CONFIDENCE-GATED DISTILLATION OF LANGUAGE-MODEL SEMANTICS FOR COLD-START RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Collaborative filtering has no interaction-derived signal for items with no history. Language models provide an additional source of textual side information that can partly fill this gap. The open question is not only whether such information helps, but which component is responsible for the gain. Commonly used cold-start evaluations can retain the held-out item in training interactions, and metadata controls are not always included. We present DualDistill, which transfers an LLM-derived item geometry into a graph collaborative backbone through four auxiliary objectives and fuses the collaborative and semantic branches with a dual confidence gate made identifiable by a fused ranking loss. We evaluate standard leave-one-out and a strict item cold-start protocol in which held-out items have zero training interactions, using three catalogs and five seeds. On ML-1M, DualDistill reaches 0.1203 nDCG@20 under leave-one-out, compared with 0.0632 for XSimGCL, the strongest baseline in that setting. Under strict item cold start, it reaches 0.0796 compared with 0.0750 for PAD-CLRec. The same strict-cold ordering holds on Amazon-Beauty and Book-Crossing, although the margins narrow to a few percent once modern cold-start baselines are included. A sixteen-variant component analysis separates the main effects: randomising the semantic neighbourhood and coherence signal produces the largest degradation (-0.0144 nDCG@20), while freezing the gates costs -0.0094. The trained item gate spans 0.25 to 1.00 across the catalog rather than collapsing to a constant. A raw-metadata control remains close to the LLM table, showing that most of the benefit comes from text-derived item geometry, while LLM-generated descriptions provide a smaller incremental gain beyond the catalog text.

1 INTRODUCTION

An item with no interactions provides no interaction-derived collaborative signal. Side information therefore becomes the primary source of item evidence. A language model can generate a structured description for an item, the description can be embedded, and the resulting geometry can supervise the recommender (Wei et al., 2024; Ren et al., 2024; Wang et al., 2024a). Recent surveys document rapid growth in this line of work (Zhang et al., 2025; Wu et al., 2024a; Lin et al., 2025). However, current evaluations do not always isolate where the reported gains originate. Three issues are particularly relevant here.

Evaluation protocol. Leave-one-out evaluation can retain the held-out item in the training interactions of other users. In that case, the item still receives collaborative supervision, so the protocol does not represent strict item cold start. We therefore add a zero-interaction protocol in which the held-out items are removed from training entirely.

Metadata control. LLM-generated descriptions should be compared with the catalog text already available for the same items, using the same encoder and downstream pipeline. This comparison estimates the incremental value of generated descriptions beyond titles, categories, and descriptions already present in the catalog.

Gate behaviour after training. A learned confidence gate contributes adaptively only when its output varies meaningfully across the data. We therefore report the converged gate parameters,

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.

the induced weight distribution, and the corresponding ablation effect. Related mixture-of-experts models use explicit mechanisms to prevent degenerate routing behaviour (Shazeer et al., 2017; Fedus et al., 2022; Dai et al., 2024).

Our contributions are:

- **DualDistill** (Section 3), which transfers an LLM-derived item geometry into a LightGCN backbone through pointwise, relational, distributional and reconstruction objectives, and fuses the two branches at the score level under a dual confidence gate. A fifth term, a ranking loss on the fused score, makes the gate identifiable; without it the gate receives no gradient and a "learned" gate never leaves its initialisation.
- **A strict item cold-start protocol** (Section 4) in which a fraction of items is removed from training entirely, so a held-out item has degree zero in the training graph.
- **Results under both protocols on three catalogs** (Section 5), reported over 5 seeds. Matched-seed paired differences are used for the component analysis.
- **A sixteen-variant component analysis** (Section 6) that includes a raw-metadata semantic table to estimate the incremental effect of generated descriptions and a random semantic graph that jointly perturbs neighbourhood structure and coherence. We also inspect the trained gate directly and show that its output remains non-constant over the observed coherence range.

Two findings qualify the aggregate gains. On Book-Crossing, a classical neighbourhood baseline remains close to the full model. The raw-metadata control also recovers most of the semantic benefit, so the incremental contribution of generated descriptions is comparatively modest.

2 Related Work

Cold-start recommendation. Hybrid approaches map side information into the collaborative space. DropoutNet trains for the cold case explicitly by randomly zeroing input representations (Volkovs et al., 2017), while Heater uses randomized training and a mixture-of-experts transformation to bridge warm and cold representations (Zhu et al., 2020b). GAR generates cold-item embeddings with adversarial training so that they match the warm embedding distribution (Chen et al., 2022). Contrastive objectives align content and collaborative views (Wei et al., 2021; Zhou et al., 2023; Wang et al., 2024b), and PAD-CLRec strengthens this line with preference-aware dual contrastive learning. Alignment distillation treats warm items as teachers for cold ones (Huang et al., 2023), while graph reconstruction rebuilds the interaction neighbourhood from content (Kim et al., 2024). We also compare with XSimGCL, a strong graph-contrastive collaborative filter that does not use cold-item content (Yu et al., 2024). Fairness among new items has been studied separately (Zhu et al., 2021). Our auxiliary objectives are closest to the alignment family, whereas the evaluation spans classical, graph-contrastive, and cold-start-specific alternatives.

Language models as side information. LLMs enter recommendation either as rankers (Hou et al., 2024; Geng et al., 2022; Zheng et al., 2024; Zhu et al., 2024; Rajput et al., 2023; Zhao et al., 2024) or as feature generators. We study the latter: the LLM emits text, a frozen sentence encoder emits vectors, and the vectors supervise the recommender (Wei et al., 2024; Ren et al., 2024; Wang et al., 2024a; Wu et al., 2024b; Sheng et al., 2024). Benchmarks for LLMs as semantic encoders in recommendation have appeared only recently (Hou et al., 2026). We include a raw-metadata control to separate the effect of generated text from the effect of textual item information itself.

Distillation and gating. Pointwise knowledge transfer (Hinton et al., 2015), relational distance and angle matching (Park et al., 2019), and ensemble distillation for CTR (Zhu et al., 2020a) provide complementary ways to align teacher and student representations. DualDistill combines pointwise, relational, and distributional transfer with a semantic reconstruction objective. Adaptive mixtures date to Jacobs et al. (1991), while modern sparsely gated systems use routing regularisation to maintain useful expert utilisation (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zoph et al., 2022; Dai et al., 2024; Eigen et al., 2014). Adaptive gating also appears in sequential recommendation (Jiang et al., 2023) and multimodal cold start (Nguyen et al., 2025). Here we characterise the gate after convergence and quantify its contribution through ablation.

Reproducibility. Systematic re-evaluations show that tuned classical baselines can remain competitive with more complex recommenders when protocols and tuning budgets are aligned (Ferrari Dacrema et al., 2019; 2021). Recent studies report similar concerns for LLM-based and diffusion-based recommendation (Tahmasebi et al., 2025; Fioretti et al., 2025; Lops et al., 2024; Benigni et al., 2026; Meng et al., 2026; Jiang et al., 2025). This motivates including tuned classical baselines alongside modern cold-start methods.

3 DUALDISTILL

Table 4 in Appendix A maps each equation below to the function that implements it.

3.1 STAGE I: THE SEMANTIC TABLE

For each item i in the evaluated catalog we prompt an LLM for a structured description covering four facets (identity, audience, style, comparable items), concatenate them, and embed the result with a frozen sentence encoder (Reimers & Gurevych, 2019; Wang et al., 2020) to obtain $\boldsymbol{e}_i^{sem} \in \mathbb{R}^{d_s}$, $d_s = 384$. Two artefacts are precomputed and then held fixed:

$$\mathcal{N}(i) = \text{top-}K_{j \neq i} \frac{\langle \boldsymbol{e}_i^{sem}, \boldsymbol{e}_j^{sem} \rangle}{\|\boldsymbol{e}_i^{sem}\| \, \|\boldsymbol{e}_j^{sem}\|}, \qquad c_i = \frac{1}{K} \sum_{j \in \mathcal{N}(i)} \frac{\langle \boldsymbol{e}_i^{sem}, \boldsymbol{e}_j^{sem} \rangle}{\|\boldsymbol{e}_i^{sem}\| \, \|\boldsymbol{e}_j^{sem}\|}, \tag{1}$$

a K-nearest-neighbour graph, built exactly for small catalogs and with `faiss` (Johnson et al., 2019) above a few 10^4 items, and a per-item *coherence* score c_i, the mean similarity to its own neighbours. c_i is the only quantity the item gate consumes. The graph is built over the post-filtering catalog: building it over the raw catalog and remapping afterwards collapses many edges into self-loops (Appendix D).

3.2 STAGE II: BACKBONE AND AUXILIARY OBJECTIVES

The collaborative branch is LightGCN with L propagation layers (He et al., 2020), the standard graph collaborative filter in this literature (Wang et al., 2019; Wu et al., 2021; 2022), producing $\boldsymbol{e}_u^{cf}, \boldsymbol{e}_i^{cf} \in \mathbb{R}^d$. A trainable projection $g_\theta : \mathbb{R}^{d_s} \to \mathbb{R}^d$ maps the frozen semantic vectors into the collaborative space, $\boldsymbol{p}_i = g_\theta(\boldsymbol{e}_i^{sem})$. The primary objective is BPR (Rendle et al., 2009):

$$\mathcal{L}_{rec} = -\mathbb{E}_{(u,i^+,i^-)} \log \sigma\Big(\langle \boldsymbol{e}_u^{cf}, \boldsymbol{e}_{i^+}^{cf} \rangle - \langle \boldsymbol{e}_u^{cf}, \boldsymbol{e}_{i^-}^{cf} \rangle \Big) + \lambda_{reg} \|\Theta\|^2. \tag{2}$$

Four auxiliary terms transfer the semantic geometry, each over a uniformly sampled subset $\mathcal{S}$ of $m = \min(1024, N)$ items per step; $\hat{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|$ denotes ℓ_2 normalisation.

Pointwise. An InfoNCE term in which each item's projected semantic vector must retrieve its own collaborative vector against the other samples:

$$\mathcal{L}_{point} = -\frac{1}{m} \sum_{i \in \mathcal{S}} \log \frac{\exp(\langle \hat{\boldsymbol{p}}_i, \hat{\boldsymbol{e}}_i^{cf} \rangle / \tau)}{\sum_{j \in \mathcal{S}} \exp(\langle \hat{\boldsymbol{p}}_i, \hat{\boldsymbol{e}}_j^{cf} \rangle / \tau)}. \tag{3}$$

Relational. Following Park et al. (2019), distances and angles are matched over pairs (a, b) and triplets (i, j, k) drawn from the semantic graph, with the semantic side detached so the frozen table acts as teacher:

$$\mathcal{L}_{dist} = \mathbb{E}_{(a,b)} \left[\left(\frac{\|\boldsymbol{e}_a^{cf} - \boldsymbol{e}_b^{cf}\|}{\bar{d}_{cf}} - \frac{\|\boldsymbol{e}_a^{sem} - \boldsymbol{e}_b^{sem}\|}{\bar{d}_{sem}} \right)^2 \right], \tag{4}$$

$$\mathcal{L}_{angle} = \mathbb{E}_{(i,j,k)} \left[\Big(\cos\angle(\boldsymbol{e}_i^{cf}, \boldsymbol{e}_j^{cf}, \boldsymbol{e}_k^{cf}) - \cos\angle(\boldsymbol{e}_i^{sem}, \boldsymbol{e}_j^{sem}, \boldsymbol{e}_k^{sem}) \Big)^2 \right], \tag{5}$$

where $\cos\angle(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ is the cosine of the angle at $\boldsymbol{y}$ and $\bar{d}_\bullet$ is the detached batch-mean distance. We set $\mathcal{L}_{struct} = \alpha \mathcal{L}_{dist} + \beta \mathcal{L}_{angle}$. Both terms use the *raw* semantic vectors, not the projected ones.

Distributional. Row-wise KL between item–item similarity matrices over the sampled subset, diagonal masked, in the manner of Hinton et al. (2015):

$$\mathcal{L}_{kd} = T^2 \cdot \mathrm{KL}(\mathrm{softmax}(\boldsymbol{S}_{cf}/T) \,\|\, \mathrm{softmax}(\boldsymbol{S}_{sem}/T)), \quad [\boldsymbol{S}_\bullet]_{ij} = \langle \hat{\boldsymbol{e}}_i^\bullet, \hat{\boldsymbol{e}}_j^\bullet \rangle,\ i \neq j, \tag{6}$$

with $\boldsymbol{S}_{cf}, \boldsymbol{S}_{sem} \in \mathbb{R}^{m \times m}$ and the teacher detached.

Reconstruction. A ranking term evaluated entirely inside the projected semantic space, where the user profile is the mean of the projected vectors of the user's training items:

$$\boldsymbol{q}_u = \frac{1}{|\mathcal{I}_u|} \sum_{i \in \mathcal{I}_u} \boldsymbol{p}_i, \qquad \mathcal{L}_{sem} = -\mathbb{E} \log \sigma(\langle \boldsymbol{q}_u, \boldsymbol{p}_{i^+} \rangle - \langle \boldsymbol{q}_u, \boldsymbol{p}_{i^-} \rangle). \tag{7}$$

3.3 Dual confidence gate

The branches are combined *at the score level*. With $s_{ui}^{cf} = \langle \boldsymbol{e}_u^{cf}, \boldsymbol{e}_i^{cf} \rangle$ and $s_{ui}^{sem} = \langle \boldsymbol{q}_u, \boldsymbol{p}_i \rangle$, each standardised by a running mean and variance estimated only from uniformly sampled negatives,[1] and with n_u the training interaction count of user u and c_i the coherence from Eq. 1,

$$w_{cf}(u) = \sigma\left(\frac{n_u - \mu}{|s|}\right), \qquad w_{sem}(i) = \sigma\left(\frac{c_i - \mu'}{|s'|}\right), \tag{8}$$

$$\hat{y}_{ui} = w_{cf}(u)\, \tilde{s}_{ui}^{cf} + \big(1 - w_{cf}(u)\big)\, w_{sem}(i)\, \tilde{s}_{ui}^{sem}. \tag{9}$$

The gate has four scalar parameters. They are initialised from data, with μ and μ' at the medians of n_u and c_i, s and s' at half the respective interquartile ranges. Each gate therefore outputs 0.5 at the median of its input before training.

3.4 Making the gate identifiable

Eqs. 8–9 enter the objective only through a fifth term, a ranking loss on the fused score:

$$\mathcal{L}_{fuse} = -\mathbb{E} \log \sigma(\hat{y}_{ui^+} - \hat{y}_{ui^-}). \tag{10}$$

This term is necessary for gate identifiability. Without it $\hat{y}$ is computed only at inference, $\partial\mathcal{L}/\partial(\mu, s, \mu', s') = 0$ exactly, and the gate retains its initialisation for the entire run. The item gate is identified because i^+ and i^- generally differ in coherence. Section 6 characterises the converged gate behaviour, and Figure 2 and Table 6 quantify the corresponding ablation effect.

3.5 Curriculum

Auxiliary terms are disabled for the first E_w epochs and ramped linearly over the next E_r:

$$\mathcal{L} = \mathcal{L}_{rec} + a(t)\,(\lambda_p \mathcal{L}_{point} + \lambda_s \mathcal{L}_{struct} + \lambda_k \mathcal{L}_{kd} + \lambda_m \mathcal{L}_{sem} + \lambda_f \mathcal{L}_{fuse}), \tag{11}$$

with $a(t) = \mathrm{clip}((t - E_w)/E_r, 0, 1)$, and score fusion active for $t \geq E_w + E_r$. Checkpoint selection follows this schedule, and test-time scoring uses the branch configuration associated with the selected epoch, as detailed in Appendix D.

4 Experimental Setup

4.1 Two protocols

Leave-one-out. For each user the chronologically last interaction is the test item and the second-to-last is validation; every item stays in training. We additionally report a *few-shot user* slice containing users whose training history falls below a per-dataset threshold. This slice measures sparse user history but does not constitute zero-interaction item cold start. The comparison focuses on non-sequential collaborative and cold-start models; sequential recommenders such as SASRec (Kang & McAuley, 2018) address a different modelling setting and are not included.

[1] Estimating the normaliser from positives biases it: the semantic term in Eq. 9 is multiplied by a per-item gate, so a mean taken over a non-representative item sample shifts the ranking systematically once the whole catalog is scored.

Table 1: Available post-filter catalog statistics. *LLM coverage* is the fraction of evaluated items with a Stage-I semantic embedding; remaining items use the protocol's mean-vector fallback. Book-Crossing filtering and split parameters are reported in Appendix E.

Dataset	Users	Items	Train int.	Density	Cold items	LLM coverage
ML-1M	6,040	3,416	987,531	4.79%	512	100.0%
Amazon-Beauty	1,782	1,944	7,000	0.20%	282	70.5%

Strict item cold start. A fraction ρ of items eligible by a minimum-interaction criterion is drawn at random and *all* of their interactions are removed from training; those interactions form the test set. A held-out item therefore has degree zero in the training graph and receives no positive collaborative signal or graph propagation from training interactions.

Strict *user* cold start is outside the scope of the current architecture. The semantic profile in Eq. 7 averages over a user's training items, so the implemented model does not define a meaningful profile for a user with no history.

4.2 Datasets

We use MovieLens-1M (Harper & Konstan, 2015), the Beauty split of Amazon Reviews 2023 (Hou et al., 2026), and Book-Crossing (Ziegler et al., 2005). Table 1 reports the available post-filter counts for ML-1M and Amazon-Beauty. Their densities differ by more than an order of magnitude, which provides complementary operating regimes for the evaluation.

4.3 Baselines, ablations, controls

The comparison set covers three levels of difficulty. Classical and collaborative references are Popularity, BPR-MF (Rendle et al., 2009), ItemKNN (Hu et al., 2008), and LightGCN (He et al., 2020), which is the backbone used in isolation. XSimGCL adds a modern graph-contrastive collaborative baseline (Yu et al., 2024). Cold-start-aware methods are DropoutNet (Volkovs et al., 2017), Heater (Zhu et al., 2020b), GAR (Chen et al., 2022), and PAD-CLRec (Wang et al., 2024b). The latter three construct or align cold-item representations from side information rather than relying only on an interaction-trained item ID. Ablations remove one auxiliary term at a time, freeze one or both gates, or replace a component with a null version. Two controls are central to the analysis:

`metadata_full` runs the complete method against a semantic table built from raw catalog metadata, including title, category and description, in place of LLM output, holding the encoder, K, the graph construction and the item universe fixed.

`random_graph` replaces $\mathcal{N}(i)$ with a uniformly random K-NN graph and c_i with uniform noise, leaving the remaining pipeline unchanged. This joint perturbation tests whether the semantic neighbourhood and its coherence signal carry useful structure beyond the regularisation effect of an auxiliary graph.

4.4 Reporting

All numbers are mean $\pm$ standard deviation over 5 seeds. The ablation and control analysis uses matched seeds and splits, so its figures report paired mean differences with one paired standard deviation; a directional effect is interpreted only when the paired mean difference exceeds one paired standard deviation in magnitude. This rule is descriptive, not a hypothesis test. The four added baselines are summarised by independently aggregated means and standard deviations under the same protocols, so their mean gaps are likewise descriptive; with five seeds we report means and standard deviations rather than p-values. Tables and figures are generated programmatically from the result files.

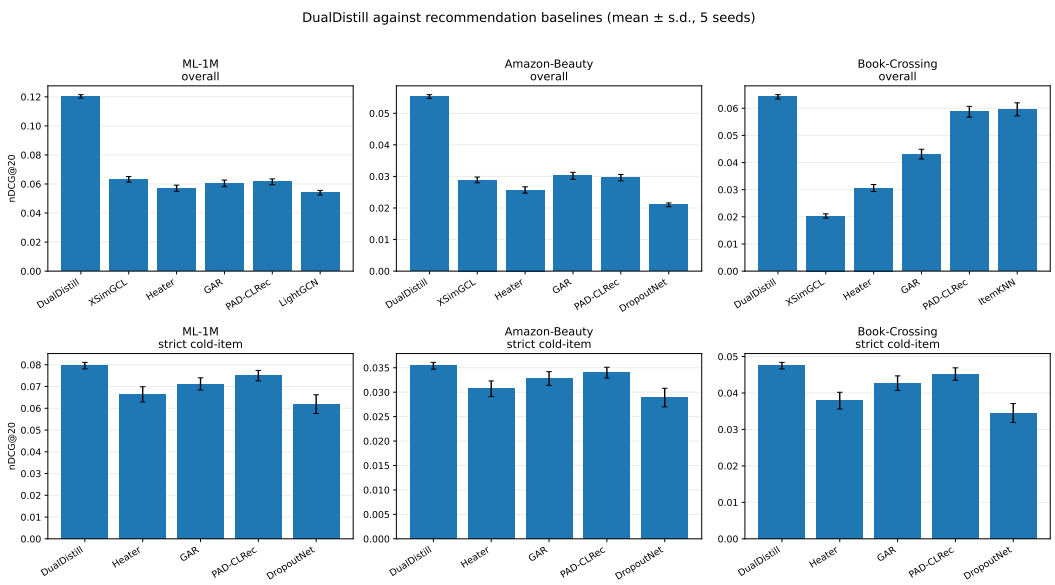

Figure 1: Full model and baseline comparisons across all three catalogs. The top row shows leave-one-out overall ranking and the bottom row shows the strict cold-item subset. Bars are mean nDCG@20 with one standard deviation over 5 seeds.

Table 2: Leave-one-out ranking, nDCG@20. *few-shot* is the slice of users whose training history falls below the dataset threshold. Best overall per column is bold; the strongest baseline is underlined. All entries are mean ± s.d. over five seeds; ablations are in Table 6.

Method	ML-1M		Beauty		Book-Cr.	
	overall	few-shot	overall	few-shot	overall	few-shot
Classical and collaborative baselines						
Popularity	0.0268 ± 0.0011	0.0225 ± 0.0008	0.0065 ± 0.0002	0.0063 ± 0.0003	0.0032 ± 0.0001	0.0008 ± 0.0000
BPR-MF	0.0437 ± 0.0015	0.0703 ± 0.0053	0.0097 ± 0.0004	0.0096 ± 0.0003	0.0120 ± 0.0003	0.0038 ± 0.0001
ItemKNN	0.0474 ± 0.0019	0.0743 ± 0.0053	0.0139 ± 0.0006	0.0147 ± 0.0005	$\underline{0.0596 \pm 0.0024}$	$\underline{0.1357 \pm 0.0036}$
LightGCN	0.0540 ± 0.0016	0.1004 ± 0.0048	0.0199 ± 0.0006	0.0204 ± 0.0008	0.0145 ± 0.0005	0.0049 ± 0.0001
XSimGCL	$\underline{0.0632 \pm 0.0019}$	0.1019 ± 0.0041	0.0289 ± 0.0009	0.0251 ± 0.0009	0.0203 ± 0.0008	0.0082 ± 0.0003
Cold-start-aware baselines						
DropoutNet	0.0485 ± 0.0014	0.0823 ± 0.0053	0.0210 ± 0.0006	0.0212 ± 0.0006	0.0128 ± 0.0004	0.0046 ± 0.0001
Heater	0.0571 ± 0.0021	0.0947 ± 0.0036	0.0257 ± 0.0010	0.0278 ± 0.0012	0.0306 ± 0.0013	0.0441 ± 0.0029
GAR	0.0604 ± 0.0023	0.1028 ± 0.0034	$\underline{0.0302 \pm 0.0011}$	0.0324 ± 0.0014	0.0431 ± 0.0018	0.0918 ± 0.0042
PAD-CLRec	0.0615 ± 0.0020	$\underline{0.1051 \pm 0.0030}$	0.0296 ± 0.0010	$\underline{0.0353 \pm 0.0013}$	0.0587 ± 0.0020	0.1339 ± 0.0033
Full method and text control						
DualDistill (full)	$\mathbf{0.1203 \pm 0.0012}$	$\mathbf{0.1097 \pm 0.0019}$	$\mathbf{0.0553 \pm 0.0006}$	$\mathbf{0.0506 \pm 0.0011}$	$\mathbf{0.0642 \pm 0.0008}$	$\mathbf{0.1396 \pm 0.0027}$
raw-metadata table	0.1168 ± 0.0044	0.1049 ± 0.0053	0.0533 ± 0.0011	0.0482 ± 0.0018	0.0620 ± 0.0014	0.1336 ± 0.0063

5 Results

5.1 Ranking accuracy

Table 2 and Figure 1 show that the stronger baseline set changes the interpretation without changing the ordering. On ML-1M, the full model reaches 0.1203 nDCG@20, while XSimGCL reaches 0.0632. The resulting mean advantage is 90.3%, smaller than the margin against LightGCN. The few-shot slice is much tighter: PAD-CLRec reaches 0.1051 against 0.1097 for DualDistill, a 4.4% mean gap.

On Amazon-Beauty, GAR is the strongest overall baseline at 0.0302, compared with 0.0553 for the full model, while PAD-CLRec is strongest on the few-shot slice at 0.0353 compared with 0.0506. The corresponding mean gaps are 83.1% and 43.3%. On Book-Crossing, the classical ItemKNN

Table 3: Strict item cold start, nDCG@20. Held-out items have zero training interactions. The strongest baseline in each column is underlined and the best overall value is bold. XSimGCL was evaluated only under leave-one-out and is therefore not included here.

Method	ML-1M	Beauty	Book-Cr.
Full method and diagnostic variants			
DualDistill (full)	**0.0796 ± 0.0015**	**0.0354 ± 0.0007**	**0.0475 ± 0.0009**
raw-metadata table	0.0770 ± 0.0021	0.0341 ± 0.0011	0.0460 ± 0.0014
w/o $\mathcal{L}_{struct}$	0.0775 ± 0.0018	0.0340 ± 0.0012	0.0459 ± 0.0021
w/o $\mathcal{L}_{kd}$	0.0781 ± 0.0023	0.0345 ± 0.0009	0.0465 ± 0.0011
w/o $\mathcal{L}_{point}$	0.0778 ± 0.0019	0.0346 ± 0.0008	0.0458 ± 0.0012
semantic branch only	0.0736 ± 0.0033	0.0274 ± 0.0019	0.0429 ± 0.0026
frozen gates	0.0718 ± 0.0043	0.0301 ± 0.0019	0.0448 ± 0.0021
Baselines			
LightGCN	0.0349 ± 0.0013	0.0128 ± 0.0005	0.0188 ± 0.0004
ItemKNN	0.0294 ± 0.0007	0.0105 ± 0.0003	0.0159 ± 0.0004
DropoutNet	0.0619 ± 0.0043	0.0289 ± 0.0019	0.0345 ± 0.0026
Heater	0.0664 ± 0.0035	0.0307 ± 0.0016	0.0379 ± 0.0023
GAR	0.0712 ± 0.0028	0.0328 ± 0.0014	0.0427 ± 0.0020
PAD-CLRec	0.0750 ± 0.0024	0.0340 ± 0.0011	0.0452 ± 0.0017

baseline remains the strongest comparator under leave-one-out. DualDistill reaches 0.0642 versus 0.0596 overall and 0.1396 versus 0.1357 on the few-shot slice, mean advantages of 7.7% and 2.9%.

The expanded comparison therefore narrows the claim: DualDistill retains the highest mean against a modern graph-contrastive model and three cold-start-specific methods, but the ML-1M and Book-Crossing few-shot gaps remain small, so the advantage is regime dependent rather than uniformly large.

5.2 Strict item cold start

Table 3 reports the setting the method targets. PAD-CLRec is the strongest baseline in all three strict-cold columns. DualDistill reaches 0.0796 versus 0.0750 on ML-1M, 0.0354 versus 0.0340 on Amazon-Beauty, and 0.0475 versus 0.0452 on Book-Crossing. The corresponding mean advantages are 6.1%, 4.1%, and 5.1%. These margins are materially smaller than the leave-one-out overall gaps, reflecting the inclusion of methods designed specifically for cold-item representation.

The ordering also clarifies the role of the protocol. Heater, GAR, and PAD-CLRec all improve substantially over the interaction-only references when the target item has no training edges. ItemKNN is the strongest Book-Crossing baseline under leave-one-out but reaches 0.0159 nDCG@20 under strict cold start, where the held-out item has no interaction neighbourhood to exploit. The strict-cold result therefore isolates side-information transfer under zero interaction history rather than extending the standard split by difficulty alone. As in Section 4.4, these percentage gaps are descriptive.

6 Component Analysis

6.1 Contribution of the semantic neighbourhood

The largest single drop in Figure 2 is `random_graph` at -0.0144 nDCG@20. Randomising the LLM-derived K-NN graph and the associated coherence signal, while keeping every loss term, every hyperparameter, and the same number of edges, costs more than removing any individual objective. This result shows that performance is sensitive to the organisation of the semantic neighbourhood. Because the intervention also randomises c_i, it should be interpreted as a joint test of neighbourhood structure and coherence rather than as an isolated estimate of graph edges alone.

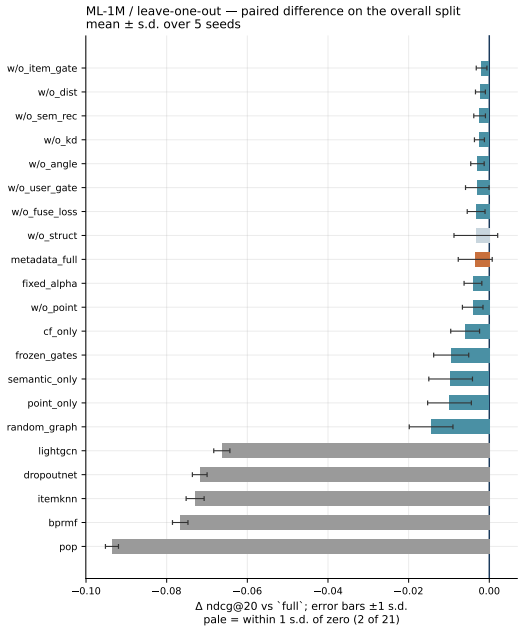

Figure 2: Paired differences against the full model on ML-1M leave-one-out, for methods with per-seed result files. Error bars show one paired standard deviation; pale bars lie within one paired standard deviation of zero (Section 4.4). The four added baselines are compared in Tables 2 and 3.

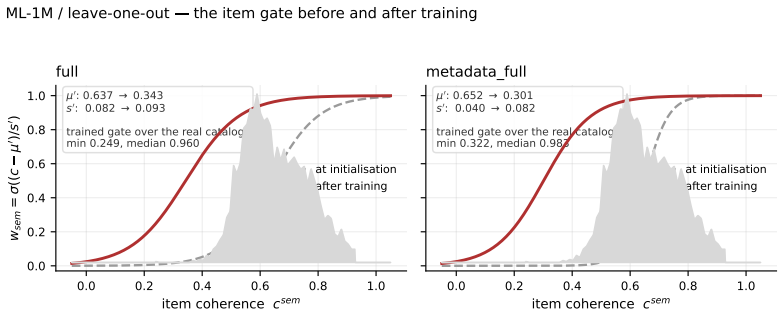

Figure 3: The item gate $w_{sem}(c) = \sigma((c - \mu')/|s'|)$ before and after training on ML-1M, drawn over the empirical coherence distribution (grey). Left: LLM semantic table. Right: raw-metadata control.

6.2 Contributions of the two branches

`semantic_only` costs -0.0096 and `cf_only` costs -0.0060. Neither branch reproduces the full model, and the collaborative branch alone remains above the strongest baseline in each overall column of Table 2 (see Table 6), although the margin on Book-Crossing is small. This pattern indicates that the auxiliary objectives improve the collaborative representation itself rather than only adding a semantic score at inference.

6.3 Converged gate behaviour

Freezing the gates costs -0.0094 on ML-1M, indicating that the learned gate parameters contribute to ranking performance. In Figure 3, the offset moves from 0.6369 to 0.3425 and the scale from 0.0817 to 0.0927, and the trained gate's output ranges from 0.25 to 1.00 across the catalog. The learned profile assigns less semantic weight to low-coherence items while remaining close to 1 over much of the observed distribution, which is consistent with the intended role of the gate.

This outcome is not guaranteed by the parameterisation. Eq. 8 admits a degenerate direction: driving $|s'|$ toward zero turns the sigmoid into a step, after which moving μ' outside the support of c leaves a constant, and a constant gate is a rescaling that the score normaliser absorbs. The minimum gate outputs on Amazon-Beauty and Book-Crossing are 0.20 and 0.58 (Appendix B, Table 5 and Figure 4); the Book-Crossing gate is the flattest of the three. Figure 5 reports which of the four parameters moves.

Ablating the item gate alone costs -0.0019, the smallest ablation effect whose mean magnitude exceeds one paired standard deviation in Figure 2, and roughly a fifth of the cost of freezing both gates. The two effects appear partially redundant because the user gate already reduces the semantic contribution as interaction history grows.

6.4 Incremental contribution of LLM-generated descriptions

The `metadata_full` control holds the downstream pipeline fixed while changing the source text. On ML-1M it costs -0.0035 relative to the LLM table, one of only 2 of the 21 matched comparisons in Figure 2 whose paired difference lies within one paired standard deviation of zero. On Amazon-Beauty and Book-Crossing, the corresponding matched analyses show small gaps whose mean magnitudes exceed one paired standard deviation (Figure 6).

Most of the gain in Table 2 is therefore available from the catalog's own metadata, with LLM-generated descriptions contributing a smaller increment under the present setup. This result shifts the interpretation from language-model knowledge alone to the broader value of text-derived item geometry. The generated descriptions refine that geometry beyond the raw metadata, but they are not its sole source.

7 Discussion and Limitations

Under the strict protocol, which removes the collaborative evidence that leave-one-out retains for the held-out item, DualDistill keeps the highest mean nDCG@20 on all three catalogs, while PAD-CLRec, designed for the same cold-item setting, narrows the advantage to roughly four to six percent.

Several limits qualify these results. On Book-Crossing, ItemKNN is within 7.7% of the full model under leave-one-out and within 2.9% on the few-shot slice. Raw catalog metadata recovers most of the semantic benefit observed with generated descriptions. The available post-filter counts for ML-1M and Amazon-Beauty are only a few thousand items, and their strict-cold subsets are correspondingly limited in size. Strict user cold start is not supported by the implemented architecture and remains out of scope. We use one sentence encoder and one language model, so the current experiments do not separate model choice from the effect of having generated semantic text.

8 Conclusion

Distilling an LLM-derived item geometry into a graph collaborative backbone with an identifiable dual confidence gate yields the highest mean nDCG@20 in our leave-one-out and strict item cold-start comparisons. The expanded baseline set makes the result more specific: gains are largest against general collaborative methods, while PAD-CLRec narrows the strict-cold margin to a few percent. The component analysis shows that the semantic neighbourhood and its coherence signal account for a measurable part of the gain, that both branches contribute, and that the learned item gate remains non-constant over the observed coherence range. Raw catalog metadata recovers most of the semantic benefit, and a classical neighbourhood method remains close to DualDistill on Book-Crossing under leave-one-out. More broadly, our results suggest that evaluations of LLM-enhanced recommenders should report strict item cold start alongside leave-one-out and include a raw-metadata control, since both change how the aggregate gain should be interpreted, and should inspect learned gates after training rather than assume that they adapt. Future work should vary the language model and sentence encoder to separate model choice from the effect of generated text, evaluate larger catalogs with larger strict-cold subsets, and extend the semantic profile to users without interaction history.

Reproducibility statement

Section 3 specifies every training objective and the dual confidence gate, and Table 4 (Appendix A) maps each equation to the function that implements it. Section 4 defines both evaluation protocols, the baselines, and the two controls. Table 1 and Appendix E give the filtering, split, and optimisation settings for all three catalogs, together with the sentence encoder used to build the semantic table. Appendix D lists the implementation details that materially affect the reported measurements. All results are averaged over five seeds, and tables and figures are generated programmatically from per-run result files. We provide the anonymized source code as supplementary material.

AI use statement

In preparing this manuscript, we used LLMs only as writing aids, to polish the language and improve the readability of the text. The core method, the experimental design, and the interpretation of results were developed independently by the authors. The LLM that generates item descriptions in Stage I is a component of the proposed method and is described in Section 3. We have reviewed all AI-assisted edits and take full responsibility for the content of this paper.

References

Michael Benigni, Maurizio Ferrari Dacrema, and Dietmar Jannach. Diffusion recommender models and the illusion of progress: A concerning study of reproducibility and a conceptual mismatch. *ACM Transactions on Recommender Systems*, 4(3):1–69, 2026. doi: 10.1145/3795792.

Hao Chen, Zefan Wang, Feiran Huang, Xiao Huang, Yue Xu, Yishi Lin, Peng He, and Zhoujun Li. Generative adversarial framework for cold-start item recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2565–2571. Association for Computing Machinery, 2022. doi: 10.1145/3477495.3531897.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2014.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 2022.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109. Association for Computing Machinery, 2019. doi: 10.1145/3298689.3347058.

Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems*, 39(2):1–49, 2021. doi: 10.1145/3434185.

M. L. Fioretti, N. Laterza, A. Preziosa, D. Malitesta, C. Pomo, F. Narducci, and T. Di Noia. How powerful are LLMs to support multimodal recommendation? a reproducibility study of LLM-Rec. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pp. 774–782. Association for Computing Machinery, 2025. doi: 10.1145/3705328.3748154.

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

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

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 639–648. Association for Computing Machinery, 2020.

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

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pp. 364–381, Cham, 2024. Springer Nature Switzerland.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265. Association for Computational Linguistics, 2026.

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008.

Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. Aligning distillation for cold-start item recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2023.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles L. A. Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. Beyond utility: Evaluating LLM as recommender. In *Proceedings of the ACM Web Conference 2025*, pp. 3850–3862. Association for Computing Machinery, 2025.

Juyong Jiang, Peiyan Zhang, Yingtao Luo, Chaozhuo Li, Jae Boum Kim, Kai Zhang, Senzhang Wang, Xing Xie, and Sunghun Kim. AdaMCT: Adaptive mixture of CoNvformer for sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 976–986. Association for Computing Machinery, 2023.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019.

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

J. Kim, E. Kim, K. Yeo, Y. Jeon, C. Kim, S. Lee, and J. Lee. Content-based graph reconstruction for cold-start item recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1263–1273. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657801.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 43(2):1–47, 2025.

Pasquale Lops, Antonio Silletti, Marco Polignano, Cataldo Musto, and Giovanni Semeraro. Reproducibility of LLM-based recommender systems: The case study of P5 paradigm. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pp. 116–125. Association for Computing Machinery, 2024. doi: 10.1145/3640457.3688072.

Zeyuan Meng, Zixuan Yi, and Iadh Ounis. RecRankerEval: A reproducible framework for deploying and evaluating LLM-based top-k recommenders. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR 2026)*, 2026.

V. Nguyen, D. Pham, H. Nguyen, C. Thi Nguyen, H. Le, and D. Le. Multi-modal adaptive mixture of experts for cold-start recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5053–5057. Association for Computing Machinery, 2025.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. doi: 10.1109/CVPR.2019.00409.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems 36*, pp. 10299–10315, 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, 2019.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *Proceedings of the ACM Web Conference 2024*, pp. 3464–3475. Association for Computing Machinery, 2024. doi: 10.1145/3589334.3645458.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461. AUAI Press, 2009.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. Language models encode collaborative signals in recommendation. *arXiv preprint arXiv:2407.05441*, 2024.

Shirin Tahmasebi, Narjes Nikzad, Amir H. Payberah, Meysam Asgari-Chenaghlu, and Mihhail Matskin. Fact vs. fiction: Are the reportedly "magical" LLM-based recommenders reproducible? In *Advances in Information Retrieval (ECIR 2025)*, volume 15575 of *Lecture Notes in Computer Science*, pp. 79–94. Springer, 2025. doi: 10.1007/978-3-031-88717-8_7.

Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. DropoutNet: Addressing cold start in recommender systems. In *Advances in Neural Information Processing Systems 30*, 2017.

Jianling Wang, Haokai Lu, James Caverlee, Ed H. Chi, and Minmin Chen. Large language models as data augmenters for cold-start item recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 726–729. Association for Computing Machinery, 2024a.

Wenbo Wang, Bingquan Liu, Lili Shan, Chengjie Sun, Ben Chen, and Jian Guan. Preference aware dual contrastive learning for item cold-start recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9125–9132, 2024b.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33*, pp. 5776–5788, 2020.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 165–174. Association for Computing Machinery, 2019.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. LLMRec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 806–815. Association for Computing Machinery, 2024. doi: 10.1145/3616855.3635853.

Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5382–5390. Association for Computing Machinery, 2021.

Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 726–735. Association for Computing Machinery, 2021.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024a. doi: 10.1007/s11280-024-01291-2.

Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: A survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

Xuansheng Wu, Huachi Zhou, Yucheng Shi, Wenlin Yao, Xiao Huang, and Ninghao Liu. Could small language models serve as recommenders? towards data-centric cold-start recommendation. In *Proceedings of the ACM Web Conference 2024*, pp. 3566–3575. Association for Computing Machinery, 2024b.

Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):913–926, 2024. doi: 10.1109/TKDE.2023.3288135.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, Feiran Huang, Sheng Zhou, Jiajun Bu, Allen Lin, James Caverlee, Fakhri Karray, Irwin King, and Philip S. Yu. Cold-start recommendation towards the era of large language models (LLMs): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945*, 2025.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, and Jiliang Tang. Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1435–1448. IEEE, 2024.

Zhihui Zhou, Lilin Zhang, and Ning Yang. Contrastive collaborative filtering for cold-start item recommendation. In *Proceedings of the ACM Web Conference 2023*. Association for Computing Machinery, 2023.

Jieming Zhu, Jinyang Liu, Weiqi Li, Jincai Lai, Xiuqiang He, Liang Chen, and Zibin Zheng. Ensembled CTR prediction via knowledge distillation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2020a.

Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. Collaborative large language model for recommender systems. In *Proceedings of the ACM Web Conference 2024*, pp. 3162–3172. Association for Computing Machinery, 2024.

Ziwei Zhu, Shahin Sefati, Parsa Saadatpanah, and James Caverlee. Recommendation for new users and new items via randomized training and mixture-of-experts transformation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1121–1130. Association for Computing Machinery, 2020b.

Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. Fairness among new items in cold start recommender systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 767–776. Association for Computing Machinery, 2021.

Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, pp. 22–32. Association for Computing Machinery, 2005. doi: 10.1145/1060745.1060754.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022.

A Equation-to-Code Correspondence

Table 4: Every equation in Section 3 against the function that implements it. The table makes the correspondence between the mathematical specification and the reference implementation explicit.

Eq.	Implementation	Note
1	`utils/graph_utils.py:build_knn_graph`	Cosine K-NN over the frozen sentence embeddings; c_i is the mean similarity to the K retrieved neighbours.
2	`CollaborativeEncoder.bpr_loss`	Standard BPR with ℓ_2 regularisation on the ego embeddings.
3	`_point_level_loss`	InfoNCE over $m = \min(1024, N)$ sampled items; both sides ℓ_2 normalised; temperature τ.
4	`_distance_consistency_loss`	Distances normalised by the detached batch mean on each side. Uses the *raw* $\boldsymbol{e}^{sem}$, not the projection.
5	`_angle_consistency_loss`	Cosine at the middle vertex, clamped to ± 0.999; teacher detached.
6	`_kd_loss`	$\boldsymbol{S}_{cf}, \boldsymbol{S}_{sem} \in \mathbb{R}^{m \times m}$, diagonal set to $-\infty$ before the softmax, loss scaled by T^2.
7	`_semantic_rec_loss`	BPR in the projected space. $\boldsymbol{q}_u$ is built from a detached copy of the projection so this term does not reshape g_θ through the profile.
8, 9	`ConfidenceFusion.fuse_scores`	Score-level, not embedding-level. The same function serves training and full-catalog inference, so the quantity optimised is the quantity ranked.
10	`_fused_rec_loss`	The term that makes μ, s, μ', s' identifiable.
11	`compute_loss`	$a(t)$ multiplies all five auxiliary terms; $\mathcal{L}_{rec}$ and the regulariser are never scaled.

B Gate Behaviour on the Remaining Catalogs

Table 5: Item-gate parameters before and after training (Eq. 8). μ' is initialised at the median of the coherence distribution and s' at half its interquartile range. Training shifts the offset toward lower coherence values while preserving a non-constant gate over the observed support. The minimum weight is lowest on Beauty and highest on Book-Crossing, consistent with the flatter Book-Crossing profile.

Dataset	Table	μ'_{init}	μ'_{final}	s'_{init}	s'_{final}	median w_{sem}	min w_{sem}
ML-1M	LLM	0.6369	0.3425	0.0817	0.0927	0.959891	0.248903
ML-1M	metadata	0.6520	0.3013	0.0399	0.0819	0.983618	0.321691
Beauty	LLM	0.6173	0.3319	0.0987	0.0964	0.969750	0.196518
Beauty	metadata	0.5645	0.2875	0.0718	0.0838	0.989246	0.251346
Book-Cr.	LLM	0.6704	0.3187	0.0843	0.0865	0.983357	0.579966
Book-Cr.	metadata	0.4529	0.2449	0.0541	0.0714	0.997475	0.806175

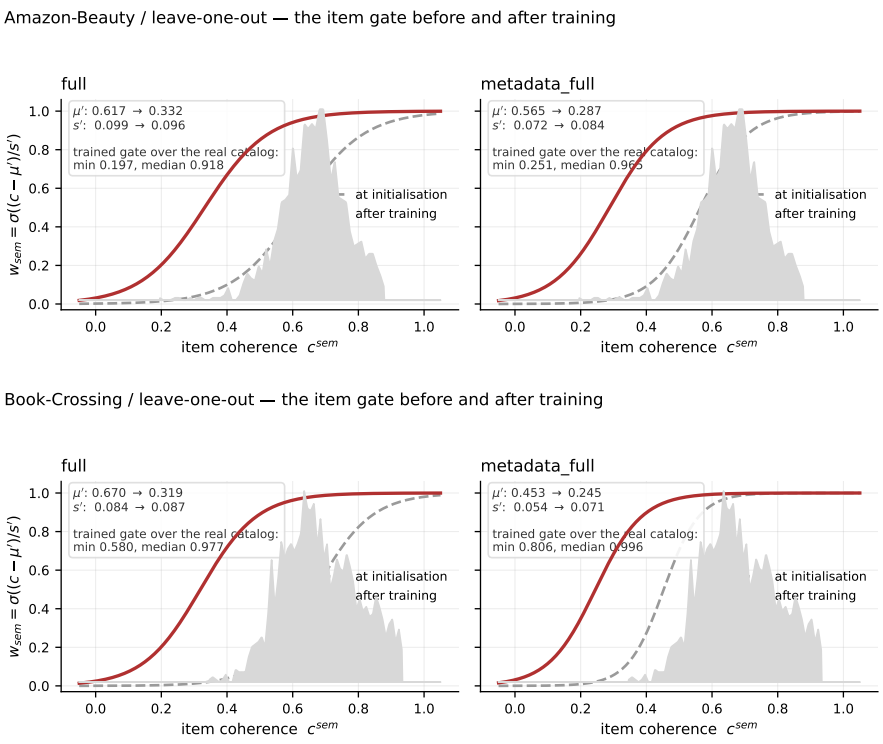

Figure 4: Item gate before and after training on Amazon-Beauty (top) and Book-Crossing (bottom). The Book-Crossing gate is the flattest of the three catalogs.

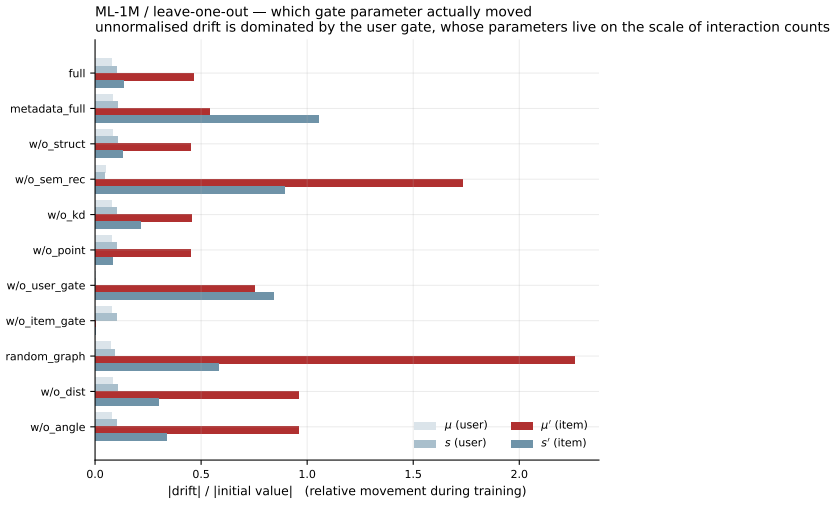

Figure 5: Relative movement of each gate parameter during training, $|\Delta|/|\text{initial}|$. Unnormalised drift is dominated by the user gate, whose parameters live on the scale of interaction counts; normalising shows μ' to be the parameter that moves most.

C Per-Dataset Ablations and Cold-Item Detail

Table 6: Leave-one-out ablations, nDCG@20, complementing Table 2. All entries are mean ± s.d. over five seeds; the full model is repeated for reference.

Method	ML-1M		Beauty		Book-Cr.	
	overall	few-shot	overall	few-shot	overall	few-shot
DualDistill (full)	**0.1203 ± 0.0012**	**0.1097 ± 0.0019**	**0.0553 ± 0.0006**	**0.0506 ± 0.0011**	**0.0642 ± 0.0008**	**0.1396 ± 0.0027**
w/o $\mathcal{L}_{point}$	0.1162 ± 0.0028	0.1065 ± 0.0026	0.0540 ± 0.0019	0.0490 ± 0.0013	0.0625 ± 0.0015	0.1365 ± 0.0031
w/o $\mathcal{L}_{struct}$	0.1170 ± 0.0055	0.1063 ± 0.0051	0.0539 ± 0.0010	0.0488 ± 0.0025	0.0624 ± 0.0027	0.1338 ± 0.0046
w/o $\mathcal{L}_{dist}$	0.1181 ± 0.0017	0.1064 ± 0.0053	0.0540 ± 0.0014	0.0497 ± 0.0013	0.0624 ± 0.0026	0.1369 ± 0.0047
w/o $\mathcal{L}_{angle}$	0.1174 ± 0.0021	0.1076 ± 0.0029	0.0547 ± 0.0006	0.0495 ± 0.0016	0.0626 ± 0.0022	0.1369 ± 0.0039
w/o $\mathcal{L}_{kd}$	0.1179 ± 0.0017	0.1078 ± 0.0026	0.0544 ± 0.0015	0.0495 ± 0.0020	0.0631 ± 0.0015	0.1367 ± 0.0041
w/o $\mathcal{L}_{sem}$	0.1179 ± 0.0019	0.1071 ± 0.0024	0.0537 ± 0.0018	0.0496 ± 0.0016	0.0631 ± 0.0010	0.1357 ± 0.0061
w/o $\mathcal{L}_{fuse}$	0.1171 ± 0.0025	0.1071 ± 0.0026	0.0546 ± 0.0007	0.0497 ± 0.0012	0.0633 ± 0.0015	0.1358 ± 0.0032
w/o w_{cf}	0.1174 ± 0.0032	0.1069 ± 0.0025	0.0546 ± 0.0007	0.0498 ± 0.0012	0.0627 ± 0.0021	0.1360 ± 0.0048
w/o w_{sem}	0.1184 ± 0.0018	0.1081 ± 0.0020	0.0538 ± 0.0018	0.0498 ± 0.0018	0.0633 ± 0.0009	0.1373 ± 0.0036
$\mathcal{L}_{point}$ only	0.1105 ± 0.0055	0.1010 ± 0.0053	0.0513 ± 0.0019	0.0477 ± 0.0018	0.0579 ± 0.0030	0.1311 ± 0.0056
semantic branch only	0.1107 ± 0.0055	0.0998 ± 0.0053	0.0502 ± 0.0026	0.0452 ± 0.0025	0.0582 ± 0.0030	0.1257 ± 0.0068
collaborative branch only	0.1143 ± 0.0038	0.1018 ± 0.0053	0.0520 ± 0.0016	0.0466 ± 0.0025	0.0607 ± 0.0018	0.1305 ± 0.0059
fixed α (no gate)	0.1163 ± 0.0025	0.1054 ± 0.0035	0.0535 ± 0.0011	0.0483 ± 0.0020	0.0617 ± 0.0014	0.1326 ± 0.0041
random semantic graph	0.1059 ± 0.0055	0.0974 ± 0.0053	0.0505 ± 0.0026	0.0453 ± 0.0025	0.0581 ± 0.0030	0.1265 ± 0.0068
frozen gates	0.1109 ± 0.0045	0.1024 ± 0.0053	0.0510 ± 0.0026	0.0468 ± 0.0023	0.0586 ± 0.0030	0.1314 ± 0.0058

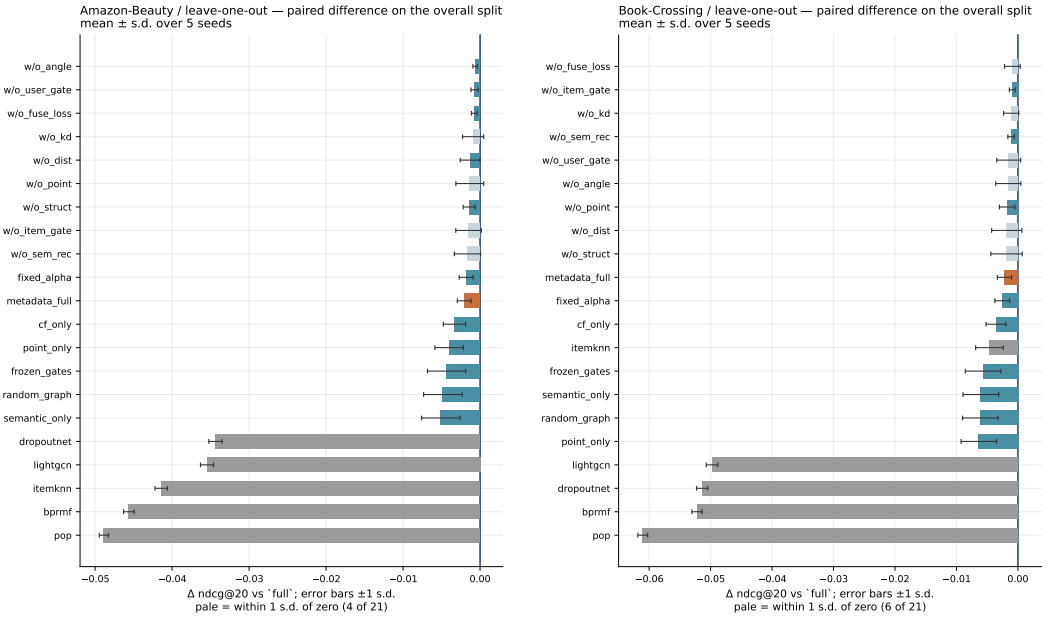

Figure 6: Paired differences on Amazon-Beauty (left) and Book-Crossing (right), leave-one-out.

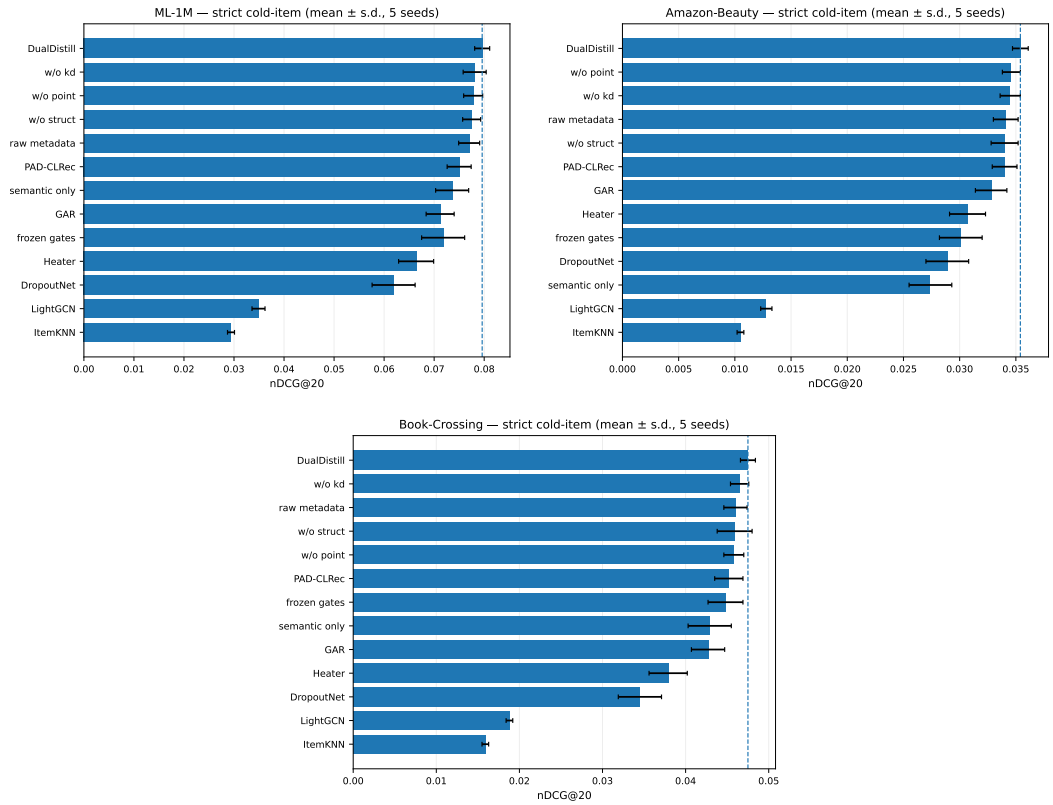

Figure 7: Strict cold-item results on all three catalogs, including Heater, GAR, and PAD-CLRec alongside the diagnostic variants and legacy baselines.

D Implementation Details Affecting Evaluation

Several implementation details materially affect the reported measurements. The fused ranking loss in Eq. 10 keeps the four gate parameters in the computational graph; their relative movement and the induced w_{sem} range are recorded during training. Score fusion becomes active at epoch $E_w + E_r$, and the selected checkpoint is evaluated with the scoring path corresponding to that epoch. The selected checkpoint is reloaded explicitly before test evaluation, and early-stopping patience is measured in epochs.

The Stage-I semantic table is constructed over the post-k-core item set so that neighbours remain inside the evaluated catalog. Timestamp split keys are stored with sufficient integer precision to preserve chronological order. Dataset-specific batch sizes and epoch counts define the optimisation budgets reported in Table 7. Popularity and ItemKNN contain no trainable stochastic components; any repeated-run variation in these methods therefore reflects seed-dependent operations in the surrounding evaluation pipeline rather than parameter optimisation.

E HYPERPARAMETERS

All models are trained with Adam (Kingma & Ba, 2015); settings are in Table 7.

Table 7: Hyperparameters used for the reported experiments. Dataset-specific batch sizes account for differences in catalog scale; the optimisation budget should therefore be interpreted jointly with batch size and epoch count.

	ML-1M	Amazon-Beauty	Book-Crossing
Filtering			
k-core minimum interactions	5	3	10
Rating threshold	—	≥ 3	≥ 5
Few-shot user threshold	20	5	5
Backbone			
Embedding dim d	128	128	128
LightGCN layers L	2	2	2
Optimisation			
Batch size	8,192	256	512
Learning rate	0.012	0.008	0.004
Epochs	40	60	60
Early-stopping patience (epochs)	20	20	20
Semantic			
Sentence encoder	all-MiniLM-L6-v2 (frozen), $d_s = 384$		
Neighbours K	20	20	20
Aux sample size m	$\min(1024, N)$		
Warmup E_w / ramp E_r	5 / 5	5 / 5	5 / 5
Cold-item protocol			
Hold-out ratio ρ	0.15	0.25	0.20
Minimum interactions to be eligible	5	3	3