LEARNING MULTIRESOLUTION RELEVANCE FOR HIERARCHICAL GENERATIVE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative retrieval with semantic identifiers (SIDs) makes successive decisions over a document hierarchy. Relevant documents for the same query may share coarse prefixes and diverge at finer depths, with branching patterns varying across queries. These paths reveal how relevance is distributed across successive refinements, yet standard full-SID supervision treats them as separate training targets. To make this allocation explicit, we formulate multiresolution relevance as consistent conditional distributions induced by a single document-level relevance measure across the SID hierarchy. We introduce **RARS**, **R**esolution-**A**ligned **R**elevance **S**upervision, which uses the resulting refinement-level distributions to supervise a shared query representation. RARS aggregates document relevance over prefixes and trains a prefix-conditioned predictor to allocate relevance among sibling branches. All relevance-bearing children participate in local competition, and each local loss is weighted by the relevance mass reaching its parent. This objective trains the query encoder to capture both the coarse structure shared by relevant documents and their finer branch allocations. The predictor is discarded after training, preserving standard autoregressive retrieval at inference. Experiments on three multilingual ESCI locales show consistent improvements over matched full-SID training under autoregressive decoding. RARS also outperforms grouped soft-target, decoder soft-target, and sampled-tree supervision under a common retrieval rule. The gains persist across alternative identifier structures and relevance definitions. Code is available at: https://anonymous.4open.science/r/RARS-A868.

1 INTRODUCTION

Generative retrieval casts retrieval as sequence generation over document identifiers (Tay et al., 2022; Wang et al., 2022; Bevilacqua et al., 2022). Semantic identifiers (SIDs) further organize documents into structured codes whose prefixes capture progressively finer distinctions (Wang et al., 2022; Sun et al., 2023; Rajput et al., 2023; Zhang et al., 2025a; Liu et al., 2026). A shallow prefix identifies a coarse region of the corpus, and subsequent tokens progressively refine that region toward an individual document. SID-based retrieval is therefore inherently multiresolution: each generation step resolves the query over a finer corpus partition, and an incorrect prefix can exclude an entire relevant subtree from constrained search (Wang et al., 2022; Zeng et al., 2024). Because the complete SID also serves as the retrieval address, full-SID supervision is a natural training target.

This supervision is limited for queries with multiple relevant documents. Relevant documents may share a prefix and diverge only at a later SID depth, so a single query can support several competing child branches under the same parent. Figure 1 illustrates this case with the query "Qi-compatible power bank." The query specifies charging compatibility while leaving capacity unspecified, allowing several Exact products to share coarse SID structure and separate at finer retrieval resolutions. At one depth, relevance may be concentrated on a single branch; at a subsequent depth, the same relevance mass may divide among several children. Standard full-SID training presents these products as separate root-to-leaf targets (Tang et al., 2024a; Li et al., 2024a; Tang et al., 2024b), so the same query is associated with different child decisions at a shared parent. At the query level, these paths induce a relevance distribution over competing refinements, but this distribution is not represented explicitly during training. Full-path supervision therefore leaves the local relevance structure of intermediate SID decisions implicit. We refer to this granularity gap as a supervision–resolution mismatch.

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

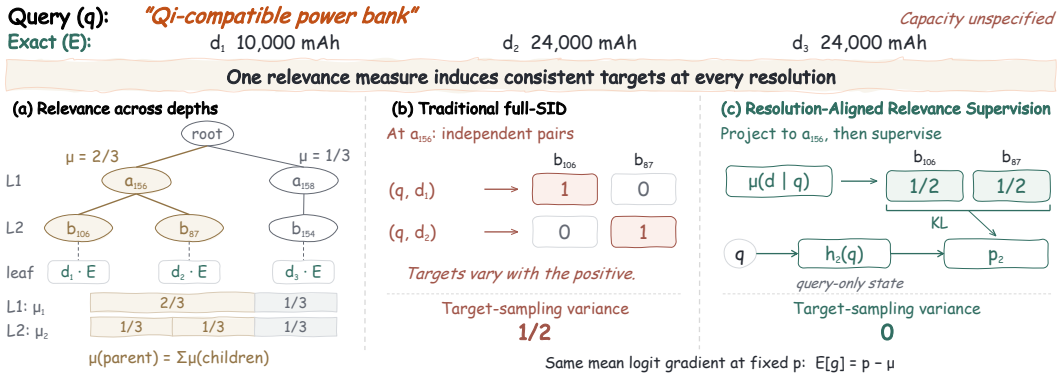


Figure 1: **Relevance across SID resolutions.** (a) Aggregating document relevance over prefixes preserves total mass. (b) Full-SID supervision presents each positive as a separate path. (c) RARS learns the conditional allocation $[1/2, 1/2]$ at the shared prefix from a shared query representation.

Recent work has incorporated ranking objectives, prefix-level preferences, graded relevance, relevance-aware identifiers, and multi-positive supervision (Zhou et al., 2023; Zhang et al., 2025b; Zeng et al., 2024; Zhang et al., 2026b). These directions establish the value of relevance-aware learning in generative retrieval, but leave open how document-level relevance should be represented when retrieval proceeds through nested corpus partitions and local branch decisions. We address this question with a multiresolution view of relevance. For each query, document-level relevance defines a distribution over relevant leaves of the SID hierarchy. Projecting this distribution onto each depth yields a relevance measure over the corresponding corpus partition. Because the partitions are nested, parent mass equals the total mass of its children, and conditioning child mass on the parent gives a local relevance distribution for each refinement. The targets at different depths are therefore consistent views of the same query relevance at progressively finer hierarchical retrieval resolutions.

To learn these multiresolution relevance targets, we introduce RARS, Resolution-Aligned Relevance Supervision. RARS uses a shared query representation to predict the relevance allocation at each supported refinement through depth-specific projections and prefix-conditioned child scores. Each local target retains all relevance-bearing branches, including those that diverge only at finer SID depths, thereby preserving the query's multiresolution relevance structure. The loss weights each refinement by the relevance mass reaching its parent and includes competing legal siblings in local normalization. This design couples supervision across the query's relevant paths through a single encoder representation, requiring the representation to encode branch allocations across successive resolutions. The multiresolution objective is optimized alongside the full-SID objective. The hierarchical predictor is discarded after training, preserving standard autoregressive retrieval.

We evaluate RARS on the English, Spanish, and Japanese ESCI product-search benchmarks (Reddy et al., 2022). Controlled experiments separate the effects of distributional supervision, hierarchical prediction, and retrieval scoring. RARS improves over matched full-SID training under standard autoregressive decoding across all three locales and remains stronger than grouped soft-target, decoder soft-target, and sampled-tree controls. The gains persist across alternative SID constructions and relevance definitions. Further analyses show that relevance branching is common at coarse and intermediate SID depths, while incomplete judgments can obscure this structure.

Our contributions are threefold:

(i) We formulate multiresolution relevance for generative retrieval. A single document-level relevance measure induces consistent conditional targets across SID refinements, and a refinement-entropy profile identifies where observed relevance supports competing branches.

(ii) We propose RARS to learn branch allocations through a shared query representation. Prefix-conditioned prediction and parent-mass weighting supervise relevance-bearing refinements, while local sibling competition encourages the encoder to distinguish competing child branches. The predictor is used only during training, without modifying autoregressive retrieval at inference.

(iii) Controlled experiments on three multilingual ESCI locales establish gains over grouped soft-target, decoder soft-target, and sampled-tree supervision. Further comparisons verify improvements under autoregressive retrieval and across alternative SID constructions and relevance definitions.

2 Related Work

Generative retrieval and structured identifiers. Generative retrieval maps queries directly to document or entity identifiers (De Cao et al., 2021; Tay et al., 2022). NCI employs hierarchical identifiers with prefix-aware decoding, while SEAL generates document substrings (Wang et al., 2022; Bevilacqua et al., 2022). Learned tokenization, semantic trees, and residual quantization provide alternative constructions of structured identifiers with progressively finer semantic organization (Sun et al., 2023; Si et al., 2024; Rajput et al., 2023; Jin et al., 2024). For product search, Hi-Gen, MERGE, and CAT-ID^2 incorporate category structure or query relevance into identifier construction (Wu et al., 2024; Zhang et al., 2025b; Liu et al., 2026). CaLIR further introduces category-guided latent reasoning and query-conditioned tries (Zhang et al., 2026b). We build on CaLIR and focus on a complementary question: how document-level relevance should directly supervise local decisions across successive refinements of the SID hierarchy at multiple retrieval resolutions.

Relevance-aware supervision. Listwise ranking methods model relative relevance across documents: ListNet matches score-induced distributions, while ListMLE models ranking likelihood (Cao et al., 2007; Xia et al., 2008). In generative retrieval, LTRGR learns from document rankings, RIPOR from pairwise preferences over prefixes, and GR^2 from graded relevance (Li et al., 2024a; Zeng et al., 2024; Tang et al., 2024a). Relevance feedback and rank distillation align identifier generation with retrieval quality (Zhou et al., 2023; Li et al., 2024b). Multi-positive objectives supervise representations using multiple relevant items (Khosla et al., 2020). In particular, CaLIR averages over positive category prototypes at each level (Zhang et al., 2026b). RARS aggregates document relevance over SID prefixes to characterize how a query's relevance is distributed at each successive refinement. The projection preserves relevance mass across depths, while the conditional targets retain the relative relevance among child branches. RARS learns these allocations through a shared query representation across relevant paths, using prefix-conditioned prediction and parent-mass weighting.

Hierarchical prediction. Label trees factorize predictions along root-to-leaf paths (Bengio et al., 2010; Wydmuch et al., 2018), while hierarchical classifiers incorporate this structure into probabilities and losses (Ramaswamy et al., 2015; Bertinetto et al., 2020). Selective hierarchical classification can return internal nodes when predictions are uncertain (Goren et al., 2024). RARS uses the SID hierarchy to define query-conditioned relevance targets for refinement decisions. Its predictor supervises the query encoder across relevant branches during training, while retrieval retains the full-SID decoder. The refinement-entropy profile characterizes how observed relevance is distributed across the hierarchy. Confidence calibration concerns the reliability of predictions (Guo et al., 2017).

3 Preliminaries

Semantic identifiers. Let $\mathcal{D}$ denote a document corpus, and let $s(d) = (s_1(d), \ldots, s_L(d))$ be the length-L semantic identifier (SID) of document $d \in \mathcal{D}$ (Wang et al., 2022; Sun et al., 2023; Rajput et al., 2023). We denote its depth-ℓ prefix by $\pi_\ell(d) = (s_1(d), \ldots, s_\ell(d))$, with $\pi_0(d) = \varnothing$. Let $\mathcal{P}_\ell = \{\pi_\ell(d) : d \in \mathcal{D}\}$ be the set of valid prefixes at depth ℓ, with $\mathcal{P}_0 = \{\varnothing\}$, and let $\mathrm{Ch}(u)$ denote the valid children of $u \in \mathcal{P}_{\ell-1}$. The resulting SID trie induces a sequence of nested corpus partitions: shallow prefixes represent coarse document groups, while deeper prefixes progressively refine them toward individual documents.

Autoregressive generative retrieval. Given a query q, a generative retriever models a document SID as a sequence of refinement decisions and factorizes its probability along the SID path as

$$p_\theta(s(d) \mid q) = \prod_{\ell=1}^{L} p_\theta(s_\ell(d) \mid q, s_{<\ell}(d)), \tag{1}$$

where $s_{<\ell}(d) = (s_1(d), \ldots, s_{\ell-1}(d))$ denotes the previously generated prefix. Each factor in Eq. (1) corresponds to a local refinement in the SID trie: conditioned on the query and current prefix, the model selects among valid child branches. Under trie-constrained decoding, invalid continuations are excluded, so each prefix decision determines which subset of documents remains reachable.

Full-SID supervision. Let $\mathcal{R}$ denote the set of relevant query–document pairs. Standard full-SID supervision treats each pair as a target path and minimizes the maximum-likelihood objective

$$\mathcal{L}_{\text{leaf}} = -\mathbb{E}_{(q,d)\sim\mathcal{R}} \log p_\theta(s(d) \mid q). \tag{2}$$

Each relevant pair contributes a deterministic root-to-leaf path. For a query with multiple relevant documents, these paths may share a prefix and therefore a common coarse decision, then diverge at a later refinement, thereby specifying different child branches below the same parent. The next section develops a query-conditioned relevance distribution over these refinements.

4 RARS: Relevance Across Retrieval Resolutions

RARS trains a query representation to predict how relevance is allocated across successive SID refinements (Figure 2). Document relevance defines the target mass on the leaves, and the SID partitions determine its allocation to prefixes and their children. A prefix-conditioned predictor learns these local allocations from the shared query encoder. Parent-mass weighting determines the contribution of each refinement to training, and local sibling competition supplies the candidate alternatives. The full-SID decoder retains its original objective and retrieval procedure.

4.1 Relevance Across Identifier Resolutions

For a query q, let $\mathcal{D}_q^+$ denote the set of observed relevant documents, and let $\rho(q, d) \geq 0$ denote the relevance evidence associated with $d \in \mathcal{D}_q^+$. We define the normalized document-level relevance measure as $\mu(d \mid q) = \rho(q, d) / \sum_{d' \in \mathcal{D}_q^+} \rho(q, d')$. Binary relevance assigns equal mass to all observed positives, while graded judgments or continuous relevance scores induce non-uniform masses.

The SID hierarchy induces a corpus partition at each depth. At depth ℓ, we obtain the relevance mass of a prefix by aggregating the masses of its relevant descendant documents:

$$\mu_\ell(z \mid q) = \sum_{d \in \mathcal{D}_q^+} \mu(d \mid q)\, \mathbf{1}[\pi_\ell(d) = z], \qquad z \in \mathcal{P}_\ell. \tag{3}$$

Hence, μ_ℓ gives the relevance distribution induced by the same document-level measure at SID resolution ℓ. Since SID partitions are nested, the projected distributions are consistent across depths. For each parent prefix $u \in \mathcal{P}_{\ell-1}$, its relevance mass equals the total mass of its children. When $\mu_{\ell-1}(u \mid q) > 0$, the normalized child masses define the conditional distribution $\mu_\ell(v \mid u, q) = \mu_\ell(v \mid q) / \mu_{\ell-1}(u \mid q)$. This distribution specifies the allocation of relevance among the child branches of u. Consequently, the supervision at different depths is induced by a common document-level relevance measure and remains consistent with the hierarchical structure.

We quantify how relevance is distributed across competing refinements. Let $D \sim \mu(\cdot \mid q)$ and $Z_\ell = \pi_\ell(D)$ denote the SID prefix of D at depth ℓ. The refinement ambiguity at depth ℓ is defined as

$$\Delta_\ell(q) = H_\mu(Z_\ell \mid Z_{\ell-1}, q) = \sum_{u \in \mathcal{P}_{\ell-1}} \mu_{\ell-1}(u \mid q)\, H(\mu_\ell(\cdot \mid u, q))\,. \tag{4}$$

The quantity $\Delta_\ell(q)$ is zero when every relevance-bearing parent assigns all of its mass to a single child, and increases as relevance is distributed over multiple sibling branches. It is determined solely by the observed relevance evidence and the SID hierarchy, and does not depend on model predictions. The profile $\{\Delta_\ell(q)\}_{\ell=1}^{L}$ therefore characterizes the SID depths at which observed relevance remains distributed across competing refinements.

4.2 Query-Level Hierarchical Prediction

The projected targets specify what should be predicted at each refinement. We parameterize these predictions using a representation shared across relevant documents of the query. Let $\bar{h}(q)$ be the masked mean of the query encoder's final token representations. A depth-specific projection produces $h_\ell(q) = W_\ell \bar{h}(q) \in \mathbb{R}^r$. For a parent prefix $u = (u_1, \ldots, u_{\ell-1})$, we encode its path as $c(u) = \sum_{j<\ell} b_{j,u_j}$ using position-specific prefix embeddings, with $c(\varnothing) = 0$. The score of child v is

$$a_\ell(v \mid u, q) = \frac{e_{\ell,v}^\top \left[h_\ell(q) \odot (1 + \tanh c(u))\right]}{\sqrt{r}} + \beta_{\ell,v}. \tag{5}$$

Conditioning on $c(u)$ allows the same token value to receive different scores under different prefixes, while $h_\ell(q)$ preserves a common query-level state across the relevant paths.

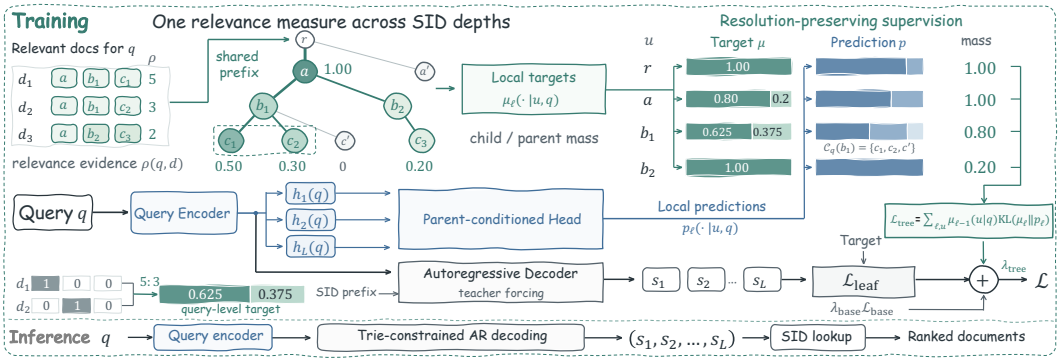


Figure 2: **RARS overview.** Document relevance is projected onto SID partitions to form conditional targets at each refinement. A query-level predictor learns these distributions with parent-mass-weighted supervision, while the full-SID decoder is retained and used alone at inference.

The local predictive distribution is obtained by normalizing over a candidate set $\mathcal{C}_q(u)$:

$$p_\ell(v \mid u, q) = \frac{\exp(a_\ell(v \mid u, q)/\tau_p)}{\sum_{c \in \mathcal{C}_q(u)} \exp(a_\ell(c \mid u, q)/\tau_p)}. \tag{6}$$

The candidate set at each relevance-bearing parent includes every relevance-bearing child and competing legal siblings. This local normalization maintains competition among children of the same parent without requiring a softmax over the full identifier space. We use $r = 64$ and $\tau_p = 1$, with up to 32 high-scoring legal sibling negatives included alongside all relevance-bearing children.

4.3 Resolution-Preserving Objective

We train the hierarchical predictor to match the projected relevance distribution at every relevance-bearing refinement, aligning predictions with the relevance structure induced by the SID hierarchy:

$$\mathcal{L}_{\text{tree}}(q) = \sum_{\ell=1}^{L} \sum_{u \in \mathcal{P}_{\ell-1}} \mu_{\ell-1}(u \mid q)\, D_{\text{KL}}(\mu_\ell(\cdot \mid u, q) \,\|\, p_\ell(\cdot \mid u, q)). \tag{7}$$

The parent mass $\mu_{\ell-1}(u \mid q)$ measures how much query relevance reaches the corresponding refinement. It weights each local loss by the relevance carried through that parent, preserving the contribution of each branch to the query-level objective. All relevance-bearing children remain in $\mathcal{C}_q(u)$, so each local loss supervises the allocation among supported branches in the presence of selected legal sibling negatives. The shared query representation receives these constraints from every relevance-bearing parent. Appendix B.5 derives the full-path consistency of the weighting under full-child normalization and its relation to grouped supervision. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{leaf}} + \lambda_{\text{tree}}\mathcal{L}_{\text{tree}} + \lambda_{\text{base}}\mathcal{L}_{\text{base}}. \tag{8}$$

Our reference retriever is CaLIR (Zhang et al., 2026b), which combines full-SID likelihood with category classification and multi-positive category alignment. We retain its original auxiliary objective $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{con}}$ with $\lambda_{\text{base}} = 0.1$, and set $\lambda_{\text{tree}} = 0.5$. The matched base optimizes $\mathcal{L}_{\text{leaf}} + 0.1(\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{con}})$. RARS adds $\mathcal{L}_{\text{tree}}$ to supervise the query encoder at the local refinement decisions. Gradients from $\mathcal{L}_{\text{tree}}$ update the hierarchical predictor and the shared query encoder. The autoregressive decoder and the remaining retriever parameters continue to be optimized by their original objectives. Since the hierarchical predictor is absent at inference, its effect on retrieval is mediated entirely through the representations learned during training.

4.4 Supervision at Branching Prefixes

For a query with a single relevant document $d^\star$ and $\mu(d^\star \mid q) = 1$, $\mu_\ell(z \mid q) = \mathbf{1}[z = \pi_\ell(d^\star)]$ for every depth ℓ. All local targets therefore collapse to the same deterministic root-to-leaf path. For queries with multiple relevant documents, relevance can branch below a shared prefix. Consider one such parent, with target distribution $\boldsymbol{\mu}$ over its children and model prediction $\boldsymbol{p}$. Sampling $Y \sim \boldsymbol{\mu}$

produces the one-hot logit gradient $\boldsymbol{g}_Y = \boldsymbol{p} - \boldsymbol{e}_Y$, whereas direct distributional supervision produces $\boldsymbol{g}_\mu = \boldsymbol{p} - \boldsymbol{\mu}$. For the same temperature-scaled local logits,

$$\mathbb{E}_Y[\boldsymbol{g}_Y] = \boldsymbol{g}_\mu, \qquad \mathbb{E}_Y\left[\|\boldsymbol{g}_Y - \boldsymbol{g}_\mu\|_2^2\right] = 1 - \|\boldsymbol{\mu}\|_2^2. \tag{9}$$

Relevance-weighted target sampling and explicit distributional supervision have the same expected local gradient. The latter removes variance from drawing a target, while leaving other sources of optimization stochasticity unchanged. The variance term in Eq. (9) is the Gini impurity, whereas Eq. (4) uses Shannon entropy. With natural logarithms, their parent-weighted quantities satisfy

$$\mathcal{V}_\ell(q) := \sum_u \mu_{\ell-1}(u \mid q)\left[1 - \|\mu_\ell(\cdot \mid u, q)\|_2^2\right] \le \Delta_\ell(q). \tag{10}$$

Both quantities vanish for deterministic refinements, but they capture different properties: $\Delta_\ell(q)$ characterizes the dispersion of relevance across siblings, whereas $\mathcal{V}_\ell(q)$ measures the variance introduced by sampling a one-hot target. At a branching prefix, the explicit target supplies a gradient for the entire observed relevance allocation in each update. Through Eq. (5), these local gradients update the same query representation across parents and depths. RARS thus couples the learning of shared coarse prefixes with the learning of their finer relevant branches.

4.5 Inference and Retrieval Scoring

The hierarchical predictor is not used during retrieval. Let $\mathcal{S}_C(q)$ denote the set of identifiers admitted by the top-C category predictions of the retriever. Category probabilities determine this candidate set but do not enter the identifier score. Writing $s_{L+1} = \mathrm{EOS}$, the autoregressive score is

$$S_{\mathrm{AR}}(s \mid q) = \sum_{t=1}^{L+1} \log p_\theta(s_t \mid q, s_{<t}), \qquad s \in \mathcal{S}_C(q). \tag{11}$$

Trie constraints assign score $-\infty$ to illegal continuations. Equation (11) is the retrieval rule used for the matched AR-only comparison. We evaluate a compatibility-augmented scoring rule. Let $f(q, d)$ measure query–catalog compatibility using catalog attributes and lexical evidence, without relevance judgments. For a prefix u at depth t, define its subtree compatibility as $F_t(u, q) = \max_{d:\pi_t(d)=u} f(q, d)$. With a depth mask $m \in \{0, 1\}^L$, the beam score is

$$S_{\mathrm{LT}}(s \mid q) = S_{\mathrm{AR}}(s \mid q) + \lambda_{\mathrm{comp}} \sum_{t=1}^{L} m_t F_t(s_{\le t}, q), \qquad s \in \mathcal{S}_C(q). \tag{12}$$

Compatibility is added only to legal SID continuations; EOS receives no compatibility term. Matched comparisons under these rules distinguish training gains from compatibility-based decoding.

5 Experiments

We study seven research questions: **RQ1**: How does RARS compare with baselines, and how do training and retrieval scoring contribute? **RQ2**: How does RARS compare with matched training controls under a common retrieval rule? **RQ3**: How is relevance distributed across SID resolutions, and how does incomplete evidence affect its fidelity? **RQ4**: Are the gains robust to identifier geometry, training evidence, and objective design? **RQ5**: Does multiresolution scoring improve local prediction alignment and prefix coverage in final rankings? **RQ6**: Which retrieval resolutions improve effectiveness? **RQ7**: How can we intuitively understand the advantage of RARS?

5.1 Experimental Setup

(1) Datasets. We use the category-filtered US (English), ES (Spanish), and JP (Japanese) subsets of the ESCI product-search benchmark (Reddy et al., 2022). Query–product pairs are labeled Exact, Substitute, Complement, or Irrelevant. The first three labels receive weights 3, 2, and 1, normalized per query. Evaluation retains the original graded judgments. **(2) Baselines.** Published baselines include sparse retrieval (BM25), dense retrieval (DPR, MPNet, Sentence-T5, Sentence-mT5, BGE-M3, and LaSER), and generative retrieval ($\mathrm{DSI}_{\mathrm{naive}}$, $\mathrm{DSI}_{\mathrm{semantic}}$, TIGER, Hi-Gen, LTRGR, RIPOR,

Model	ESCI-US					ESCI-ES					ESCI-JP				
	R@5	R@10	R@100	N@10	N@100	R@5	R@10	R@100	N@10	N@100	R@5	R@10	R@100	N@10	N@100
Sparse retrieval															
BM25 (FnT IR'09)	4.02	5.84	14.08	5.64	8.11	4.13	6.15	16.21	7.45	10.20	4.70	7.96	16.73	8.63	11.25
Dense retrieval															
DPR (EMNLP'20)	5.54	8.93	29.30	8.17	15.55	5.24	7.08	25.27	8.52	14.33	4.18	7.26	23.84	8.98	14.73
MPNet (NeurIPS'20)	2.76	4.58	15.30	4.12	9.83	2.53	4.02	13.27	5.87	8.91	1.13	1.97	5.58	1.79	2.82
Sentence-T5 (ACL'22)	4.63	6.97	24.59	6.84	11.58	–	–	–	–	–	–	–	–	–	–
Sentence-mT5 (COLING'24)	–	–	–	–	–	5.44	8.68	30.04	8.42	15.09	5.69	8.73	29.01	9.19	15.70
BGE-M3 (ACL'24)	6.59	9.82	32.29	9.91	16.77	5.71	9.07	28.75	9.76	16.02	5.14	8.92	27.93	10.12	17.02
LaSER (SIGIR'26)	6.52	10.23	35.37	8.19	15.74	5.28	8.40	32.22	8.69	16.16	5.88	8.78	28.66	9.28	15.53
Generative retrieval															
$\text{DSI}_{\text{naive}}$ (NeurIPS'22)	0.42	1.74	2.03	0.32	0.99	0.28	0.94	1.83	0.77	1.60	0.19	0.24	1.35	0.21	1.11
$\text{DSI}_{\text{semantic}}$ (NeurIPS'22)	3.74	6.02	20.69	6.24	10.60	4.27	7.13	21.63	9.92	14.23	4.08	7.40	23.90	9.15	14.58
TIGER (NeurIPS'23)	4.79	7.84	25.98	7.24	12.68	3.83	7.31	25.03	8.72	14.51	4.67	7.98	25.89	9.67	15.52
Hi-Gen (ICDM'24)	3.13	4.97	15.91	5.25	8.55	2.97	5.65	20.23	7.06	11.85	3.36	6.40	21.66	7.73	12.86
LTRGR (AAAI'24)	2.88	4.71	13.70	4.48	7.82	5.13	8.23	27.92	9.26	15.42	5.24	8.37	26.79	10.01	16.93
RIPOR (WWW'24)	5.05	8.35	26.80	7.92	13.75	3.49	7.12	23.03	7.94	13.82	5.02	8.18	27.87	9.92	17.28
MERGE (ACL'25)	5.68	9.26	29.74	9.05	15.17	5.80	9.25	30.86	10.20	17.45	5.21	8.90	27.64	10.95	17.60
CaLIR (arXiv'26)	7.22	11.75	36.15	10.75	18.14	6.06	10.43	34.83	12.15	20.03	5.57	9.64	31.71	11.74	18.85
CAT-ID^2 (WSDM'26)	6.11	8.89	29.03	8.96	14.53	5.70	9.34	31.44	10.68	18.01	5.36	8.87	28.82	10.34	17.14
FORGE (KDD'26)	5.59	9.26	35.11	8.50	16.21	4.87	8.29	28.42	9.99	16.66	3.88	6.58	24.06	8.27	14.22
HGRec (ICML'26)	6.70	11.41	34.61	10.19	17.32	5.76	9.60	31.55	11.37	18.32	3.78	6.18	24.37	8.40	13.83
Controlled comparisons: training objective / retrieval rule															
Base / AR only	7.50	12.69	37.40	11.12	18.79	6.03	10.48	34.15	12.43	19.88	6.05	9.77	30.89	11.77	18.42
RARS / AR only	8.57	13.39	38.98	12.36	20.11	6.40	11.38	34.86	13.65	20.84	6.51	10.29	31.59	12.34	19.44
Base / all levels	8.05	13.63	38.24	12.00	19.38	6.47	10.93	35.00	12.91	20.69	6.29	10.42	31.16	12.69	18.99
RARS / all levels	**9.29**	**14.36**	**39.62**	**13.00**	**20.82**	**6.87**	**11.80**	**35.55**	**14.19**	**21.53**	**6.71**	**10.89**	**32.02**	**13.23**	**19.80**

Table 1: Retrieval results on multilingual ESCI (%). R@K and N@K denote Recall@K and NDCG@K. Controlled comparisons use matched backbones and report means across five seeds.

CAT-ID^2, MERGE, FORGE, HGRec, and CaLIR). Controlled comparisons evaluate matched Base and RARS models under AR-only and all-level retrieval, fixing the backbone, data, SIDs, and decoding settings. **(3) Implementation Details.** Following CaLIR (Zhang et al., 2026b), we use T5-base for US and mT5-base for ES and JP. Default SIDs comprise four residual quantization levels with 256 codewords each. We train matched models for 300 epochs using AdamW with a learning rate of 5×10^{-4} and a per-device batch size of 512. We report five-seed means from final checkpoints. Retrieval uses trie-constrained beam search with beam size 100 and $C = 3$. All-level scoring uses $\lambda_{\text{comp}} = 2$, $K_{\text{lex}} = 2$, $\eta_{\text{lex}} = 0.25$, and $m = (1, 1, 1, 1)$. We report Recall@K ($K = 5, 10, 100$) and NDCG@K ($K = 10, 100$), with Recall@100 and NDCG@10 as the primary metrics. Training settings are fixed beforehand, and retrieval hyperparameters are selected on validation data and fixed before testing. Subsequent sweeps assess sensitivity without revising these settings. Dataset statistics, baseline protocols, and implementation and statistical details appear in Appendices A, C.2, and C.3.

5.2 Experimental Results and Discussion

Retrieval Effectiveness (RQ1). Table 1 compares RARS with sparse, dense, and generative retrieval methods, including matched variants that distinguish the contributions of training and retrieval scoring. AR-only retrieval follows Eq. (11); all-level retrieval adds the prefix compatibility terms in Eq. (12). Under AR-only retrieval, RARS improves Recall@100 and NDCG@10 over the matched base in each locale. Since the hierarchical predictor is discarded at inference, these gains support the benefit of multiresolution relevance supervision under standard autoregressive decoding. All-level scoring further improves both training variants while preserving this advantage. The proposed objective improves autoregressive retrieval, while compatibility scoring yields additional gains under the augmented retrieval rule. These improvements remain consistent across locales and primary metrics. The resulting configuration also compares favorably with sparse, dense, and generative baselines. The matched comparisons use a common training and evaluation protocol with the backbone and decoding settings fixed, allowing the contribution of the proposed objective to be assessed directly.

Contributions of the Training Objective (RQ2). Table 2 compares five training objectives under the same all-level retrieval rule. This isolates differences in training supervision. The matched base uses full-SID supervision. Grouped soft targets weight complete identifiers by relevance. AR-decoder soft targets apply projected conditional distributions directly to the decoder. Sampled tree and RARS share the hierarchical predictor in Eq. (5), using sampled one-hot and explicit conditional targets, respectively. These comparisons separate the effects of target aggregation, prediction site, and target stochasticity. The predictor is used only during training and discarded at inference. The AR-decoder control shows that projected local distributions are beneficial even when applied directly to the decoder. The sampled-tree comparison isolates the benefit of explicit conditional supervision within the same predictor, with distributional targets consistently outperforming sampled one-hot targets. Improvements over grouped soft targets indicate that supervising relevance allocation at SID

Training objective	ESCI-US					ESCI-ES					ESCI-JP				
	R@5	R@10	R@100	N@10	N@100	R@5	R@10	R@100	N@10	N@100	R@5	R@10	R@100	N@10	N@100
Matched base	8.05	13.63	38.24	12.00	19.38	6.47	10.93	35.00	12.91	20.69	6.29	10.42	31.16	12.69	18.99
AR-decoder soft target	8.74	13.93	39.04	12.46	20.36	6.68	11.33	35.17	13.77	21.06	6.57	10.60	31.70	12.88	19.42
Grouped soft-target	8.61	13.69	38.88	12.28	20.03	6.55	11.27	35.08	13.48	20.91	6.46	10.43	31.61	12.71	19.26
Sampled tree	8.89	14.01	39.18	12.55	20.29	6.63	11.51	35.28	13.69	21.19	6.52	10.55	31.81	12.79	19.55
RARS	**9.29**	**14.36**	**39.62**	**13.00**	**20.82**	**6.87**	**11.80**	**35.55**	**14.19**	**21.53**	**6.71**	**10.89**	**32.02**	**13.23**	**19.80**

Table 2: Training-objective comparison across all three ESCI locales under a common all-level retrieval rule (%). Values are means over five training seeds for each locale.

refinements through the shared query representation provides additional benefit beyond relevance weighting over complete identifier paths. Paired five-seed comparisons show positive gains over grouped soft-target supervision on both primary metrics across locales, as reported in Appendix C.3.

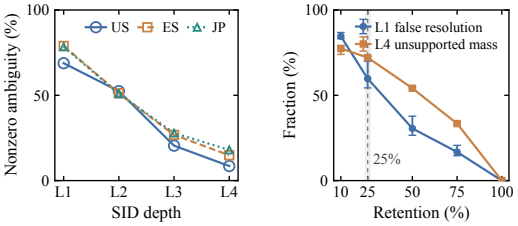


Figure 3: Left: ambiguity by SID depth. Right: evidence sparsity, averaged over US, ES and JP. Bars span the minimum and maximum across locales.

Condition	Recall@100 Base	Recall@100 RARS	Recall@100 Δ	NDCG@10 Base	NDCG@10 RARS	NDCG@10 Δ
(a) Identifier geometry						
RQ-VAE, 4×256	38.24	39.62	+1.38	12.00	13.00	+1.00
Hier. k-means, 20^3	36.84	37.65	+0.81	11.21	11.95	+0.74
RQ-VAE, 5×256	37.55	38.60	+1.05	11.54	12.42	+0.88
(b) Training positives						
E only	35.91	36.48	+0.57	10.86	11.35	+0.49
E + S	37.56	38.44	+0.88	11.48	12.21	+0.73
E + S + C	38.24	39.62	+1.38	12.00	13.00	+1.00

Table 3: Robustness on ESCI-US across identifier geometries and training positives. Five-seed means (%); Δ is the gain over Base in percentage points.

Relevance Structure and Evidence Fidelity (RQ3). Figure 3 characterizes observed relevance across SID resolutions. Refinement ambiguity is concentrated at coarse and intermediate depths and becomes less prevalent toward the leaves. Across the three locales, roughly 70% to 80% of queries eligible for retrieval are ambiguous at L1, about half at L2, and fewer than 20% at L4. We assess evidence fidelity by uniformly subsampling judgments for queries with at least two observed positives. At 25% nominal retention, more than half of these queries appear resolved at L1 despite spanning multiple branches in the full observed set, while roughly 70% of L4 relevance mass has no retained support. The right panel reports unweighted means across locales, with bars spanning the minimum and maximum at each retention rate. Incomplete judgments can make coarse refinements appear deterministic while leaving relevance at finer depths unsupported. RARS preserves the multiresolution structure supported by observed evidence, so its target fidelity depends on judgment coverage. Additional hierarchy controls and analyses of evidence subsampling appear in Appendix D.

Identifier Geometry and Training Relevance (RQ4). Table 3 evaluates the robustness of RARS across SID constructions and training relevance definitions. We compare the default RQ-VAE hierarchy with hierarchical k-means and a deeper RQ-VAE, retraining the matched base and the proposed method under the same training protocol for each identifier space. The method improves both primary metrics across all three constructions, showing that the gains are not tied to a particular SID geometry. We next vary the training positives from Exact-only to E+S and E+S+C while keeping the evaluation protocol and test judgments fixed. The method improves over the matched base under every relevance definition, including Exact-only training, so the gains do not rely on treating Complement items as positive evidence. The larger improvements with richer relevance sets are descriptive and do not isolate their source. On ESCI-US, every tested nonzero tree-loss weight improves both primary metrics over the matched base. Using 32 high-scoring sibling negatives outperforms random negatives and approaches full legal-sibling normalization. Using only relevance-bearing children also improves over the matched base. These results show robustness to hierarchy design, relevance definition, and local candidate construction across the tested configurations.

Prediction Alignment and Prefix Coverage (RQ5). We examine whether compatibility scoring improves prediction at ambiguous SID refinements. For each query-parent pair with multiple relevant children, we compare all-level and AR-only scoring using the same model, candidate set, retrieval budget, and normalized graded-relevance targets. Figure 4 shows that all-level scoring reduces Jensen–Shannon divergence and assigns more probability mass to relevant children across all locale–depth comparisons. The largest improvements occur at L2, where AR-only scoring deviates most from the relevance targets; all differences remain significant after Holm correction. We next assess whether these local gains persist in the final ranking. Prefix coverage measures the fraction of relevant prefixes represented among the top-100 retrieved SIDs. Figure 4 shows coverage gains at L2–L4 in all three

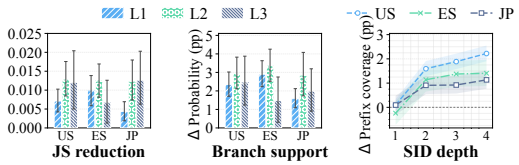


Figure 4: Prediction alignment and final-ranking prefix coverage: all-level vs. AR-only scoring at fixed model parameters. Paired query-bootstrap 95% intervals.

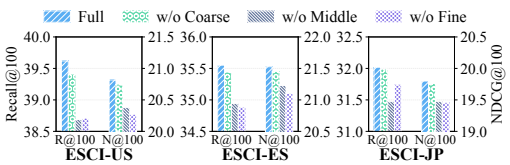


Figure 5: Depth ablation with fixed model weights. Full uses all four compatibility terms; variants remove L1, L2 or L3. Both metrics are percentages.

locales; L1 intervals include zero. These L2–L4 gains indicate better retention of relevant branches through successive SID refinements. At L4, prefix coverage equals SID Recall@100 and is not a separate retrieval endpoint; all coverage statistics are computed from final rankings.

Contributions of Retrieval Resolutions (RQ6). Figure 5 examines the contribution of retrieval resolutions by removing one compatibility term while leaving the remaining scoring terms unchanged for the same trained model. Across locales, the full scoring function achieves the highest Recall@100 and NDCG@100. Removing the intermediate and fine-level terms produces the largest drops, particularly on ESCI-US, whereas the contribution of the coarsest level is smaller and statistically significant only on ESCI-US. This indicates that useful retrieval evidence is distributed across multiple stages of the SID hierarchy. Intermediate and fine resolutions contribute complementary retrieval information not captured by the leaf-level score alone. Resolution-aware scoring improves both the matched base and RARS. Using all compatibility levels increases NDCG@10 across locales under both training conditions, showing that the benefit is not specific to resolution-aligned training. The AR-only comparisons further show that hierarchical supervision and multiresolution scoring improve different parts of the retrieval process and remain consistently effective when applied in combination.



Figure 6: ESCI retrieval examples. Arrows show CaLIR → RARS ranks and NDCG@100 (%); >100 denotes absence from the top-100 results. Branch fractions give conditional relevance targets.

Qualitative Analysis (RQ7). Figure 6 shows that relevant products separate at different depths of the SID hierarchy. In the US example, RARS promotes the Exact zippered tote despite its shared coarse prefix with a Substitute shoulder tote. The ES example spans distinct coarse branches: RARS improves the rankings of the Exact phone, a Substitute, and a Complement accessory. In the JP example, the two Exact bottles share a prefix until the final refinement; RARS retrieves a previously missed Exact bottle and moves the smaller Substitute down. These cases illustrate why relevance supervision must account for both shared prefixes and competing refinements. RARS learns how relevance is distributed among branches at each depth, preserving support for multiple relevant products while distinguishing their relevance grades. Appendix E provides additional examples.

6 Conclusion

RARS learns multiresolution relevance by supervising branch allocations in the SID hierarchy. A shared query representation captures consistent conditional targets induced by document relevance, and the auxiliary predictor is discarded after training to preserve standard autoregressive retrieval. Experiments on three ESCI locales show consistent gains over matched training controls across SID constructions and relevance definitions, including Exact-only supervision. A controlled NQ320K comparison further demonstrates the effectiveness of retrieval scoring on Wikipedia documents, with strong gains for documents absent from the training annotations. RARS connects document relevance to retrieval decisions throughout the SID hierarchy, from coarse prefixes to individual documents.

AI USE STATEMENT

We used generative AI tools to improve the readability and language of the manuscript. We did not use generative AI tools to develop the research methodology, design experiments, formulate hypotheses, or interpret the experimental results. All AI-assisted content and code were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We have considered the potential ethical implications of this work and do not identify any specific ethical concerns that require further discussion.

REPRODUCIBILITY STATEMENT

We have made efforts to support the reproducibility of this work by providing detailed descriptions of the proposed methodology, experimental setup, and evaluation protocol in the main paper and appendix. We have also released the source code and implementation details through an anonymous repository to facilitate reproduction of our results.

REFERENCES

Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In *NeurIPS*, 2010.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *CVPR*, 2020.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. In *NeurIPS*, 2022.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *ICML*, 2007.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *ACL Findings*, 2024.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *ICLR*, 2021.

Kairui Fu, Tao Zhang, Shuwen Xiao, Ziyang Wang, Xinming Zhang, Chenchi Zhang, Yuliang Yan, Junjun Zheng, Xiangheng Kong, Shengyu Zhang, Kun Kuang, and Yuning Jiang. FORGE: Forming semantic identifiers for generative retrieval in industrial datasets. In *KDD*, 2026.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *NeurIPS*, 2024.

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

Bowen Jin, Hansi Zeng, Guoyin Wang, Xiusi Chen, Tianxin Wei, Ruirui Li, Zhengyang Wang, Zheng Li, Yang Li, Hanqing Lu, Suhang Wang, Jiawei Han, and Xianfeng Tang. Language models as semantic indexers. In *ICML*, 2024.

Jiajie Jin, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, Yutao Zhu, and Zhicheng Dou. Internalizing explicit reasoning into latent space for dense retrieval. In *SIGIR*, 2026.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Learning to rank in generative retrieval. In *AAAI*, 2024a.

Yongqi Li, Zhen Zhang, Wenjie Wang, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Distillation enhanced generative retrieval. In *ACL Findings*, 2024b.

Xiaoyu Liu, Fuwei Zhang, Yiqing Wu, Xinyu Jia, Zenghua Xia, Fuzhen Zhuang, Zhao Zhang, Fei Jiang, and Wei Lin. CAT-ID^2: Category-tree integrated document identifier learning for generative retrieval in e-commerce. In *WSDM*, 2026.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. In *ACL Findings*, 2022.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *NeurIPS*, 2023.

Harish Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Convex calibrated surrogates for hierarchical classification. In *ICML*, 2015.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *arXiv e-prints*, 2022.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *FnTIR*, 2009.

Zihua Si, Zhongxiang Sun, Jiale Chen, Guozhang Chen, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, Jun Xu, and Kun Gai. Generative retrieval with semantic tree-structured identifiers and contrastive learning. In *SIGIR-AP*, 2024.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *NeurIPS*, 2020.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. Learning to tokenize for generative retrieval. In *NeurIPS*, 2023.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, and Xueqi Cheng. Generative retrieval meets multi-graded relevance. In *NeurIPS*, 2024a.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, and Xueqi Cheng. Listwise generative retrieval models via a sequential learning process. *TOIS*, 2024b.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In *NeurIPS*, 2022.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Allen Sun, Weiwei Deng, Qi Zhang, and Mao Yang. A neural corpus indexer for document retrieval. In *NeurIPS*, 2022.

Yanjing Wu, Yinfu Feng, Jian Wang, Wenji Zhou, Yunan Ye, Rong Xiao, and Jun Xiao. Hi-Gen: Generative retrieval for large-scale personalized e-commerce search. In *ICDM*, 2024.

Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, and Krzysztof Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In *NeurIPS*, 2018.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *ICML*, 2008.

Chihiro Yano, Akihiko Fukuchi, Shoko Fukasawa, Hideyuki Tachibana, and Yotaro Watanabe. Multilingual sentence-t5: Scalable sentence encoders for multilingual applications. In *LREC-COLING*, 2024.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. Scalable and effective generative information retrieval. In *WWW*, 2024.

Aoran Zhang, Yu-Bin Yang, and Yonghong Yu. Hyperbolic RQ-VAE enhanced generative recommendation with differential-length codebook strategy. In *ICML*, 2026a.

Fuwei Zhang, Xiaoyu Liu, Xinyu Jia, Yingfei Zhang, Zenghua Xia, Fei Jiang, Fuzhen Zhuang, Wei Lin, and Zhao Zhang. HierGR: Hierarchical semantic representation enhancement for generative retrieval in food delivery search. In *ACL Industry*, 2025a.

Fuwei Zhang, Xiaoyu Liu, Xinyu Jia, Yingfei Zhang, Shuai Zhang, Xiang Li, Fuzhen Zhuang, Wei Lin, and Zhao Zhang. Multi-level relevance document identifier learning for generative retrieval. In *ACL*, 2025b.

Fuwei Zhang, Xiaoyu Liu, Jiajie Jin, Jiale Mao, Wei Chen, Dongbo Xi, Yifan Yang, Peng Yan, Zichao Hao, Zhao Zhang, and Fuzhen Zhuang. Beyond matching: Category-guided latent intent reasoning for generative retrieval in e-commerce. *arXiv e-prints*, 2026b.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. Enhancing generative retrieval with reinforcement learning from relevance feedback. In *EMNLP*, 2023.

Appendix Contents

A Experimental Setup and Implementation **13**
A.1 Datasets and evaluation metrics . . . 13
A.2 Baselines . . . 14
A.3 Implementation and optimization . . . 15
A.4 Sparse target construction . . . 16
A.5 Retrieval and compatibility scoring . . . 16

B Proofs and Derivations **17**
B.1 Hierarchical consistency . . . 17
B.2 Target-induced gradient variance . . . 17
B.3 Entropy and target-sampling variance . . . 17
B.4 Cross-entropy decomposition . . . 17
B.5 Full-path consistency and grouped supervision . . . 17

C Supplementary Experimental Results **18**
C.1 Retrieval effectiveness on NQ320K . . . 18
C.2 Retrieval effectiveness and statistical uncertainty . . . 19
C.3 Training objectives and retrieval controls . . . 19
C.4 Identifier and training sensitivity . . . 21
C.5 Prediction alignment and prefix coverage . . . 23
C.6 Ablation of retrieval resolutions . . . 24

D Relevance Resolution and Evidence Analyses **24**
D.1 Relevance across identifier resolutions . . . 24
D.2 Incomplete evidence and resolution limits . . . 27

E Real-Product Case Studies **32**

A Experimental Setup and Implementation

All controlled comparisons use the same data partitions, backbone family, and evaluation protocol. The settings below specify the reference training and retrieval configurations.

A.1 Datasets and evaluation metrics

We use the category-filtered US, ES, and JP subsets of ESCI. Table 4 gives catalog and query–product statistics. The eligible test sets contain 6,014 US, 1,656 ES, and 1,883 JP queries.

		Train		Test	
Dataset	Products	Queries	Q–P	Queries	Q–P
ESCI-US	288,372	20,109	292,354	6,137	29,971
ESCI-ES	101,957	5,421	108,587	1,877	14,554
ESCI-JP	119,052	6,543	127,788	2,163	15,861

Table 4: ESCI statistics after category filtering. Q–P denotes query–product pairs.

Retrieval metrics. Let G_q contain the evaluated Exact, Substitute, and Complement products for query q. Recall@K is $|\pi_q^{1:K} \cap G_q|/|G_q|$, where $\pi_q^{1:K}$ lists the top-K products. NDCG assigns gains $g(q,d) = 3, 2, 1, 0$ to Exact, Substitute, Complement, and Irrelevant:

$$\mathrm{DCG@}K(q) = \sum_{j=1}^{K} \frac{g(q, \pi_q(j))}{\log_2(j+1)}, \qquad \mathrm{NDCG@}K(q) = \frac{\mathrm{DCG@}K(q)}{\mathrm{IDCG@}K(q)}.$$

IDCG applies the same discount to gains in decreasing order. We report query-averaged recall and NDCG as percentages and keep the evaluation weights fixed across training-evidence ablations.

A.2 Baselines

Table 1 includes sparse, dense, and generative retrieval baselines. The matched **Base** models are retrained under the common ESCI protocol, providing controlled comparisons with RARS. The methods are summarized below.

Sparse retrieval.

- **BM25** (Robertson & Zaragoza, 2009) combines term frequency, inverse document frequency, and length normalization to score query–product relevance.

Dense retrieval.

- **DPR** (Karpukhin et al., 2020) trains separate query and passage encoders with a contrastive objective. Products are ranked by the inner product of their representations with the query representation.
- **MPNet** (Song et al., 2020) combines masked and permuted language-model pretraining. Its sentence-embedding variant represents queries and products as vectors for similarity-based retrieval.
- **Sentence-T5** (Ni et al., 2022) learns sentence embeddings by pooling T5 encoder states and training semantically related texts to have similar representations. The resulting embeddings support dense retrieval based on pretrained sequence-to-sequence representations.
- **Sentence-mT5** (Yano et al., 2024) extends Sentence-T5 with multilingual mT5 encoders and sentence-embedding data, providing a dense baseline for ESCI-ES and ESCI-JP.
- **BGE-M3** (Chen et al., 2024) supports multilingual dense, multi-vector, and sparse retrieval, with self-knowledge distillation across these signals. We report the dense retrieval configuration.
- **LaSER** (Jin et al., 2026) distills explicit reasoning into dense retrieval representations. Dual-view self-distillation aligns explicit and latent reasoning through output and intermediate-trajectory objectives, removing the need to generate a rationale at inference.

Generative retrieval.

- **DSI**$_{\text{naive}}$ (Tay et al., 2022) directly generates non-semantic document identifiers. The retriever learns query-to-identifier associations without a semantic hierarchy over the identifiers.
- **DSI**$_{\text{semantic}}$ (Tay et al., 2022) assigns identifiers by hierarchical semantic clustering. Related documents share prefixes, which constrained beam search follows to retrieve documents.
- **TIGER** (Rajput et al., 2023) represents items with tuples of discrete semantic codewords. An autoregressive Transformer predicts the next item identifier from interaction context.
- **Hi-Gen** (Wu et al., 2024) combines metric learning for semantic and efficiency-aware item representations with hierarchical clustering guided by product categories. A position-aware objective supervises identifier generation for personalized e-commerce search.
- **LTRGR** (Li et al., 2024a) adds ranking-oriented training to generative retrieval. A margin-based objective over retrieved passages supplements token-level identifier likelihood.
- **RIPOR** (Zeng et al., 2024) combines relevance-oriented identifiers with progressive pairwise ranking. Its objective favors relevant over non-relevant identifiers at every prefix depth.
- **MERGE** (Zhang et al., 2025b) aligns documents and identifiers through multi-relevance query–document supervision. Outer-level contrastive learning models binary relevance, while inner-level learning distinguishes relevance grades and preserves identifier uniqueness.

- **CaLIR** (Zhang et al., 2026b) learns category-guided latent intent states before SID decoding. Hierarchical semantic reasoning and query-wise multi-positive reasoning supervise category paths, and a query-specific prefix trie restricts generation to category-consistent identifiers.
- **CAT-ID**^{2} (Liu et al., 2026) incorporates the product category tree into identifier learning. Its hierarchical class constraint injects category structure during quantization, while cluster-scale and dispersion losses balance token usage and distinguish reconstructed product representations.
- **FORGE** (Fu et al., 2026) examines semantic-identifier construction at industrial scale. It compares design choices and proposes efficient identifier-quality metrics associated with retrieval performance, with offline and online evaluations.
- **HGRec** (Zhang et al., 2026a) uses a hyperbolic residual vector-quantized autoencoder for item hierarchies. Differential-length codebooks adapt identifier capacity before autoregressive generation.

Matched training reference.

- **Base** is the matched CaLIR retriever used for our controlled comparisons. Its loss consists of full-SID likelihood, category classification, and multi-positive category alignment. The "AR only" and "all levels" rows use autoregressive scoring and compatibility-augmented scoring, respectively.

A.3 Implementation and optimization

Component	Setting	Scope
Reference architecture	CaLIR	Category-guided latent reasoning and SID decoding
English backbone	T5-base	ESCI-US
Spanish/Japanese backbone	mT5-base	ESCI-ES and ESCI-JP
Default SID	RQ-VAE, 4×256	Alternative geometries evaluated separately
RQ-VAE optimizer	AdamW	300 epochs; batch size 2048
GR optimizer	AdamW	Learning rate 5×10^{-4}
GR batch size	512 per device	Main training comparisons
GR duration	300 epochs	Fixed training length
Main checkpoint	Final checkpoint at epoch 300	Fixed training duration
Training repetitions	Five seeds	Main comparisons and added sensitivity experiments
Reference auxiliary loss	$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{con}}$	Coefficient λ_{base}=0.1
Tree-loss weight	λ_{tree}=0.5	Fixed main setting; sensitivity reported separately
Tree-head rank / temperature	64 / 1	Query-level auxiliary predictor
Sibling-negative cap	32 high-scoring legal siblings	Relevant children always included
Relevance mass	E/S/C weights 3/2/1	Normalized within each query
Beam size	100	Trie-constrained complete-SID decoding
Category width	C=3	Top three predicted category tries
Compatibility weight	λ_{comp}=2	All-level retrieval
Lexical features / weight	K_{lex}=2, η_{lex}=0.25	All-level retrieval
Primary metrics	Recall@100 and NDCG@10	NDCG gains: E/S/C/I =3/2/1/0
Query uncertainty	Paired bootstrap 95% interval	Query variation at fixed model parameters
Significance	Paired randomization; Holm correction	Diagnostic comparisons specified in each experiment

Table 5: Training, retrieval, and statistical settings for the ESCI experiments.

The reference retriever follows CaLIR (Zhang et al., 2026b). Its decoder forms category-supervised latent states before SID prediction. Let $c_j(d)$ be the category of product d at valid level j, and let $r_j(c \mid q, c_{j-1}(d))$ be the category softmax with invalid children masked. The classification loss is

$$\mathcal{L}_{\text{cls}}(q, d) = -\sum_{j \le J_d} \log r_j(c_j(d) \mid q, c_{j-1}(d)),$$

where J_d is the product's observed category depth. For query q_i, let P_{ij} be its positive categories at level j, B_j the category candidates represented in the batch, and s_{ijc} the temperature-scaled cosine similarity between the projected latent state and category prototype c. The multi-positive term is

$$\mathcal{L}_{\text{con}} = -\sum_{i,j:\, P_{ij}\neq\varnothing} \frac{1}{|P_{ij}|} \sum_{c\in P_{ij}} \log \frac{\exp s_{ijc}}{\sum_{c'\in B_j} \exp s_{ijc'}}.$$

These terms correspond to the HSR and QRE objectives of CaLIR, each with coefficient 0.1. RARS adds supervision over the SID hierarchy through the shared query encoder.

Table 5 lists the shared settings. RQ-VAE is trained with AdamW for 300 epochs at batch size 2048. The generative retriever uses AdamW with learning rate 5×10^{-4}, per-device batch size 512, and 300 epochs. Main and sensitivity comparisons use five training seeds and evaluate the final epoch-300 checkpoints. The tree-loss weight $\lambda_{\text{tree}} = 0.5$ was fixed before training. Retrieval parameters $C = 3$, $\lambda_{\text{comp}} = 2$, $K_{\text{lex}} = 2$, and $\eta_{\text{lex}} = 0.25$ were selected exclusively on validation data and fixed before test evaluation. Sensitivity analyses vary individual components around this reference configuration.

Table 1 reports five-seed means, with training-run standard deviations in Appendix C.3. Diagnostic comparisons estimate query-level uncertainty by paired bootstrap and randomization tests, conditional on fixed trained models. Scoring variants share the query set, relevance judgments, SID assignments, category trie, beam size, category width, and single-pass constrained decoding procedure. Outcomes are paired by query within each locale.

A.4 Sparse target construction

For each query, we merge duplicate query–document pairs and normalize the relevance masses. Summing each leaf's mass at every ancestor gives the nonzero projected targets. At each relevance-bearing parent, the candidate set includes all relevant children and the selected high-scoring sibling negatives. A shared query representation at each depth predicts the local child distributions, which contribute to the parent-weighted objective in Eq. (7). A parent with one relevant child has zero target ambiguity. Sibling negatives supply competing alternatives in its local normalization.

Given the candidate sets, auxiliary prediction has cost linear in the number of active query-tree edges. The cost of hard-negative selection depends on the number of legal siblings scored at each relevance-bearing parent. The resolution head shares the query encoder and is discarded after training.

A.5 Retrieval and compatibility scoring

Top-three category predictions define the legal SID set. Equations (11) and (12) score these admitted sequences. For a given model, AR-only and compatibility-augmented retrieval share the legal SID set. Matched models share category width and trie construction but may predict different categories.

Queries and catalog text are normalized by Unicode NFKC and case folding. The compatibility function uses color aliases, brands, numbers, and normalized measurements. Let $\mathcal{A}(q)$ be the extracted attribute constraints and $I_a(d)$ indicate a catalog match. The structured component is

$$f_{\text{attr}}(q,d) = \frac{1}{|\mathcal{A}(q)|} \sum_{a\in\mathcal{A}(q)} \big[(1+\gamma_a) I_a(d) - \gamma_a\big],$$

where $\gamma_a = 0.25$ for brands and 0.20 for other attributes; the component is zero if no attribute is extracted. Attribute constraints encode affirmative mentions in the query. From the query's word and CJK character features, we retain at most two features in decreasing IDF order, each occurring in at least two and at most 2% of catalog items. We use $\text{IDF}(k) = \log[(|\mathcal{D}|+1)/(\text{df}(k)+1)]$. For this set $\mathcal{K}(q)$,

$$f_{\text{lex}}(q,d) = \frac{\sum_{k\in\mathcal{K}(q)} \text{IDF}(k) I_k(d)}{\sum_{k\in\mathcal{K}(q)} \text{IDF}(k)}, \qquad f(q,d) = f_{\text{attr}}(q,d) + 0.25 f_{\text{lex}}(q,d),$$

with zero lexical contribution when the denominator is zero. Prefix compatibility takes the maximum of $f(q,d)$ across items sharing that SID prefix. Exact computation scores catalog items and propagates their maximum scores from the leaves to the root. For N items, SID depth L, and T trie nodes, this requires $O(N(|\mathcal{A}(q)| + |\mathcal{K}(q)|) + T)$ work per query and up to $O(NL)$ stored item-prefix associations. Catalog text normalization and feature postings are query-independent, and prefix

maxima depend on the query. Feature postings provide an index that reduces the number of candidate items evaluated. Exact prefix compatibility maximizes the joint item score over each subtree after combining attribute and lexical evidence. Compatibility values are added to legal token scores with weight $\lambda_{\text{comp}} = 2$ at all four identifier depths. The decoder generates up to 100 complete identifiers with beam 100; duplicate identifiers are removed while retaining their first occurrence. For a trained retriever, inference uses the query and catalog. Test relevance judgments define evaluation targets and diagnostic strata.

B Proofs and Derivations

B.1 Hierarchical consistency

The child subtrees of $u \in \mathcal{P}_{\ell-1}$ partition its descendant leaves. Hence

$$\sum_{v \in \mathrm{Ch}(u)} \mu_\ell(v \mid q) = \sum_{v \in \mathrm{Ch}(u)} \sum_{d:\pi_\ell(d)=v} \mu(d \mid q) \tag{13}$$

$$= \sum_{d:\pi_{\ell-1}(d)=u} \mu(d \mid q) = \mu_{\ell-1}(u \mid q). \tag{14}$$

Projecting leaves directly to depth j equals successive marginalization from L to j.

B.2 Target-induced gradient variance

For a target $Y \sim \boldsymbol{\mu}$, the logit gradient is $\boldsymbol{g}_Y = \boldsymbol{p} - \boldsymbol{e}_Y$. Since $\mathbb{E}[\boldsymbol{e}_Y] = \boldsymbol{\mu}$, we have $\mathbb{E}[\boldsymbol{g}_Y] = \boldsymbol{p} - \boldsymbol{\mu} = \boldsymbol{g}_\mu$. The centered gradient is $\boldsymbol{g}_Y - \boldsymbol{g}_\mu = \boldsymbol{\mu} - \boldsymbol{e}_Y$, and hence

$$\mathbb{E}\|\boldsymbol{g}_Y - \boldsymbol{g}_\mu\|_2^2 = \mathbb{E}\|\boldsymbol{e}_Y\|_2^2 - 2\boldsymbol{\mu}^\top \mathbb{E}[\boldsymbol{e}_Y] + \|\boldsymbol{\mu}\|_2^2 \tag{15}$$

$$= 1 - 2\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\mu}\|_2^2 = 1 - \|\boldsymbol{\mu}\|_2^2. \tag{16}$$

At fixed local predictions, one-hot target sampling induces the logit-gradient variance above.

B.3 Entropy and target-sampling variance

For a categorical target $\boldsymbol{\mu}$, Jensen's inequality gives $-\sum_i \mu_i \log \mu_i \geq -\log \sum_i \mu_i^2$. Combining this entropy bound with the scalar inequality $-\log x \geq 1 - x$ yields

$$1 - \|\boldsymbol{\mu}\|_2^2 \leq -\log \|\boldsymbol{\mu}\|_2^2 \leq H(\boldsymbol{\mu}).$$

Weighting by each parent's relevance mass and summing proves Eq. (10). The bound uses entropy in nats. Applying it to the normalized empirical entropies requires the corresponding scale factor. The target-sampling variance is the Gini impurity, bounded above by Shannon entropy.

B.4 Cross-entropy decomposition

For target distribution μ_ℓ and prediction p_ℓ with matching support,

$$\mathrm{CE}(\mu_\ell, p_\ell) = H(\mu_\ell) + D_{\mathrm{KL}}(\mu_\ell \| p_\ell). \tag{17}$$

The target entropy is constant with respect to the model. Cross-entropy and the KL divergence in Eq. (7) have the same minimizer, $p_\ell = \mu_\ell$, on the target support.

B.5 Full-path consistency and grouped supervision

When each local distribution is normalized over all legal children, the conditionals define the complete-path distribution $p_{\text{tree}}(s(d) \mid q) = \prod_{\ell=1}^{L} p_\ell(s_\ell(d) \mid \pi_{\ell-1}(d), q)$. The chain rule of KL divergence yields

$$D_{\mathrm{KL}}(\mu_L(\cdot \mid q) \,\|\, p_{\text{tree}}(\cdot \mid q)) = \sum_{\ell=1}^{L} \sum_{u \in \mathcal{P}_{\ell-1}} \mu_{\ell-1}(u \mid q)\, D_{\mathrm{KL}}(\mu_\ell(\cdot \mid u, q) \,\|\, p_\ell(\cdot \mid u, q)). \tag{18}$$

Parent-mass weighting recovers full-path matching from the local objectives.

Grouped full-SID supervision optimizes $\mathcal{L}_{\text{group}}(q) = -\sum_d \mu(d \mid q) \log p_\theta(s(d) \mid q)$. With matched path parameterization $p_\theta = p_{\text{tree}}$ and full-child normalization, Eq. (18) gives

$$\mathcal{L}_{\text{group}}(q) = \mathcal{L}_{\text{tree}}(q) + H(\mu_L(\cdot \mid q)).$$

RARS uses a query-level predictor and candidate sets containing all relevance-bearing children and selected legal sibling negatives. Restricted candidate sets define a local discriminative objective. Equation (18) applies under full legal-child normalization.

A set-mass loss $-\log \sum_{d \in \mathcal{D}_q^+} p_\theta(s(d) \mid q)$ constrains the total probability of relevant documents. RARS specifies how this mass is allocated among children at each refinement.

C Supplementary Experimental Results

We report supplementary NQ320K retrieval results and examine training objectives, SID and supervision choices, local prediction alignment, and retrieval scoring. Training-run variability is reported separately from query-level uncertainty.

C.1 Retrieval effectiveness on NQ320K

Dataset and evaluation. We evaluate RARS's retrieval scoring on NQ320K, extending the evaluation from product search to Wikipedia retrieval. The benchmark contains 109,739 Wikipedia documents and 307,373 annotated training pairs, with one relevant document per query. Following Sun et al. (2023), we divide the 7,830 test queries into 6,075 seen queries and 1,755 unseen queries according to whether their relevant documents appear in the training annotations. Both subsets are evaluated over the same document collection. The full test set measures overall retrieval effectiveness, while the partitioned results distinguish performance by document exposure during training.

Training and retrieval. We compare RARS with TIGER, MERGE, and CaLIR under a common evaluation protocol. The baselines use T5-base, with all query–document supervision drawn from the original annotations. Training runs for 100 epochs with AdamW, a learning rate of 5×10^{-4}, and a batch size of 512 per device. We evaluate the final models using constrained beam search with a beam size of 100. CaLIR and RARS share trained parameters, SID constraints, and beam size, isolating the contribution of retrieval scoring. Recall@10 and Recall@100 measure retrieval coverage at two ranking depths; MRR@100 measures how early the relevant document appears.

	Full test			Seen test			Unseen test		
Method	R@10	R@100	MRR	R@10	R@100	MRR	R@10	R@100	MRR
TIGER	53.72	60.20	47.22	68.86	74.65	60.71	1.31	10.20	0.54
MERGE	54.23	61.28	47.07	69.30	75.28	60.44	<u>2.05</u>	12.82	<u>0.77</u>
CaLIR	<u>56.54</u>	<u>63.75</u>	<u>51.51</u>	<u>72.72</u>	<u>76.21</u>	<u>66.24</u>	0.51	<u>20.63</u>	0.51
RARS	**61.19**	**66.25**	**53.30**	**74.09**	**76.51**	**67.03**	**16.52**	**30.71**	**5.77**
Improv. (Δ)	+4.65	+2.50	+1.79	+1.37	+0.30	+0.79	+14.47	+10.08	+5.00

Table 6: Retrieval performance on NQ320K (%). R@K/MRR denote Recall@K/MRR@100. CaLIR and RARS share trained parameters and SID constraints. Bold and underlining mark the best and second-best scores; Δ is the gain over the strongest baseline in each column (percentage points).

Retrieval effectiveness. RARS achieves the strongest retrieval performance across the full, seen, and unseen test sets (Table 6), leading all three baselines on every reported metric. On the full test set, its gains over the strongest baseline for each metric are 4.65 points in Recall@10, 2.50 in Recall@100, and 1.79 in MRR@100. Seen queries also benefit across all three metrics, with Recall@10 reaching 74.09 and MRR@100 reaching 67.03. The largest Recall@10 gain occurs on the unseen test subset. RARS retrieves the relevant document within the top ten results for 16.52% of unseen queries, achieving $8.06\times$ the strongest baseline's Recall@10 of 2.05%. Its Recall@100 reaches 30.71, exceeding the strongest baseline by 10.08 points, while MRR@100 rises from 0.77 to 5.77. The improvements span both retrieval coverage and reciprocal rank, reaching relevant documents beyond the training annotations. The controlled comparison with CaLIR establishes retrieval scoring as an effective source of these gains on NQ320K.

C.2 Retrieval effectiveness and statistical uncertainty

Table 1 reports five-seed means, and Table 7 gives sample standard deviations for all five metrics. Figure 7 summarizes within-seed differences against grouped soft-target supervision. Appendix C.3.2 evaluates Base and RARS under AR-only and all-level scoring.

Training-run statistics quantify variation across seeds under the fixed training protocol. Diagnostic bootstrap intervals and randomization tests quantify query-level variation conditional on the evaluated models. Diagnostic checkpoints and scoring weights were selected through exploratory analysis; the reported intervals condition on these choices. The five-seed comparisons use training duration and parameters fixed before training. Retrieval significance tests compare matched conditions using paired predictions for the same queries.

C.3 Training objectives and retrieval controls

C.3.1 Training objectives and variation across runs

We compare five training objectives under the common all-level retrieval rule in Eq. (12). The matched base uses full-SID supervision. Grouped soft targets weight complete identifiers by their query-conditioned relevance mass. AR-decoder soft targets apply the projected local distributions directly to the autoregressive decoder. Sampled tree supervision and RARS use the same auxiliary predictor, trained with sampled one-hot targets and explicit local distributions, respectively.

Table 7 reports means and sample standard deviations over five runs for all objectives and metrics. Figure 7 summarizes paired gains over grouped soft-target supervision. The Base and RARS means correspond to the all-level results in Table 1.

Local targets on the autoregressive decoder. Let $z_\theta(c \mid q, u)$ denote the decoder logit for child token c after prefix u. AR-decoder soft-target supervision defines

$$p_\theta^{\mathrm{AR}}(c \mid q, u) = \frac{\exp z_\theta(c \mid q, u)}{\sum_{c' \in \mathrm{Ch}(u)} \exp z_\theta(c' \mid q, u)},$$

and minimizes the parent-weighted conditional KL in Eq. (7) in addition to the reference losses. This control applies the targets directly to the decoder. Grouped supervision averages complete-SID negative log-likelihoods using weights $\mu(d \mid q)$. Appendix B.5 gives the formal relation between grouped and conditional objectives. The controls assess where the targets are applied: to complete SID likelihoods, to the decoder's local predictions, or to the shared query representation through the RARS predictor. The decoder normalizes over all legal children; RARS uses Eq. (6).

Training condition	R@5	R@10	R@100	N@10	N@100
ESCI-US					
Matched base	8.05 ± 0.08	13.63 ± 0.13	38.24 ± 0.15	12.00 ± 0.17	19.38 ± 0.11
AR-decoder soft target	8.74 ± 0.06	13.93 ± 0.12	39.04 ± 0.21	12.46 ± 0.15	20.36 ± 0.13
Grouped soft-target	8.61 ± 0.09	13.69 ± 0.14	38.88 ± 0.16	12.28 ± 0.08	20.03 ± 0.20
Sampled tree	8.89 ± 0.11	14.01 ± 0.07	39.18 ± 0.14	12.55 ± 0.12	20.29 ± 0.18
RARS	**9.29** ± 0.10	**14.36** ± 0.11	**39.62** ± 0.15	**13.00** ± 0.07	**20.82** ± 0.20
ESCI-ES					
Matched base	6.47 ± 0.07	10.93 ± 0.07	35.00 ± 0.09	12.91 ± 0.13	20.69 ± 0.09
AR-decoder soft target	6.68 ± 0.08	11.33 ± 0.05	35.17 ± 0.11	13.77 ± 0.16	21.06 ± 0.07
Grouped soft-target	6.55 ± 0.05	11.27 ± 0.09	35.08 ± 0.14	13.48 ± 0.10	20.91 ± 0.12
Sampled tree	6.63 ± 0.06	11.51 ± 0.10	35.28 ± 0.18	13.69 ± 0.06	21.19 ± 0.15
RARS	**6.87** ± 0.05	**11.80** ± 0.04	**35.55** ± 0.11	**14.19** ± 0.13	**21.53** ± 0.11
ESCI-JP					
Matched base	6.29 ± 0.10	10.42 ± 0.09	31.16 ± 0.19	12.69 ± 0.11	18.99 ± 0.13
AR-decoder soft target	6.57 ± 0.08	10.60 ± 0.13	31.70 ± 0.11	12.88 ± 0.05	19.42 ± 0.17
Grouped soft-target	6.46 ± 0.12	10.43 ± 0.07	31.61 ± 0.15	12.71 ± 0.09	19.26 ± 0.14
Sampled tree	6.52 ± 0.07	10.55 ± 0.06	31.81 ± 0.20	12.79 ± 0.14	19.55 ± 0.09
RARS	**6.71** ± 0.05	**10.89** ± 0.09	**32.02** ± 0.09	**13.23** ± 0.06	**19.80** ± 0.10

Table 7: Training objectives under all-level retrieval on ESCI (%). Values are mean ± sample SD over five runs. AR-decoder soft target is a training objective; AR only in Table 1 is a retrieval setting. Retrieval follows Eq. (12).

Paired differences. RARS improves the five-run means of both primary metrics over the matched base, grouped soft targets, and sampled tree (Table 7). Mean NDCG@10 gains over grouped soft targets are 0.72, 0.71, and 0.52 percentage points on US, ES, and JP; the corresponding Recall@100 gains are 0.74, 0.47, and 0.41 points. Figure 7 summarizes these paired comparisons. Diamonds show mean gains, and horizontal bars show one sample SD of the five within-seed differences for each locale and metric.

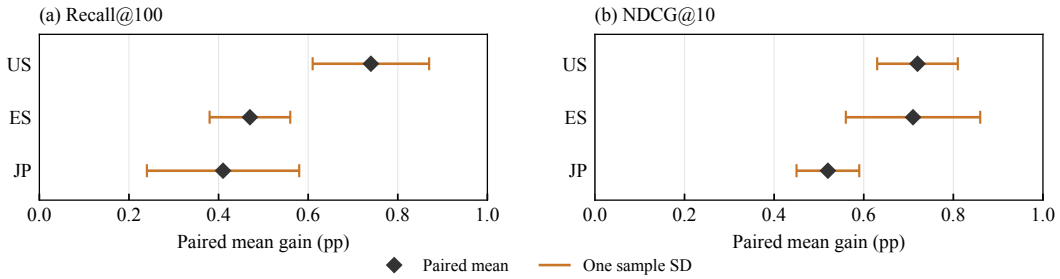


Figure 7: RARS minus grouped soft-target supervision. Diamonds show the mean of five within-seed differences; horizontal bars show one sample SD.

C.3.2 Training and retrieval resolution

We evaluate both training conditions under four retrieval settings. The masks 0000, 0001, 1110, and 1111 enable no compatibility terms, the leaf term only, the intermediate terms only, and all four terms, respectively. The auxiliary prediction head is absent throughout retrieval. Comparisons within a trained model measure the contribution of scoring; comparisons between training conditions at the same mask measure the training contrast under a common retrieval rule.

Retrieval rule	Mask	Base R@100	Base N@10	RARS R@100	RARS N@10
ESCI-US					
AR only	0000	37.40	11.12	38.98	12.36
Leaf only	0001	37.76	11.39	38.87	12.43
Intermediate only	1110	38.08	11.58	39.21	12.79
All levels	1111	38.24	12.00	39.62	13.00
ESCI-ES					
AR only	0000	34.15	12.43	34.86	13.65
Leaf only	0001	34.33	12.71	35.03	13.58
Intermediate only	1110	34.55	12.92	35.32	13.88
All levels	1111	35.00	12.91	35.55	14.19
ESCI-JP					
AR only	0000	30.89	11.77	31.59	12.34
Leaf only	0001	31.03	11.96	31.61	12.69
Intermediate only	1110	31.27	12.17	31.85	13.02
All levels	1111	31.16	12.69	32.02	13.23

Table 8: Training objectives evaluated under four retrieval settings (%). Masks indicate active compatibility terms at L1–L4; autoregressive scores are always included. All-level results agree with Table 7, and the AR-only and all-level results also appear in Table 1.

Under AR-only retrieval, RARS improves Recall@100 by $1.58/0.71/0.70$ points and NDCG@10 by $1.24/1.22/0.57$ points on US/ES/JP. All-level scoring adds $0.88/0.48/0.92$ NDCG@10 points for Base and $0.64/0.54/0.89$ for RARS. The differences in these increments are $-0.24/0.06/-0.03$ points. Training and compatibility scoring both improve retrieval under matched conditions.

For a metric evaluated under AR-only and all-level retrieval, let B_{AR} denote the base-training score and let L_{AR} and L_F denote the corresponding RARS scores. The total gain is

$$L_F - B_{\mathrm{AR}} = \underbrace{L_{\mathrm{AR}} - B_{\mathrm{AR}}}_{\text{training contrast under AR}} + \underbrace{L_F - L_{\mathrm{AR}}}_{\text{retrieval increment within RARS}}. \tag{19}$$

The decomposition follows an ordered contrast: changing the training objective under AR-only retrieval, then adding compatibility scoring to the RARS model.

C.4 IDENTIFIER AND TRAINING SENSITIVITY

We vary SID construction, training positives, tree-loss weight, temperature, sibling candidates, and head dimension. Each comparison changes one component of the fixed reference configuration.

C.4.1 IDENTIFIER GEOMETRY AND RELEVANCE DEFINITIONS

Alternative identifier geometries. We evaluate two alternative SID constructions on ESCI-US. Hierarchical k-means recursively partitions the product representations into a three-level tree with branching factor 20 at each level, following the CaLIR construction (Zhang et al., 2026b). Each product receives the sequence of cluster indices along its path through this $20 \times 20 \times 20$ hierarchy. When products share a terminal cluster, we append a deterministic suffix solely to ensure identifier uniqueness. The hierarchy contains three semantic clustering levels. The alternative RQ-VAE uses five residual codebooks with 256 entries each (5×256), compared with the default four-codebook construction (4×256). The two variants use the same RQ-VAE settings at different residual depths. For each geometry, both Base and RARS are trained from scratch with the corresponding identifiers, using the same data partitions, optimization settings, and five random seeds.

Training-evidence sensitivity. The E-only and E+S variants use different sets of training positives under the same evaluation protocol. E-only treats Exact-labeled products as positives, while E+S uses Exact and Substitute products; the full setting includes Exact, Substitute, and Complement. The original $3/2/1$ relevance weights are retained and renormalized over the positives available for each query. All variants retain the same test queries, product corpus, semantic identifiers, retrieval settings, and ESCI relevance judgments. The evaluation target is fixed across training variants.

SID geometry	Training	R@5	R@10	R@100	N@10	N@100
RQ-VAE, 4×256	Base	8.05	13.63	38.24	12.00	19.38
	RARS	**9.29**	**14.36**	**39.62**	**13.00**	**20.82**
Hier. k-means, 20^3	Base	7.81	12.88	36.84	11.21	18.83
	RARS	**8.36**	**13.57**	**37.65**	**11.95**	**19.54**
RQ-VAE, 5×256	Base	8.02	13.11	37.55	11.54	19.17
	RARS	**8.84**	**13.85**	**38.60**	**12.42**	**20.13**

Table 9: Retrieval under three SID constructions on ESCI-US. Base and RARS are trained from scratch for each construction. The 20^3 hierarchy uses three clustering levels plus an identifier-uniqueness suffix; the RQ-VAE variants differ only in residual depth. Values are five-seed means (%), with the better value in each pair in bold.

Training positives	Training	R@5	R@10	R@100	N@10	N@100
E only	Base	7.52	12.31	35.91	10.86	18.35
	RARS	**7.90**	**12.75**	**36.48**	**11.35**	**18.86**
E + S	Base	7.97	13.00	37.56	11.48	19.15
	RARS	**8.64**	**13.64**	**38.44**	**12.21**	**19.94**
E + S + C	Base	8.05	13.63	38.24	12.00	19.38
	RARS	**9.29**	**14.36**	**39.62**	**13.00**	**20.82**

Table 10: Sensitivity to training relevance on ESCI-US. E/S/C denote Exact/Substitute/Complement labels. Each pair compares Base and RARS under the same training relevance definition and fixed evaluation judgments. Values are five-seed means (%); bold indicates the better value in each pair.

RARS improves all five metrics for each SID construction. NDCG@10 gains in percentage points are 0.74 for hierarchical k-means, 0.88 for the five-level RQ-VAE, and 1.00 for the reference RQ-VAE.

With Exact-only positives, the gains are 0.57 points in Recall@100 and 0.49 points in NDCG@10. The corresponding gains are $0.88/0.73$ with Exact and Substitute, and $1.38/1.00$ when the training evidence includes Exact, Substitute, and Complement products.

C.4.2 TRAINING HYPERPARAMETERS AND SIBLING CANDIDATES

Every tested nonzero tree weight improves both primary metrics over $\lambda_{\text{tree}} = 0$. NDCG@10 changes from 13.00 to 12.76 between weights 0.5 and 1.0, while Recall@100 changes from 39.62 to 39.38. At weight 2.0, Recall@100 is 39.71% and NDCG@10 is 12.91%.

λ_{tree}	R@5	R@10	R@100	N@10	N@100
0	8.05	13.63	38.24	12.00	19.38
0.10	8.72	13.81	39.00	12.51	20.21
0.25	9.08	14.13	39.43	12.85	20.60
0.50	**9.29**	**14.36**	39.62	**13.00**	20.82
1.00	9.25	14.31	39.38	12.76	**20.95**
2.00	8.96	14.06	**39.71**	12.91	20.63

Table 11: Sensitivity to tree-loss weight on ESCI-US (%). The reference weight 0.50 was specified before training. Bold indicates the maximum in each metric.

Local candidate construction	R@5	R@10	R@100	N@10	N@100
Relevant children only	8.48	13.77	38.65	12.18	19.89
+ 32 random sibling negatives	8.79	13.93	39.07	12.57	20.28
+ 8 high-scoring siblings	9.05	14.17	39.39	12.83	20.58
+ 32 high-scoring siblings	**9.29**	**14.36**	**39.62**	**13.00**	**20.82**
All legal siblings	9.26	14.34	39.24	12.85	20.73

Table 12: Sibling-negative construction on ESCI-US (%). All candidate sets include every relevance-bearing child. Random and high-scoring negatives are legal siblings. Bold indicates the best value in each metric.

Relevant-child-only normalization improves NDCG@10 from the base value of 12.00 to 12.18. Adding 32 random sibling negatives raises it to 12.57; 32 high-scoring negatives reach 13.00. This setting exceeds full legal-sibling normalization by 0.38 points in Recall@100 and 0.15 points in NDCG@10. Gains are present with relevant-child normalization and increase when high-scoring sibling negatives are added. At a parent with one relevant child, this restricted softmax has zero local gradient. Sibling negatives provide competing alternatives at such parents.

Hyperparameter sensitivity across locales. Figure 8 evaluates the sensitivity of RARS to its training hyperparameters under all-level retrieval. Each parameter is varied in turn, holding the others at their reference values: $\lambda_{\text{tree}} = 0.5$, $\tau_p = 1$, $K_{\text{neg}} = 32$, and $r = 64$. Each parent's candidate set contains all relevance-bearing children and up to K_{neg} high-scoring legal sibling negatives.

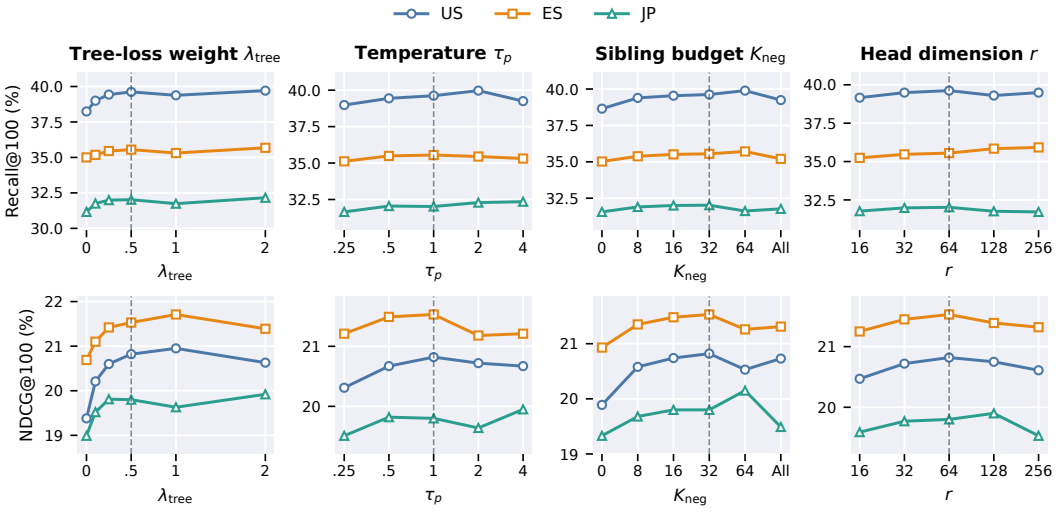


Figure 8: Sensitivity to training hyperparameters under all-level retrieval on ESCI-US, ES, and JP. The top and bottom rows report Recall@100 and NDCG@100 (%). Vertical dashed lines mark the reference parameter values.

Loss weight and prediction temperature. Relative to $\lambda_{\text{tree}} = 0$, the reference weight 0.5 improves Recall@100 and NDCG@100 in all three locales. Recall@100 peaks at weight 2.0 in every locale, while NDCG@100 peaks at 1.0 in US and ES and at 2.0 in JP. The temperature sweep also gives different optima: US Recall@100 peaks at $\tau_p = 2$, US and ES NDCG@100 peak at $\tau_p = 1$, and both JP metrics peak at $\tau_p = 4$.

Sibling budget and head dimension. Adding eight high-scoring sibling negatives improves both metrics in every locale. The reference budget $K_{\text{neg}} = 32$ exceeds full legal-child normalization on both metrics in all three locales. Increasing the head dimension from $r = 16$ to $r = 64$ also improves both metrics in every locale. US reaches its highest values on both metrics at $r = 64$. ES Recall@100 peaks at $r = 256$, while its NDCG@100 peaks at $r = 64$; JP NDCG@100 peaks at $r = 128$.

C.5 Prediction alignment and prefix coverage

Local prediction diagnostics. Local prediction is evaluated on 512 held-out queries per locale using fixed trained models. Queries with applicable constraints are stratified by ambiguity and positive count. Of these, 457 US, 447 ES, and 451 JP queries have at least one parent with multiple relevant children. For each such parent, relevance grades E/S/C receive weights 3/2/1 and are normalized to a conditional distribution over catalog-legal children. The comparison varies compatibility scores at fixed model parameters, query, and parent prefix.

Within each query and depth, parent-level measurements are weighted by relevance mass and normalized over the included ambiguous parents. The figure averages queries separately at each depth, retaining all nine locale-depth comparisons. Query counts by depth appear in Table 13. Confidence intervals use 10,000 paired query-bootstrap samples. Two-sided paired sign-randomization uses 10,000 draws; Holm correction covers each metric's nine locale-depth comparisons.

(a) Jensen–Shannon divergence

Locale	Depth	Queries	AR-only	All-level	JS reduction [95% CI]
US	L1	394	0.3197	0.3126	0.0071 [0.0042, 0.0103]*
	L2	297	0.4531	0.4402	0.0129 [0.0084, 0.0176]*
	L3	160	0.3233	0.3112	0.0120 [0.0050, 0.0204]*
ES	L1	432	0.4054	0.3956	0.0098 [0.0059, 0.0139]*
	L2	280	0.4228	0.4104	0.0124 [0.0081, 0.0169]*
	L3	160	0.2972	0.2905	0.0067 [0.0013, 0.0126]*
JP	L1	431	0.4210	0.4166	0.0044 [0.0019, 0.0069]*
	L2	268	0.4367	0.4243	0.0124 [0.0074, 0.0180]*
	L3	170	0.2667	0.2540	0.0126 [0.0063, 0.0203]*

(b) Probability assigned to relevant children (%)

Locale	Depth	Queries	AR-only	All-level	ΔSupport [95% CI]
US	L1	394	50.74	53.09	+2.35 [1.71, 3.02]*
	L2	297	31.04	33.98	+2.94 [2.11, 3.82]*
	L3	160	59.69	62.12	+2.43 [1.24, 3.88]*
ES	L1	432	35.11	38.01	+2.90 [2.23, 3.63]*
	L2	280	36.02	39.38	+3.36 [2.52, 4.26]*
	L3	160	61.93	63.41	+1.48 [0.31, 2.75]*
JP	L1	431	34.98	36.57	+1.58 [1.09, 2.12]*
	L2	268	36.04	38.89	+2.85 [1.73, 4.08]*
	L3	170	68.85	70.83	+1.99 [0.91, 3.19]*

Table 13: Prediction at ambiguous refinements under AR-only and all-level scoring, by locale and SID depth. JS measures divergence from the relevance target; Support is the probability assigned to relevant children (%). Intervals are paired query-bootstrap 95% intervals for the improvement. * denotes Holm-adjusted $p < 0.05$.

Ambiguity and prediction error. Among ambiguous parents, one standard deviation of normalized refinement ambiguity is associated with a 0.0267 increase in baseline JS divergence (query-clustered 95% CI $[0.0078, 0.0456]$). The regression controls for relevant-child count, legal-child count, depth, and locale. The coefficient for JS reduction is -0.0073 (95% CI $[-0.0120, -0.0026]$).

Prefix coverage on the complete evaluation sets. Retrieval diagnostics use one fixed trained model per locale and the full evaluation sets, separately from the five-seed means in Table 1. AR-only

and all-level retrieval evaluate the same trained model under their respective scoring rules. At each SID depth, prefix coverage is the fraction of relevant prefixes represented among the final top-100 retrieved SIDs. A loss event occurs when a query has positive prefix coverage at one depth and zero at the next. Both statistics are computed from the final top-100 rankings.

Figure 4 reports coverage changes for 6,014 US, 1,656 ES, and 1,883 JP queries. Intervals use 10,000 paired query-bootstrap samples and tests use 10,000 paired sign randomizations. Holm correction covers four coverage and three loss-event comparisons within each locale. Resampling estimates query-level variation for each fixed model and scoring rule.

Prefix coverage improves significantly at L2–L4 in all three locales. The L4 differences are 2.22, 1.40, and 1.13 percentage points on US, ES, and JP, respectively. The corresponding L1 intervals include zero. L4 coverage equals SID Recall@100.

C.6 Ablation of retrieval resolutions

We remove one prefix compatibility term at a time, holding model parameters and the remaining scoring terms fixed. Training-objective comparisons are reported in Appendix C.3.

Locale	Variant	R@100	NDCG@100
US	Full RARS	**39.6187**	**20.8169**
	w/o coarse (L1)	39.4102	20.7389
	w/o middle (L2)	38.6885	20.3803
	w/o fine (L3)	38.7108	20.2723
ES	Full RARS	**35.5515**	**21.5271**
	w/o coarse (L1)	35.4424	21.4499
	w/o middle (L2)	34.9420	21.2282
	w/o fine (L3)	34.8847	21.1042
JP	Full RARS	**32.0187**	**19.8000**
	w/o coarse (L1)	31.9904	19.7607
	w/o middle (L2)	31.4732	19.4764
	w/o fine (L3)	31.7407	19.4557

Table 14: Ablation of compatibility scoring by SID depth, corresponding to Figure 5. L1–L3 denote coarse, middle, and fine prefix terms. All conditions use the same trained model, decoder, beam 100, and query set, with the leaf term retained. Bold indicates the best value within each locale.

Table 14 shows that full scoring achieves the highest R@100 and NDCG@100 in all three locales. Removing L2 or L3 produces the largest decreases in both metrics. The paired differences in Table 15 are statistically significant for both resolutions across locales. L1 yields positive differences in every locale, with significant gains on both metrics in US. The L2 and L3 ablations retain the leaf term and all other prefix terms. The decrease in retrieval effectiveness after removing either term shows that both resolutions contribute to the full scoring function. Paired tests quantify query-level variation for the fixed models and scoring settings (Appendix C.2).

D Relevance Resolution and Evidence Analyses

We analyze relevance profiles across SID hierarchies and evidence subsets.

D.1 Relevance across identifier resolutions

Relevance ambiguity is concentrated at coarse SID depths, with substantial variation across queries and identifier assignments. Table 16(a) reports nonzero refinement entropy in 68.8%–78.9% of queries at L1, falling to 8.5%–18.0% at L4. Semantic SIDs allocate more entropy to L3–L4 than the controls in Table 16(b). Their mean entropy-weighted depth exceeds the L1 + full-category control by 0.022–0.032 ($p_{\mathrm{H}} = 0.024$ after Holm correction). This comparison holds tree structure, relevance judgments, L1 prefixes, and category paths fixed. Queries with identical E/S/C counts and total SID entropy exhibit distinct depth profiles. Under graded relevance, matched profiles in Table 16(c) differ by 0.234–0.347 in mean total variation, and the depth of maximum ambiguity differs in 16.7%–38.1% of pairs. In Table 16(d), graded profiles from disjoint subsets of the same query have 18.2%–28.8%

Locale	Term	R@100 Δ	R@100 95% CI	R@100 p	NDCG@100 Δ	NDCG@100 95% CI	NDCG@100 p
US	L1	+0.2085	$[0.0613, 0.3602]$	0.0046	+0.0780	$[0.0247, 0.1322]$	0.0034
	L2	+0.9302	—	< 0.001	+0.4367	—	< 0.001
	L3	+0.9079	—	< 0.001	+0.5446	—	< 0.001
ES	L1	+0.1090	—	0.4955	+0.0772	—	0.2324
	L2	+0.6095	$[0.2608, 0.9867]$	0.0013	+0.2990	$[0.1607, 0.4447]$	< 0.001
	L3	+0.6668	$[0.4093, 0.9714]$	< 0.001	+0.4229	$[0.2867, 0.5585]$	< 0.001
JP	L1	+0.0283	—	0.6955	+0.0393	—	0.3839
	L2	+0.5455	$[0.3199, 0.8099]$	< 0.001	+0.3236	$[0.1923, 0.4548]$	< 0.001
	L3	+0.2780	$[0.1361, 0.4424]$	< 0.001	+0.3443	$[0.2361, 0.4735]$	< 0.001

Table 15: Paired differences for the retrieval-resolution ablations. Δ is full RARS minus the ablation of the indicated term, in percentage points. Brackets give reported 95% paired query-bootstrap confidence intervals; p-values use 10,000 two-sided paired sign randomizations.

	ESCI-US			ESCI-ES			ESCI-JP		
(a) Depth	Δ_ℓ	Q (%)	M (%)	Δ_ℓ	Q (%)	M (%)	Δ_ℓ	Q (%)	M (%)
L1	0.700	68.8	68.8	1.128	78.9	78.9	1.127	78.4	78.4
L2	0.399	52.4	38.1	0.320	51.2	30.1	0.321	50.9	30.6
L3	0.065	20.5	8.1	0.071	26.8	8.2	0.072	27.9	8.6
L4	0.020	8.5	2.5	0.029	14.9	3.5	0.038	18.0	4.6
(b) Hierarchy	Δ_2	$\Delta_{3:4}$	τ	Δ_2	$\Delta_{3:4}$	τ	Δ_2	$\Delta_{3:4}$	τ
Global shuffle	0.012	<0.001	1.008	0.058	0.001	1.032	0.037	<0.001	1.020
Within-L1 shuffle	0.477	0.006	1.374	0.411	0.010	1.239	0.421	0.010	1.249
L1 + category	0.433	0.050	1.413	0.347	0.073	1.283	0.348	0.083	1.302
Semantic	0.399	**0.084**	**1.445**	0.320	**0.100**	**1.305**	0.321	**0.110**	**1.328**
Δ *vs. category*	**↓0.035**	**↑0.035**	**↑0.032**	**↓0.027**	**↑0.027**	**↑0.022**	**↓0.027**	**↑0.027**	**↑0.026**
(c) Evidence	Pairs	TV	Peak diff.	Pairs	TV	Peak diff.	Pairs	TV	Peak diff.
Binary	2,134	**0.344**	**37.7%**	467	**0.232**	**16.7%**	579	**0.253**	**19.0%**
Graded	2,134	**0.347**	**38.1%**	467	**0.234**	**16.7%**	579	**0.255**	**19.7%**
(d) Replication	Within	Cross	Gap	Within	Cross	Gap	Within	Cross	Gap
Binary	**0.315**	0.386	**+0.071**	**0.184**	0.242	**+0.058**	**0.194**	0.273	**+0.079**
Graded	**0.317**	0.388	**+0.071**	**0.186**	0.245	**+0.058**	**0.195**	0.274	**+0.079**

Table 16: Relevance ambiguity across SID resolutions: (a) graded entropy and ambiguous mass; (b) hierarchy controls; (c) heterogeneity under exact E/S/C matching; and (d) replication across disjoint positive subsets.

smaller distances from matched queries ($p_{\mathrm{H}} < 0.001$). Binary relevance yields the same ordering. Definitions and sensitivity analyses appear in Appendix D.1.1.

D.1.1 Measures and controls

Population and measures. The analysis uses the retrieval-eligible test sets: 6,014 US, 1,656 ES, and 1,883 JP queries. Duplicate products are merged before normalization, retaining the highest observed grade. Binary evidence gives every E/S/C positive unit weight; graded evidence uses weights 3, 2, and 1. The subsets with multiple positives contain 4,649, 1,348, and 1,534 queries, all with nonzero observed SID entropy under the semantic hierarchy.

Refinement entropies are measured in nats. Q is the percentage of queries with nonzero refinement entropy; M is the mean percentage of relevance mass at parents with multiple positive children. For nonzero total entropy, $r_\ell(q) = \Delta_\ell(q) / \sum_j \Delta_j(q)$ gives its allocation across depths and $\tau(q) = \sum_\ell \ell\, r_\ell(q)$ its mean location. $\Delta_{3:4} = \Delta_3 + \Delta_4$ summarizes fine-resolution ambiguity.

Arrows compare semantic assignments with the L1 + full-category control. TV measures profile distance, and Peak diff. denotes pairs with disjoint maximizing depths. Within and Cross compare disjoint positive subsets from the same query and from matched queries, with Gap equal to Cross minus Within under identical grade counts and total SID entropy.

The capacity-normalized statistic is

$$\bar{\Delta}_\ell(q) = \frac{\Delta_\ell(q)}{\sum_{u \in \mathcal{P}_{\ell-1}} \mu_{\ell-1}(u \mid q) \log |\mathrm{Ch}(u)|}, \tag{20}$$

with value zero when the denominator is zero. The denominator includes every catalog-legal child of each relevance-bearing parent. Raw entropy measures unresolved relevance, while normalized entropy expresses it relative to the available branching capacity. Prefix projection preserves parent mass and satisfies the entropy chain rule.

The relevance mass in Table 16(a) is the query-mean quantity $M_\ell(q) = \sum_u \mu_{\ell-1}(u \mid q)\mathbf{1}[|\{v \in \mathrm{Ch}(u) : \mu_\ell(v \mid q) > 0\}| > 1]$. It measures relevance mass at parents with multiple positive children. Q counts queries with at least one branching prefix, and M weights branching prefixes by their relevance mass. Both statistics include all retrieval-eligible queries.

Prevalence and magnitude. Table 16(a) reports ambiguity prevalence and raw entropy, which decreases with depth in all three locales. The rise in capacity-normalized entropy from L3 to L4 reflects reduced mean logarithmic branching capacity: from 1.858 to 0.306 nats in US, 1.474 to 0.348 in ES, and 1.464 to 0.331 in JP.

Topology-preserving randomization. Global shuffling reassigns the catalog SID multiset to items while preserving tree topology and SID multiplicities. Within-L1 shuffling performs the same reassignment separately within each original first-level prefix. The second control preserves every query's L1 relevance distribution exactly while disrupting finer semantic organization. Both controls preserve query positive counts, item relevance frequencies, and evidence weights. Each control uses 999 assignments per locale. The primary statistic is mean entropy-weighted depth $\tau(q)$ among queries with positive total SID entropy. SID collisions after reassignment can produce zero total entropy. Relative to the semantic assignment, within-L1 shuffling changes the eligible query count by at most one per locale and assignment; global shuffling preserves this count.

Table 17 reports observed query-bootstrap confidence intervals alongside the central 95% of the random-assignment distribution. The permutation envelopes describe variation across identifier assignments; query-bootstrap intervals describe variation across queries. Two-sided permutation p-values use twice the smaller empirical tail probability, with the plus-one correction. Holm correction covers twelve contrasts across three locales and four shuffle controls: global, within-L1, within-L1 with the first three category levels fixed, and within-L1 with the complete category path fixed. The reported depth contrasts have $p_{\mathrm{H}} = 0.024$. Relative to within-L1 shuffling, semantic SIDs reduce L2 entropy by 0.0783, 0.0906, and 0.1000 nats in US, ES, and JP, respectively, and increase entropy at L3 and L4. Semantic organization places more of the relevance branching at finer refinements.

Hierarchy	ESCI-US	ESCI-ES	ESCI-JP
Semantic	1.4449 [1.4327, 1.4569]	1.3054 [1.2850, 1.3259]	1.3276 [1.3097, 1.3477]
Global shuffle	1.0082 [1.0070, 1.0097]	1.0325 [1.0286, 1.0369]	1.0204 [1.0177, 1.0234]
Within-L1 shuffle	1.3745 [1.3735, 1.3756]	1.2388 [1.2375, 1.2405]	1.2490 [1.2478, 1.2507]
L1 + full category	1.4128 [1.4101, 1.4158]	1.2831 [1.2793, 1.2870]	1.3021 [1.2982, 1.3064]

Table 17: Entropy-weighted depth τ under topology-preserving controls. Semantic brackets give query-bootstrap 95% confidence intervals; control brackets give the central 95% of 999 identifier assignments. The full-category control preserves both L1 and the complete category path. Means of τ are computed conditional on positive total SID entropy.

Positive-count and evidence sensitivity. For each multi-positive query, we uniformly sample two distinct positives, preserve their relevance weights, and average the depth profile over 30 subsamples. Under binary and graded evidence, mean raw entropy ranges from 0.4835 to 0.5916 nats at L1 and from 0.0052 to 0.0076 at L4. The concentration at early depths persists with positive count fixed.

On identical query sets, binary and graded relevance differ by at most 0.0031 in mean normalized ambiguity, with query-level global ambiguity correlations of 0.9986–0.9991.

Observed mean intervals use 2,000 query-bootstrap draws; matched-pair intervals resample disjoint pairs. Randomization and subsampling are fixed independently of their results.

Controlling catalog taxonomy. We further restrict item-to-SID shuffling to groups sharing an L1 prefix and the complete catalog category path. This control preserves category occupancy of SID slots, the query's category composition, its first-level relevance distribution, and the catalog SID multiset. The fractions of positive item occurrences eligible for reassignment are 88.6%, 76.5%, and 78.2% in US, ES, and JP, respectively. Semantic entropy-weighted depth exceeds this control in every locale (Table 17), with differences of 0.0320, 0.0224, and 0.0256, respectively.

Replication across disjoint positive subsets. For each query and grade, we split $2\lfloor N_{\text{grade}}/2 \rfloor$ products into two disjoint halves of equal size. Eligible queries have at least two positives per half and distinct positive SIDs. Disjoint query pairs match the half-grade counts. Within-query and cross-query comparisons have identical E/S/C composition and total SID entropy. The analysis includes 1,244 US pairs (2,488 queries), 379 ES pairs (758 queries), and 479 JP pairs (958 queries).

For a pair (q, q'), the within-query statistic averages $\mathrm{TV}(r_{q,A}, r_{q,B})$ and $\mathrm{TV}(r_{q',A}, r_{q',B})$; the cross-query statistic averages $\mathrm{TV}(r_{q,A}, r_{q',B})$ and $\mathrm{TV}(r_{q',A}, r_{q,B})$. We average 100 annotation splits within each disjoint query pair. Confidence intervals use 2,000 pair-bootstrap draws, and two-sided tests use 10,000 sign randomizations. Within-query distances are lower under binary and graded evidence in all locales (Table 16(d); six-test Holm-adjusted $p < 0.001$). The graded reductions relative to cross-query distance are 18.2%, 23.8%, and 28.8%. The smaller within-query distances show consistent depth structure across disjoint subsets of each query's observed judgments.

Exact evidence composition. Table 16(c) pairs queries with identical E, S, and C counts. We exclude queries with positive-SID collisions (5 US, 12 ES, and 1 JP query) and form disjoint pairs within each count stratum. The analysis contains 2,134 US, 467 ES, and 579 JP pairs. Equal grade counts and distinct positive SIDs give equal total SID entropy under both evidence mappings. At fixed relevance composition and total SID entropy, queries exhibit distinct depth profiles under both binary and graded relevance. Profile TV compares normalized depth profiles. Dominant depths differ when the sets of entropy-maximizing depths are disjoint, with a tolerance of 10^{-12} for tied maxima.

Restricting the relevance definition. On fixed subsets with at least two Exact positives (2,698 US, 858 ES, and 1,063 JP queries), mean refinement entropy remains nonzero at every depth under E-only, E+S, and E+S+C evidence, with the original relevance weights renormalized over the retained label set. With E alone, mean entropy ranges from 0.7941 to 1.2157 nats at L1 and from 0.0226 to 0.0381 at L4, showing that relevance branching persists under Exact-only evidence. Appendix C.4 reports retrieval results for training with these relevance definitions.

Refinement entropy and supervision. Refinement entropy locates branching in the observed relevance distribution. At a zero-entropy refinement, observed positive mass is concentrated on one child. For a fixed local predictor, relevance-weighted one-hot sampling and explicit conditional targets have the same expected gradient (Eq. (9)). Explicit targets remove the variance due to target sampling. Training and scoring comparisons appear in Appendices C.3, C.5, and C.6.

D.2 Incomplete evidence and resolution limits

Using all observed positives as a reference, we examine how subsampling changes target total variation (TV) and coverage of withheld relevance mass. Figure 9 compares semantic identifiers with global and within-L1 reassignment at 25% retention. With candidate counts matched per query at 25% retention, Table 18(a) reports expected L2 coverage of 2.56%–3.12% for semantic assignments and 0.17%–0.24% after shuffling within L1. The same ordering holds at L3, with differences at both depths significant after Holm correction ($p_{\mathrm{H}} < 0.004$). Relative to uniform masking, preferentially retaining products with higher relevance grades reduces mean TV, whereas concentrating evidence in fewer branches increases it, as shown in Table 18(b). Under uniform masking at 25% retention, Table 18(c) shows that 54.3%–70.0% of queries appear resolved at L1 even though their full observed positive sets span multiple branches. At L4, 70.7%–72.9% of relevance mass lacks retained support. Incomplete judgments can produce apparent certainty at coarse depths while missing relevant branches at finer depths. The subsampling design and uncertainty estimates are detailed in Appendix D.2.1.

D.2.1 Evidence subsampling and statistical analysis

Population and reference distribution. The evidence-recovery analysis uses the ESCI seen-query test partitions from the CaLIR evaluation. It includes all queries with at least two distinct positive

Evidence control	US	ES	JP
(a) Coverage at matched candidate counts (%)			
L2: Semantic	2.64	2.56	3.12
L2: Within-L1	0.17	0.23	0.24
L3: Semantic	0.46	0.69	0.98
L3: Within-L1	0.01	0.02	0.02
(b) Missingness: mean target TV			
Uniform	0.637	0.673	0.671
Grade-biased	0.621	0.652	0.653
Branch-concentrated	0.658	0.701	0.698
(c) Resolution limits (%)			
False resolution at L1	70.0	54.3	55.0
Unobserved L4 mass	70.7	72.8	72.9

Table 18: Relevance recovery at 25% positive retention with linear weights. Coverage is measured at matched candidate counts; target TV is averaged over L1–L4.

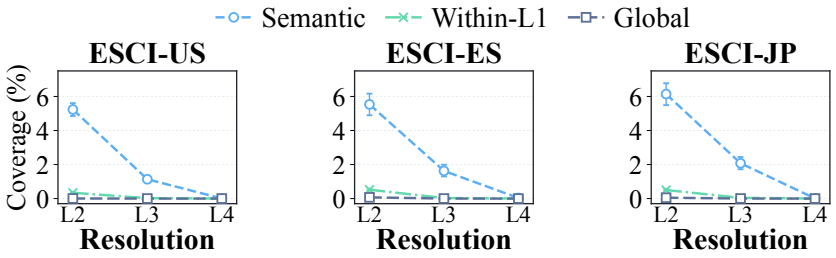


Figure 9: Coverage of withheld relevance mass by SID depth at 25% retention under semantic identifiers, within-L1 reassignment, and global reassignment. Coverage is evaluated over all products in supported prefixes. Table 18(a) uses equal candidate counts across identifier assignments. Error bars show 95% confidence intervals.

products in the catalog. Eligibility depends on positive count, independent of category routing, retrieval success, or ambiguity. This population differs from the retrieval-eligible effectiveness sets. Table 19 accounts for all 10,177 test queries. The population accounting lists single-positive queries separately; retained–withheld comparisons use the multi-positive subset. Duplicate product judgments retain the highest grade. Distinct products sharing a SID contribute their summed mass after sampling. The benchmark's test grades define the reference relevance distribution.

Locale	Source	Multi-pos.	Single-pos.	No positive	Catalog	Median n_q	Retained	Actual (%)
US	6,137	4,742	1,395	0	288,372	4	1.47	29.3
ES	1,877	1,418	321	138	101,957	7	2.37	27.2
JP	2,163	1,594	367	202	119,052	7	2.29	27.1

Table 19: Population for the evidence-subsampling analysis. All queries with at least two catalog-mapped positives are included, without category-routing filters. No positive denotes queries without a catalog-mapped positive. The final columns give the mean retained count and realized retention at the nominal 25% level.

The reference distribution normalizes relevance mass over all observed positives for each query. Linear evidence assigns C/S/E weights $(1, 2, 3)$, binary evidence assigns weight one to every positive, and exponential evidence assigns $(1, e, e^2)$. For each nominal fraction f, the mask retains $\max(1, \lfloor fn_q \rfloor)$ products. The 25% budget consequently retains a mean of $1.47/2.37/2.29$ products and realized mean fractions $29.3\%/27.2\%/27.1\%$ in US/ES/JP. The realized fraction reflects integer rounding and the minimum of one retained positive per query.

Experimental design. The exploratory design crosses three locales, five retention levels, and three identifier assignments, with 20 paired replicates per condition. Each condition is evaluated at four SID depths and at the individual-product level. Primary comparisons use linear relevance weights, uniform masking, and 25% retention. Within each replicate, masking and identifier reassignment are randomized independently. Uniform masks sample nested positive sets without replacement.

Global and within-L1 controls permute entire SID paths over catalog products. The multiset of paths is preserved, as are prefix sizes, numbers of legal children, and SID collisions. Within-L1 assignment

also preserves every product's L1 prefix. The retained product set is identical across hierarchy and evidence-weight comparisons. Partitioning by product identity is invariant to SID reassignment.

Measures and statistical inference. We measure target error by TV and Jensen–Shannon divergence in nats, and quantify reference mass in prefix groups absent from the retained evidence. Withheld coverage is the fraction of withheld positive mass in groups containing at least one retained positive. Retained and withheld products are disjoint. Withheld coverage is evaluated below 100% retention, where the withheld set is nonempty. Catalog expansion is the fraction of catalog products contained in supported groups, measuring the size of the resulting candidate set. Figure 10 plots withheld relevance coverage against catalog expansion across SID depths and identifier assignments.

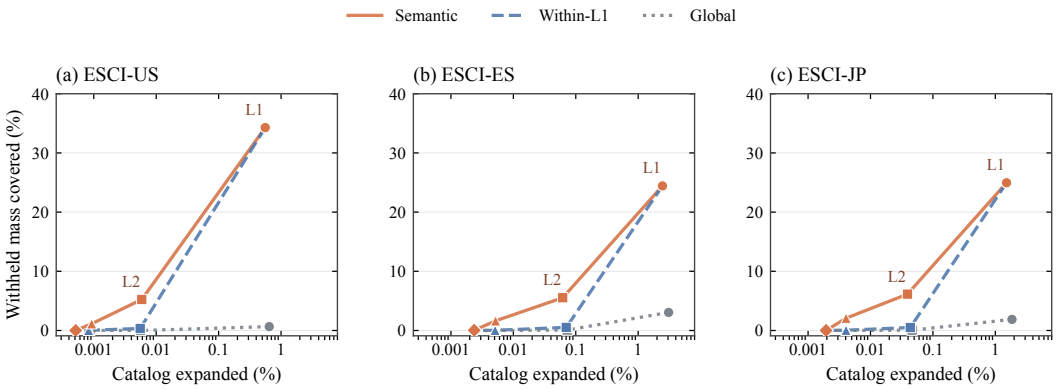


Figure 10: **Withheld coverage and catalog expansion at 25% retention.** Curves connect L1–L4 under uniform masking and linear relevance weights. Circles, squares, triangles, and diamonds denote successive depths. The horizontal axis is logarithmic. Semantic and within-L1 assignments coincide at L1 by construction. Coverage counts withheld relevance mass sharing a prefix with retained positives.

Queries receive equal weight within each locale. The 20 paired replicates are averaged within query before query-cluster bootstrap (4,000 draws) and paired sign randomization (10,000 draws). The primary test family compares semantic and global assignments on TV reduction and coverage gain at all four SID depths in all three locales, using uniform masking, linear weights, and 25% retention. Holm correction is applied jointly to these 24 tests. Crossed-bootstrap intervals additionally resample queries and replicates independently to account for variation in the finite set of random assignments and masks. Table 20 reports the complete set of primary comparisons. L4 coverage gains are $0.0055/0.0349/0.0065$ points on US/ES/JP, with corresponding Holm-adjusted p-values of $0.2476/0.0042/0.9969$.

Locale	Depth	TV reduction [95% CI]	p_{H}	Coverage gain (pp) [95% CI]	p_{H}
US	L1	0.2320 [0.2256, 0.2382]	0.0024	33.6331 [32.6809, 34.5729]	0.0024
US	L2	0.0373 [0.0348, 0.0398]	0.0024	5.2315 [4.8586, 5.6240]	0.0024
US	L3	0.0083 [0.0073, 0.0095]	0.0024	1.1337 [0.9796, 1.2986]	0.0024
US	L4	0.0000 [0.0000, 0.0001]	0.2476	0.0055 [0.0005, 0.0140]	0.2476
ES	L1	0.1458 [0.1369, 0.1547]	0.0024	21.3771 [20.0118, 22.7367]	0.0024
ES	L2	0.0388 [0.0346, 0.0433]	0.0024	5.4617 [4.8383, 6.1346]	0.0024
ES	L3	0.0116 [0.0095, 0.0140]	0.0024	1.6248 [1.2946, 1.9876]	0.0024
ES	L4	0.0003 [0.0001, 0.0005]	0.0042	0.0349 [0.0146, 0.0609]	0.0042
JP	L1	0.1601 [0.1507, 0.1694]	0.0024	23.0821 [21.7149, 24.4512]	0.0024
JP	L2	0.0433 [0.0391, 0.0476]	0.0024	6.0914 [5.4547, 6.7778]	0.0024
JP	L3	0.0149 [0.0127, 0.0173]	0.0024	2.0735 [1.7408, 2.4520]	0.0024
JP	L4	0.0001 [0.0000, 0.0002]	0.9969	0.0065 [0.0000, 0.0195]	0.9969

Table 20: Primary comparisons of semantic and global identifier assignments under linear relevance weights, uniform masking, and 25% retention. Positive differences favor semantic assignments. Brackets give pointwise crossed query-and-replicate bootstrap 95% intervals (4,000 draws). Adjusted p_{H} values use 10,000 paired query sign randomizations with Holm correction over all 24 tests.

False certainty and absent support. A parent is falsely resolved when its full observed positive set occupies at least two children but its retained positives occupy exactly one. We sum the reference

mass of such parents. Missing support is the reference mass in prefix groups with no retained positives (Figure 11). These two quantities measure collapsed branching and unobserved groups, respectively. At 25% retention, 71%–73% of reference L4 mass belongs to groups with no retained positives. Table 21 reports results by positive-count stratum.

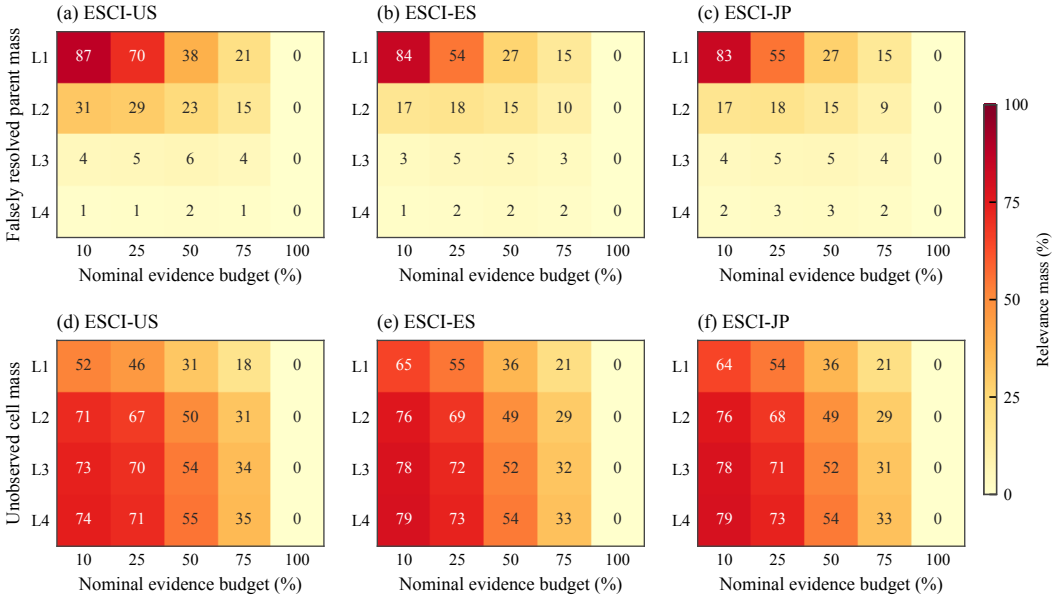


Figure 11: **Apparent resolution and missing support.** Results use semantic identifiers, uniform masking, and linear relevance weights. Top: reference mass on ambiguous parents that appear deterministic after subsampling. Bottom: reference mass in prefix groups with no retained positives. Percentages are rounded to integers.

Locale	n_q	Queries	TV L1	TV L2	TV L3	TV L4	Cov. L1	Cov. L2
US	2-3	1,810	0.417	0.551	0.570	0.573	27.16	3.81
US	4-7	1,756	0.571	0.770	0.794	0.801	28.82	3.89
US	8-15	917	0.431	0.717	0.763	0.777	48.97	7.94
US	16+	259	0.314	0.659	0.740	0.765	69.11	14.74
ES	2-3	371	0.496	0.557	0.564	0.572	12.89	2.62
ES	4-7	393	0.691	0.777	0.796	0.801	13.78	3.06
ES	8-15	349	0.574	0.729	0.763	0.778	28.80	6.69
ES	16+	305	0.450	0.688	0.745	0.764	47.18	10.93
JP	2-3	423	0.494	0.552	0.564	0.574	13.66	3.73
JP	4-7	452	0.667	0.778	0.795	0.805	17.32	3.46
JP	8-15	394	0.562	0.727	0.765	0.778	30.30	6.87
JP	16+	325	0.469	0.679	0.733	0.764	43.83	12.12

Table 21: Evidence recovery by positive-count stratum under the semantic hierarchy, uniform masking, linear relevance weights, and 25% retention. Coverage is the percentage of withheld relevance mass recovered. Results are shown for all four prespecified strata.

Properties of relevance projection. Prefix aggregation preserves relevance mass and contracts TV and JS. At 100% retention, the estimated and reference distributions coincide. SID permutations preserve catalog paths, within-L1 permutations preserve L1 relevance distributions, and product-level quantities are invariant to SID reassignment.

Invariance under restricted reassignment. A within-L1 permutation preserves $\pi_1(d)$ for every product. Both the reference and masked L1 distributions, their support, and the expanded catalog are identical to the semantic assignment. L1 serves as a control for the preserved resolution. At a one-to-one document partition, the TV error is $1 - \sum_{d \in S_q} \mu(d \mid q)$, where S_q is the retained positive set. This quantity is invariant to renaming unique document identifiers. SID collisions account for small departures at L4. Table 22 exhibits the corresponding equality at L1 and near-equality at L4.

Locale	Assignment	Target TV ↓ L1	L2	L3	L4	L1 support (%) Coverage	Expansion
US	Semantic	0.4709	0.6702	0.6992	0.7075	34.28	0.56
	Within-L1	0.4709	0.7050	0.7073	0.7075	34.28	0.56
	Global	0.7028	0.7075	0.7075	0.7075	0.65	0.65
ES	Semantic	0.5595	0.6884	0.7161	0.7275	24.43	2.45
	Within-L1	0.5595	0.7238	0.7275	0.7278	24.43	2.45
	Global	0.7053	0.7272	0.7278	0.7278	3.05	3.07
JP	Semantic	0.5547	0.6851	0.7138	0.7287	24.96	1.51
	Within-L1	0.5547	0.7250	0.7285	0.7288	24.96	1.51
	Global	0.7148	0.7284	0.7288	0.7288	1.88	1.85

Table 22: Target error and L1 support at 25% retention. Semantic and within-L1 assignments have identical L1 error, coverage, and expansion by construction. Table 18(a) compares their affected fine resolutions. L1 coverage and catalog expansion are percentages.

Equal catalog budgets. To equalize candidate counts, let $C_h(q, r)$ be the supported catalog size under assignment h in replicate r, and set $B(q, r) = \min_h C_h(q, r)$ across semantic, within-L1, and global assignments. Uniformly sampling $B(q, r)$ products from each supported set gives expected withheld coverage $\widetilde{R}_h(q, r) = R_h(q, r)B(q, r)/C_h(q, r)$. The common count depends only on catalog geometry and retained evidence; withheld labels enter the coverage evaluation. The expectation is taken over uniform samples from each supported candidate set.

At L2, the shared cap averages 8.12, 19.56, and 15.33 candidates per query in US, ES, and JP. Semantic expected coverage is 2.64%, 2.56%, and 3.12%, compared with 0.17%, 0.23%, and 0.24% for within-L1 reassignment. Semantic identifiers retain higher fine-level coverage at identical candidate counts within each query and replicate.

E REAL-PRODUCT CASE STUDIES

Figure 12 presents six ESCI queries with their judged products and SID prefixes. The examples highlight two aspects of multiresolution relevance: the depth at which relevant products separate in the SID hierarchy, and the distribution of relevance mass among competing children. Across queries, relevant products may diverge immediately at coarse levels or remain grouped until the final refinement, producing distinct supervision patterns along the same type of identifier hierarchy.

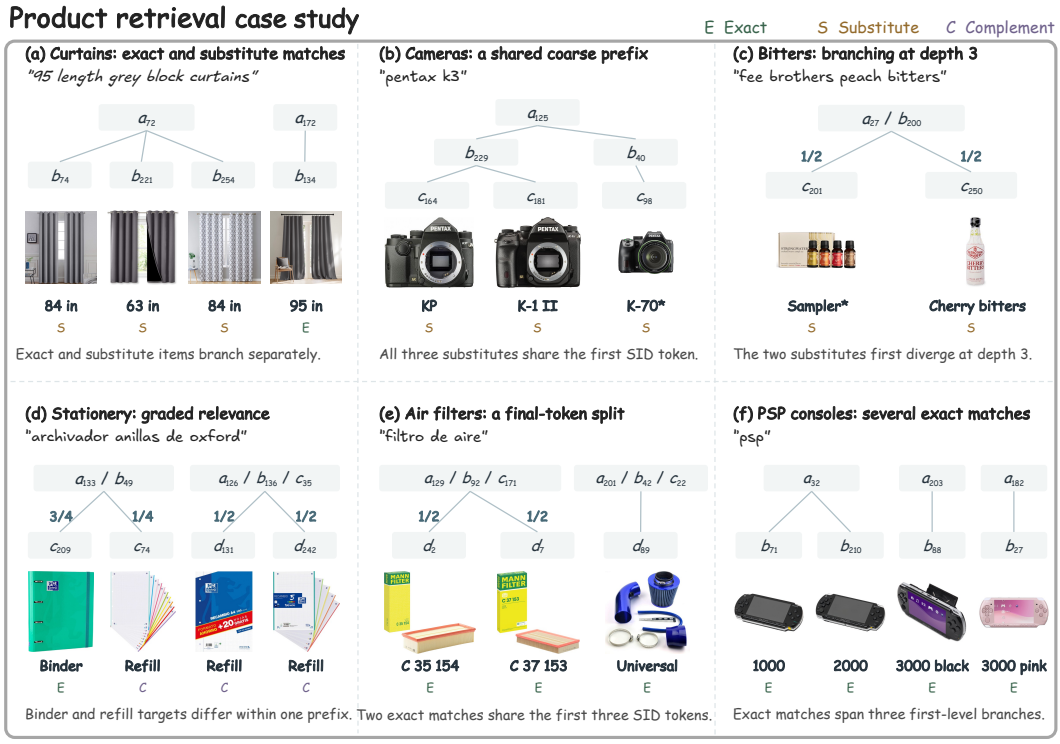


Figure 12: **Relevance structure in product retrieval.** Six ESCI queries exhibit different branching depths and local relevance distributions in the SID hierarchy. E, S, and C denote Exact, Substitute, and Complement judgments. Fractions indicate conditional relevance mass over the displayed children. Unary paths are compressed, and identifier-uniqueness suffixes are omitted. For starred products, the K-70 image includes a lens although the catalog listing is body-only, and the sampler image shows a four-bottle pack for a historical five-bottle listing.

Branching depth. The examples show branching structures. For the curtain query in (a), the shorter Substitute products fall under a_{72}, while the Exact 95-inch product belongs to a_{172}, so relevance separates at the first SID level. The four Exact PSP matches in (f) likewise span three first-level branches, showing that products with the same relevance grade need not share a coarse prefix. Other queries retain shared structure for several refinements. The camera substitutes in (b) share a_{125} before separating, while the two bitters products in (c) share the first two SID tokens and diverge at depth 3. The C 35 154 and C 37 153 air filters in (e) share three tokens and separate only at the final SID token, while the Universal filter occupies another coarse branch. Thus, the resolution at which relevance branches is query-dependent, ranging from the first decision to the final refinement.

Conditional relevance mass. Branching depth alone does not determine the supervision target. The displayed targets use E/S/C weights of $3/2/1$, normalized over positive children at each parent. In (d), the Exact binder and Complement refill below a_{133}/b_{49} receive conditional masses $3/4$ and $1/4$, reflecting their different relevance grades within the same local decision. The two Complement refills below $a_{126}/b_{136}/c_{35}$ each receive mass $1/2$. Equal local targets also arise for the two Substitute products in (c) and the two Exact filters in (e). SID structure determines which relevant products compete at a refinement, while relevance grades determine their relative mass. The same grade may occur in different branches, as in the PSP and curtain examples, while different grades may share a parent, as in the stationery example. Thus, document-level relevance induces different local target distributions across SID resolutions. Table 1 evaluates the effect of learning this structure at scale.