx 2.1\times10^{-4}\,\gamma$ (base-relative) or $\approx 1.7\times10^{-4}\,\gamma$ (base-neutral b=1/2). All three rows above

Table 32: Pre-registered regularizer predictions and pilot observations at λ=1.15 (fixed operating point, 3-epoch, 1.7B×4B). Closed-form $\lambda^\star_{\text{reg}}$ from Thm. 4.3 at p_{eff}=0.9993, reference-policy modal confidence q_{B}=0.93, c=5. Pilot columns (β=0.01, γ=0.001) probe small-β behavior; remaining rows are deferred to a future revision per App. I.

regularizer	setting	predicted $\lambda^\star$	observed parse	observed NDCG@1
none (baseline, 5-seed)	—	1.22	0.921 ± 0.019	0.930
KL-to-base (pilot)	β=0.01	$\approx$ 1.22	0.887	0.929
KL-to-base	β=0.05	$\approx$ 1.24	deferred	—
KL-to-base	β=0.20	$\approx$ 1.28	deferred	—
entropy bonus (pilot)	γ=0.001	$\approx$ 1.22	0.731	0.926
entropy bonus	γ=0.01	$\approx$ 1.22	deferred	—
entropy bonus	γ=0.05	$\approx$ 1.22	deferred	—
λ warmup	T_{w}=10/N=42	$\approx$ 1.60	deferred	—
λ warmup (pilot)	T_{w}=20/N=42	$\approx$ 2.33	0.929	0.935

($\gamma \in \{0.001, 0.01, 0.05\}$) therefore predict $\lambda^\star_\gamma \approx \lambda^\star_0$ to four decimals: the entropy bonus does not perceptibly shift the cliff at the Fashion regime, regardless of γ within the deployable range.

Pilot observations. Two small-β regularizer runs completed: KL-to-base at β=0.01 and entropy bonus at γ=0.001, both at fixed λ=1.15, 3 epochs, 1.7B×4B Fashion (single seed, 42 optimizer steps). The theorem predicts tiny $\lambda^\star$ shifts at these regularizer strengths (≈ 0.003 and ≈ 0.01 respectively), well below our λ-grid resolution, so λ=1.15 should remain sub-critical under both. Observed parse rate drops from the strict 5-seed baseline of 0.921 ± 0.019 to 0.887 (KL) and 0.731 (entropy), outside the baseline CI in both cases. NDCG@1 on the parseable subset is unchanged ($0.929, 0.926$ vs. baseline 0.930), so the effect is concentrated in parse rate, not rank quality. A separate 20-step λ-warmup pilot stays inside the baseline seed band (parse 0.929, NDCG@1 0.935, USEFUL=0.869), consistent with its predicted right-shift rather than a new positive lift. The theorem predicts the cliff location, not regularizer-induced parse-rate degradation within the sub-critical regime; with the entropy bonus, this is now quantitative: Eq. (19) gives $\delta\lambda \approx 2\times10^{-7}$ at γ=0.001, eight orders of magnitude below the λ-grid resolution, so λ=1.15 remains sub-critical and the observed parse collapse is orthogonal to the $\lambda^\star$-shift prediction. The drop instead surfaces a third scope boundary of the 2-token reduction: small regularizer strengths interact non-trivially with sequence-level dynamics in a way the equilibrium analysis does not capture. We list this as a rebuttal-tier follow-up alongside the c-axis sweep (Sec. 5.3) and the GSM8K cross-task scope boundary (App. F.6).

The pre-registered protocol: (i) Qwen3-1.7B student × Qwen3-4B teacher (same as main body); 3 epochs (N=42); Fashion PL-K8 train set; c=5; same vLLM rollout, AdamW, learning rate, and batch-size configuration as the baseline ListOPD runs. (ii) KL-to-base penalty $\beta \in \{0.05, 0.20\}$ at $\lambda \in \{1.20, 1.30, 1.40\}$. (iii) Entropy bonus $\gamma \in \{0.01, 0.05\}$ at the same three λ values. (iv) λ warmup at fixed λ=1.50 with $T_{\text{w}} \in \{10, 20\}$. (v) Per-configuration observation: end-of-training parse rate on the 212-prompt Fashion val set; aggregate the 3-point λ grid into an observed onset and compare to the predicted $\lambda^\star_{\text{reg}}$ of Tab. 32. Total budget: 18 runs × ∼10 min each on 8×B200 with vLLM TP=4 rollout = ∼3 GPU-hours, well within a single-node overnight window in any environment without the cgroup multi-tenancy issue.