PREFERENCE OPTIMIZATION IS NOT ENOUGH: HARNESSING CATALOG SUPPORT FOR GENERATIVE RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) and its multi-negative extension, Softmax-DPO, align generative recommenders through relative item comparisons. However, unconstrained generation operates over the full autoregressive sequence space, where better preference margins do not guarantee that probability mass remains on valid catalog items. We term this failure mode *catalog-support drift*: a model can become better at preference discrimination while generating item strings that cannot be resolved against the catalog, undermining the practical usefulness of its recommendations. Addressing this mismatch requires learning not only which items users prefer, but also how to preserve support for valid outputs. To this end, we introduce DEHALL-SDPO, a dehallucination-oriented extension of Softmax-DPO that integrates sequence- and prefix-level support control with preference learning. At the sequence level, a full-slate ranking objective improves within-slate ordering, while a primal–dual constraint regulates the slate's total probability mass against an SFT-calibrated budget. At the prefix level, an on-policy trie loss penalizes probability assigned to catalog-invalid continuations along the evolving policy's generation paths. Together, these objectives shape the policy itself without imposing a catalog mask at inference. Under full-policy ancestral sampling, our analysis relates slate probability mass to exact outside-slate risk and bounds finite-horizon catalog-exit probability by the expected prefix-support loss. Experiments across five real-world recommendation datasets reveal the practical severity of catalog-support drift. Extensive results show that DeHall achieves superior performance on both catalog validity and policy-ranking accuracy, reducing nucleus-sampling catalog-string violation rates. Component analyses further support the complementary roles of prefix- and sequence-level support control. Code is available at .

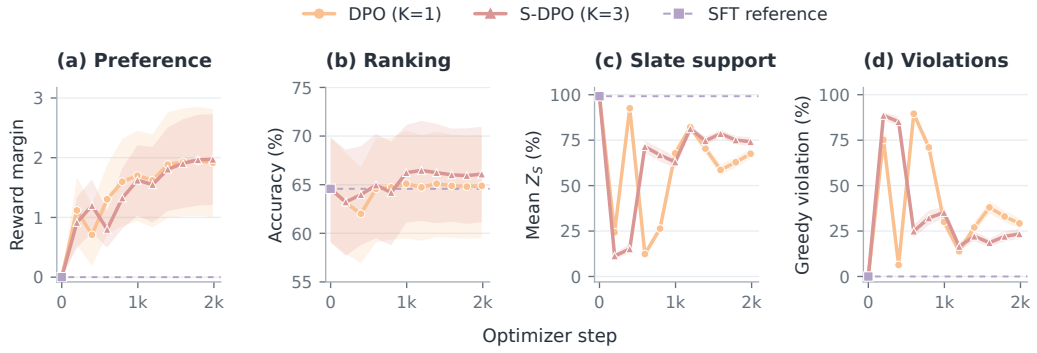

Figure 1: **Catalog-support drift: preference gains do not guarantee valid recommendations.** DPO and S-DPO improve SFT-relative margins and retain ranking accuracy while losing slate mass and generating catalog-invalid strings on MovieLens. Bands: pointwise 95% user-cluster intervals; slate support and observed-union catalog validity are distinct.

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

1 INTRODUCTION

Generative recommendation recasts item prediction as language generation, allowing user histories and item semantics to be modeled within a shared textual framework (Geng et al., 2022; Li et al., 2024). This formulation makes language models a natural interface between behavioral evidence and item suggestions, extending their role beyond understanding recommendation requests to producing recommendations directly. Realizing this potential, however, requires outputs that are both relevant to the user and resolvable to items in the catalog: a plausible description alone is not actionable.

Preference optimization offers a direct way to address the relevance requirement as recommendation is inherently comparative—the goal is to favor items that better match a user's interests over competing alternatives. Direct Preference Optimization (DPO) learns such preferences from chosen–rejected pairs (Rafailov et al., 2023), while Softmax-DPO (S-DPO) extends this formulation to multiple rejected items (Chen et al., 2024b). By converting these comparisons into policy updates, both methods provide a principled mechanism for adapting generative models to item preferences.

Yet learning relative preferences and generating valid recommendations are not the same objective. DPO and S-DPO optimize reference-relative scores within finite comparison sets, whereas unconstrained decoding operates over the model's full autoregressive sequence space, including strings that identify no catalog item. This objective–generation mismatch raises central question: *does learning which item is preferable also preserve probability of generating a valid recommendation?*

Relative preference objectives alone cannot guarantee this preservation. At the distribution level, uniformly reducing the probabilities of catalog strings leaves chosen–rejected log-ratio margins unchanged if the released mass is reassigned to noncatalog sequences. We term this failure mode *catalog-support drift*: relative preference can improve without preserving valid-item probability, connecting likelihood displacement in preference alignment to a concrete recommendation failure—outputs that cannot be resolved against the catalog (Razin et al., 2025; Chen et al., 2024a).[1]

Our checkpoint study reveals this disconnect in practice (Figure 1). On MovieLens (Harper & Konstan, 2015), DPO and S-DPO improve reward margins relative to supervised fine-tuning (SFT) and retain policy-ranking accuracy close to SFT, yet lose substantial probability mass on the displayed slate and generate many catalog-invalid strings under unconstrained decoding. These synchronized, prompt-matched measurements distinguish support for displayed candidates from validity against the broader catalog, while their non-monotonic trajectories show why neither preference progress nor a final ranking score adequately captures generation reliability throughout training.

These observations motivate three complementary requirements for preference learning. The ordering objective should compare all displayed candidates using the raw policy scores employed in recommendation ranking, rather than relying on reference-relative rewards over sampled negatives. Because accurate ordering remains compatible with shrinking candidate probabilities, this motivates explicit control of complete-sequence support; because aggregate mass does not reveal where invalid continuations arise, support control should operate along the evolving policy's generation paths.

The matched component study in Figure 2 provides empirical motivation for combining these sequence- and prefix-level signals. On MovieLens, prefix regularization reduces nucleus-sampling catalog violations from 46.40% to 1.90%, and adding sequence-mass control further reduces them to 0.67%, with policy accuracy remaining near 65%. These results support complementary mechanisms for restoring valid generation without modifying the inference-time decoder, while full-slate ordering addresses the distinct question of which valid candidate should receive preference.

Building on this reasoning, we introduce DEHALL-SDPO, a dehallucination-oriented extension of Softmax-DPO that coordinates preference ordering, sequence mass, and prefix legality within one policy (Figure 3).[2] At the sequence level, full-slate listwise policy ranking (LPR) supervises the chosen item's ordering against every displayed candidate, while a primal–dual constraint regulates their total probability mass against an SFT-calibrated budget. At the prefix level, an on-policy trie loss penalizes probability assigned to catalog-invalid continuations at valid prefixes visited by the current policy, using catalog structure beyond the displayed slate to regularize free-generation behavior.

[1] We use catalog violations, or catalog hallucinations, to denote catalog-string nonmembership under the specified matching protocol, rather than broader factual hallucination.

[2] DeHall denotes dehallucination; SDPO denotes Softmax Direct Preference Optimization.

Crucially, these objectives shape the policy itself without imposing a catalog mask at inference. Under full-policy ancestral sampling, our analysis identifies the complement of EOS-inclusive slate mass as exact outside-slate probability and bounds finite-horizon catalog-exit risk by the expected prefix-support loss. These sequence- and path-level interpretations connect the proposed training signals to generation reliability, while evaluation under unconstrained decoding tests whether the learned support translates into catalog-grounded recommendations aligned with user preferences.

Our contributions are threefold:

- **Characterizing catalog-support drift.** We identify the gap between relative preference optimization and catalog-grounded generation in practical recommendation settings. Synchronized checkpoint measurements of preference margins, policy ranking, slate mass, and catalog validity expose a systematic failure that ranking metrics alone can conceal.
- **Joint preference and support optimization.** We introduce DEHALL-SDPO, integrating full-slate ranking, SFT-calibrated primal–dual mass control, and on-policy prefix regularization within a unified framework. Our theoretical analysis connects sequence mass and prefix-support loss to distinct generation risks under full-policy ancestral sampling.
- **Recommendation gains through learned support.** Experiments across five real-world recommendation datasets demonstrate that DEHALL-SDPO improves both policy-ranking accuracy and catalog validity, translating learned preferences into more reliable recommendations without inference-time catalog masking. Component studies further support the complementary roles of sequence- and prefix-level control in preserving valid generation.

2 LEARNING SETUP AND PRELIMINARIES

We study preference optimization for LLM-based generative recommendation (Deldjoo et al., 2024; Qu et al., 2025), where an autoregressive policy generates an item string conditioned on a user's interaction history and a displayed candidate slate. Let $\mathcal{C}$ denote the item catalog and $\mathcal{S}(x) \subseteq \mathcal{C}$ the candidate slate included with the history in the input prompt x. The observed target $y^+ \in \mathcal{S}(x)$ provides the positive item for preference learning (Ji et al., 2024; Ye et al., 2026).

DPO contrasts the target with one negative, whereas Softmax-DPO (Chen et al., 2024b) jointly compares it against K negatives selected from a frozen pool. These preference comparisons differ from our full-slate objectives, which use every distinct candidate in $\mathcal{S}(x)$ to supervise policy ranking and regulate sequence probability mass. Accordingly, K specifies the number of preference negatives: our default $K = 3$ retains the full 19–20-item deduplicated slate for ranking and support control.

EOS-inclusive policy scores. Following standard autoregressive sequence modeling (Sutskever et al., 2014), we represent each item by its canonical tokenization $\bar{y} = (y_1, \ldots, y_L, \mathrm{EOS})$, where EOS distinguishes a completed output from an unfinished prefix. To retain the probability interpretation required by mass control, we use the total conditional log probability without length normalization. Candidate ranking uses the same policy score rather than a reference-relative reward:

$$s_\theta(x, y) = \log \pi_\theta(\bar{y} \mid x) = \sum_{t=1}^{L+1} \log \pi_\theta(\bar{y}_t \mid x, \bar{y}_{<t}). \tag{1}$$

Relative preference optimization. Let π_{ref} be the frozen SFT policy (Ouyang et al., 2022) and $r_\theta(x, y) = s_\theta(x, y) - s_{\mathrm{ref}}(x, y)$ the log-ratio score; βr_θ is the implicit reward. S-DPO raises the target's reward relative to its K negatives within each finite comparison set. This score measures change from SFT rather than the raw probability used for policy ranking, giving the objective

$$\mathcal{L}_{\mathrm{SDPO},K} = \log\left[1 + \sum_{k=1}^{K} \exp\big(\beta[r_\theta(x, y_k^-) - r_\theta(x, y^+)]\big)\right]. \tag{2}$$

At $K = 1$, Equation 2 reduces to binary DPO; DPO and S-DPO baselines use exactly one and three negatives. Neither explicitly controls the probability of completing a valid catalog item. Reward margins, raw policy ranking, and generation support therefore measure distinct aspects of learning.

Absolute support. Distinct canonical EOS-terminated item sequences define mutually exclusive outcomes, so their probabilities can be summed over either the displayed slate or the catalog (Loula

et al., 2025). The slate mass is directly available from the same scores used for ranking, whereas full-catalog mass may be expensive to enumerate. Since $\mathcal{S}(x) \subseteq \mathcal{C}$, slate mass provides a tractable lower bound on catalog mass, connecting candidate scoring to a directly measurable support quantity:

$$Z_{\mathcal{S},\theta}(x) = \sum_{y \in \mathcal{S}(x)} e^{s_\theta(x,y)}, \qquad Z_{\mathcal{C},\theta}(x) = \sum_{y \in \mathcal{C}} e^{s_\theta(x,y)}, \qquad 0 \le Z_{\mathcal{S},\theta} \le Z_{\mathcal{C},\theta} \le 1. \tag{3}$$

Under full-policy ancestral sampling, the complement of each mass is the probability of leaving its corresponding set of completed sequences (Park et al., 2024; Loula et al., 2025). An output outside the slate can still name a valid catalog item, so outside-slate probability is not the same as catalog violation. Their decomposition makes explicit which quantity the sequence-level constraint controls and why outside-slate probability upper-bounds outside-catalog probability:

$$1 - Z_{\mathcal{S},\theta} = \underbrace{Z_{\mathcal{C},\theta} - Z_{\mathcal{S},\theta}}_{\text{valid, outside slate}} + \underbrace{1 - Z_{\mathcal{C},\theta}}_{\text{outside catalog}}. \tag{4}$$

These identities concern canonical token events under the original policy. Empirical CV instead checks whitespace-stripped strings under each decoder; it is reported separately, not equated with mass. Appendix B defines these events and the ancestral consistency audit.

3 Analysis of Catalog-Support Drift

3.1 Relative preference leaves absolute support unidentified

Relative objectives identify score differences but not a common shift of compared log probabilities. The construction below makes this known ambiguity explicit for a catalog.

Proposition 1 (Rank-equivalent support scaling)**.** *Fix x, the reference policy, and a positive distribution a over $\mathcal{S}(x)$. Let b be supported outside $\mathcal{C}$, and define $\pi_q = qa$ on the slate and $\pi_q = (1-q)b$ elsewhere for $q \in (0,1)$. All within-slate rankings and DPO/S-DPO log-ratio differences are independent of q, but $Z_{\mathcal{S},q} = Z_{\mathcal{C},q} = q$.*

A common $\log q$ cancels from every comparison (Appendix A). The corresponding decomposition separates absolute support from conditional ordering:

$$s_\theta(x,y) = \log Z_{\mathcal{S},\theta}(x) + \log q_\theta(y \mid x, \mathcal{S}), \qquad q_\theta(y \mid x, \mathcal{S}) = \frac{\pi_\theta(\bar{y} \mid x)}{Z_{\mathcal{S},\theta}(x)}. \tag{5}$$

3.2 Checkpoint evidence: ranking conceals changes in validity

Figure 1 tests whether preference progress and free-generation validity evolve together. At the final DPO and S-DPO checkpoints, full-slate reward margins reach 1.92 and 1.97, while policy accuracy is 64.88% and 66.09%, compared with 64.58% at SFT. These preference and ranking signals alone suggest little reason for concern, making them a useful test of whether apparently successful preference learning can conceal changes in generation reliability.

The support measurements reveal a different picture on the same test prompts. Mean slate mass falls from 99.17% at SFT to 67.52% for DPO and 74.14% for S-DPO, while greedy catalog violation rises from zero to 29.23% and 23.41%. The paired measurements expose the failure hidden by relative margins: apparently successful preference learning can lose sequence support and produce invalid recommendations.

Intermediate checkpoints reveal why endpoint-only evaluation is insufficient. At step 597, DPO matches SFT's 64.58% ranking accuracy but retains only 12.37% slate mass and produces 89.50% catalog-invalid greedy outputs. Its subsequent recovery motivates checkpoint-level monitoring rather than inferring support from ranking or loss; these terminal checkpoints are distinct from the selected legacy adapters in the main result table (Appendix F.8).

3.3 Motivation: control ordering, mass, and continuation legality

Equation 5 separates ordering from absolute support. To align learning with policy ranking, the ordering signal should compare all candidates using raw policy scores rather than only reference-

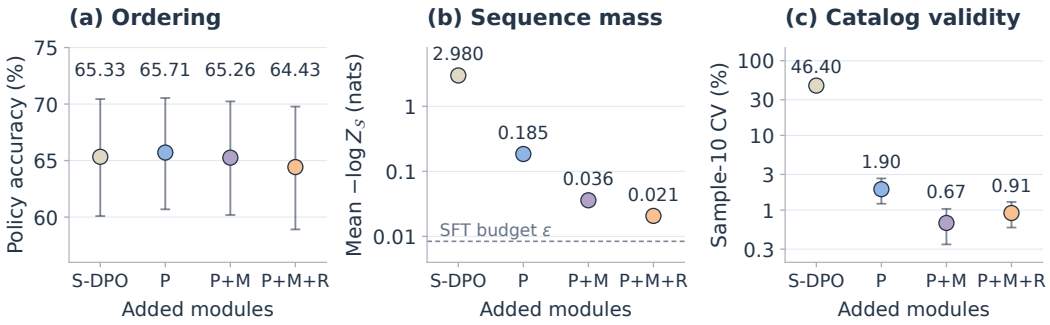

Figure 2: **Three signals motivate coordinated support control.** Matched MovieLens models: P, M, R add Prefix, Dual Mass, and LPR ($\lambda_r = 0.1$) to S-DPO; points are separate runs. Panels (b,c) use log scales; (a,c) show 95% user-cluster intervals and (b) point estimates. The dashed SFT budget is a target, not achieved feasibility; CV uses the observed-union diagnostic catalog.

adjusted rewards; a separate term must regulate their total probability. Prefix-level supervision complements this sequence signal by locating invalid transitions along current-policy paths and using the catalog beyond the displayed slate.

An EOS-aware trie identifies legal next tokens $A_{\mathcal{C}}(p)$ after prefix p. Their aggregate probability measures local support; its complement is full-policy next-transition risk. Evaluating this signal at current-policy prefixes targets visited generation paths:

$$m_\theta(p) = \sum_{v \in A_{\mathcal{C}}(p)} \pi_\theta(v \mid x, p), \qquad h_\theta(p) = 1 - m_\theta(p). \tag{6}$$

Figure 2 tests the support side of this separation in a matched component study. Prefix reduces sample CV from 46.40% to 1.90%, and adding Dual Mass lowers it to 0.67% while accuracy stays near 65%. This complementary repair motivates coupling local continuation legality with complete-sequence support, leaving full-slate ordering to guide which valid item receives that support.

The resulting design supervises item order, regularizes complete-candidate mass, and penalizes invalid continuations jointly. Section 5 implements these controls without a catalog mask; Section 6 tests both support repair and recommendation utility, including the effect of the ranking weight.

4 Background and Related Work

Generative recommendation and preference learning. Text-to-text recommendation and generative retrieval formulate item prediction as sequence generation (Geng et al., 2022; Rajput et al., 2023; Li et al., 2024), with DPO and S-DPO supplying binary and multi-negative preference objectives (Rafailov et al., 2023; Chen et al., 2024b). We use a one-positive, top-one ListNet-style objective for full-slate LPR (Cao et al., 2007) and integrate this ordering signal with explicit sequence- and prefix-level support control.

Likelihood displacement and absolute-likelihood control. Razin et al. (2025) analyze likelihood displacement in DPO; NCA explicitly distinguishes relative and absolute likelihood, and BDPO bounds rejected-response influence (Chen et al., 2024a; Cho et al., 2025). We extend this perspective to catalog-grounded recommendation through separate slate-mass, prefix-exit, and decoded-string measurements, linking the alignment objective to recommendation validity. Chosen-NLL and absolute-likelihood alignment provide relevant comparisons for isolating these mechanisms.

Catalog grounding and constrained generation. Unlikelihood training penalizes unwanted continuations (Welleck et al., 2019); LCFT and RecLM address hallucination and out-of-domain recommendation (Deng et al., 2025; Liao et al., 2026). Our prefix term uses the trie-defined invalid-token set at visited prefixes, combined with calibrated primal–dual slate control (Nandwani et al., 2019).

Unlike trie-constrained retrieval (De Cao et al., 2020), this combination learns catalog support in the policy while retaining unmasked inference; Appendix C.5 details deployment comparisons.

5 DeHall-SDPO: Harnessing Sequence and Prefix Support

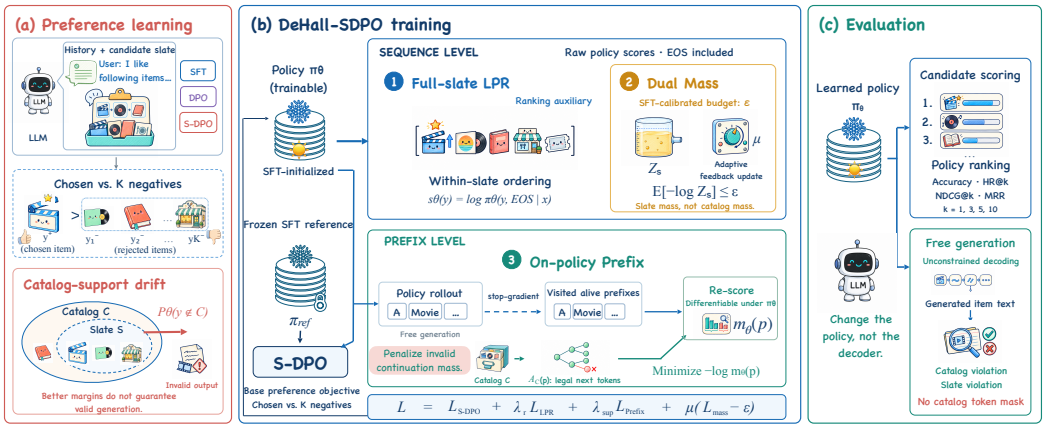

Figure 3: **DeHall-SDPO learns preference ordering and generation support.** (a) Relative comparison leaves catalog support unspecified; (b) the policy combines S-DPO with LPR, Dual Mass, and on-policy Prefix while the SFT reference stays frozen. (c) Ranking and free-generation validity are evaluated separately, without a catalog token mask; Appendix H provides details.

Three signals, one policy. Figure 3 combines S-DPO against a frozen SFT reference with LPR ordering, complete-slate mass control, and on-policy prefix regularization. Algorithm 1 in Appendix C.4 gives the complete calibration, training, and inference procedure. Let $\rho_{\bar{\theta}}$ denote the rollout decoder of the current policy, held fixed during each update, and define $\overline{\mathcal{L}}_{\text{sup}}(x;\theta,\bar{\theta}) = \mathbb{E}_{\widetilde{Y}\sim\rho_{\bar{\theta}}(\cdot|x)}\mathcal{L}_{\text{sup}}(x,\widetilde{Y};\theta)$. The training surrogate, with $\bar{\theta}$ refreshed from θ, is

$$\min_{\theta} \quad \mathbb{E}_{x\sim\mathcal{D}_{\text{tr}}}[\mathcal{L}_{\text{SDPO},K} + \lambda_r\mathcal{L}_{\text{LPR}} + \lambda_{\text{sup}}\overline{\mathcal{L}}_{\text{sup}}], \tag{7}$$

$$\text{subject to} \quad \mathbb{E}_{x\sim\mathcal{D}_{\text{tr}}}[\mathcal{L}_{\text{mass}}] \le \epsilon. \tag{8}$$

5.1 Full-slate listwise policy ranking

To supervise the ordering used at evaluation, LPR compares the chosen item with all real slate items:

$$\mathcal{L}_{\text{LPR}}(x) = -\log \frac{\exp(s_\theta(x,y^+)/\tau)}{\sum_{y\in\mathcal{S}(x)}\exp(s_\theta(x,y)/\tau)}. \tag{9}$$

Unlike S-DPO, it uses raw policy scores instead of reference-adjusted rewards and the full slate instead of K frozen preference negatives. LPR supervises the conditional ordering used by policy ranking, while remaining invariant to a common score shift. This motivates the separate mass term: preference identifies where support should go, and mass control regulates how much the slate retains.

5.2 Validation-calibrated dual mass control

The mass objective supplies direct supervision for the absolute slate coordinate left unidentified by both relative losses, reusing the full-slate policy scores:

$$\mathcal{L}_{\text{mass}}(x) = -\log Z_{\mathcal{S},\theta}(x) = -\log \sum_{y\in\mathcal{S}(x)} \exp s_\theta(x,y). \tag{10}$$

We calibrate ϵ once as the mean validation mass loss of the dataset- and seed-matched Mistral SFT policy, using identical scoring precision and full-slate batching. Training minimizes the Lagrangian

and updates the multiplier after each successful optimizer step:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SDPO},K} + \lambda_r \mathcal{L}_{\mathrm{LPR}} + \lambda_{\mathrm{sup}} \mathcal{L}_{\mathrm{sup}} + \mu(\mathcal{L}_{\mathrm{mass}} - \epsilon), \tag{11}$$

$$\mu \leftarrow \Pi_{[0,\mu_{\max}]} \left[\mu + \eta_\mu (\overline{\mathcal{L}}_{\mathrm{mass}} - \epsilon) \right]. \tag{12}$$

The multiplier adapts support pressure to the globally averaged training-batch budget residual. At $\tau = 1$, equally weighted fixed versions of both sequence objectives reduce to chosen-sequence NLL:

$$\mathcal{L}_{\mathrm{LPR}} + \mathcal{L}_{\mathrm{mass}} = -s_\theta(x, y^+). \tag{13}$$

At temperature $\tau = 1$, the sequence penalty in ordering and mass coordinates is $-\lambda_r \log q_\theta(y^+ \mid x, \mathcal{S}) - \mu \log Z_{\mathcal{S},\theta} - \mu\epsilon$. The dual update therefore adapts support pressure without tying it to the ranking weight; it is not a certificate of held-out feasibility. Appendix C derives the weighted identity and specifies comparisons that isolate adaptive support control from chosen-NLL regularization.

5.3 On-policy catalog-prefix support

The policy generates a trajectory without gradient, and the trie identifies alive prefixes before termination or the first invalid transition, including the prefix immediately preceding an exit. A differentiable re-scoring pass minimizes

$$\mathcal{L}_{\mathrm{sup}}(x, \widetilde{Y}) = - \sum_{p \in \mathcal{P}_{\mathrm{alive}}(\widetilde{Y})} \log m_\theta(p). \tag{14}$$

This targets catalog-invalid continuation probability at visited prefixes without prescribing which valid title should be generated. Resampling directs supervision toward prefixes visited by the evolving policy, and differentiable re-scoring implements an on-policy *semi-gradient* with detached trajectory selection. The recorded experiments use nucleus rollouts ($p = 0.9$, temperature one, horizon 64); $\rho_{\bar{\theta}}$ is therefore not the full-policy sampling law assumed below. The local logit gradient increases support for legal continuations (refert to Appendix A).

5.4 Risk interpretation and inference

Under full-policy ancestral sampling, the canonical EOS events give

$$R_{\mathcal{S}}(x) = 1 - Z_{\mathcal{S},\theta}(x) = 1 - e^{-\mathcal{L}_{\mathrm{mass}}(x)}, \qquad \mathbb{E}R_{\mathcal{S}} \le 1 - e^{-\mathbb{E}\mathcal{L}_{\mathrm{mass}}}. \tag{15}$$

The second relation follows from Jensen's inequality, and outside-slate probability upper-bounds outside-catalog probability because $\mathcal{S} \subseteq \mathcal{C}$. This gives sequence-mass control a direct risk interpretation under the stated sampling law, distinct from empirical frequencies under other decoders.

Proposition 2 (Finite-horizon catalog exit)**.** *Let $P_t = Y_{<t}$ under full-policy ancestral sampling and I_t indicate that generation has neither terminated nor exited the catalog trie before t. Let $\tau_{\mathcal{C}}$ be the first invalid transition, or infinity after valid termination; summands vanish off alive prefixes. Then*

$$\Pr(\tau_{\mathcal{C}} \le T) = \mathbb{E} \sum_{t=1}^{T} I_t[1 - m_\theta(P_t)] \le \mathbb{E} \sum_{t=1}^{T} I_t[-\log m_\theta(P_t)]. \tag{16}$$

The proof sums disjoint first-exit events and applies $1 - u \le -\log u$ (Appendix A). An unfinished but still-valid prefix at the horizon is not a first exit. The bound interprets the loss statistic under ancestral sampling, not the optimization trajectory of the detached-rollout update.

Learned support versus constrained decoding. A constrained decoder replaces $\pi_\theta(v \mid x, p)$ by $\pi_\theta(v \mid x, p)\mathbf{1}\{v \in A_{\mathcal{C}}(p)\}/m_\theta(p)$, masking invalid tokens at inference. Our prefix loss instead penalizes the original policy's invalid-token mass during training, leaving its inference distribution unmasked. We therefore evaluate learned grounding and recommendation utility jointly; closed-set selection and constrained decoding remain complementary deployment options (Appendix C.5).

6 Experiments

We test support drift, its repair, and recommendation utility on MovieLens (Harper & Konstan, 2015), Yelp (Zhang et al., 2015), Amazon Book, Amazon Movie, and Amazon CD (Ni et al., 2019). Table 1 reports seven core metrics, with full-method outcomes available on these five datasets. Table 2 reports ranking metrics; Appendix E gives two detailed tables per dataset.

Table 1: **Recommendation utility and catalog validity across five datasets.** CV denotes catalog-string violation; Sample uses ten nucleus outputs.

		Policy ranking ↑			Greedy		Sample-10	
Dataset	Method	Acc.	N@10	MRR	Hit ↑	CV ↓	Hit ↑	CV ↓
MovieLens	SFT	63.97	81.00	75.80	64.35	0.00	57.47	0.05
	DPO	62.61	80.71	75.35	22.05	75.15	21.69	75.45
	Softmax-DPO	64.73	81.78	76.62	20.85	76.36	20.42	76.41
	DEHALL-SDPO	65.56	82.25	77.22	65.56	0.83	64.16	1.07
Yelp	SFT	38.52	58.85	52.19	38.42	0.10	29.90	0.21
	DPO	38.71	59.64	52.82	34.88	20.75	32.38	31.08
	Softmax-DPO	38.58	59.88	52.87	21.33	66.69	20.54	67.84
	DEHALL-SDPO	39.02	60.19	53.14	38.91	0.27	37.94	0.36
A. Book	SFT	46.96	66.99	60.76	46.00	1.17	37.86	1.62
	DPO	49.50	69.83	63.28	22.29	70.54	21.58	70.62
	Softmax-DPO	51.13	70.93	64.45	34.29	48.17	33.75	48.52
	DEHALL-SDPO	51.96	71.12	64.74	51.37	1.21	51.09	1.53
A. Movie	SFT	46.10	66.65	60.32	45.70	1.05	36.40	1.44
	DPO	49.80	71.03	64.32	41.45	25.35	40.25	28.73
	Softmax-DPO	50.15	71.52	64.82	28.65	61.10	27.56	61.69
	DEHALL-SDPO	50.65	71.68	65.14	50.20	1.45	49.60	1.72
A. CD	SFT	31.63	52.90	46.18	31.38	1.44	22.66	1.93
	DPO	36.81	59.44	51.83	10.38	85.00	10.50	83.81
	Softmax-DPO	38.00	60.30	52.98	22.19	59.19	21.63	59.56
	DEHALL-SDPO	39.44	60.71	53.67	38.69	1.56	37.52	1.68

6.1 EVALUATION DESIGN

Data, training, and comparison groups. The generative methods use Mistral-7B-Instruct-v0.3, seed 1958, and three preference epochs; DPO/S-DPO use $K = 1/3$. Recorded SFT exposure is three epochs on MovieLens/Yelp and one on the Amazon datasets; full-method weights are $\lambda_r = 0.03$ on MovieLens and 0.01 on Movie/CD. Appendix E.1 records split details and optimizer settings.

Baselines. We compare DEHALL-SDPO with conventional recommendation models, content- and language-based recommenders, and preference-optimization methods; baseline descriptions and implementation details are provided in Appendix E.1.

Ranking and generation validity. Full-slate ranking uses total EOS-inclusive log probability; Accuracy=HR@1=NDCG@1. Generation Hit and catalog violation (CV) count correct and catalog-invalid outputs, respectively, under greedy and nucleus sample-10 ($p = 0.9$, temperature one). Appendix B defines all cutoffs, strict string matching, and per-output denominators.

Paired evaluation and catalog scope. Generative comparisons share prompts, catalogs, and SFT initialization; legacy MovieLens batching is marked separately. Appendix E.1 documents evaluation sources and settings.

6.2 CATALOG-SUPPORT DRIFT AND USEFUL FREE RECOMMENDATIONS

Preference gains do not ensure valid outputs. On Amazon Movie, S-DPO raises Accuracy from SFT's 46.10% to 50.15%, yet greedy CV rises from 1.05% to 61.10%. CD shows the same gap: Accuracy improves from 31.63% to 38.00% while CV increases from 1.44% to 59.19%. Yelp/Book also exhibit high S-DPO greedy CV despite competitive ranking, Figure 1 beyond MovieLens.

Support control improves delivered recommendations. On Amazon Movie, DEHALL-SDPO reaches 50.65% Accuracy and 50.20/49.60% greedy/sample Hit, versus S-DPO's 50.15% and 28.65/27.56%. CV falls to 1.45/1.72%, versus 61.10/61.69% for S-DPO and 25.35/28.73% for DPO. Better ranking thus accompanies a much larger recovery in delivered recommendations.

The benefit extends to music recommendation. On CD, DEHALL-SDPO exceeds DPO/S-DPO in Accuracy, NDCG@10, and MRR, with greedy/sample Hit of 38.69/37.52%. CV is 1.56/1.68%, compared with S-DPO's 59.19/59.56%; MovieLens likewise retains low CV of 0.83/1.07%.

Table 2: **Ranking across five datasets.** Percentages (↑). H@1: HR@1; N@10: NDCG@10.

	MovieLens		Yelp		A. Book		A. Movie		A. CD	
Method	H@1	N@10	H@1	N@10	H@1	N@10	H@1	N@10	H@1	N@10
MoRec	25.53	34.80	7.40	16.85	12.50	21.78	13.05	22.04	9.93	18.42
ChatRec	33.48	45.81	8.04	17.36	13.39	22.75	14.07	23.27	10.83	19.60
TALLRec	39.80	54.62	10.48	25.23	16.81	29.91	17.81	30.71	13.32	25.68
LLaRA	43.60	59.91	11.26	27.18	18.33	32.67	19.47	33.63	14.31	27.63
SPRec	48.23	67.00	12.51	31.18	21.92	40.07	24.20	42.36	17.36	34.18
R-CE	51.99	71.65	21.13	38.38	6.57	27.30	6.87	24.75	7.75	24.96
T-CE	51.89	71.57	21.19	38.11	5.29	26.59	6.68	25.80	8.93	27.51
GRU4Rec	51.23	70.89	25.66	42.75	2.39	25.26	9.03	20.34	4.19	24.91
Caser	55.68	73.80	22.10	48.60	15.70	49.55	12.30	41.92	12.02	43.39
SASRec	57.29	74.37	25.69	42.83	19.65	46.20	13.61	35.49	17.47	40.37
BERT4Rec	54.17	72.81	27.92	49.77	9.86	36.29	9.03	32.67	10.43	32.85
LightSANs	59.00	74.94	25.93	47.47	20.04	48.11	13.29	36.51	16.76	40.94
LPO4Rec	52.65	72.54	22.31	39.34	5.23	23.29	4.12	22.16	5.14	22.64
SFT	63.97	81.00	38.52	58.85	46.96	66.99	46.10	66.65	31.63	52.90
DPO	62.61	80.71	38.71	59.64	49.50	69.83	49.80	71.03	36.81	59.44
Softmax-DPO	64.73	81.78	38.58	59.88	51.13	70.93	50.15	71.52	38.00	60.30
DEHALL-SDPO	65.56	82.25	39.02	60.19	51.96	71.12	50.65	71.68	39.44	60.71

Blocks separate source-reported references, local ranking baselines, and generative policies. Metric definitions and provenance: Appendix E.1.

6.3 ABLATION STUDY: COMPLEMENTARY SUPPORT REPAIR

Prefix support repairs the largest observed validity loss. At matched batch and exposure, Prefix lowers greedy/sample CV from 44.56/46.40% to 1.74/1.90%, while Accuracy changes from 65.33% to 65.71% (Figure 2). The substantial validity recovery without a corresponding ranking loss supports regularizing continuations along policy-visited paths. Details report in Table 15.

Dual Mass strengthens support beyond local repair. Adding Dual Mass raises mean slate mass from 97.53% to 99.15%, lowers sample CV to 0.67%, with 65.26% Accuracy. The additional 1.23-point CV reduction has a paired 95% interval of $[0.64, 1.88]$ (Table 28). This improvement supports combining local continuation repair with explicit control of complete-sequence probability mass.

LPR weight balances ordering and support. In the validation sweep, $\lambda_r = 0.01/0.03/0.05$ yield 65.18/65.56/64.73% MovieLens Accuracy. Weight 0.03 also gives the highest observed NDCG@10 and MRR, although 0.01 retains greater slate mass. This separation motivates evaluating ordering and support jointly; Appendix F.2, Figure 4, and Table 27 provide the configurations and paired uncertainty.

7 CONCLUSION

We identify catalog-support drift as a disconnect between relative preference learning and valid recommendation generation. DEHALL-SDPO addresses this mismatch through full-slate ordering, SFT-calibrated sequence-mass control, and on-policy prefix regularization, while retaining unmasked inference. Five-dataset experiments expose the practical consequences of this drift, and component analyses support complementary sequence- and prefix-level repair. Together, the findings highlight a principle for generative recommendation: learning which item to prefer should be accompanied by learning to generate valid items.

AI USE STATEMENT

In this work, we used generative AI tools to draft parts of the manuscript, edit the paper for readability, and identify and retrieve relevant literature. The authors reviewed and revised all AI-assisted text and independently verified retrieved sources and associated claims against the original publications. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Equations 2–16 specify the objectives and risk interpretations; Appendix A supplies the proofs, and Algorithm 1 in Appendix C.4 specifies the complete update protocol. Appendices B and C.5 document metric events, catalog provenance, checkpoint selection, and verification steps. We also make our codes publicly available.

ETHICS STATEMENT

Catalog membership alone does not establish user benefit, fairness, factual accuracy, or safety. Observed-union inventories can incorporate held-out metadata and must not be presented as independently available deployment catalogs without provenance checks. Any release of user histories, item metadata, or qualitative examples must respect the originating datasets' licenses and privacy requirements.

REFERENCES

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, pp. 1007–1014, 2023.

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*, pp. 129–136, 2007.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards. *Advances in neural information processing systems*, 37:117784–117812, 2024a.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. *Advances in Neural Information Processing Systems*, 37:27463–27489, 2024b.

Jae Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking DPO: The role of rejected responses in preference misalignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8159–8176, 2025.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*, 2020.

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD conference on Knowledge Discovery and Data Mining*, pp. 6448–6458, 2024.

Jianfeng Deng, Qingfeng Chen, Debo Cheng, Jiuyong Li, and Lin Liu. Logit space constrained fine-tuning for mitigating hallucinations in LLM-based recommender systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29311–29324, 2025. doi: 10.18653/v1/2025.emnlp-main.1491.

Xinyan Fan, Zheng Liu, Jianxun Lian, Wayne Xin Zhao, Xing Xie, and Ji-Rong Wen. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1733–1737, 2021.

Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. Sprec: Self-play to debias llm-based recommendation. In *Proceedings of the ACM on Web Conference 2025*, pp. 5075–5084, 2025.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*, 2023.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *Proceedings of the 16th ACM conference on recommender systems*, pp. 299–315, 2022.

F Maxwell Harper and Joseph A Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Jianchao Ji, Zelong Li, Shuyuan Xu, Wenyue Hua, Yingqiang Ge, Juntao Tan, and Yongfeng Zhang. Genrec: Large language model for generative recommendation. In *European Conference on Information Retrieval*, pp. 494–502. Springer, 2024.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the 18th IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.

Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. Large language models for generative recommendation: A survey and visionary discussions. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pp. 10146–10159, 2024.

Zihao Li, Chao Yang, Tong Zhang, Yakun Chen, Xianzhi Wang, Guandong Xu, and Daoyi Dong. Listwise preference alignment optimization for tail item recommendation. *IEEE Transactions on Computational Social Systems*, 2025.

Hao Liao, Jiwei Zhang, Jianxun Lian, Wensheng Lu, Mingqi Wu, Shuowangg, Yong Zhang, Yitian Huang, Mingyang Zhou, and Rui Mao. Eliminating out-of-domain recommendations in LLM-based recommender systems: A unified view. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6251–6271, 2026.

Jiayu Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yanan Wang, Xiang Wang, and Xiangnan He. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2024.

João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, et al. Syntactic and semantic control of large language models via sequential monte carlo. In *International Conference on Learning Representations*, volume 2025, pp. 64758–64791, 2025.

Yatin Nandwani, Abhishek Pathak, Mausam, and Parag Singla. A primal dual formulation for deep learning with constraints. *Advances in neural information processing systems*, 32, 2019.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. *Advances in Neural Information Processing Systems*, 37:24547–24568, 2024.

Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. Tokenrec: Learning to tokenize id for llm-based generative recommendations. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, 2025.

Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. Neural collaborative filtering vs. matrix factorization revisited. In *Proceedings of the 14th ACM conference on recommender systems*, pp. 240–248, 2020.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1441–1450, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 565–573, 2018.

Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. Denoising implicit feedback for recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 373–381, 2021.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*, 2019.

Wencai Ye, Mingjie Sun, Shuhang Chen, Wenjin Wu, and Peng Jiang. Align^{3}gr: Unified multi-level alignment for llm-based generative recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16154–16162, 2026.

Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2639–2649, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Appendix Contents

A Proofs and Interpretation . . . 14

B Evaluation and Implementation Details . . . 14

B.1 Metric definitions . . . 14

B.2 Decoding protocols and ancestral consistency . . . 15

B.3 Training and checkpoint provenance . . . 15

B.4 Uncertainty and planned comparison controls . . . 16

C Objective Geometry and Optimization Details . . . 17

C.1 Ordering and support coordinates . . . 17

C.2 Relation to likelihood regularization . . . 17

C.3 Detached rollouts and the risk interpretation . . . 18

C.4 Complete algorithm and update protocol . . . 18

C.5 Catalog and numerical assumptions . . . 20

D Scope of the Empirical Claims . . . 20

E Five-Dataset Mistral Results . . . 21

E.1 Result provenance and completion status . . . 21

E.2 MovieLens: detailed results . . . 22

E.3 Yelp: detailed results . . . 24

E.4 Amazon Book: detailed results . . . 25

E.5 Amazon Movie: detailed results . . . 26

E.6 Amazon CD: detailed results . . . 27

F Extended Mechanism and Historical Results . . . 28

F.1 Selected-adapter baselines and the prefix-only predecessor . . . 28

F.2 All eight module arms . . . 29

F.3 Ranking-weight sensitivity . . . 29

F.4 Amazon CD: improved recommendation utility with low sampled violation . . . 31

F.5 Paired uncertainty for utility and operating-point comparisons . . . 32

F.6 Paired evidence for support repair . . . 32

F.7 Additional view of the matched component study . . . 33

F.8 Training-log evidence and checkpoint trajectories . . . 33

G Further Work . . . 35

H Enlarged Framework and Reading Guide . . . 36

A Proofs and Interpretation

Proof of Proposition 1. For any $y \in \mathcal{S}(x)$, $\log \pi_q(\bar{y} \mid x) = \log q + \log a(y)$. The common $\log q$ cancels from all within-slate differences and hence, with the reference fixed, from all differences of r_θ in Equation 2. Summing $qa(y)$ over the slate gives q.

Since b has no mass in $\mathcal{C}$, the catalog mass is also q. This is a construction over complete sequence distributions; it need not be realizable by one parameter update of a particular neural network. One may additionally improve chosen–rejected margins by changing a while decreasing q; this possibility, not its inevitability, is the identified blind spot. □

Exact slate risk and its average bound. Canonical candidates include EOS, so even if one title is a prefix of another, their completed events are disjoint. Their union therefore has probability $Z_{\mathcal{S},\theta}$. The complement includes all other completions and any nonterminating paths, yielding $R_{\mathcal{S}} = 1 - Z_{\mathcal{S},\theta}$.

Concavity of $f(u) = 1 - e^{-u}$ gives $\mathbb{E}f(\mathcal{L}_{\text{mass}}) \le f(\mathbb{E}\mathcal{L}_{\text{mass}})$. When $\mathbb{E}\mathcal{L}_{\text{mass}} \le \epsilon$ actually holds on the distribution of interest, outside-slate risk is at most $1 - e^{-\epsilon}$. Calibrating ϵ on validation does not establish that the trained model satisfies this condition there.

Proof of Proposition 2. Define $\tau_{\mathcal{C}} = \infty$ for paths that terminate with valid EOS without a prior exit. Given an alive prefix P_t, the next token exits the trie with conditional probability $1 - m_\theta(P_t)$. The first-exit events at different steps are disjoint, so iterated expectation yields

$$\Pr(\tau_{\mathcal{C}} \le T) = \sum_{t=1}^{T} \mathbb{E}[I_t(1 - m_\theta(P_t))].$$

Applying $1 - u \le -\log u$ term by term proves the bound. All summands are defined as zero when $I_t = 0$; no probability is evaluated after valid termination or after leaving the trie. The result requires the same full-policy distribution for trajectory sampling and hazard evaluation. It is not an importance-corrected statement about nucleus rollouts. □

Gradient of the prefix objective. For legal token set A and logits z, let $p_j = \text{softmax}(z)_j$ and $m = \sum_{a \in A} p_a$. Differentiating $-\log m$ gives

$$\frac{\partial[-\log m]}{\partial z_j} = p_j - \frac{\mathbf{1}\{j \in A\} p_j}{m}.$$

At fixed logits, the gradient moves probability toward legal continuations. Parameter sharing need not preserve valid-item ratios, and the implemented update differentiates the loss at sampled prefixes but not their sampling probabilities. The risk bound is an interpretation of the loss statistic, not a proof that this semi-gradient monotonically decreases risk.

B Evaluation and Implementation Details

B.1 Metric definitions

For prompt i, let r_i be the rank of its single ground-truth item under descending EOS-inclusive total policy log probability; exact ties follow the frozen slate order. Then

$$\text{HR}\,@k = \frac{1}{n}\sum_i \mathbf{1}\{r_i \le k\}, \qquad \text{NDCG}\,@k = \frac{1}{n}\sum_i \frac{\mathbf{1}\{r_i \le k\}}{\log_2(1 + r_i)}, \tag{17}$$

$$\text{MRR} = \frac{1}{n}\sum_i \frac{1}{r_i}, \qquad \text{Accuracy} = \text{HR}\,@1 = \text{NDCG}\,@1. \tag{18}$$

MRR is over the whole real slate, not truncated at ten. The probabilities used to rank items are not renormalized by response length. Reference-relative rewards are diagnostic quantities and do not replace the policy ranking score; these definitions apply to our evaluator, not necessarily to imported reference metrics (Appendix E.1).

For Figure 1, the per-prompt margin is $\Delta_{\min}(x) = \min_{y \in \mathcal{S}(x) \setminus \{y^+\}} \beta[r_\theta(x, y^+) - r_\theta(x, y)]$, with $\beta = 0.1$. The displayed value averages this statistic over prompts; it is not the margin against only the K training negatives.

Let g_{ij} be the text of generated output j for prompt i after stripping surrounding whitespace only, with B outputs per prompt. The primary catalog-string violation rate is

$$\mathrm{CV}_B = \frac{1}{nB} \sum_{i=1}^{n} \sum_{j=1}^{B} \mathbf{1}\{g_{ij} \notin \mathcal{C}\}.$$

Greedy uses $B = 1$; beam-10 and sample-10 use $B = 10$. For sampling, duplicates remain outputs rather than being deduplicated before estimating a rate. Top-one CV and the fraction of prompts with *any* invalid output are separate statistics. For each output, the mutually exclusive events are

$$g = y^+, \quad g \in \mathcal{S} \setminus \{y^+\}, \quad g \in \mathcal{C} \setminus \mathcal{S}, \quad g \notin \mathcal{C}.$$

They sum to one when all prompts have a valid ground truth and slate within the evaluation catalog. The last event is catalog-string violation; the last two together are outside-slate violation. Catalog hallucination in this paper refers to catalog-string nonmembership, not user preference error or a claim of factual nonexistence. An outside-slate output that belongs to the catalog is therefore not a catalog hallucination.

B.2 Decoding protocols and ancestral consistency

For M generated outputs, generation Hit is the mean of $\mathbf{1}\{g_{im} = y_i^+\}$ over prompts and outputs; for greedy, $M = 1$. Any Hit instead averages $\mathbf{1}\{\exists m : g_{im} = y_i^+\}$ over prompts, and Unique averages the number of distinct output strings divided by M. These statistics measure different properties: concentrated samples can increase per-output correctness while reducing the chance that a ten-output set covers the target.

Free greedy and beam search have no catalog token mask. The sample-10 protocol uses temperature one and nucleus probability 0.9; other generation settings, including any inherited top-k filter, must remain tied to the recorded generation configuration. The response horizon is 256 tokens for evaluation, distinct from the 64-token training rollout horizon.

To test the probability identity rather than a decoder heuristic, the sample-100 audit uses temperature one, top-$p = 1$, top-$k = 0$, and no additional logit truncation or catalog masking. For each prespecified validation-subset prompt, it compares $1 - Z_{\mathcal{S},\theta}(x)$ with the fraction of 100 trajectories that do not equal a canonical EOS-terminated slate event. The horizon must accommodate all candidate events; unfinished outputs, first trie exits, and noncanonical string aliases are counted separately.

Any context, precision, padding, or batching discrepancy between scoring and generation is also audited, since numerical mismatch is not evidence against the probability identity. Report per-prompt binomial uncertainty, a prompt-paired aggregate gap with user-cluster uncertainty, and the distribution of residuals, not merely a correlation.

B.3 Training and checkpoint provenance

Table 3 separates the experimental groups used in the argument. Matching effective batch and training exposure is not a claim of equal wall-clock cost, and a selected legacy adapter is not the terminal checkpoint plotted in Figure 1. All comparisons below retain their recorded initialization and endpoint rule.

The preference reference is the original SFT adapter, never a continued-SFT or preference-trained adapter. LPR/mass use all real candidate strings even when the preference pool has only three negatives. Full-slate scoring disables dropout and uses fixed capacity 20: a missing slot is filled for the forward pass and then excluded from every sum, loss, and ranking metric.

Padding is an implementation control, not an extra candidate or a second probability event. Reference scores may be cached; primal–dual state must be restored with training checkpoints. The implemented trie admits both canonical item tokenizations and leading-space variants.

Table 3: **Experimental groups and the questions they answer.** The generative-model groups listed here use Mistral, seed 1958, and validation. Batch denotes effective preference-training batch size.

Group	Evaluation endpoint	Supported comparison
MovieLens legacy trajectories, batch 16	SFT start and saved DPO/S-DPO checkpoints through step 1986	Whether preference progress tracks support and validity.
MovieLens legacy adapters, batch 16	Selected saved adapters; selection can use validation preference loss	Recorded endpoint behavior, separately from terminal trajectories.
MovieLens v2 components, batch 16	Terminal adapters for S-DPO, P, $P + M$, and $P + M + R$	Cumulative additions at fixed training exposure; not equal compute.
MovieLens full-method sweep, batch 16	Terminal adapters; $\lambda_r = 0.03$ selection can use validation preference loss	Descriptive operating-point comparison, not held-out significance.
Amazon Movie/CD full methods, batch 16	Terminal adapters at $\lambda_r = 0.01$, with their one-epoch SFT initializations	End-to-end utility gains, not isolated module effects.

Its first-exit event therefore refers to that augmented token language, whereas $Z_{\mathcal{C}}$ in Equation 3 sums one canonical EOS event per title. These are distinct events; the first-exit bound applies to the chosen trie language and is not asserted to equal $1 - Z_{\mathcal{C}}$ for the single-tokenization definition.

The matched MovieLens ablation uses $\lambda_{\text{sup}} = 0.03$, and $\lambda_r = 0.1$ when LPR is enabled. The separate sweep uses $\lambda_r \in \{0.01, 0.03, 0.05\}$ and effective batch 16. Recorded dual defaults are $\mu_0 = 0$, $\eta_\mu = 0.01$, and $\mu_{\max} = 100$; the MovieLens mass budget of 0.008396913930410983 nats must be recalibrated for each dataset and SFT initialization.

MovieLens uses three SFT epochs and three preference epochs; the completed Amazon CD metadata instead records one SFT epoch followed by three preference epochs. The CD full method uses $\lambda_r = 0.01$, $\lambda_{\text{sup}} = 0.03$, effective batch 16, and its own SFT-calibrated mass budget of 0.05825049 nats. Each SFT/full-method pair shares data, catalog, evaluation IDs, scoring settings, and the initial SFT adapter, but the CD study is not a matched-component ablation or a replication of MovieLens's selected weight.

The observed-union catalogs contain 9,737 MovieLens titles and 27,516 CD titles. All 1,324 MovieLens prompts have 20 real candidates; on CD, 1,599 prompts have 20 and one has 19, with the padded slot excluded from scoring aggregates. The CD full adapter retains the recorded post-hoc archived/user-attested source binding rather than being described as an atomic provenance snapshot.

Checkpoint plots include saved training checkpoints and a measured SFT initialization. Legacy DPO/S-DPO selected adapters can correspond to the best validation-loss checkpoint, whereas the plotted step 1986 is the final actual training checkpoint. V2 endpoint selection follows its recorded final-checkpoint configuration.

Accordingly, old selected-adapter results, terminal-checkpoint trajectories, and separate compute-matched runs are not interchangeable endpoints. Intermediate DeHall checkpoint measurements use a post-hoc diagnostic provenance ledger and are not silently promoted to formally bound final-adapter evaluations.

B.4 Uncertainty and planned comparison controls

For a user-cluster bootstrap, users are sampled with replacement and all prompts belonging to each sampled user are retained. Metrics are aggregated over prompts in the resulting resample; this is prompt-weighted, not an equal-user average. Paired method differences resample the same users for both methods.

Pointwise intervals in a trajectory are not simultaneous confidence bands and do not account for selecting a hyperparameter on the same validation set. Independent training seeds are needed to estimate training variability. The paired intervals in Table 27 likewise condition on the trained models and are not corrected for multiple comparisons or validation selection.

Continued SFT controls additional chosen-response training; a chosen-NLL auxiliary tests the identity in Equation 13; a fixed mass penalty tests the need for adaptive dual weighting. A constrained-decoding control should report utility, validity, and inference cost under the same inventory, without treating its by-construction validity as evidence of learned support. An authoritative ex-ante catalog control separates distribution repair from access to held-out item metadata.

C Objective Geometry and Optimization Details

The sequence objectives act on conditional item preference and absolute slate mass, while the prefix objective acts on visited continuation distributions. The following derivations specify these roles and the update actually used by the implementation. They characterize the losses, without assuming independent neural parameters for each role.

C.1 Ordering and support coordinates

For $\tau = 1$, write $\pi_\theta(\bar{y} \mid x) = Zq_y$ on the slate, where $Z = Z_{\mathcal{S},\theta}(x)$ and $q_y = q_\theta(y \mid x, \mathcal{S})$. Holding the reference fixed, S-DPO depends on differences of $\log q_y$ and contains no $\log Z$ term; LPR equals $-\log q_{y^+}$, whereas the mass loss equals $-\log Z$. Their weighted sequence penalty therefore satisfies

$$\lambda_r \mathcal{L}_{\mathrm{LPR}} + \mu(\mathcal{L}_{\mathrm{mass}} - \epsilon) = -\lambda_r \log q_{y^+} - \mu \log Z - \mu\epsilon. \tag{19}$$

This exposes the unregulated support coordinate of preference learning and explains the distinct role of μ. For $\tau \neq 1$, the common $\log Z/\tau$ still cancels from LPR, but its conditional distribution is proportional to $q_y^{1/\tau}$ rather than q_y. These coordinate identities do not imply gradient orthogonality in a shared-parameter language model.

The response-score derivatives also make the roles explicit:

$$\frac{\partial \mathcal{L}_{\mathrm{LPR}}}{\partial s_j} = \frac{\widetilde{q}_j - \mathbf{1}\{j = +\}}{\tau}, \qquad \frac{\partial \mathcal{L}_{\mathrm{mass}}}{\partial s_j} = -q_j, \quad \widetilde{q}_j = \frac{e^{s_j/\tau}}{\sum_{\ell \in \mathcal{S}} e^{s_\ell/\tau}}. \tag{20}$$

The LPR derivatives sum to zero, leaving a common score shift unchanged; the mass derivatives sum to -1. Because increasing all slate scores requires reallocating probability from its complement, the second direction addresses absolute support, not a different ranking normalization. Actual policy updates couple these score directions through the model Jacobian.

C.2 Relation to likelihood regularization

At $\tau = 1$, Equation 9 gives $\mathcal{L}_{\mathrm{LPR}} = -s_\theta(x, y^+) - \mathcal{L}_{\mathrm{mass}}$. Consequently, at a fixed multiplier the weighted sequence part of the objective satisfies

$$\lambda_r \mathcal{L}_{\mathrm{LPR}} + \mu(\mathcal{L}_{\mathrm{mass}} - \epsilon) = -\lambda_r s_\theta(x, y^+) + (\mu - \lambda_r)\mathcal{L}_{\mathrm{mass}} - \mu\epsilon. \tag{21}$$

This is an algebraic reparameterization, not an independence claim about the effects of the two losses. At equal fixed weights their sum is a chosen-sequence NLL term plus a constant; an adaptive multiplier changes that weighting over training.

The coefficient $\mu - \lambda_r$ can be negative without making LPR an explicit negative support penalty: the chosen score also contains the mass coordinate. Substituting $s_\theta = \log Z_{\mathcal{S},\theta} + \log q_\theta$ into the original objective leaves $-\lambda_r \log q_\theta - \mu \log Z_{\mathcal{S},\theta} - \mu\epsilon$. Holding conditional ordering fixed thus retains a nonnegative mass penalty; the decomposition alone neither diagnoses destructive gradients nor explains the observed ranking trade-off.

At an alive prefix, $-\log m_\theta = -\log(1 - h_\theta)$ is a set-level unlikelihood penalty on aggregate invalid-token probability, related to the unwanted-continuation principle of Welleck et al. (2019). The catalog-defined legal set and current-policy prefix distribution specify its application here; the logarithmic penalty itself is not a new mathematical primitive. Evolving-policy, teacher-forced, and frozen-policy prefixes differ in their visitation distributions.

C.3 Detached rollouts and the risk interpretation

The implementation samples one continuation per training prompt and differentiates only the re-scored support loss at the selected alive prefixes. In Equation 8, $\bar{\theta}$ is a frozen copy in the mathematical expectation, not an additional trainable model; the current policy is used to draw these detached samples. Thus the update is a Monte Carlo semi-gradient of a refreshed surrogate, not the full gradient of its changing rollout distribution.

For a differentiable ancestral policy and fixed horizon, let $J(\theta) = \mathbb{E}_{\widetilde{Y}\sim\pi_\theta}\mathcal{L}_{\mathrm{sup}}(\theta, \widetilde{Y})$. The identity

$$\nabla_\theta J = \mathbb{E}_{\pi_\theta}[\nabla_\theta \mathcal{L}_{\mathrm{sup}}] + \mathbb{E}_{\pi_\theta}[\mathcal{L}_{\mathrm{sup}} \nabla_\theta \log \pi_\theta(\widetilde{Y} \mid x)] \tag{22}$$

distinguishes the implemented first term from the score-function term that is omitted. The recorded nucleus rollout additionally differs from π_θ by its decoder filtering. Proposition 2 bounds a statistic under full-policy ancestral sampling; it neither makes these sampling laws identical nor proves monotone risk improvement from the implemented update.

C.4 Complete algorithm and update protocol

Algorithm 1 makes the three support-and-preference signals operational in a single training loop. Here $\bar{y}$ denotes the canonical item-token sequence followed by EOS, $s_\theta(x, y) = \log \pi_\theta(\bar{y} \mid x)$, and sg denotes stop-gradient. The frozen preference pool supplies $\mathcal{N}_K(x)$, whereas $\mathcal{S}(x)$ contains every real displayed candidate, including y^+.

Sequence-level ordering and mass. The reference-relative S-DPO term uses the chosen item and the first K negatives in the frozen pool; LPR and Mass instead use all distinct, complete slate responses. Policy scores include EOS, exclude prompt tokens, and are not length-normalized; the matched calibration and training scorers disable dropout and use fixed-capacity, prompt-wise slate forwards. For a 19-item slate, the implementation pads to 20 slots with a chosen-response copy for the forward pass only, then excludes that copy from every loss and metric.

Visited-prefix support. PrefixSupport sums the legal-token mass penalty along one detached rollout per prompt, including the prefix immediately before the first invalid token, and the outer loss averages over the original prompt batch. Its differentiable re-scoring uses the original policy softmax, not a catalog-masked or nucleus-renormalized distribution; it stops after the first exit or legal EOS, without declaring an unfinished legal prefix invalid. The trie includes the declared item tokenizations and their leading-space variants, while sequence mass uses canonical EOS events; a genuinely generated PAD before EOS is retained so that an invalid transition is not hidden.

Calibration and update order. The fixed budget is the matched SFT validation mean of $-\log Z_{\mathcal{S}}$, not a target equal to the mean slate probability, and its dataset, adapter, tokenizer, and scoring provenance are preserved. Within each optimizer step, μ stays fixed; after a successful primal update, dual ascent uses the detached, pre-update mass-loss sum divided by the actual prompt count across all accumulated microbatches and devices. A skipped primal update leaves μ unchanged and discards the pending mass statistics, while resumable checkpoints retain optimizer, scheduler, random, data-progress, and dual state together with the policy.

Sampling and inference. The recorded training rollout uses the current policy with temperature one, nucleus $p = 0.9$, and horizon 64, with the run's recorded or inherited top-k setting; trajectory selection carries no gradient. At inference, ranking sorts raw EOS-inclusive policy scores, and free generation uses the learned policy without a catalog mask or the frozen reference. The full-policy ancestral identities in Section 5.4 interpret the support statistics under their stated sampling law, rather than identifying the nucleus rollout with ancestral sampling.

Validation and logs. Validation's returned loss is preference loss, not the full training Lagrangian. The implementation evaluates prefix diagnostics on chosen continuations during validation, whereas training uses current-policy rollouts when on-policy support is enabled. These curves cannot be compared as if they were the same risk estimator; generated-prefix measurements come from the separate checkpoint inference protocol.

Algorithm 1 DEHALL-SDPO: full-slate preference learning with sequence and prefix support

Inputs: training tuples $(x, y^+, \mathcal{N}_K(x), \mathcal{S}(x))$; calibration prompts $\mathcal{D}_{\text{cal}}$; SFT policy θ_0; declared catalog $\mathcal{C}$; epochs E; rollout decoder ρ, horizon T.
Parameters: $\beta, \tau, \lambda_r, \lambda_{\text{sup}}$; dual step size η_μ, cap $\mu_{\max}$; optimizer and accumulation schedule.

I. Initialize and calibrate once

$\theta \leftarrow \theta_0$; freeze $\theta_{\text{ref}} \leftarrow \theta_0$; $\mu \leftarrow 0$; disable dropout for coherent scoring

$$\epsilon \leftarrow -\frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{x \in \mathcal{D}_{\text{cal}}} \log \sum_{y \in \mathcal{S}(x)} \exp s_{\theta_0}(x, y)$$

Build EOS-aware trie $\mathcal{T}_\mathcal{C}$; fix tokenizer, response boundaries, and scoring settings

II. Joint primal–dual training

for each effective training batch, over E epochs **do**
  Zero policy gradients; $U \leftarrow 0$, $N \leftarrow 0$; hold μ fixed during accumulation
  for each local microbatch b **do**
    for each prompt $x \in b$ with its chosen item y^+ **do**
      $s_y \leftarrow s_\theta(x, y)$ for every real $y \in \mathcal{S}(x)$ ▷ EOS-inclusive full-slate scores
      $r_y \leftarrow s_y - s_{\theta_{\text{ref}}}(x, y)$ for $y \in \{y^+\} \cup \mathcal{N}_K(x)$ ▷ reference may be cached
      $$\ell_{\text{pref},x} \leftarrow \log\left[1 + \sum_{y^- \in \mathcal{N}_K(x)} \exp\big(\beta[r_{y^-} - r_{y^+}]\big)\right]$$
      $$\ell_{\text{rank},x} \leftarrow -s_{y^+}/\tau + \log \sum_{y \in \mathcal{S}(x)} \exp(s_y/\tau)$$
      $$\ell_{\text{mass},x} \leftarrow -\log \sum_{y \in \mathcal{S}(x)} \exp(s_y)$$ ▷ exclude all padding copies
      Draw one $\widetilde{Y}_x \sim \rho_\theta(\cdot \mid x)$ up to EOS or T, without gradient or catalog masking
      $\ell_{\text{sup},x} \leftarrow \textsc{PrefixSupport}(x, \text{sg}(\widetilde{Y}_x), \theta, \mathcal{T}_\mathcal{C})$
      $\ell_x \leftarrow \ell_{\text{pref},x} + \lambda_r \ell_{\text{rank},x} + \lambda_{\text{sup}} \ell_{\text{sup},x} + \mu(\ell_{\text{mass},x} - \epsilon)$
    end for
    Backpropagate $|b|^{-1} \sum_{x \in b} \ell_x$ with the configured gradient-accumulation scaling
    $U \leftarrow U + \sum_{x \in b} \text{sg}(\ell_{\text{mass},x})$; $N \leftarrow N + |b|$
  end for
  Apply the accumulated policy optimizer step and corresponding learning-rate schedule
  if the policy optimizer step succeeded **then**
    $\bar{\ell}_{\text{mass}} \leftarrow \text{AllReduceSum}(U) / \text{AllReduceSum}(N)$
    $\mu \leftarrow \min\{\mu_{\max}, \max\{0, \mu + \eta_\mu(\bar{\ell}_{\text{mass}} - \epsilon)\}\}$
  end if
  Log component losses, support statistics, and μ; save complete training state at checkpoint steps
end for

III. Differentiable support at detached, alive prefixes

procedure $\textsc{PrefixSupport}(x, \widetilde{Y}, \theta, \mathcal{T}_\mathcal{C})$
  $p \leftarrow ()$; $\ell \leftarrow 0$
  for each generated token a in $\widetilde{Y}$, in order **do**
    $A \leftarrow A_\mathcal{C}(p)$ from $\mathcal{T}_\mathcal{C}$ ▷ EOS is legal only at an item endpoint
    $\ell \leftarrow \ell - \log \sum_{v \in A} \pi_\theta(v \mid x, p)$ ▷ re-score with gradient; no token mask
    if $a \notin A$ or $a = \text{EOS}$ **then**
      break ▷ the exit-preceding or terminating prefix was already penalized
    end if
    $p \leftarrow p \,\|\, a$
  end for
  return ℓ
end procedure

Output and inference: learned policy π_θ; rank candidates by s_θ or generate freely, without the reference model, dual variable, or a catalog token mask.

The logged components separate unweighted preference, LPR, mass, and prefix losses from their weighted contributions and the multiplier. Optimizer-step and prompt counts describe data exposure; candidate scoring, rollout, and re-scoring also incur token and device costs. Matching effective batch and epoch count controls data exposure, but does not imply equal computation when additional candidate forwards or rollouts are used.

C.5 Catalog and numerical assumptions

Inventory availability. The support catalog's construction and split coverage determine which inventory information is supplied during training. Held-out item metadata can be legitimate when independently available at deployment; a union reconstructed from evaluation records does not by itself establish such availability.

Scoring coherence. The mass interpretation assumes that candidate and generated-continuation probabilities come from one policy, with identical prompt boundaries, EOS conventions, precision, and evaluation mode. Fixed-slate batching and a check for mass above one provide implementation safeguards, but do not replace an independent comparison with single-sequence or non-quantized scoring. The historical decoder records share explicit temperature, nucleus, horizon, and library settings; older records omit an explicit inherited top-k value, which remains a numerical-audit consideration.

String validity and beam failures. A strict string mismatch can reflect a noncatalog name, an alias, formatting, a special token, or an incomplete response. The complete generation tables retain both per-output and highest-scored-beam measurements where available, rather than identifying list validity with first-output validity. A method-blinded error audit is needed to attribute the strict violation rate to semantic item errors rather than surface-form mismatches.

Ancestral risk consistency. The planned consistency audit uses 100 full-softmax trajectories per prespecified prompt at temperature one, top-$p = 1$, and top-$k = 0$, without other probability filtering. A horizon covering every candidate's canonical EOS event permits comparison of token-event nonmembership with $1 - Z_{\mathcal{S},\theta}$, separately from normalized-string CV and trie exits. Agreement requires a small paired probability gap with suitable uncertainty, not merely correlation.

For trie events at a finite horizon T, failure to complete a valid item decomposes as

$$\Pr(\text{no valid trie completion by } T) = \Pr(\tau_{\mathcal{C}} \le T) + \Pr(\text{alive and unfinished after } T). \tag{23}$$

The events are disjoint: the first-exit bound controls the first term, not unfinished legal prefixes. The augmented trie language, canonical EOS events, and whitespace-normalized string equality remain distinct throughout the audit. Nucleus sampling does not test the full-policy identity and is not interchangeable with this ancestral experiment.

D Scope of the Empirical Claims

The main text emphasizes two measured benefits: complementary support repair on MovieLens and improved ranking and per-output recommendation utility on Amazon Movie and Amazon CD. Its generation claims concern greedy decoding and the recorded nucleus protocol; all measured beam-search outcomes remain in the extended result tables, for every method with available outputs. This organization distinguishes the main operating regime from additional decoder diagnostics without treating unreported main-text metrics as successful or absent experiments.

Single-output utility and search-expanded validity. On CD, the full method increases highest-scored-beam correctness from 23.81% to 38.19%, while catalog violation across all ten beams increases from 19.96% to 54.69%. On MovieLens, the selected full model has 0.60% highest-scored-beam CV but 68.27% CV across all ten outputs. These results distinguish a useful first recommendation from validity throughout a returned list; lower nucleus CV does not establish full-distribution or decoder-independent dehallucination.

Catalog provenance and statistical scope. The current evaluations use one trained seed and observed-union catalogs, and the CD comparison uses the recorded one-epoch SFT initialization. Strict catalog-string equality is not an adjudication of factual item nonexistence, and user-cluster intervals do not capture training-seed variability or adjust for validation selection.

Control objectives and achieved performance. An SFT-calibrated training target does not guarantee that the final validation mass loss satisfies that budget: MovieLens Prefix+Mass has 0.03591 versus 0.00840 nats, and CD has 0.09462 versus 0.05825 nats. The complete method's utility gains do not, by themselves, identify the marginal effect of LPR or establish that all components are necessary; the Movie/CD DPO/S-DPO comparisons appear in Tables 11 and 13.

E Five-Dataset Mistral Results

This section expands the two main result tables into a ranking table and a generation table for each dataset. The ranking tables include sequential and language-based baselines together with LPO4Rec; the generation tables report SFT, DPO, Softmax-DPO, and DeHall. Shading identifies DeHall rather than a significance claim; missing entries are unavailable measurements, not zeros.

E.1 Result provenance and completion status

SFT/preference results retain the designated archives, with completed tasks checked for prompt, catalog, and initialization identity. Missing tasks are not inferred from other outputs.

Table 4: **Recorded training settings and result availability.** All SFT/preference rows use Mistral-7B-Instruct-v0.3, seed 1958, and observed-union validation. Preference training runs for three epochs.

Dataset	Prompts	Users	SFT epochs	DeHall λ_r
MovieLens	13,240	610	3	0.03
Yelp	48,000	3,000	3	0.03
Amazon Book	24,000	3,000	1	0.03
Amazon Movie	20,000	2,500	1	0.03
Amazon CD	16,000	2,000	1	0.01

These are the epochs recorded by the evaluated adapters, not a claim that all SFT runs used three epochs. The full-method runs are not treated as matched-module ablations against legacy baselines.

The generative runs use LoRA rank 32, 4-bit loading, learning rate 10^{-5}, and preference $\beta = 0.1$. Effective batches are 16 for SFT/DPO/S-DPO and the three full-method endpoints. The 10,000 user-cluster bootstrap replicates quantify sampling uncertainty, not seed variability.

MovieLens legacy baselines lack fixed-capacity scoring metadata; other datasets and full methods use fixed-20-slot scoring. MovieLens endpoint ranking is therefore not a batching-matched comparison. Historical fixed-pad and checkpoint studies retain their separate sources.

Conventional recommendation baselines. We include GRU4Rec (Hidasi et al., 2016), Caser (Tang & Wang, 2018), SASRec (Kang & McAuley, 2018), BERT4Rec (Sun et al., 2019), and LightSANs (Fan et al., 2021), covering recurrent, convolutional, and attention-based approaches to sequential recommendation. We additionally evaluate R-CE and T-CE (Wang et al., 2021), two denoising objectives implemented on an NCF-based NeuMF model (He et al., 2017; Rendle et al., 2020): R-CE reweights cross-entropy losses, whereas T-CE progressively discards high-loss positive interactions. These non-sequential baselines complement the sequence models by examining preference estimation under noisy implicit feedback.

Content- and language-based recommendation baselines. We include MoRec (Yuan et al., 2023), ChatRec (Gao et al., 2023), TALLRec (Bao et al., 2023), and LLaRA (Liao et al., 2024) as representative approaches that exploit item content or language models for recommendation.

Their ranking results provide a broader reference for assessing the utility of language-informed recommendation. SFT serves as the direct supervised baseline for our generative policy and as the common initialization for the corresponding DPO, Softmax-DPO, and DEHALL-SDPO runs.

Preference-optimization baselines. DPO (Rafailov et al., 2023) and Softmax-DPO (Chen et al., 2024b) are our primary comparisons, using one and three preference negatives, respectively. They isolate the effect of relative preference learning without our auxiliary ranking, sequence-mass, and prefix-support objectives. We also include SPRec (Gao et al., 2025) and LPO4Rec (Li et al., 2025) as additional preference-learning baselines, with LPO4Rec providing a complementary listwise-alignment approach for recommendation.

Local implementations and model selection. GRU4Rec, Caser, SASRec, BERT4Rec, and LightSANs use the corresponding RecBole 1.2.1 implementations with cross-entropy training. R-CE and T-CE use the authors' NeuMF implementation and denoising objectives, while LPO4Rec uses its public implementation adapted to the frozen data interface. The eight locally evaluated baselines use seed 1958 and validation NDCG@10 for model selection, followed by evaluation on the held-out test split.

Result provenance and evaluation scope. MoRec, ChatRec, TALLRec, LLaRA, and SPRec use the same validation NDCG@10 for model selection, followed by evaluation on the held-out test split. Item-ID recommenders are evaluated through candidate ranking; their closed-set outputs are not treated as evidence of learned catalog grounding under free generation. Accordingly, the free-generation comparisons focus on SFT, DPO, Softmax-DPO, and DEHALL-SDPO, and ranking results from different evaluation protocols are distinguished in the accompanying provenance records.

E.2 MOVIELENS: DETAILED RESULTS

The saved validation split contains 13,240 prompts from 610 users. The first table reports the complete set of ranking metrics; the second separates generation utility, catalog violation, and slate violation by decoder.

Table 5: **MovieLens: full ranking results.** Percentages (↑).

Method	H@1	H@3	H@5	H@10	N@3	N@5	N@10	MRR
MoRec	25.53	35.00	38.98	45.65	31.02	32.66	34.80	31.87
ChatRec	33.48	46.08	51.36	60.24	40.78	42.97	45.81	41.92
TALLRec	39.80	54.95	61.31	71.99	48.58	51.20	54.62	49.94
LLaRA	43.60	60.27	67.26	79.01	53.26	56.15	59.91	54.76
SPRec	48.23	67.41	75.46	88.99	59.35	62.67	67.00	61.07
R-CE	51.99	74.72	84.09	91.76	65.26	69.14	71.65	65.76
T-CE	51.89	74.05	83.52	91.76	64.94	68.87	71.57	65.67
GRU4Rec	51.23	73.20	84.28	91.00	64.08	68.66	70.89	65.04
Caser	55.68	77.65	85.32	91.57	68.60	71.75	73.80	68.66
SASRec	57.29	77.65	85.23	91.10	69.31	72.45	74.37	69.56
BERT4Rec	54.17	77.75	84.00	90.44	68.11	70.72	72.81	67.73
LightSANs	59.00	77.94	85.13	91.19	69.98	72.96	74.94	70.32
LPO4Rec	52.65	75.38	85.13	92.52	66.06	70.12	72.54	66.63
SFT	63.97	84.44	91.69	97.66	76.02	79.04	81.00	75.80
DPO	62.61	85.35	91.77	97.73	76.06	78.72	80.71	75.35
Softmax-DPO	64.73	85.80	91.84	98.19	77.23	79.72	81.78	76.62
DEHALL-SDPO	65.56	86.48	93.05	98.19	77.86	80.55	82.25	77.22

H@K and N@K denote HR@K and NDCG@K. Policy MRR spans the full slate; source-reference definitions are not assumed identical (Appendix E.1).

Table 6: **MovieLens: free-generation results.** Percentages. Hit is mean per-output correctness; Any Hit/CV count prompts with at least one such output. Beam rows average all ten outputs.

Decoder	Method	Hit ↑	CV ↓	SV ↓	Any Hit ↑	Any CV ↓	Unique
Greedy	SFT	64.35	0.00	0.00	64.35	0.00	100.00
	DPO	22.05	75.15	75.23	22.05	75.15	100.00
	Softmax-DPO	20.85	76.36	76.44	20.85	76.36	100.00
	DEHALL-SDPO	65.56	0.83	0.83	65.56	0.83	100.00
Sample-10	SFT	57.47	0.05	0.07	82.85	0.38	24.22
	DPO	21.69	75.45	75.54	29.08	83.99	72.74
	Softmax-DPO	20.42	76.41	76.47	35.57	91.24	69.52
	DEHALL-SDPO	64.16	1.07	1.09	72.43	3.02	13.68
Beam-10	SFT	9.32	23.36	24.46	82.55	97.36	98.68
	DPO	5.42	93.03	93.06	43.13	100.00	98.59
	Softmax-DPO	5.17	92.53	92.61	50.15	100.00	99.80
	DEHALL-SDPO	14.40	68.27	68.43	83.38	100.00	88.70

CV: outside the observed-union catalog; SV: outside the prompt slate. Unique is a diversity diagnostic, not a correctness measure.

E.3 YELP: DETAILED RESULTS

The saved validation split contains 48,000 prompts from 3,000 users. The first table reports the complete set of ranking metrics; the second separates generation utility, catalog violation, and slate violation by decoder.

Table 7: **Yelp: full ranking results.** Percentages (↑).

Method	H@1	H@3	H@5	H@10	N@3	N@5	N@10	MRR
MoRec	7.40	13.86	19.15	30.25	11.11	13.30	16.85	14.30
ChatRec	8.04	14.41	19.63	30.58	11.70	13.86	17.36	14.84
TALLRec	10.48	20.56	28.84	46.18	16.27	19.69	25.23	21.25
LLaRA	11.26	22.14	31.08	49.79	17.51	21.20	27.18	22.88
SPRec	12.51	25.27	35.75	57.69	19.84	24.17	31.18	26.14
R-CE	21.13	36.82	45.32	60.10	30.16	33.64	38.38	34.21
T-CE	21.19	35.67	44.78	59.75	29.54	33.30	38.11	34.02
GRU4Rec	25.66	42.05	50.42	63.21	35.16	38.63	42.75	38.60
Caser	22.10	44.09	60.45	82.52	34.69	41.41	48.60	39.40
SASRec	25.69	41.75	50.28	63.42	35.09	38.59	42.83	38.65
BERT4Rec	27.92	47.63	58.59	76.94	39.42	43.91	49.77	43.10
LightSANs	25.93	47.01	58.25	72.89	38.11	42.73	47.47	41.31
LPO4Rec	22.31	37.89	46.29	60.39	31.32	34.78	39.34	35.29
SFT	38.52	56.73	67.38	84.23	49.06	53.43	58.85	52.19
DPO	38.71	58.19	68.25	85.27	50.01	54.13	59.64	52.82
Softmax-DPO	38.58	58.27	68.90	85.94	49.99	54.37	59.88	52.87
DEHALL-SDPO	39.02	58.71	69.35	86.38	50.43	54.86	60.19	53.14

H@K and N@K denote HR@K and NDCG@K. Policy MRR spans the full slate; source-reference definitions are not assumed identical (Appendix E.1).

Table 8: **Yelp: free-generation results.** Percentages. Hit is mean per-output correctness; Any Hit/CV count prompts with at least one such output. Beam rows average all ten outputs.

Decoder	Method	Hit ↑	CV ↓	SV ↓	Any Hit ↑	Any CV ↓	Unique
Greedy	SFT	38.42	0.10	0.15	38.42	0.10	100.00
	DPO	34.88	20.75	20.79	34.88	20.75	100.00
	Softmax-DPO	21.33	66.69	66.75	21.33	66.69	100.00
	DEHALL-SDPO	38.91	0.27	0.29	38.91	0.27	100.00
Sample-10	SFT	29.90	0.21	0.24	63.79	1.31	44.18
	DPO	32.38	31.08	31.13	39.67	57.27	40.04
	Softmax-DPO	20.54	67.84	67.92	28.00	81.44	65.33
	DEHALL-SDPO	37.94	0.36	0.39	63.91	1.25	13.84
Beam-10	SFT	7.44	13.72	14.17	73.65	81.19	99.88
	DPO	4.11	88.48	88.54	39.63	100.00	99.79
	Softmax-DPO	2.83	94.75	94.78	25.90	100.00	99.71
	DEHALL-SDPO	11.58	12.94	13.73	74.26	81.05	97.58

CV: outside the observed-union catalog; SV: outside the prompt slate. Unique is a diversity diagnostic, not a correctness measure.

E.4 AMAZON BOOK: DETAILED RESULTS

The saved validation split contains 24,000 prompts from 3,000 users. The first table reports the complete set of ranking metrics; the second separates generation utility, catalog violation, and slate violation by decoder.

Table 9: **Amazon Book: full ranking results.** Percentages (↑).

Method	H@1	H@3	H@5	H@10	N@3	N@5	N@10	MRR
MoRec	12.50	20.09	24.74	34.28	16.80	18.72	21.78	19.15
ChatRec	13.39	21.04	25.74	35.37	17.73	19.67	22.75	20.10
TALLRec	16.81	27.52	34.09	47.56	22.88	25.59	29.91	26.20
LLaRA	18.33	30.05	37.24	51.98	24.97	27.94	32.67	28.61
SPRec	21.92	35.69	45.58	64.82	29.84	33.90	40.07	34.99
R-CE	6.57	18.37	30.46	58.69	13.34	18.27	27.30	20.81
T-CE	5.29	17.37	29.06	59.47	12.10	16.85	26.59	19.62
GRU4Rec	2.39	7.13	17.71	67.76	4.99	9.26	25.26	15.36
Caser	15.70	48.78	71.71	88.75	34.45	43.90	49.55	37.98
SASRec	19.65	47.77	64.81	75.72	35.60	42.65	46.20	38.36
BERT4Rec	9.86	30.12	47.38	71.83	21.27	28.35	36.29	27.38
LightSANs	20.04	50.39	65.87	79.29	37.32	43.73	48.11	39.59
LPO4Rec	5.23	16.26	26.17	50.67	11.40	15.48	23.29	18.48
SFT	46.96	67.96	77.21	89.25	59.29	63.09	66.99	60.76
DPO	49.50	71.29	81.25	92.38	62.14	66.24	69.83	63.28
Softmax-DPO	51.13	71.58	81.29	93.13	63.07	67.07	70.93	64.45
DEHALL-SDPO	51.96	72.05	81.83	93.28	63.53	67.49	71.12	64.74

H@K and N@K denote HR@K and NDCG@K. Policy MRR spans the full slate; source-reference definitions are not assumed identical (Appendix E.1).

Table 10: **Amazon Book: free-generation results.** Percentages. Hit is mean per-output correctness; Any Hit/CV count prompts with at least one such output. Beam rows average all ten outputs.

Decoder	Method	Hit ↑	CV ↓	SV ↓	Any Hit ↑	Any CV ↓	Unique
Greedy	SFT	46.00	1.17	1.29	46.00	1.17	100.00
	DPO	22.29	70.54	70.67	22.29	70.54	100.00
	Softmax-DPO	34.29	48.17	48.63	34.29	48.17	100.00
	DEHALL-SDPO	51.37	1.21	1.27	51.37	1.21	100.00
Sample-10	SFT	37.86	1.62	1.75	71.29	7.92	39.41
	DPO	21.58	70.62	70.84	31.79	86.04	70.00
	Softmax-DPO	33.75	48.52	49.02	42.63	69.00	46.70
	DEHALL-SDPO	51.09	1.53	1.57	72.02	5.84	19.53
Beam-10	SFT	6.05	32.15	32.91	59.67	99.79	99.84
	DPO	2.70	95.52	95.59	25.21	100.00	99.67
	Softmax-DPO	7.03	85.42	85.73	53.46	100.00	97.49
	DEHALL-SDPO	9.53	31.94	31.97	60.19	99.65	93.27

CV: outside the observed-union catalog; SV: outside the prompt slate. Unique is a diversity diagnostic, not a correctness measure.

E.5 AMAZON MOVIE: DETAILED RESULTS

The saved validation split contains 20,000 prompts from 2,500 users. The first table reports the complete set of ranking metrics; the second separates generation utility, catalog violation, and slate violation by decoder.

Table 11: **Amazon Movie: full ranking results.** Percentages (↑).

Method	H@1	H@3	H@5	H@10	N@3	N@5	N@10	MRR
MoRec	13.05	20.40	24.83	34.09	17.24	19.07	22.04	19.54
ChatRec	14.07	21.60	26.14	35.62	18.37	20.24	23.27	20.72
TALLRec	17.81	28.37	34.72	48.01	23.83	26.46	30.71	27.13
LLaRA	19.47	31.06	38.03	52.61	26.08	28.96	33.63	29.70
SPRec	24.20	39.05	48.00	66.70	32.67	36.37	42.36	37.31
R-CE	6.87	16.56	26.37	52.75	12.32	16.34	24.75	19.74
T-CE	6.68	17.60	27.42	55.24	12.93	16.94	25.80	20.19
GRU4Rec	9.03	17.74	23.95	35.73	14.03	16.59	20.34	19.62
Caser	12.30	38.87	58.90	78.93	27.26	35.49	41.92	31.91
SASRec	13.61	35.34	46.34	62.04	25.93	30.44	35.49	29.67
BERT4Rec	9.03	28.53	42.67	63.42	20.11	25.98	32.67	25.72
LightSANs	13.29	35.14	47.97	64.99	25.71	31.01	36.51	29.98
LPO4Rec	4.12	15.18	25.39	49.08	10.43	14.60	22.16	17.57
SFT	46.10	68.30	77.00	89.15	59.11	62.68	66.65	60.32
DPO	49.80	73.70	82.45	93.70	63.76	67.38	71.03	64.32
Softmax-DPO	50.15	74.05	82.90	94.00	64.22	67.90	71.52	64.82
DEHALL-SDPO	50.65	74.55	83.30	93.70	64.69	68.28	71.68	65.14

H@K and N@K denote HR@K and NDCG@K. Policy MRR spans the full slate; source-reference definitions are not assumed identical (Appendix E.1).

Table 12: **Amazon Movie: free-generation results.** Percentages. Hit is mean per-output correctness; Any Hit/CV count prompts with at least one such output. Beam rows average all ten outputs.

Decoder	Method	Hit ↑	CV ↓	SV ↓	Any Hit ↑	Any CV ↓	Unique
Greedy	SFT	45.70	1.05	1.15	45.70	1.05	100.00
	DPO	41.45	25.35	26.10	41.45	25.35	100.00
	Softmax-DPO	28.65	61.10	61.20	28.65	61.10	100.00
	DEHALL-SDPO	50.20	1.45	1.55	50.20	1.45	100.00
Sample-10	SFT	36.40	1.44	1.63	71.20	6.40	39.04
	DPO	40.25	28.73	29.35	47.65	46.60	29.61
	Softmax-DPO	27.56	61.69	61.92	37.40	80.05	44.43
	DEHALL-SDPO	49.60	1.72	1.83	57.95	4.10	14.38
Beam-10	SFT	6.49	26.45	28.80	64.40	98.40	99.91
	DPO	6.86	84.51	85.12	56.10	100.00	98.15
	Softmax-DPO	6.77	86.54	86.86	53.40	100.00	97.53
	DEHALL-SDPO	11.17	65.93	66.40	71.55	99.90	91.17

CV: outside the observed-union catalog; SV: outside the prompt slate. Unique is a diversity diagnostic, not a correctness measure.

E.6 AMAZON CD: DETAILED RESULTS

The saved validation split contains 16,000 prompts from 2,000 users. The first table reports the complete set of ranking metrics; the second separates generation utility, catalog violation, and slate violation by decoder.

Table 13: **Amazon CD: full ranking results.** Percentages (↑).

Method	H@1	H@3	H@5	H@10	N@3	N@5	N@10	MRR
MoRec	9.93	16.17	20.32	30.39	13.49	15.21	18.42	16.26
ChatRec	10.83	17.28	21.58	31.99	14.51	16.28	19.60	17.37
TALLRec	13.32	22.42	28.46	43.13	18.51	21.01	25.68	22.54
LLaRA	14.31	24.11	30.62	46.43	19.90	22.59	27.63	24.24
SPRec	17.36	29.73	37.96	57.92	24.42	27.81	34.18	29.90
R-CE	7.75	15.89	26.01	52.89	12.23	16.36	24.96	20.05
T-CE	8.93	18.97	27.75	56.68	14.67	18.26	27.51	21.96
GRU4Rec	4.19	11.70	21.11	60.79	8.45	12.27	24.91	17.31
Caser	12.02	38.74	62.37	82.13	27.14	36.89	43.39	32.55
SASRec	17.47	41.03	53.91	66.72	30.93	36.23	40.37	34.14
BERT4Rec	10.43	29.09	41.42	62.29	21.08	26.14	32.85	26.36
LightSANs	16.76	43.08	54.62	68.22	31.77	36.52	40.94	34.41
LPO4Rec	5.14	14.39	24.03	49.88	10.41	14.39	22.64	17.94
SFT	31.63	50.75	61.00	79.50	42.70	46.90	52.90	46.18
DPO	36.81	57.50	69.63	87.13	48.84	53.82	59.44	51.83
Softmax-DPO	38.00	58.75	71.44	87.00	50.09	55.30	60.30	52.98
DEHALL-SDPO	39.44	60.00	70.44	86.69	51.23	55.50	60.71	53.67

H@K and N@K denote HR@K and NDCG@K. Policy MRR spans the full slate; source-reference definitions are not assumed identical (Appendix E.1).

Table 14: **Amazon CD: free-generation results.** Percentages. Hit is mean per-output correctness; Any Hit/CV count prompts with at least one such output. Beam rows average all ten outputs.

Decoder	Method	Hit ↑	CV ↓	SV ↓	Any Hit ↑	Any CV ↓	Unique
Greedy	SFT	31.38	1.44	1.81	31.38	1.44	100.00
	DPO	10.38	85.00	85.06	10.38	85.00	100.00
	Softmax-DPO	22.19	59.19	60.38	22.19	59.19	100.00
	DEHALL-SDPO	38.69	1.56	1.81	38.69	1.56	100.00
Sample-10	SFT	22.66	1.93	2.17	59.50	10.44	53.48
	DPO	10.50	83.81	83.87	20.06	96.31	72.28
	Softmax-DPO	21.63	59.56	60.97	31.25	83.25	64.40
	DEHALL-SDPO	37.52	1.68	2.05	47.56	5.50	17.09
Beam-10	SFT	6.31	19.96	21.15	62.88	94.44	99.95
	DPO	4.07	89.48	89.75	34.56	100.00	99.01
	Softmax-DPO	5.35	83.66	85.50	41.44	100.00	97.74
	DEHALL-SDPO	9.29	54.69	55.67	63.56	99.56	93.51

CV: outside the observed-union catalog; SV: outside the prompt slate. Unique is a diversity diagnostic, not a correctness measure.

F Extended Mechanism and Historical Results

The following results expand the main-text comparisons without mixing training groups. Legacy endpoints, v2 module arms, the weight sweep, and the Amazon CD SFT-to-full-method comparison are reported separately with their settings intact. Rates are percentages, mass losses are in nats, and shaded rows identify DeHall-SDPO configurations rather than statistical superiority. DeHall abbreviates the method name in compact tables; legacy and component labels identify distinct configurations, not interchangeable results. The fixed-pad MovieLens re-evaluations below are retained as a separate historical analysis and are not substituted for the user-designated legacy records in Table 1.

Table 15: **Prefix and Dual Mass provide complementary support repair.** Matched MovieLens validation ablation, effective batch 16; P, M, and R denote Prefix, Dual Mass, and LPR ($\lambda_r = 0.1$). Rates are percentages; shaded rows are our variants. Complete decoding results and all eight arms appear in Tables 19–20.

	Policy ranking ↑			Support ↑	Catalog violation ↓	
Configuration	Acc.	NDCG@10	MRR	Mean Z_S	Greedy	Sample-10
S-DPO	65.33	82.26	77.23	52.40	44.56	46.40
DeHall (P)	65.71	82.39	77.43	97.53	1.74	1.90
DeHall ($P + M$)	65.26	82.10	77.15	99.15	0.60	0.67
DeHall ($P + M + R$)	64.43	81.79	76.67	98.50	0.68	0.91

F.1 Selected-adapter baselines and the prefix-only predecessor

These endpoints use the selected legacy adapters, not necessarily their terminal training checkpoints. Tables 17 and 18 give all ranking cutoffs and decoding protocols behind Table 16. DeHall's legacy prefix-only row is kept separate from the v2 component study because its run and checkpoint selection differ.

Table 16: **Ranking alone conceals the support failure.** MovieLens validation, legacy selected-adapter endpoints re-scored under one EOS-inclusive protocol. All values are percentages; CV counts every returned output, and blue shading identifies our prefix-only predecessor. These endpoints are separate from the final checkpoints in Figure 1; complete cutoffs and mass losses appear in Tables 17–18.

	Policy ranking ↑			Support ↑	Catalog violation ↓		
Method	Acc.	NDCG@10	MRR	Mean Z_S	Greedy	Sample-10	Beam-10
SFT	64.58	81.21	76.06	99.17	0.00	0.05	23.36
Continued SFT	63.82	81.44	76.20	98.99	0.00	0.06	22.79
DPO ($K = 1$)	62.39	80.58	75.23	24.32	75.15	75.45	93.03
S-DPO ($K = 3$)	64.95	81.86	76.75	23.78	76.36	76.41	92.53
DeHall (legacy P)	64.80	81.76	76.72	98.44	0.60	0.48	72.76

Table 17: MovieLens legacy selected-adapter endpoints re-scored with fixed-20-slot ranking. Rates are percentages; H and N denote HR and NDCG, and Accuracy equals HR@1/NDCG@1.

Configuration	Batch	Acc.	H@3	H@5	H@10	N@3	N@5	N@10	MRR
SFT	16	64.58	84.29	91.69	97.73	76.14	79.23	81.21	76.06
Continued SFT	16	63.82	86.56	92.15	98.04	77.19	79.49	81.44	76.20
DPO ($K = 1$)	16	62.39	85.57	91.62	97.58	76.08	78.58	80.58	75.23
S-DPO ($K = 3$)	16	64.95	85.73	91.84	98.11	77.30	79.82	81.86	76.75
Legacy Prefix	16	64.80	85.57	92.15	97.81	77.14	79.87	81.76	76.72

Table 18: MovieLens legacy selected-adapter support and decoding results. Slate mass and per-output catalog violations are percentages; negative-log mass is in nats.

Configuration	Batch	$\overline{Z_S}$	$\overline{-\log Z_S}$	Greedy	Beam-10	Sample-10
SFT	16	99.17	0.00840	0.00	23.36	0.05
Continued SFT	16	98.99	0.01031	0.00	22.79	0.06
DPO ($K = 1$)	16	24.32	8.74701	75.15	93.03	75.45
S-DPO ($K = 3$)	16	23.78	3.58264	76.36	92.53	76.41
Legacy Prefix	16	98.44	0.07350	0.60	72.76	0.48

F.2 All eight module arms

Tables 19 and 20 retain every available module arm, with P, M, and R denoting Prefix, Dual Mass, and LPR. The two no-Prefix mass arms use effective batch 16, whereas the four rows selected for the main matched comparison use batch 16. The batch column is therefore part of the comparison protocol, not a cosmetic annotation.

Table 19: MovieLens module-ablation ranking: P denotes Prefix, R denotes LPR (weight 0.1), and M denotes Dual Mass; additions are relative to S-DPO, and Batch is effective batch size. Rates are percentages; H and N denote HR and NDCG, and Accuracy equals HR@1/NDCG@1.

Configuration	Batch	Acc.	H@3	H@5	H@10	N@3	N@5	N@10	MRR
S-DPO	16	65.33	87.01	92.75	98.11	78.06	80.45	82.26	77.23
+R	16	62.76	86.86	93.28	97.96	76.98	79.63	81.19	75.84
+M	16	65.48	87.08	92.67	97.89	78.12	80.43	82.18	77.22
+R+M	16	63.44	85.95	92.67	98.04	76.66	79.42	81.21	75.88
+P	16	65.71	86.78	93.13	98.04	78.13	80.74	82.39	77.43
+P+R	16	63.44	86.86	93.20	97.89	77.27	79.87	81.44	76.22
+P+M	16	65.26	86.56	92.98	97.81	77.86	80.50	82.10	77.15
+P+R+M	16	64.43	86.71	92.82	97.96	77.56	80.09	81.79	76.67

Table 20: MovieLens ablation support and free-generation results, with module labels as in Table 19. Slate mass and per-output catalog violations are percentages; negative-log mass is in nats.

Configuration	Batch	$\overline{Z_S}$	$\overline{-\log Z_S}$	Greedy	Beam-10	Sample-10
S-DPO	16	52.40	2.97968	44.56	89.24	46.40
+R	16	93.34	0.13532	3.70	70.88	4.85
+M	16	99.07	0.04798	0.83	78.13	0.79
+R+M	16	98.08	0.03076	0.83	63.98	1.19
+P	16	97.53	0.18476	1.74	80.51	1.90
+P+R	16	95.79	0.07860	2.27	68.61	2.82
+P+M	16	99.15	0.03591	0.60	79.86	0.67
+P+R+M	16	98.50	0.02064	0.68	81.57	0.91

F.3 Ranking-weight sensitivity

The batch-16 sweep changes only the full method's ranking weight among the recorded configurations. Figure 4 and Tables 21–22 show that the best observed accuracy at 0.03 does not coincide with the largest mean slate mass.

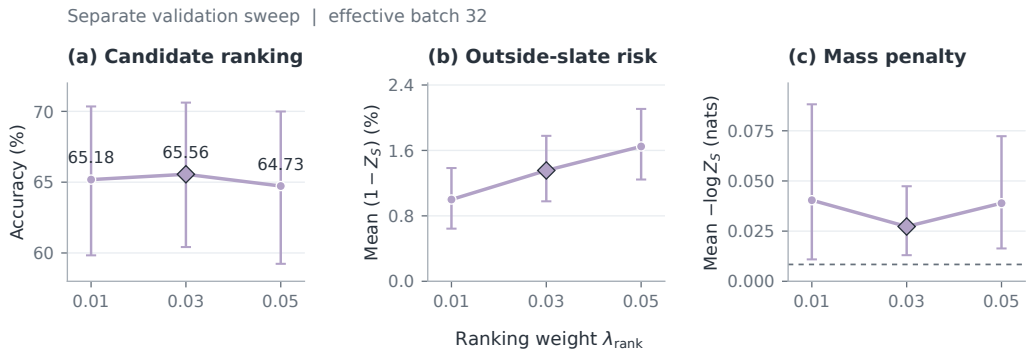

Figure 4: **Ranking-weight sensitivity.** Full-method MovieLens validation sweep, effective batch 16, seed 1958. Panels show policy accuracy, mean outside-slate probability, and mean negative-log slate mass; diamonds identify the validation-selected weight 0.03. Intervals are user-cluster bootstrap intervals, not evidence of significance after selection; outside-slate probability is not catalog violation.

Table 21: Separate MovieLens validation sweep of the full method (effective batch 16). Rates are percentages; H and N denote HR and NDCG, and Accuracy equals HR@1/NDCG@1.

Configuration	Batch	Acc.	H@3	H@5	H@10	N@3	N@5	N@10	MRR
$\lambda_{\text{rank}} = 0.01$	16	65.18	86.40	92.67	97.89	77.67	80.26	82.01	77.00
$\lambda_{\text{rank}} = 0.03$	16	65.56	86.48	93.05	98.19	77.86	80.55	82.25	77.22
$\lambda_{\text{rank}} = 0.05$	16	64.73	86.25	93.05	98.04	77.37	80.17	81.83	76.70

Table 22: Support and available final-adapter decoding results for the batch-16 LPR sweep. Slate mass and per-output catalog violations are percentages; negative-log mass is in nats.

Configuration	Batch	$\overline{Z_S}$	$\overline{-\log Z_S}$	Greedy	Beam-10	Sample-10
$\lambda_{\text{rank}} = 0.01$	16	99.00	0.04042	0.53	72.62	0.74
$\lambda_{\text{rank}} = 0.03$	16	98.65	0.02729	0.83	68.27	1.07
$\lambda_{\text{rank}} = 0.05$	16	98.35	0.03892	0.75	69.03	0.86

The 0.03 final adapter produces 11 catalog-invalid greedy outputs out of 1,324 and 142 invalid nucleus outputs out of 13,240, compared with 9,039 invalid strings among 13,240 beam returns. Its highest-scored beam has only 0.60% CV, illustrating the difference between a single delivered recommendation and the full search-expanded list. Mean negative-log slate mass is 0.02729 nats, above the 0.00840 SFT budget despite a mean slate mass of 98.65%; neither statistic is silently substituted for the other.

Within the matched sweep, 0.03 has higher NDCG@10 than 0.05 by 0.43 points (paired 95% interval $[0.07, 0.79]$), whereas its advantage over 0.01 is less certain (Table 27). Relative to 0.01, its conditional slate entropy increases from 0.105 to 0.223 nats and the count of incorrectly ranked prompts with maximum conditional item probability above 0.99 falls from 250 to 120. These are descriptive changes in within-slate concentration, not estimates of full-policy entropy or proof of improved probability calibration.

Table 23: MovieLens final-adapter generation utility and coverage. All values are percentages; Sample Mean is per-output correctness, Any is at least one correct output in ten, and Unique is distinct strings divided by ten. Beam-1 columns concern only the highest-scored returned beam.

	Greedy	Sample-10			Beam-1	
Configuration	Hit	Mean Hit	Any Hit	Unique	Hit	CV
SFT	64.35	57.47	82.85	24.22	57.48	0.30
Full, $\lambda_r = 0.01$	65.33	65.29	68.96	11.62	65.18	0.60
Full, $\lambda_r = 0.03$	65.56	64.16	72.43	13.68	65.03	0.60

F.4 Amazon CD: improved recommendation utility with low sampled violation

The Amazon CD comparison evaluates the full method and its own SFT initialization on the same 1,600 validation prompts from 200 users, with an identical catalog and scorer. The recorded SFT run uses one epoch and effective batch 16; the full method adds three preference epochs at batch 16 and $\lambda_r = 0.01$, rather than adopting MovieLens's 0.03 setting. The DeHall run retains its post-hoc archived/user-attested implementation binding; this is a single-seed observed-union study, not a frozen-test or independent-catalog benchmark.

Tables 24–25 show that higher utility extends across all measured ranking cutoffs while mean slate mass increases from 94.72% to 97.46%. HR@3, HR@5, and HR@10 improve by 9.25, 9.44, and 7.19 percentage points, and all paired ranking intervals in Table 27 lie above zero. These historical paired comparisons establish an advantage over the evaluated SFT initialization without attributing the gain to one auxiliary; the newly available CD DPO/S-DPO comparisons appear separately in Tables 13 and 14.

Table 24: Amazon CD ranking relative to the method's SFT initialization (1,600 validation prompts). The full method uses LPR weight 0.01; Batch records each training stage's effective batch. Rates are percentages; H and N denote HR and NDCG, and Accuracy equals HR@1/NDCG@1.

Configuration	Batch	Acc.	H@3	H@5	H@10	N@3	N@5	N@10	MRR
SFT	16	31.63	50.75	61.00	79.50	42.70	46.90	52.90	46.18
DeHall-SDPO	16	39.44	60.00	70.44	86.69	51.23	55.50	60.71	53.67

Table 25: Amazon CD support and decoding under the same evaluation protocol (full-method LPR weight 0.01). Slate mass and per-output catalog violations are percentages; negative-log mass is in nats. Beam CV counts all ten outputs rather than only the highest-scored beam.

Configuration	Batch	$\overline{Z_S}$	$\overline{-\log Z_S}$	Greedy	Beam-10	Sample-10
SFT	16	94.72	0.05825	1.44	19.96	1.93
DeHall-SDPO	16	97.46	0.09462	1.56	54.69	1.68

Table 26: Amazon CD generation utility and coverage, using the same definitions as Table 23. All entries are percentages; Sample Mean Hit is not a top-k ranking metric.

	Greedy	Sample-10			Beam-1	
Configuration	Hit	Mean Hit	Any Hit	Unique	Hit	CV
SFT	31.38	22.66	59.50	53.48	23.81	1.38
DeHall-SDPO	38.69	37.52	47.56	17.09	38.19	1.69

The free-generation gains concern the probability of delivering the correct item, not greater diversity of a ten-output list. Correct greedy outputs increase from 502 to 619 out of 1,600, and correct

nucleus outputs from 3,626 to 6,003 out of 16,000; however, sample any-hit falls from 59.50% to 47.56% as outputs become more concentrated. This distinction explains why higher mean correctness and lower set coverage can coexist, and keeps the single-output benefit separate from exploration quality.

Observed greedy and sample CV stay around 1–2%, but this is not a statistical equivalence or non-inferiority claim. Across all ten beams, CV increases from 19.96% to 54.69%, although highest-scored-beam correctness improves by 14.38 points; mean mass NLL also rises above the SFT-calibrated budget despite higher mean mass. The method therefore improves utility in the measured operating regime without establishing full-distribution validity, held-out budget feasibility, or a uniformly better recommendation list.

F.5 Paired uncertainty for utility and operating-point comparisons

Table 27 resamples users jointly across the compared models with 10,000 bootstrap replicates and random seed 1958. All prompts of each sampled user are retained and metrics are prompt-weighted; the 61 MovieLens and 200 CD users define the resampling units. The intervals quantify paired evaluation-sample uncertainty for one trained seed, not training variability, post-selection significance, or catalog-validity equivalence.

Table 27: **Paired utility gains and operating-point differences.** Differences are percentage points (first configuration minus second); intervals are 95% paired user-cluster bootstrap intervals, unadjusted for multiple comparisons. Positive utility differences are favorable; negative CV differences are favorable.

Metric	Difference	95% interval
MovieLens: full method ($\lambda_r = 0.03$) minus SFT		
Accuracy	+0.98	$[-0.46, 2.50]$
NDCG@10	+1.04	$[0.32, 1.80]$
MRR	+1.15	$[0.30, 2.05]$
Amazon CD: full method ($\lambda_r = 0.01$) minus SFT		
Accuracy	+7.81	$[5.94, 9.75]$
HR@3	+9.25	$[7.25, 11.31]$
HR@5	+9.44	$[7.19, 11.69]$
HR@10	+7.19	$[5.19, 9.19]$
NDCG@10	+7.81	$[6.57, 9.11]$
MRR	+7.50	$[6.14, 8.86]$
Mean slate mass	+2.75	$[2.31, 3.16]$
Greedy Hit	+7.31	$[5.31, 9.31]$
Sample-10 mean Hit	+14.86	$[13.15, 16.60]$
Beam-first Hit	+14.38	$[11.94, 16.81]$
Greedy CV	+0.13	$[-0.56, 0.81]$
Sample-10 CV	-0.25	$[-0.68, 0.19]$
MovieLens: full method, $\lambda_r = 0.03$ minus 0.01		
Accuracy	+0.38	$[-0.68, 1.37]$
NDCG@10	+0.24	$[-0.25, 0.74]$
Sample-10 mean Hit	-1.13	$[-1.89, -0.37]$
Sample-10 any Hit	+3.47	$[2.42, 4.57]$
Sample-10 CV	+0.33	$[0.02, 0.66]$
All-beam CV	-4.35	$[-5.03, -3.71]$

F.6 Paired evidence for support repair

Table 28 quantifies the cumulative contrasts used in the main component study. The analysis pairs all 1,324 prompts, verifies their user and protocol identifiers, and resamples the 61 users with replacement 10,000 times (seed 1958), retaining all prompts per sampled user. These exploratory

percentile intervals condition on one trained seed and are unadjusted for multiple comparisons; a ranking interval containing zero does not establish equivalence or noninferiority.

Table 28: **Paired support-repair effects.** MovieLens validation, effective batch 16. Differences are percentage points (first configuration minus second). Negative CV changes are favorable; all rows use the same paired user-cluster procedure.

Metric	Difference	95% interval
Prefix minus S-DPO		
Accuracy	+0.38	$[-0.75,\ 1.51]$
NDCG@10	+0.13	$[-0.40,\ 0.67]$
Greedy CV	−42.82	$[-46.47,\ -39.28]$
Sample-10 CV	−44.49	$[-47.54,\ -41.50]$
Prefix+Mass minus Prefix		
Accuracy	−0.45	$[-1.59,\ 0.68]$
NDCG@10	−0.28	$[-0.83,\ 0.27]$
Greedy CV	−1.13	$[-1.88,\ -0.46]$
Sample-10 CV	−1.23	$[-1.88,\ -0.64]$
Prefix+Mass minus S-DPO		
Accuracy	−0.08	$[-1.21,\ 1.06]$
NDCG@10	−0.15	$[-0.69,\ 0.39]$
Greedy CV	−43.96	$[-47.61,\ -40.39]$
Sample-10 CV	−45.73	$[-48.89,\ -42.64]$

F.7 Additional view of the matched component study

Figure 5 preserves the earlier multi-decoder view of the same batch-16 comparison used for the motivation figure. These are independent training runs with cumulative module configurations, not successive states of one model. The additional decoder traces test whether an apparent support benefit is shared by greedy, nucleus, and beam generation; they do not identify one decoder's frequency with full-policy ancestral risk.

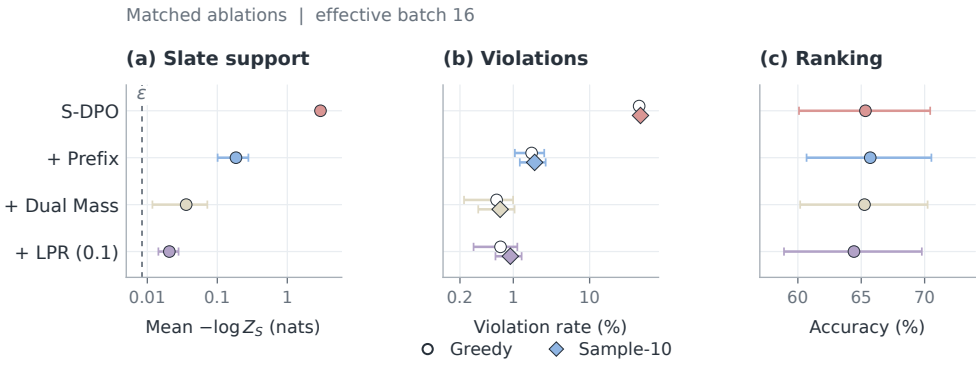

Figure 5: **Extended component diagnostics.** MovieLens validation, effective batch 16, observed-union catalog. Prefix, Dual Mass, and LPR ($\lambda_r = 0.1$) are added cumulatively across separate runs. The dashed mass budget is a target rather than achieved held-out feasibility; the intervals quantify user-cluster sampling uncertainty, not training-seed variability.

F.8 Training-log evidence and checkpoint trajectories

Figure 6 supplies optimization context for the research-problem figure. Chosen and rejected rewards can both decrease while their difference grows because rejected responses lose more probability,

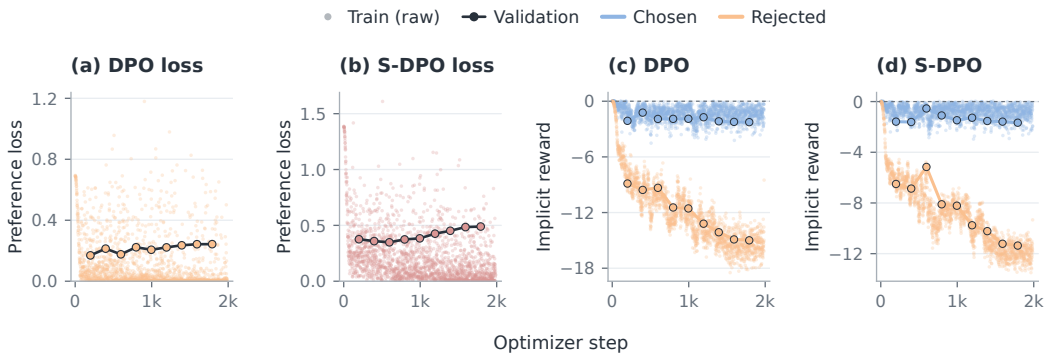

Figure 6: **Original optimization signals.** MovieLens, seed 1958. Raw training entries are not smoothed; lines connect logged validation evaluations. Preference losses use different negative counts ($K = 1$ versus $K = 3$), and rejected rewards summarize those different comparison sets. They must not be interpreted as a shared full-slate support measurement. Logged validation ends at step 1791; no terminal validation-log point is imputed.

and training and validation losses need not evolve identically. These teacher-forced signals motivate direct support measurement but do not themselves estimate free-generation catalog risk.

The additional full-method $\lambda_r = 0.03$ trajectory uses effective batch 16 and reaches three epochs at step 993, compared with step 1986 for the legacy batch-16 runs. Its ten post-initialization measurements have mean slate mass between 97.70% and 99.39% and greedy CV between 0.23% and 1.44%, consistent with stable support repair but not compute-matched superiority. The terminal step-993 adapter and the completed 0.03 final adapter have identical weight-file hashes, and their ranking and greedy results agree. Table 22 now uses the independently completed final-adapter evaluations; intermediate points retain their exploratory provenance rather than inheriting the endpoint's binding.

Table 29: All measured legacy checkpoint observations behind Figure 1. Rates are percentages. The reward margin is a shared full-slate diagnostic, not the method-specific training loss.

Method	Step	Reward margin	Accuracy	Mean Z_S	Greedy CV
DPO (K=1)	0	0.000	64.58	99.17	0.00
DPO (K=1)	199	1.117	63.14	24.35	75.15
DPO (K=1)	398	0.708	62.01	92.55	6.34
DPO (K=1)	597	1.306	64.58	12.37	89.50
DPO (K=1)	796	1.598	64.73	26.32	71.00
DPO (K=1)	995	1.700	65.11	67.79	30.06
DPO (K=1)	1194	1.621	64.73	82.17	13.90
DPO (K=1)	1393	1.885	65.11	70.34	26.96
DPO (K=1)	1592	1.931	64.88	58.65	37.99
DPO (K=1)	1791	1.948	64.80	62.85	33.01
DPO (K=1)	1986	1.917	64.88	67.52	29.23
S-DPO (K=3)	0	0.000	64.58	99.17	0.00
S-DPO (K=3)	199	0.915	63.22	11.28	88.60
S-DPO (K=3)	398	1.199	63.97	15.14	85.12
S-DPO (K=3)	597	0.800	64.95	71.41	24.92
S-DPO (K=3)	796	1.325	64.20	66.85	32.18
S-DPO (K=3)	995	1.622	66.24	62.96	35.20
S-DPO (K=3)	1194	1.551	66.47	81.24	16.54
S-DPO (K=3)	1393	1.807	66.24	74.64	22.13
S-DPO (K=3)	1592	1.904	66.01	78.55	18.58
S-DPO (K=3)	1791	1.965	65.94	75.07	22.05
S-DPO (K=3)	1986	1.972	66.09	74.14	23.41

Table 30: Exploratory full-method checkpoint trajectory, $\lambda_r = 0.03$, effective batch 16. Post-hoc intermediate-checkpoint provenance is retained; these rows are not formal final-adapter decoding results.

Step	Epoch	Accuracy (%)	Mean Z_S (%)	Greedy CV (%)
0	0.000	64.58	99.17	0.00
100	0.303	63.22	97.70	1.28
200	0.605	64.50	98.70	0.91
300	0.908	65.26	99.39	0.23
400	1.209	65.03	98.66	0.83
500	1.511	64.95	99.10	0.38
600	1.814	65.03	97.73	1.44
700	2.115	65.26	98.65	0.53
800	2.417	65.18	98.77	0.60
900	2.720	65.33	98.66	0.83
993	3.000	65.56	98.65	0.83

G Further Work

Future work will extend catalog-support control to evolving recommendation environments, where new items and changing user interests require continual adaptation. We will investigate context-dependent support budgets that balance recommendation utility and generation risk across users, rather than imposing a shared constraint. Another direction is to combine learned support with selective constrained decoding, activating inference-time safeguards only when estimated violation risk is high. Evaluating these extensions under temporal distribution shifts and independently maintained catalogs would clarify how sequence- and prefix-level control can support reliable recommendation beyond a fixed inventory.

H ENLARGED FRAMEWORK AND READING GUIDE

Figure 7 enlarges the author-drawn three-panel framework of DEHALL-SDPO, omitting only its superseded naming banner. Panels (a–c) show the support blind spot, three training signals, and separate ranking and generation evaluation; the distribution-harnessed design changes the policy, not the decoder. The diagram's L_{Prefix} is $\mathcal{L}_{\text{sup}}$ in Equation 14; Z_S denotes slate mass, not full-catalog mass.

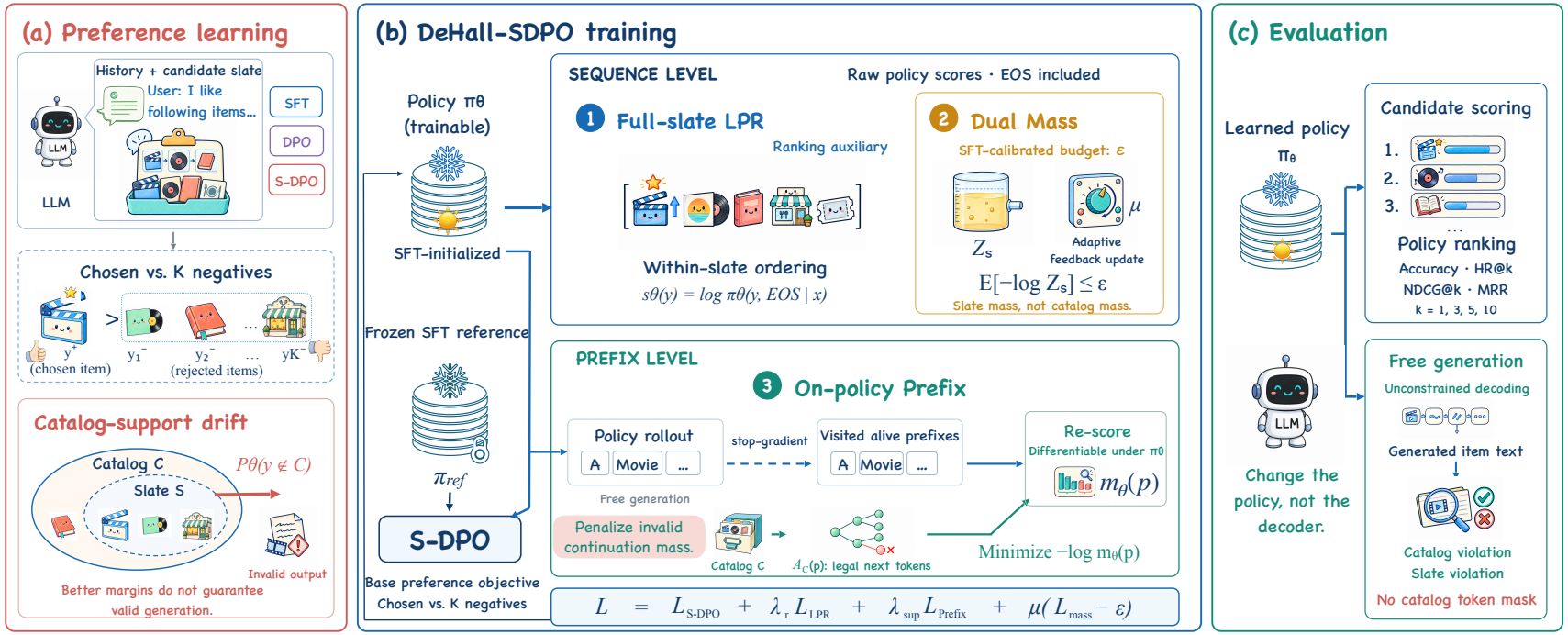

Figure 7: **Enlarged framework.** The SFT reference stays frozen, sampled prefixes are detached before re-scoring, and no catalog mask is applied during free-generation evaluation.