COMPOTE: COLLAPSE-RESISTANT MULTIPLE-HYPOTHESIS LEARNING FOR SEQUENTIAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Next-item prediction is inherently ambiguous: a single interaction history often admits multiple plausible continuations across distinct catalogue regions. Standard sequential recommenders compress these alternatives into a single prediction, conflating user intents. We introduce **CO**llapse-resistant **M**ultiple hy**POT**hesis l**E**arning (COMPOTE), a multi-hypothesis transformer trained with an annealed objective that transitions from uniform supervision to a winner-takes-all assignment. This mechanism forces hypotheses to specialize into distinct catalogue regions, while inference relies on an efficient, normalizer-free max-similarity rule. Theoretically, we show that in the population case, the optima solutionspartition the conditional next-item distribution according to a max-entropy quantization principle, while pooled scores preserve the Bayes-optimal ranking. We also prove a generalization bound scaling logarithmically with the number of hypotheses. Across three Amazon Reviews 2023 benchmarks, COMPOTE outperforms strong ID-based, generative, and multi-interest baselines using item IDs alone. Matched-capacity ablations confirm that performance gains stem directly from the annealed objective, preventing the head collapse observed in conventional multi-head architectures, rather than added capacity, ensembling, or routing.

1 INTRODUCTION

Recommender systems filter vast product catalogues where their ranking dictates item visibility, directly driving discovery, engagement, and revenue (Covington et al., 2016). As an example, *Sequential recommendation* predicts the next item from ordered interactions, closely mirroring real-world deployment where user preferences evolve and recent user's interaction history is most informative (Rendle et al., 2010). There exists essentially two paradigms for sequential recommendation, both of which learn a single conditional model of the next item. The first family encodes the history with an autoregressive sequence model and scores items against a learned embedding table, moving from recurrent (Hidasi et al., 2016) and convolutional (Tang and Wang, 2018) encoders to causal self-attention (Kang and McAuley, 2018), backbone of later variants (Sun et al., 2019; Zhou et al., 2020; Zhai et al., 2024). The second family treats retrieval as generation (Tay et al., 2022): TIGER (Rajput et al., 2023) quantizes item content embeddings into *semantic identifiers* and decodes the next item's identifier token by token, which helps on rare items at the cost of a content encoder, a learned tokenizer and beam search. Both families nonetheless learn a *single* conditional model of the next item: one query representation, one score vector, one ranked list.

Next-item prediction is ambiguous. An user's interaction history rarely determines the next item: it points to a *set* of plausible continuations, often in unrelated parts of the catalogue, so the next-item distribution is multimodal and a single query vector trained with cross-entropy settles between the modes instead of matching any one. *Multi-interest learning* addresses this by representing each user with several vectors, extracted by dynamic routing over capsules in MIND (Li et al., 2019), by multi-head attention with a diversity–accuracy aggregation step in ComiRec (Cen et al., 2020), or by activating a few concepts from a shared pool in SINE (Tan et al., 2021). REMI (Xie et al., 2023) keeps the same architecture and instead adds interest-aware hard negatives and a regularizer against *routing collapse*, where all interests attend to the same historical item. These methods already

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.

rely on a form of winner-takes-all (WTA) training, but it comes with the routing architecture and WTA is rarely examined as an objective in its own right. As a result, the optimal interests are not characterized, and diversity is usually encouraged through the architecture or additional penalties.

Our approach: COMPOTE. We frame next-item prediction as an inherently ambiguous learning problem and adopt multiple-choice learning (MCL) (Guzman-Rivera et al., 2012; Lee et al., 2016; Rupprecht et al., 2017), where a model outputs M candidate hypotheses, but optimizes only the best-performing head. While this winner-takes-all objective effectively quantizes the conditional next-item distribution (Letzelter et al., 2024), it is prone to sub-optimal local minima where heads that underperform early permanently deactivate. Annealed MCL (Perera et al., 2024) mitigates this initialization sensitivity by annealing a soft-assignment temperature during training. We adapt this formulation to discrete catalogues with **CO**llapse-resistant **M**ultiple hy**POT**hesis l**E**arning (COMPOTE; Section 2). A shared causal transformer feeds M prediction heads, each defining a distinct distribution over the catalogue, which are optimized via the objective function ℓ_τ in Equation (3) where τ is the temperature parameter. At high temperatures, all heads update uniformly; as the temperature vanishes, the objective smoothly converges to pure winner-takes-all. For inference and retrieval, we introduce max-similarity, a normalizer-free counterpart to the loss's max-pooling operator (Proposition 5), which matches max-pooling accuracy on standard benchmarks while bypassing catalogue-wide softmax operations (Section 4.1).

Contributions. Methodologically, we propose COMPOTE, extending annealed multiple-choice learning to sequential recommendation alongside an efficient, normalizer-free retrieval rule (Section 2). On the theoretical front, we first establish a generalization bound scaling logarithmically in the number of hypotheses M (Theorem 4). We further characterize the optimal hypotheses in the winner-takes-all limit, showing they partition the conditional item distribution into maximum-entropy regions (Theorem 7) and that pooled hypothesis scores preserve the Bayes ranking up to saturation ties (Proposition 6), proving head specialization emerges naturally without explicit diversity regularizers. Finally, across three Amazon Reviews 2023 benchmarks, COMPOTE consistently outperforms ID-based, generative, and multi-interest sequential baselines using item IDs alone, with capacity-matched ablations attributing these gains directly to the annealed MCL objective rather than model scaling, ensembling, or expert routing (Section 4).

2 Problem Statement and Architecture

We consider a supervised learning setting where a query $\mathbf{x} \in \mathcal{X}$ has an associated relevant item $\mathbf{y}$ in a catalog $\mathcal{Y}$ of N items, and the goal is to learn a model $g : \mathcal{X} \to \mathfrak{S}_N$ mapping a query to a ranking of the items, i.e., a permutation of $\{1, \dots, N\}$. Data pairs $(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}_{\mathbf{x},\mathbf{y}}$ are drawn i.i.d. from an unknown underlying probability distribution over $\mathcal{X} \times \mathcal{Y}$. In recommendation, queries are user histories and items are the entities being recommended (products, movies, etc.). We define $\Delta^{N\text{-}1} = \{\mathbf{p} \in [0,1]^N : \|\mathbf{p}\|_1 = 1\}$ as the $(N\text{-}1)$-dimensional probability simplex over $\mathcal{Y}$.

Datasets. We instantiate this setting on three categories from the Amazon 2023 review corpus (Hou et al., 2024): Musical Instruments, Industrial and Scientific, and Video Games. A query $\mathbf{x}$ is a user's chronologically ordered interaction history, and the target $\mathbf{y}$ is the next reviewed item. Each category contains approximately 25,000 items and between 50,000 and 95,000 users (Table 5). Because each user contributes few observations relative to the catalog size, several next items may be plausible given the same history.

Next-item Prediction with Multiple-Hypothesis. As motivated in the introduction, the model aims at performing next-item prediction via *multiple hypotheses*. It can be formalized as a function $f : \mathcal{X} \to (\Delta^{N\text{-}1})^M$ that returns, for every hypothesis $m \leq M$, a probability distribution $f_m(\mathbf{x}) \in \Delta^{N\text{-}1}$ over the catalog, where $\mathbf{p}_{m,i} \triangleq f_m(\mathbf{x})_i$ is the probability that hypothesis m assigns to item i being relevant to query $\mathbf{x}$. The exact architecture of f used in COMPOTE is detailed in Figure 1. At inference, the M per-encoder distributions $\mathbf{p} = f(\mathbf{x})$ are combined into a single score vector over the catalog, which is then sorted to output the ranking

$$g = \operatorname{argsort} \circ \omega \circ f \quad \text{where} \quad \omega(\mathbf{p}) = \Big(\max_{m \leq M} \mathbf{p}_{m,i} \Big)_{i=1}^{N} \text{ (max-pooling).} \tag{1}$$

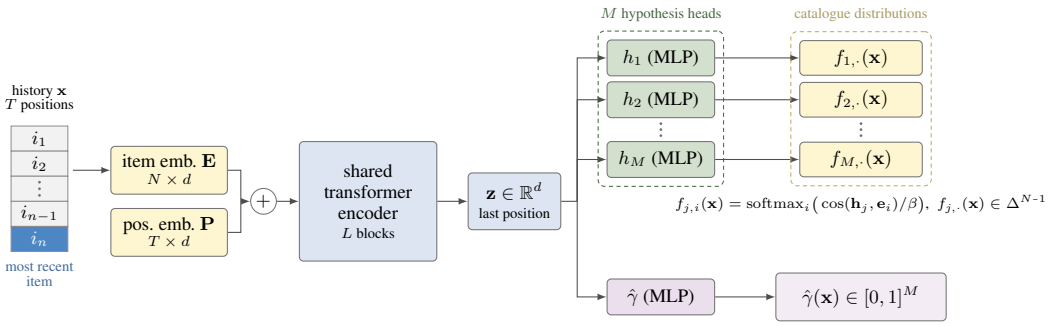


Figure 1: **Multiple-choice transformer.** An item-id history $\mathbf{x}$ is embedded, summed with a learned positional embedding, and encoded by a stack of L pre-norm causal transformer blocks shared by all heads; the representation $\mathbf{z}$ at the last (most recent) position is read off. Two modules consume $\mathbf{z}$: the M hypothesis heads $h_1, \dots, h_M$, each producing $\mathbf{h}_j = h_j(\mathbf{z}) \in \mathbb{R}^d$, and the confidence score head $\hat{\gamma}$, whose j-th output is the predicted probability that head j is the best expert for $\mathbf{x}$. Each hypothesis is turned into a distribution over the catalogue by $f_{j,i}(\mathbf{x}) = \mathrm{softmax}_i\big(\cos(\mathbf{h}_j, \mathbf{e}_i)/\beta\big)$, yielding one score vector $f_{j,\cdot}(\mathbf{x}) \in \Delta^{N-1}$ per head. Everything left of $\mathbf{z}$ is a single shared trunk; right of it the M branches are independent, which is what makes head specialisation possible.

3 Learning Problem

Ideally, to learn f avoiding hypothesis collapse, we would like to use the *winner-takes-all* loss, defined for a given prediction $\mathbf{p} = f(\mathbf{x}) \in \big(\Delta^{N-1}\big)^M$ as

$$\ell^{\mathrm{WTA}}(\mathbf{y}, \mathbf{p}) \triangleq \min_{m \le M} -\log(\mathbf{p}_{m,\mathbf{y}}) \quad \text{with associated risk} \quad \mathcal{L}^{\mathrm{WTA}}(f, \mathbb{P}_{\mathbf{x},\mathbf{y}}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ell^{\mathrm{WTA}}(\mathbf{y}, f(\mathbf{x}))]\,. \quad (2)$$

Optimizing directly ℓ^{WTA} is usually challenging because of the discontinuity introduced, which leads many people to resort to annealing methods. Following this path, we use the following loss function defined for temperature $\tau > 0$ (to be annealed towards 0) as

$$\ell_\tau(\mathbf{y}, \mathbf{p}) = -\tau \log\Big(\sum_{m \le M} \mathbf{p}_{m,\mathbf{y}}^{\frac{1}{\tau}}\Big) \quad \text{with associated risk} \quad \mathcal{L}_\tau(f, \mathbb{P}_{\mathbf{x},\mathbf{y}}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}[\ell_\tau(\mathbf{y}, f(\mathbf{x}))]\,. \quad (3)$$

As the following result illustrates, annealing on ℓ_τ leads to optimizing ℓ^{WTA}.

Proposition 1. *For any* $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \big(\Delta^{N-1}\big)^M$, $\lim_{\tau \to 0^+} \ell_\tau(\mathbf{y}, \mathbf{p}) = \ell^{\mathrm{WTA}}(\mathbf{y}, \mathbf{p})$.

Proof can be found in Appendix A.2. This allows to extend by continuity the definition of ℓ_τ to $\tau = 0$ and consider $\tau \ge 0$ in the remainder for statements and proofs.

Parallel with Annealed Multiple-Choice Learning (MCL). Methods employing a *winner-takes-all* strategy exist in other domains, such as audio source separation, under the framework of *multiple-choice learning* (MCL). In fact, our loss ℓ_τ recovers a formulation from this domain, namely annealed MCL, yet existing literature does not address how to recombine these hypotheses for inference. Our primary contribution is demonstrating, both theoretically and empirically, that ω – Equation (1) – is the principled choice for hypothesis aggregation in recommender systems. Formally, the annealed MCL (Perera et al., 2024) objective function is defined as

$$\ell_\tau^{\mathrm{aMCL}}(\mathbf{y}, \mathbf{p}) = -\sum_{m \le M} \gamma_m(\mathbf{y}, \mathbf{p}) \log(\mathbf{p}_{m,\mathbf{y}}) \quad \text{with} \quad \gamma(\mathbf{y}, \mathbf{p}) = \mathrm{softmax}\Big(\frac{1}{\tau} \log(\mathbf{p}_{\cdot,\mathbf{y}})\Big)\,. \quad (4)$$

and the following propostion establishes the link between the two losses.

Proposition 2 (Recovering aMCL). *For any* $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ *and for any* $\tau \geq 0$,

$$\ell_\tau(\mathbf{y}, \mathbf{p}) = \ell_\tau^{\text{aMCL}}(\mathbf{y}, \mathbf{p}) - \tau H(\gamma(\mathbf{y}, \mathbf{p})) \tag{5}$$

where $H(\gamma) = -\sum_{m \leq M} \gamma_m \log \gamma_m$ *is the Shannon entropy of the softmax weights* $\gamma(\mathbf{y}, \mathbf{p})$.

Proof can be found in Appendix A.3. An important detail strengthens the link: *In Perera et al. (2024), during the optimization, the gradient is only backpropagated through* $\log \mathbf{p}_{m,\mathbf{y}}$ *and not through* $\gamma_m(\mathbf{y}, \mathbf{p})$ *– i.e. there is a* `stop-gradient` *operator on the* $\gamma_j(\mathbf{y}, \mathbf{p})$ *term.* The following result shows the two learning paradigm coincide exactly in practice. Proof can be found in Appendix A.4.

Corollary 3 (Recovering the gradient of aMCL). *For any* $(\mathbf{y}, \mathbf{p}^\theta) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ *(*θ *is some parameter), for any* $\tau \geq 0$, *and* $\alpha \in \Delta^{M-1}$,

$$\nabla_\theta \ell_\tau(\mathbf{y}, \mathbf{p}^\theta) = \nabla_\theta \psi_\tau\left(\mathbf{y}, \mathbf{p}^\theta, \alpha\right)\big|_{\alpha = \gamma_m(\mathbf{y}, \mathbf{p}^\theta)} \quad \textit{with} \quad \psi_\tau(\mathbf{y}, \mathbf{p}^\theta) = -\sum_{m \leq M} \alpha_m \log(\mathbf{p}^\theta_{m,\mathbf{y}}). \tag{6}$$

Learning Guarantees. We have discussed the loss function we use to train the model and have established a connection with the annealed MCL loss. We now turn to the theoretical analysis of the learning guarantees associated with this loss function. While in practice, we learn f by minimizing the empirical risk $\widehat{\mathcal{L}}_\tau(f) = \mathcal{L}_\tau(f, \hat{\mathbb{P}}_S)$ over a training dataset. during training, τ goes to 0. Thus the learning guarantees need to be provided for the limiting loss ℓ^{WTA}. Over a training dataset $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim \mathbb{P}^n_{\mathbf{x},\mathbf{y}}$, the empirical risk is denoted $\widehat{\mathcal{L}}^{\text{WTA}}_S(f) = \mathcal{L}^{\text{WTA}}(f, \hat{\mathbb{P}}_S)$ where $\hat{\mathbb{P}}_S$ is the empirical distribution of S. In order to get insights on the impact of the number of hypotheses M on the learning problem, we are here interested in the generalization gap between the empirical WTA loss and the expected WTA loss $\mathcal{L}^{\text{WTA}}(f, \mathbb{P}_S)$ and provide a bound on this gap in terms of the Rademacher complexity of the hypothesis class $\mathcal{F}$ and the number of hypotheses M. Proof is given in Appendix D.2.

Theorem 4 (refined log-loss generalization bound). *Let* $\mathcal{F}$ *be a hypothesis class of probability-generating heads* $f_m : \mathcal{X} \to \Delta^{N-1}$. *Assume there exists a constant* $\epsilon > 0$ *such that* $f_{m,i}(\mathbf{x}) \geq \epsilon$ *for all* $m \in [M]$, $\mathbf{x} \in \mathcal{X}$, *and* $i \in [N]$. *Then, for any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ *over the sample* $S \sim P^n_{\mathbf{x},\mathbf{y}}$, *every ensemble* $f = (f_1, \ldots, f_M) \in \mathcal{F}^M$ *satisfies:*

$$\mathcal{L}^{\text{WTA}}(f, \mathbb{P}_S) \leq \widehat{\mathcal{L}}^{\text{WTA}}_S(f) + \frac{2}{\epsilon} \max_{m \in [M]} \mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}}) + 2 \log\left(\frac{1}{\epsilon}\right) \sqrt{\frac{2 \ln M}{n}} + \log\left(\frac{1}{\epsilon}\right) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{7}$$

where $\mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}})$ *is the expected Rademacher complexity of* $\mathcal{F}_{m,\mathbf{y}} = \{\mathbf{x} \mapsto f_{m,\mathbf{y}}(\mathbf{x}) : f_m \in \mathcal{F}\}$.

Insights from the bounds:

- The generalization gap is controlled by the Rademacher complexities of the individual heads, scaled by $1/\epsilon$. As the minimum predicted probability ϵ decreases, the bounds become looser, reflecting increased uncertainty in predictions. Ensuring a lower bound on predicted probabilities, via label smoothing, temperature scaling, or KL regularisation, therefore yields tighter guarantees.
- Theorem 4 shows that the number of hypotheses M impacts the bound only through a $\mathcal{O}(\sqrt{\ln M / n})$ term, on top of the complexity of the best head. Adding more hypotheses therefore does not linearly increase the risk of overfitting, provided that the individual head complexities are controlled. Hence, one can benefit from the diversity of multiple heads without a proportional increase in generalization error.

We provide an empirical validation of this theoretical insight in Appendix I in which we also discuss the practical implications of the saturation phenomenon described in Proposition 6 in Section 4.

Learning Stabilization. We observe that training f can be stabilized through the joint learning of the confidence score head via a binary cross-entropy loss over $\hat{\gamma}$ (see Figure 1). This finding

aligns with the fact that the min operator is non-smooth and can be challenging to optimize. While beneficial, this effect is not the primary driver of the overall performance gains (see ablation study in Appendix G). Consequently, the formalization and technical details of this stabilization term are deferred to the appendix (see Appendix H).

4 COMPOTE LEADS TO OPTIMAL RANKING

This section explains why the max-pooling aggregation ω defined in Equation (1) is the natural choice when learning f with ℓ_τ. In fact, the max-pooling aggregation ω can be seen as the limit for a temperature $\tau \to 0$ of some aggregation ω_τ (defined below), which cross-entropy is exactly ℓ_τ.

Proposition 5 (ℓ_τ via ω^τ)**.** *For any* $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ *and for any* $\tau \geq 0$*,*

$$\ell_\tau(\mathbf{y}, \mathbf{p}) = -\log \omega^\tau_{\mathbf{y}}(\mathbf{p}) \quad \textit{where} \quad \omega^\tau(\mathbf{p}) = \left(\left(\sum_{m \leq M} \mathbf{p}_{m,i}^{1/\tau} \right)^\tau \right)_{i=1}^N \quad \textit{and} \quad \omega^\tau \xrightarrow[\tau \to 0]{} \omega \tag{8}$$

Proof can be found in Appendix B.1. Notice that $\omega^\tau(\mathbf{p})$ does not necessarily lie in Δ^{N-1}, thus $\mathcal{L}_\tau(f) = -\mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\log \omega^\tau_{\mathbf{y}}(f(\mathbf{x}))\right]$ can be lower than the Bayes risk $\inf_{q:\mathcal{X}\to\Delta^{N-1}} -\mathbb{E}_{\mathbf{x},\mathbf{y}}[\log q_{\mathbf{y}}(\mathbf{x})]$. Because $\omega^\tau(\mathbf{p})$ and $\omega(\mathbf{p})$ do not necessarily belong to the simplex, they do not recover the Bayes predictor (they are not a probability). However, for prediction $\mathbf{p}^\star$ that minimizes ℓ^{WTA} in expectation, $\omega(\mathbf{p}^\star)$ *ranks the items in the same order as the Bayes predictor*, as formalized by the following statement.

Proposition 6 (ω induces Bayes ranking)**.** *Let* $\mathbf{q}^\star \in \Delta^{N-1}$*, and denote the optimal prediction* $\mathbf{p}^\star \in \operatorname{argmin}_{\mathbf{p} \in (\Delta^{N-1})^M} \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p})]$*. Then*

$$\forall\, a, b \in \mathcal{Y}, \qquad \mathbf{q}^\star_a > \mathbf{q}^\star_b \implies \omega_a(\mathbf{p}^\star) > \omega_b(\mathbf{p}^\star) \quad \textit{or} \quad \omega_a(\mathbf{p}^\star) = \omega_b(\mathbf{p}^\star) = 1\,. \tag{9}$$

Proof can be found in Appendix B.2. The ranking prediction associated to COMPOTE thus preserves the order of the Bayes predictor strictly, except in the single *saturation* case $\omega_a(\mathbf{p}^\star) = \omega_b(\mathbf{p}^\star) = 1$; understanding this failure case is informative for the choice of M. Intuitively, it happens when (given some $\mathbf{x}$) at least two items a and b have large probabilities $\mathbf{q}^\star_a, \mathbf{q}^\star_b \geq 1/M$: for this $\mathbf{x}$, one hypothesis m will be such that $\mathbf{p}^\star_{m,i} = \mathbb{1}_{[i=a]}$ (see Theorem 7 for why this happens), another hypothesis m' will be such that $\mathbf{p}^\star_{m',i} = \mathbb{1}_{[i=b]}$. Hence, even though $\mathbf{q}^\star_a > \mathbf{q}^\star_b$, then $\phi_a = \phi_b = 1$. This advocates for using values for M that are not too large, which is validated experimentally in Figure 5.

The performance of using the aggregation ω over the learning done via annealing of ℓ_τ is also validated empirically in the recommendation setting on several Amazon 2023 datasets.

Preprocessing and protocol. We evaluate on the three Amazon 2023 subsets introduced in Section 2. Following prior work (Rajput et al., 2023; Zhou et al., 2020; Zheng et al., 2025), we apply the benchmark *5-core* filtering (Amazon Reviews'23 Benchmark, 2023), which iteratively discards users and items with fewer than five interactions, and truncate histories to 50 items at both training and test time. We adopt the leave-one-out split of Kang and McAuley (2018): the last interaction of each user is the test target, the penultimate one is used for validation and the remainder for training. We report Hit Rate (H@K) and NDCG (N@K) for $K \in \{5, 10\}$, computed by ranking the entire catalog rather than sampled negatives; since each test instance has a single relevant item, H@K coincides with Recall@K.

Baselines. We compare against seven ID-based sequential recommenders (Caser (Tang and Wang, 2018), HGN (Ma et al., 2019), GRU4Rec (Hidasi et al., 2016), BERT4Rec (Sun et al., 2019), SASRec (Kang and McAuley, 2018), HSTU (Zhai et al., 2024) and S3-Rec (Zhou et al., 2020)) and against the generative recommender TIGER (Rajput et al., 2023), whose numbers we take from Zheng et al. (2025). Because COMPOTE predicts several hypotheses per query, we also compare against six multi-interest recommenders that represent a user by several vectors: MIND (Li et al., 2019), ComiRec-DR and ComiRec-SA (Cen et al., 2020), SINE (Tan et al., 2021), Re4 (Zhang et al.,

Table 1: Performance (%) on three Amazon 2023 subsets (H@K: Hit Rate@K, N@K: NDCG@K). First-block numbers are from Zheng et al. (2025); multi-interest models are trained by us under the same protocol (Appendix F). COMPOTE uses max-similarity aggregation (Equation (10)), equivalent to max-softmax pooling (Section 4.1). **Bold**: best, underline: second best.

	Instrument				Scientific				Game			
Method	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10
Caser (Tang and Wang, 2018)	2.41	3.86	1.51	1.97	1.59	2.57	1.01	1.32	3.30	5.53	2.09	2.81
HGN (Ma et al., 2019)	3.21	5.17	2.02	2.65	2.12	3.51	1.31	1.76	4.24	6.87	2.71	3.56
GRU4Rec (Hidasi et al., 2016)	3.24	5.01	2.09	2.66	2.02	3.38	1.29	1.73	4.99	7.99	3.20	4.16
BERT4Rec (Sun et al., 2019)	3.07	4.85	1.95	2.52	1.86	2.96	1.19	1.55	4.60	7.35	2.98	3.86
SASRec (Kang and McAuley, 2018)	3.33	5.23	2.13	2.74	2.59	4.12	1.50	1.99	5.35	8.47	3.31	4.38
HSTU (Zhai et al., 2024)	3.43	5.77	1.91	2.71	2.71	**4.29**	1.47	1.98	5.78	9.03	3.34	4.42
S3-Rec (Zhou et al., 2020)	3.17	4.96	1.99	2.57	2.63	4.18	1.71	2.19	4.85	7.69	3.15	4.06
TIGER (Rajput et al., 2023)	3.70	5.64	2.44	3.06	2.64	4.22	1.75	2.26	5.59	8.68	3.66	4.67
MIND (Li et al., 2019)	2.71	4.42	1.65	2.20	2.11	3.29	1.27	1.65	3.22	5.20	1.99	2.65
ComiRec-DR (Cen et al., 2020)	2.67	4.37	1.64	2.19	2.14	3.41	1.30	1.72	3.21	5.23	1.96	2.60
ComiRec-SA (Cen et al., 2020)	2.64	4.31	1.62	2.15	2.08	3.31	1.26	1.65	3.20	5.19	1.95	2.61
SINE (Tan et al., 2021)	3.08	5.20	1.52	2.24	2.36	3.83	1.21	1.68	5.26	8.80	2.77	3.87
Re4 (Zhang et al., 2022)	2.13	3.74	1.23	1.66	1.61	2.66	0.95	1.27	3.15	5.61	1.98	2.68
REMI (Xie et al., 2023)	2.73	4.34	1.71	2.22	2.20	3.86	1.21	1.70	4.14	7.42	2.29	3.30
COMPOTE (ours)	**3.75**	**5.83**	**2.45**	**3.10**	**2.84**	4.20	**1.92**	**2.35**	**6.06**	**9.26**	**4.01**	**5.04**

2022) and REMI (Xie et al., 2023), which we train ourselves under the protocol above. Appendix F describes each method.

In Table 1, COMPOTE is best in 11 of the 12 columns, the exception being H@10 on Scientific where HSTU (4.29) and TIGER (4.22) edge it out (4.20). Against TIGER, the strongest baseline on NDCG, COMPOTE improves N@5 by 9.7% on Scientific and 9.6% on Game, and ties on Instrument (2.44 $\rightarrow$ 2.45). Note that TIGER performs autoregressive generative retrieval over semantic identifiers derived from item text and decodes them with beam search at inference, while COMPOTE sees item identifiers only and ranks the catalogue in a single forward pass: modelling the multimodality of the next-item distribution suffices to beat a content-aware generative recommender.

Multi-interest models trade precision for coverage. Multi-interest recommenders are the closest analogue of COMPOTE, yet none reaches the second-best value of any column, and only SINE on Game H@10 (8.80) exceeds SASRec. SINE is within 5% of COMPOTE on that column but 31% below on N@5 (2.77 against 4.01): maximising over interest vectors widens the candidate pool without placing the target near the top, whereas COMPOTE improves both.

4.1 Fast Retrieval for COMPOTE

For large-scale recommender systems, it is desirable to have a model that is compatible with *vector databases* to enable a retrieval scaling sublinearly with N. As $f_{m,i}(\mathbf{x}) = \mathrm{softmax}(\cos(h_m(x), \mathbf{e}_i)/\beta)$ (see Figure 1), there is already a dependency linear in N with the computation of the softmax. This section shows that, for the purpose of ranking the items for recommendation, the softmax can be dropped at inference – it is possible to apply ω directly on the cosine similarities to obtain scores to rank the items. Formally, it means replace at inference $g = \mathrm{argsort} \circ \omega \circ f$ by

$$\tilde{g} = \mathrm{argsort} \circ \omega \circ s \quad \text{where} \quad s_{m,i}(\mathbf{x}) = \cos(h_m(\mathbf{x}), \mathbf{e}_i)/\beta) \tag{10}$$

Due to the following identity: $\log \max_{m \leq M} (\mathrm{softmax}_i(\mathbf{s}_{m,\cdot})) = \max_{m \leq M} \left(\mathbf{s}_{m,i} - \log Z_m\right)$ where $Z_m = \sum_{i=1}^{N} \exp(\mathbf{s}_{m,i})$, g and $\tilde{g}$ rank items identically if Z_m is constant across the hypotheses. This turns out to happen in practice as the standard deviation of Z_m across hypotheses is largely lower that its value (Figure 3). Experiments confirm this as using $\tilde{g}$ rather than g comes at no loss in ranking accuracy (Table 2).

This equivalence is a key deployment result: the model requires neither a catalog-wide softmax nor its normalizer, making the model directly compatible with standard vector databases. For this reason, we report most of the results with $\tilde{g}$ (ex: Table 3)

Table 2: Effect of the inference aggregator on COMPOTE performance (%) on Game, for the same trained model with $M = 3$ heads and the same checkpoint. "Normalizer" indicates whether ranking requires the per-head catalogue partition function Z_m. **Bold**: best.

Aggregator	Normalizer	H@5	H@10	N@5	N@10
g – "max softmax"	required	6.04	**9.27**	**4.01**	**5.05**
$\tilde{g}$ – "max cosine similarity"	not required	**6.06**	9.26	**4.01**	5.04

Table 3: Ablation over the prediction head and the training objective (%). Variants share the same encoder backbone and parameter budget, and rank items with max-similarity aggregation (Equation (10)): MCT-SASRec is a single-hypothesis iso-parameter head, MCT-EnsCE uses an ensemble cross-entropy objective, and MCT-Router routes queries to a single expert. **Bold**: best.

	Instrument				Scientific				Game			
Method	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10	H@5	H@10	N@5	N@10
MCT-SASRec	3.63	5.70	2.34	2.97	2.55	3.87	1.71	2.13	5.73	8.84	3.78	4.78
MCT-EnsCE	3.54	5.53	2.28	2.96	2.42	3.80	1.60	2.03	5.74	8.81	3.78	4.77
MCT-Router	3.57	5.48	2.28	2.89	2.40	3.60	1.64	2.00	5.74	8.81	3.78	4.77
COMPOTE (ours)	**3.75**	**5.83**	**2.45**	**3.10**	**2.84**	**4.20**	**1.92**	**2.35**	**6.06**	**9.26**	**4.01**	**5.04**

4.2 Ablations: Where Do the Gains Come From?

COMPOTE could outperform a single-head model for three distinct reasons, additional capacity, naive ensembling, or mixture-of-experts (MoE) routing. We isolate each with a matched-capacity variant (Table 3). COMPOTE outperforms all three on every dataset and metric.

Not a bigger model: MCT-SASRec spends the parameter budget of COMPOTE on a single hypothesis with a deeper prediction head. It is the strongest of the three variants and still trails COMPOTE everywhere (N@5 of 2.34 against 2.45 on Instrument, 1.71 against 1.92 on Scientific, 3.78 against 4.01 on Game). At equal capacity, the multi-hypothesis parameterization beats a deeper single-hypothesis head.

Not an ensemble: MCT-EnsCE keeps the architecture of COMPOTE but trains every head on every example with equal weight, without confidence supervision or annealing. It minimizes $\ell^{\text{ens}}(\mathbf{y}, f(\mathbf{x})) = -\frac{1}{M}\sum_{m\le M} \log f_m(\mathbf{x})_{\mathbf{y}}$, the $\tau \to \infty$ limit of the soft assignment in Equation (4). It falls below COMPOTE on every dataset and metric, so the gains are not explained by architectural multiplicity or naive ensembling alone.

Not a mixture of experts: MCT-Router follows the routing recipe of multi-interest recommendation (Section 1): it keeps the shared encoder and replaces the training of COMPOTE by a learned gate that selects one expert per query. It trails COMPOTE everywhere (N@5 of 2.28 against 2.45 on Instrument, 1.64 against 1.92 on Scientific, 3.78 against 4.01 on Game). The next section traces this gap to head collapse: without the annealed winner-takes-all objective, routing fails to induce distinct hypotheses (Section 5.1).

5 Hypothesis Specialization

We first show that, unlike the matched-capacity variants of Section 4.2, COMPOTE learns hypotheses that do not collapse and carry complementary rankings. We then explain this phenomenon by characterizing what each hypothesis $\mathbf{p}_m^\star$ learns at the winner-takes-all limit $\tau \to 0$.

5.1 Collapse and Oracle Evidence

We quantify collapse through the mean pairwise cosine similarity of the head outputs. For a query $\mathbf{x}$, let $\bar{\mathbf{h}}_m(\mathbf{x}) = \mathbf{h}_m(\mathbf{x})/\|\mathbf{h}_m(\mathbf{x})\|_2$ and define $S(\mathbf{x}) = \frac{1}{M(M-1)}\sum_{j,k=1}^{M} \bar{\mathbf{h}}_j(\mathbf{x})^\top \bar{\mathbf{h}}_k(\mathbf{x})$. We report its test-set average; $S = 1$ means that all heads have the same direction. MCT-EnsCE and MCT-Router reach $S = 1.00$ on all three datasets, whereas COMPOTE obtains 0.64, 0.61 and 0.58 on Instru-

MENT, SCIENTIFIC and GAME, respectively (Table 6). Thus, neither architectural multiplicity nor learned routing is sufficient to prevent collapse.

Oracle head selection. To test whether COMPOTE's distinct heads contain complementary rankings, we compare the max-similarity rule of Equation (10) with an oracle that observes the target and selects the head assigning it the highest similarity,

$$m^{\mathrm{orc}}(\mathbf{x}, \mathbf{y}) = \operatorname*{argmax}_{m \le M} s_{m,\mathbf{y}}(\mathbf{x}) \,, \tag{11}$$

then ranks every item using $s_{m^{\mathrm{orc}}(\mathbf{x},\mathbf{y}),i}(\mathbf{x})$. The oracle is not deployable, but it measures whether one hypothesis ranks the target substantially better than the aggregated list. For MCT-EnsCE and MCT-Router, aggregated and oracle N@5 coincide exactly on every dataset: selecting among collapsed heads cannot help (Table 7). For COMPOTE, oracle selection raises N@5 by 53% on INSTRUMENT (2.45 → 3.75), 83% on SCIENTIFIC (1.92 → 3.52) and 94% on GAME (4.01 → 7.78). The ranking gains therefore coincide with a qualitative change in the learned predictor: COMPOTE produces distinct heads whose rankings contain complementary signal.

These findings raise a structural question: why does COMPOTE's training create complementary hypotheses, whereas ensemble and routed objectives collapse?

5.2 WHAT EACH HYPOTHESIS LEARNS

For a fixed query, Theorem 7 describes the optimal hypotheses through three properties:

- (P1) *Partition.* The catalog splits into disjoint cells $\mathcal{Y}_1, \dots, \mathcal{Y}_M$, each hypothesis supported on its own cell.
- (P2) *Faithfulness within cells.* On $\mathcal{Y}_m$, $\mathbf{p}^\star_m$ is $\mathbf{q}^\star$ renormalized by Q_m: each hypothesis ranks the items of its cell exactly as the Bayes predictor does, and therefore still depends on the query through $\mathbf{q}^\star$.
- (P3) *Balance.* The partition maximizes the entropy of $(Q_m)_{m=1}^M$, so the cells split the mass of $\mathbf{q}^\star$ as evenly as the catalogue allows.

Theorem 7 (Extension of Th. 1 – Rupprecht et al. (2017))**.** *Let $\mathbf{q}^\star \in \Delta^{N-1}$, and denote the optimal prediction $\mathbf{p}^\star \in \operatorname{argmin}_{\mathbf{p}\in(\Delta^{N-1})^M} \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\mathrm{WTA}}(\mathbf{y}, \mathbf{p})]$. Then for every $m \le M$ and $y \in \mathcal{Y}$,*

$$\mathbf{p}^\star_{m,y} = \frac{\mathbf{q}^\star_y}{Q_m} \mathbb{1}_{[y\in\mathcal{Y}_m]} \,, \quad \textit{with} \quad Q_m = \sum_{y\in\mathcal{Y}_m} \mathbf{q}^\star_y \,. \tag{12}$$

where the tessalation is $\mathcal{Y}_m = \{y \in \mathcal{Y} : \forall m' \neq m, \, -\log(\mathbf{p}^\star_{m,y}) < -\log(\mathbf{p}^\star_{m',y})\}$. Moreover the tessalation $(\mathcal{Y}_m)_{m=1}^M$ is entropy-maximizing.

Proof can be found in Appendix K.1.

COMPOTE is trained at $\tau > 0$ with softmax heads, so the exact zeros of Equation (12) cannot be reached. We test each property through its observable signature. Mirroring the tessellation of Theorem 7, the empirical cell of head m on query $\mathbf{x}$ is $\mathcal{Y}_m(\mathbf{x}) = \{y \in \mathcal{Y} : m = \operatorname{argmax}_{m'} f_{m',y}(\mathbf{x})\}$, and the *winning head* of a test pair $(\mathbf{x}, \mathbf{y})$ is the head whose cell contains $\mathbf{y}$. All diagnostics are computed on GAME, with $M = 3$ unless stated otherwise, and Table 4 reports them for $M \in \{3, 4, 5\}$. The protocol is detailed in Appendix L, which also reports the same partition on SCIENTIFIC.

(P1) The hypotheses partition the catalogue. The non-collapsed head directions of Table 6 are a first signature of (P1). The partition holds for the full distributions and not only for the top of the rankings: each head keeps between 91.5% and 93.7% of its softmax mass inside its own cell, for every M. The retrieved sets are therefore nearly disjoint. The top-10 lists of two heads on the same query have a mean Jaccard overlap of 0.008, and 91.8% of the catalogue items retrieved at test time are reached by a single head only. Both sharpen with M, reaching 0.001 and 98.6% at $M = 5$. Figure 2 shows the resulting partition: projected onto the plane of the head directions, the catalogue splits into three sectors, one per hypothesis.

Table 4: Signatures of Theorem 7 on GAME as the number of hypotheses M varies. All columns are computed on the full test set of 94,762 query–target pairs. Win frequencies and sector shares are reported as the range over heads.

M	In-cell mass	Win frequency (%)	Sector share (%)	$\hat{\mathcal{L}}^{\text{WTA}}$	$\hat{\mathcal{L}}^{\text{WTA}} + \log M$	Sector agreement
3	0.929	[32.0, 34.4]	[27.9, 36.3]	7.32	8.41	98.5%
4	0.915	[24.5, 25.6]	[23.4, 27.1]	7.02	8.41	98.7%
5	0.937	[18.9, 21.1]	[15.5, 25.3]	6.71	8.32	99.4%

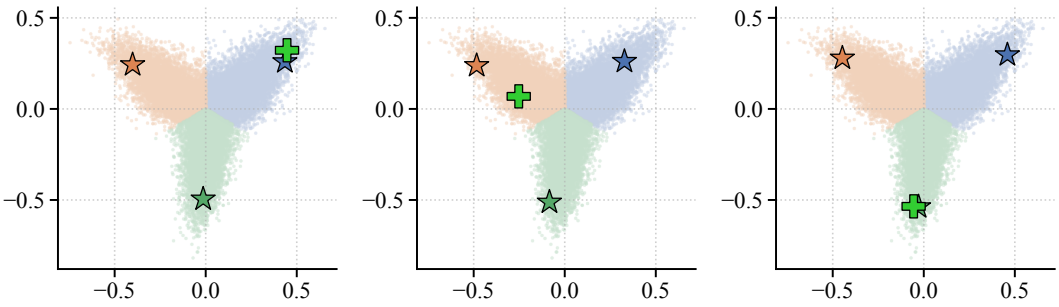


Figure 2: **Catalogue tiling on GAME ($M = 3$).** Normalized items (dots) and head outputs (stars) are projected onto the plane spanned by the static head directions $\bar{\mathbf{h}}_m$, the average normalized output of each head over the test set. Items are coloured by the direction they are most aligned with, and the green cross marks the target. Each panel is a test query won by a different head, and in each the target lies in the sector of its winning head.

(P2) Each hypothesis remains query-dependent. By (P2), a hypothesis is not a fixed list of items: its ranking follows $\mathbf{q}^\star(\mathbf{x})$. Empirically, the top-10 of a given head overlaps by only 0.017 to 0.026 Jaccard between two different histories, and in Figure 2 the head closest to the target changes from query to query. (P1) and (P2) together explain the oracle gap in Table 7. At the optimum, the target $\mathbf{y}$ belongs to a single cell $\mathcal{Y}_m$, where it competes only with the other items of $\mathcal{Y}_m$, ranked by $\mathbf{q}^\star$. Selecting that head therefore ranks $\mathbf{y}$ no worse than the Bayes predictor over the full catalogue. The cosine oracle of Equation (11) makes exactly this choice, as it picks the winning head on 98.9% to 99.7% of test pairs. This is the behavior of COMPOTE, where the oracle-selected head ranks the target substantially higher than the pooled list, while collapsed models exhibit no oracle gain.

(P3) The cells split the mass evenly. The win frequency of a head (the fraction of test targets it wins), estimates $\mathbb{E}_{\mathbf{x}}[Q_m(\mathbf{x})]$ without access to $\mathbf{q}^\star$. Win frequencies stay within 1.6 points of $1/M$ for every M, and their normalized entropy exceeds 0.999. Balance holds in mass and not in item counts: the heads own between 15.5% and 25.3% of the catalogue at $M = 5$, yet each wins close to 20% of the targets. This is the balance (P3) predicts, since Q_m weights items by $\mathbf{q}^\star$, so a head with fewer items holds items that are more often the next one. Balance also matters for aggregation: if $Q_m = 1/M$ for all m, Equation (12) gives $\omega(\mathbf{p}^\star) = M\mathbf{q}^\star$, so max-pooling recovers the Bayes ranking exactly. Balance also fixes the value of the optimal loss, and Appendix K.3 checks that the test loss follows the predicted $\log M$ law.

6 DISCUSSION AND CONCLUSION

We introduced COMPOTE, which uses annealed multiple-hypothesis learning to produce complementary rankings for sequential recommendation. Our theoretical and empirical results show that it prevents head collapse, preserves ranking quality, and improves performance across three benchmarks. We leave combining COMPOTE with semantic identifiers for future work, which could improve recommendations for rare items. It also remains to explain why its theoretically query-dependent cells form an almost static catalogue tiling and how this structure relates to item-embedding geometry (Appendix M).

AI USE STATEMENT

We used generative AI tools to polish the manuscript, primarily by correcting grammar and improving clarity, and to assist with portions of the experimental code. Parts of the theoretical proofs (the generalization bounds) were also drafted with the help of generative AI, but all proofs were verified and corrected by the authors.

REFERENCES

Amazon Reviews'23 Benchmark. Pure IDs (5-core) data processing. https://amazon-reviews-2023.github.io/data_processing/5core.html, 2023. Accessed August 2026. 5, 19

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002. 17

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002. 16

Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2942–2951. ACM, 2020. doi: 10.1145/3394486.3403344. 1, 5, 6, 20

Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 191–198. ACM, 2016. doi: 10.1145/2959100.2959190. 1

Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems 25*, pages 1799–1807, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/cfbce4c1d7c425baf21d6b6f2babe6be-Abstract.html. 2

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *4th International Conference on Learning Representations*, 2016. URL http://arxiv.org/abs/1511.06939. 1, 5, 6, 19

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian J. McAuley. Bridging language and items for retrieval and recommendation. *CoRR*, abs/2403.03952, 2024. URL https://doi.org/10.48550/ARXIV.2403.03952. 2, 19

Wang-Cheng Kang and Julian J. McAuley. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining*, pages 197–206. IEEE Computer Society, 2018. doi: 10.1109/ICDM.2018.00035. 1, 5, 6, 20

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991. 17

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, Viresh Ranjan, David J. Crandall, and Dhruv Batra. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems 29*, pages 2119–2127, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/20d135f0f28185b84a4cf7aa51f29500-Abstract.html. 2

Victor Letzelter, David Perera, Cédric Rommel, Mathieu Fontaine, Slim Essid, Gaël Richard, and Patrick Pérez. Winner-takes-all learners are geometry-aware conditional density estimators. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. 2

Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2615–2623. ACM, 2019. doi: 10.1145/3357384.3357814. 1, 5, 6, 20

Chen Ma, Peng Kang, and Xue Liu. Hierarchical gating networks for sequential recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 825–833. ACM, 2019. doi: 10.1145/3292500.3330984. 5, 6, 19

Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000. 19

Mehryar Mohri and Andres Munoz Medina. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 262–270, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/mohri14.html. 17

David Perera, Victor Letzelter, Théo Mariotte, Adrien Cortés, Mickael Chen, Slim Essid, and Gaël Richard. Annealed multiple choice learning: Overcoming limitations of winner-takes-all with annealing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 10541–10574. Curran Associates, Inc., 2024. doi: 10.52202/079017-0337. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1456560769bbc38e4f8c5055048ea712-Paper-Conference.pdf. 2, 3, 4

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 Mahesh Sathiamoorthy. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems 36*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/20dcab0f14046a5c6b02b61da9f13229-Abstract-Conference.html. 1, 5, 6, 19, 20

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 811–820. ACM, 2010. doi: 10.1145/1772690.1772773. 1

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *IEEE International Conference on Computer Vision*, pages 3611–3620. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.388. 2, 8, 23

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450. ACM, 2019. doi: 10.1145/3357384.3357895. 1, 5, 6, 20

Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. Sparse-interest network for sequential recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, 2021. doi: 10.1145/3437963.3441811. 1, 5, 6, 20

Jiaxi Tang and Ke Wang. Personalized top-N sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 565–573. ACM, 2018. doi: 10.1145/3159652.3159656. 1, 5, 6, 19

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 *Advances in Neural Information Processing Systems 35*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/892840a6123b5ec99ebaab8be1530fba-Abstract-Conference.html. 1

Yueqi Xie, Jingqi Gao, Peilin Zhou, Qichen Ye, Yining Hua, Jae Boum Kim, Fangzhao Wu, and Sunghun Kim. Rethinking multi-interest learning for candidate matching in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, 2023. doi: 10.1145/3604915.3608766. 1, 6, 20

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xye7iNsgXn. 1, 5, 6, 20

Shengyu Zhang, Lingxiao Yang, Dong Yao, Yujie Lu, Fuli Feng, Zhou Zhao, Tat-Seng Chua, and Fei Wu. Re4: Learning to re-contrast, re-attend, re-construct for multi-interest recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 2216–2226. ACM, 2022. doi: 10.1145/3485447.3512094. 5, 6, 20

Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. RecBole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 4653–4664. ACM, 2021. doi: 10.1145/3459637.3482016. 19

Bowen Zheng, Enze Liu, Zhongfu Chen, Zhongrui Ma, Yue Wang, Wayne Xin Zhao, and Ji-Rong Wen. Pre-training generative recommender with multi-identifier item tokenization. *arXiv preprint arXiv:2504.04400*, 2025. URL https://arxiv.org/abs/2504.04400. 5, 6, 19

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 1893–1902. ACM, 2020. doi: 10.1145/3340531.3411954. 1, 5, 6, 19, 20

A Proofs - Section 2

A.1 Preliminary results

Proposition 8 (ℓ_τ as a convex conjugate)**.** *For any* $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ *and for any* $\tau \geq 0$*,*

$$\ell_\tau(\mathbf{y}, \mathbf{p}) = \min_{\gamma \in \Delta^{M-1}} \sum_{m \leq M} -\gamma \log(\mathbf{p}_{m,\mathbf{y}}) - \tau H(\gamma)\,. \tag{13}$$

with optimal γ *given by*

$$\gamma(\mathbf{y}, \mathbf{p}) = \operatorname{softmax}\left(\frac{1}{\tau}\log(\mathbf{p}_{\cdot,\mathbf{y}})\right)\,. \tag{14}$$

Proof. The log-sum-exp is the convex conjugate of the negative entropy. □

A.2 Proof of Proposition 1

Proposition 1. *For any* $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$*,* $\lim_{\tau \to 0^+} \ell_\tau(\mathbf{y}, \mathbf{p}) = \ell^{\text{WTA}}(\mathbf{y}, \mathbf{p})$.

Proof of Proposition 1. The proof is a direct consequence of Proposition 8 as $\lim_{\tau \to 0^+} \operatorname{softmax}\left(\frac{1}{\tau}\log(\mathbf{p}_{\cdot,\mathbf{y}})\right) = \max_{m \leq M} \frac{1}{\tau}\log(\mathbf{p}_{m,\mathbf{y}})$. □

A.3 Proof of Proposition 2

Proposition 2 (Recovering aMCL)**.** *For any* $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ *and for any* $\tau \geq 0$*,*

$$\ell_\tau(\mathbf{y}, \mathbf{p}) = \ell_\tau^{\text{aMCL}}(\mathbf{y}, \mathbf{p}) - \tau H(\gamma(\mathbf{y}, \mathbf{p})) \tag{5}$$

where $H(\gamma) = -\sum_{m \leq M} \gamma_m \log \gamma_m$ is the Shannon entropy of the softmax weights $\gamma(\mathbf{y}, \mathbf{p})$.

Proof of Proposition 2. This is a direct consequence of Proposition 8. □

A.4 Proof of Corollary 3

Corollary 3 (Recovering the gradient of aMCL)**.** *For any $(\mathbf{y}, \mathbf{p}^\theta) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ (θ is some parameter), for any $\tau \geq 0$, and $\alpha \in \Delta^{M-1}$,*

$$\nabla_\theta \ell_\tau(\mathbf{y}, \mathbf{p}^\theta) = \nabla_\theta \psi_\tau\left(\mathbf{y}, \mathbf{p}^\theta, \alpha\right)\big|_{\alpha = \gamma_m(\mathbf{y}, \mathbf{p}^\theta)} \quad \textit{with} \quad \psi_\tau(\mathbf{y}, \mathbf{p}^\theta) = -\sum_{m \leq M} \alpha_m \log(\mathbf{p}^\theta_{m,\mathbf{y}}). \tag{6}$$

Proof of Corollary 3.

$$\nabla_\theta \psi_\tau(\mathbf{y}, \mathbf{p}^\theta, \alpha)|_{\alpha = \gamma_m(\mathbf{y}, \mathbf{p}^\theta)} = -\sum_{m \leq M} \alpha_m \frac{\nabla_\theta \mathbf{p}^\theta_{m,\mathbf{y}}}{\mathbf{p}^\theta_{m,\mathbf{y}}} = -\sum_{m \leq M} \frac{(\mathbf{p}^\theta_{m,\mathbf{y}})^{1/\tau - 1}}{\sum_{m' \leq M} (\mathbf{p}^\theta_{m',\mathbf{y}})^{1/\tau}} \nabla_\theta \mathbf{p}^\theta_{m,\mathbf{y}} \,. \tag{15}$$

On the other hand, differentiating $\ell_\tau(\mathbf{y}, \mathbf{p}^\theta) = -\tau \log\left(\sum_{m \leq M} (\mathbf{p}^\theta_{m,\mathbf{y}})^{1/\tau}\right)$ directly gives

$$\nabla_\theta \ell_\tau(\mathbf{y}, \mathbf{p}^\theta) = -\tau \frac{\sum_{m \leq M} \frac{1}{\tau} (\mathbf{p}^\theta_{m,\mathbf{y}})^{1/\tau - 1} \nabla_\theta \mathbf{p}^\theta_{m,\mathbf{y}}}{\sum_{m' \leq M} (\mathbf{p}^\theta_{m',\mathbf{y}})^{1/\tau}} = -\sum_{m \leq M} \frac{(\mathbf{p}^\theta_{m,\mathbf{y}})^{1/\tau - 1}}{\sum_{m' \leq M} (\mathbf{p}^\theta_{m',\mathbf{y}})^{1/\tau}} \nabla_\theta \mathbf{p}^\theta_{m,\mathbf{y}} \,. \tag{16}$$

This concludes the proof. □

B Proofs of Section 4

B.1 Proof of Proposition 5

Proposition 5 (ℓ_τ via ω^τ)**.** *For any $(\mathbf{y}, \mathbf{p}) \in \mathcal{Y} \times \left(\Delta^{N-1}\right)^M$ and for any $\tau \geq 0$,*

$$\ell_\tau(\mathbf{y}, \mathbf{p}) = -\log \omega^\tau_{\mathbf{y}}(\mathbf{p}) \quad \textit{where} \quad \omega^\tau(\mathbf{p}) = \left(\left(\sum_{m \leq M} \mathbf{p}^{1/\tau}_{m,i}\right)^\tau\right)_{i=1}^N \quad \textit{and} \quad \omega^\tau \xrightarrow[\tau \to 0]{} \omega \tag{8}$$

Proof of Proposition 5. Fix $(\mathbf{y}, \mathbf{p})$. For $\tau > 0$, by definition of ω^τ in Equation (8),

$$-\log\left(\omega^\tau_{\mathbf{y}}(\mathbf{p})\right) = -\log\left(\left(\sum_{m \leq M} \mathbf{p}^{\frac{1}{\tau}}_{m,\mathbf{y}}\right)^\tau\right) = -\tau \log\left(\sum_{m \leq M} \mathbf{p}^{\frac{1}{\tau}}_{m,\mathbf{y}}\right) = \ell_\tau(\mathbf{y}, \mathbf{p}) \,, \tag{17}$$

the last equality being the definition of ℓ_τ in Equation (3).

For $\tau = 0$, understood as the limit $\tau \to 0^+$ on both sides, $\omega^0 = \omega$ gives $-\log\left(\omega^0_{\mathbf{y}}(\mathbf{p})\right) = -\log \max_{m \leq M} \mathbf{p}_{m,\mathbf{y}} = \ell^{\text{WTA}}(\mathbf{y}, \mathbf{p})$, which equals $\lim_{\tau \to 0^+} \ell_\tau(\mathbf{y}, \mathbf{p})$ by Proposition 1. □

B.2 Proof of Proposition 6

Proposition 6 (ω induces Bayes ranking)**.** *Let $\mathbf{q}^\star \in \Delta^{N-1}$, and denote the optimal prediction $\mathbf{p}^\star \in \operatorname{argmin}_{\mathbf{p} \in (\Delta^{N-1})^M} \mathbb{E}_{\mathbf{y} \sim \mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p})]$. Then*

$$\forall\, a, b \in \mathcal{Y}, \qquad \mathbf{q}^\star_a > \mathbf{q}^\star_b \implies \omega_a(\mathbf{p}^\star) > \omega_b(\mathbf{p}^\star) \quad \textit{or} \quad \omega_a(\mathbf{p}^\star) = \omega_b(\mathbf{p}^\star) = 1 \,. \tag{9}$$

Proof of Proposition 6. Denote the inner risk by $L(\mathbf{p}) = \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\mathrm{WTA}}(\mathbf{y},\mathbf{p})]$. By Proposition 5,

$$L(\mathbf{p}) = -\sum_{y\in\mathcal{Y}} \mathbf{q}^\star_y \log \omega_y(\mathbf{p})\,, \tag{18}$$

so that, $L(\mathbf{p}^\star)$ being finite, $\omega_y(\mathbf{p}^\star) > 0$ for every y with $\mathbf{q}^\star_y > 0$. Fix $a, b \in \mathcal{Y}$ with $\mathbf{q}^\star_a > \mathbf{q}^\star_b$; in particular $\mathbf{q}^\star_a > 0$ and $\omega_a(\mathbf{p}^\star) > 0$.

Weak inequality. Assume $\omega_b(\mathbf{p}^\star) > 0$, the conclusion being trivial otherwise, and let Π_{ab} be the operator associated with the transposition $\pi = (a\,b)$. Then

$$L(\Pi_{ab}\mathbf{p}^\star) - L(\mathbf{p}^\star) = -\mathbf{q}^\star_a \log \omega_b(\mathbf{p}^\star) - \mathbf{q}^\star_b \log \omega_a(\mathbf{p}^\star) + \mathbf{q}^\star_a \log \omega_a(\mathbf{p}^\star) + \mathbf{q}^\star_b \log \omega_b(\mathbf{p}^\star) \tag{19}$$

$$= (\mathbf{q}^\star_a - \mathbf{q}^\star_b) \log \frac{\omega_a(\mathbf{p}^\star)}{\omega_b(\mathbf{p}^\star)}\,, \tag{20}$$

the terms at $y \notin \{a, b\}$ cancelling and the remaining logarithms being finite. Since $L(\Pi_{ab}\mathbf{p}^\star) \geq L(\mathbf{p}^\star)$ and $\mathbf{q}^\star_a > \mathbf{q}^\star_b$, this gives $\omega_a(\mathbf{p}^\star) \geq \omega_b(\mathbf{p}^\star)$.

Strictness up to saturation. Assume now $\omega_a(\mathbf{p}^\star) = \omega_b(\mathbf{p}^\star)$; we show that this common value is 1. Let $(\mathcal{Y}_m)_{m\leq M}$ be the tessellation of Theorem 7, $m(y)$ the index of the cell containing y and Q_m the mass of $\mathcal{Y}_m$, so that $\omega_y(\mathbf{p}^\star) = \mathbf{p}^\star_{m(y),y} = \mathbf{q}^\star_y / Q_{m(y)}$ whenever $Q_{m(y)} > 0$, and, by equation 18,

$$L(\mathbf{p}^\star) = -\sum_{y\in\mathcal{Y}} \mathbf{q}^\star_y \log \frac{\mathbf{q}^\star_y}{Q_{m(y)}} = H(\mathbf{q}^\star) + \sum_{m\leq M} \phi(Q_m)\,, \qquad \phi(t) = t\log t, \quad \phi(0) = 0\,. \tag{21}$$

Suppose first $\mathbf{q}^\star_b > 0$, so that $Q_A, Q_B > 0$ where $A = m(a)$ and $B = m(b)$. The items a and b lie in distinct cells: were they in the same cell, the ratio $\omega_a(\mathbf{p}^\star)/\omega_b(\mathbf{p}^\star) = \mathbf{q}^\star_a/\mathbf{q}^\star_b > 1$ would contradict the assumed equality. Exchanging a and b between the two cells produces a partition of $\mathcal{Y}$ with masses $Q_A - u$ and $Q_B + u$, where $u = \mathbf{q}^\star_a - \mathbf{q}^\star_b > 0$, the other cells being unchanged; the predictor it defines through equation 12 is admissible, so equation 21 and the optimality of $\mathbf{p}^\star$ give

$$g(u) \geq g(0)\,, \qquad \text{where} \qquad g(t) = \phi(Q_A - t) + \phi(Q_B + t)\,. \tag{22}$$

Writing $\Delta = Q_A - Q_B$, we have $g'(t) = \log \frac{Q_B+t}{Q_A-t}$, so g decreases on $(-\infty, \Delta/2]$ and increases on $[\Delta/2, Q_A)$; being symmetric about $\Delta/2$, it satisfies $g(u) \geq g(0)$ with $u > 0$ only if $u \geq \Delta$. Consequently

$$\mathbf{q}^\star_a Q_B - \mathbf{q}^\star_b Q_A \;\geq\; (\mathbf{q}^\star_b + \Delta)Q_B - \mathbf{q}^\star_b(Q_B + \Delta) \;=\; \Delta\,(Q_B - \mathbf{q}^\star_b) \;\geq\; 0\,, \tag{23}$$

using $\mathbf{q}^\star_a \geq \mathbf{q}^\star_b + \Delta$ and $Q_B \geq \mathbf{q}^\star_b$ if $\Delta > 0$, and $\mathbf{q}^\star_a > \mathbf{q}^\star_b$ with $Q_A \leq Q_B$ if $\Delta \leq 0$. The assumed equality reads $\omega_a(\mathbf{p}^\star) - \omega_b(\mathbf{p}^\star) = (\mathbf{q}^\star_a Q_B - \mathbf{q}^\star_b Q_A)/(Q_A Q_B) = 0$ and therefore forces $\Delta > 0$, $\mathbf{q}^\star_a = \mathbf{q}^\star_b + \Delta$ and $Q_B = \mathbf{q}^\star_b$, whence $\omega_b(\mathbf{p}^\star) = \mathbf{q}^\star_b / Q_B = 1$.

Suppose finally $\mathbf{q}^\star_b = 0$. If $Q_B > 0$ then $\omega_b(\mathbf{p}^\star) = \mathbf{q}^\star_b/Q_B = 0 < \omega_a(\mathbf{p}^\star)$, excluding the equality. Otherwise every $y \in \mathcal{Y}_B$ has $\mathbf{q}^\star_y = 0$; let $\mathbf{p}'$ be $\mathbf{p}^\star$ with head B replaced by e_a. For $y \notin \mathcal{Y}_B \cup \{a\}$ the maximum defining ω_y is attained by a head other than B, so $\omega_y(\mathbf{p}') = \omega_y(\mathbf{p}^\star)$; the items of $\mathcal{Y}_B$ carry no mass; and $\omega_a(\mathbf{p}') = 1 \geq \omega_a(\mathbf{p}^\star)$. Hence $L(\mathbf{p}') \leq L(\mathbf{p}^\star)$ with equality only if $\omega_a(\mathbf{p}^\star) = 1$, and the optimality of $\mathbf{p}^\star$ forces $\omega_a(\mathbf{p}^\star) = 1$. □

C Per-Head Normalizer

Protocol. For the $M = 3$ Game checkpoint of Table 2, we compute $\log Z_m(\mathbf{x}) = \log \sum_{i\leq N} \exp(s_{m,i}(\mathbf{x}))$ over all $N = 25{,}612$ items for each of the 94,762 test queries, with $s_{m,i} = \cos(h_m, \mathbf{e}_i)/\beta$ and $\beta = 0.07$.

The normalizer is head-independent. The three distributions nearly overlap (Figure 3), with means 13.67, 13.63, and 13.79. At fixed $\mathbf{x}$, the cross-head standard deviation of $\log Z_m(\mathbf{x})$ is 0.13 on average; its range has median 0.28, 95th percentile 0.62, and maximum 1.10.

Effect on the ranking. Because $\log \max_m \mathrm{softmax}_i(\mathbf{s}_{m,\cdot}) = \max_m(\mathbf{s}_{m,i} - \log Z_m)$, a ranking change requires max-cosine scores within β times the cross-head range: 0.020 for the median query, 0.043 at the 95th percentile, and 0.077 at most. Accordingly, Table 2 shows no measurable loss from dropping the normalizer.

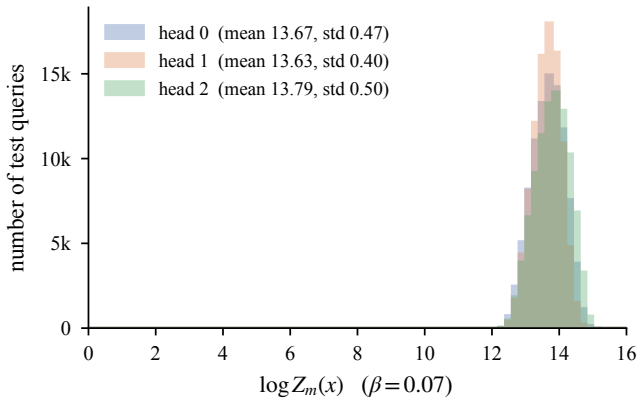


Figure 3: **Distribution of the per-head log-normalizer** $\log Z_m(\mathbf{x})$**.** Histogram of $\log Z_m(\mathbf{x}) = \log \sum_{i \le N} \exp\big(s_{m,i}(\mathbf{x})\big)$ over the 94,762 test queries of GAME, one colour per head, for the COMPOTE checkpoint with $M = 3$ of Table 2 ($\beta = 0.07$, $N = 25{,}612$). The three heads have almost superimposed distributions, concentrated between 12.2 and 15.0 on a scale where $\log Z_m$ could a priori take any value in $[\log N - 1/\beta,\ \log N + 1/\beta] = [-4.1, 24.4]$. For a fixed query, the standard deviation of $\log Z_m(\mathbf{x})$ across the three heads is 0.13 on average, against a mean value of 13.7.

D GENERALIZATION BOUND

D.1 PROOF OF THEOREM 9

We first provide here a generalization bound for the log-loss in Theorem 9, which is a direct consequence of the Rademacher complexity of the hypothesis class $\mathcal{F}^M$. We then provide a refined version of the generalization bound in Theorem 4, which has a milder dependence on the number of heads M.

Theorem 9 (log-loss generalization bound)**.** *Let $\mathcal{F}$ be a hypothesis class of probability-generating heads $f_m : \mathcal{X} \to \Delta^{N-1}$. Assume there exists a constant $\epsilon > 0$ such that $f_{m,i}(\mathbf{x}) \ge \epsilon$ for all $m \in [M]$, $\mathbf{x} \in \mathcal{X}$, and $i \in [N]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample $S \sim P^n_{\mathbf{x},\mathbf{y}}$, every ensemble $f = (f_1, \dots, f_M) \in \mathcal{F}^M$ satisfies:*

$$\mathcal{L}^{\mathrm{WTA}}(f, P_{\mathbf{x},\mathbf{y}}) \le \widehat{\mathcal{L}}^{\mathrm{WTA}}_S(f) + \frac{2}{\epsilon} \sum_{m=1}^{M} \mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}}) + \log\left(\frac{1}{\epsilon}\right) \sqrt{\frac{\ln(2/\delta)}{2n}}, \tag{24}$$

where $\mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}})$ is the expected Rademacher complexity of the set of functions $\mathcal{F}_{m,\mathbf{y}} = \{\mathbf{x} \mapsto f_{m,\mathbf{y}}(\mathbf{x}) : f_m \in \mathcal{F}\}$.

Proof of Theorem 9. **Step 1: Boundedness and Lipschitz continuity of the loss function.** Consider the scalar loss function $-\log(t)$ restricted to $t \in [\epsilon, 1]$.

- **Upper Bound (B):** Since $t \ge \epsilon$, the loss is non-negative and bounded above by:

$$B = \sup_{t \in [\epsilon,1]} (-\log t) = -\log(\epsilon) = \log\left(\frac{1}{\epsilon}\right).$$

  Thus, for every sample $(\mathbf{x}_i, \mathbf{y}_i)$, $\min_{m \in [M]}(-\log f_{m,\mathbf{y}_i}(\mathbf{x}_i)) \in [0, B]$.

- **Lipschitz Continuity (L_ϵ):** By the Mean Value Theorem, the Lipschitz constant is bounded by the supremum of the derivative:

$$L_\epsilon = \sup_{t \in [\epsilon,1]} \left| \frac{d}{dt}(-\log t) \right| = \sup_{t \in [\epsilon,1]} \frac{1}{t} = \frac{1}{\epsilon}.$$

  Therefore, for any $t_1, t_2 \ge \epsilon$, we have $|(-\log t_1) - (-\log t_2)| \le \frac{1}{\epsilon}|t_1 - t_2|$.

Step 2: Concentration via McDiarmid's Inequality.
We want to show that the empirical loss on n finite samples $\widehat{\mathcal{L}}_S^{\text{WTA}}(f)$ is close to the true expected loss $\mathcal{L}^{\text{WTA}}(f)$ uniformly across all possible models in our hypothesis class $\mathcal{F}^M$. Define the worst-case generalisation gap function:

$$\Phi(S) = \sup_{f \in \mathcal{F}^M} \left(\mathcal{L}^{\text{WTA}}(f, P_{\mathbf{x},\mathbf{y}}) - \widehat{\mathcal{L}}_S^{\text{WTA}}(f) \right).$$

Suppose we change a single data point in $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ to create a neighboring dataset $S' = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}'_i, \mathbf{y}'_i), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$. The change in $\Phi(S)$ can be bounded as:

$$|\Phi(S) - \Phi(S')| \leq \sup_{f \in \mathcal{F}^M} \frac{1}{n} \left| \min_m(-\log f_m(\mathbf{x}_i)_{\mathbf{y}_i}) - \min_m(-\log f_m(\mathbf{x}'_i)_{\mathbf{y}'_i}) \right| \leq \frac{B}{n} = \frac{\log(1/\epsilon)}{n}.$$

Because changing one sample alters $\Phi(S)$ by at most $c_i = \frac{\log(1/\epsilon)}{n}$, we can apply McDiarmid's inequality to bound the deviation of $\Phi(S)$ from its expectation (Bousquet and Elisseeff, 2002):

$$\mathbb{P}\left(\Phi(S) - \mathbb{E}_S[\Phi(S)] \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Evaluating $\sum_{i=1}^n c_i^2 = n \cdot \left(\frac{\log(1/\epsilon)}{n}\right)^2 = \frac{\log^2(1/\epsilon)}{n}$. Set the probability tail $\exp\left(\frac{-2t^2 n}{\log^2(1/\epsilon)}\right) = \frac{\delta}{2}$. Then, solving for t gives

$$t = \log\left(\frac{1}{\epsilon}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Thus, with probability at least $1 - \frac{\delta}{2}$:

$$\sup_{f \in \mathcal{F}^M} \left(\mathcal{L}^{\text{WTA}}(f, P_{\mathbf{x},\mathbf{y}}) - \widehat{\mathcal{L}}_S^{\text{WTA}}(f) \right) \leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}^M} \left(\mathcal{L}^{\text{WTA}} - \widehat{\mathcal{L}}_S^{\text{WTA}} \right) \right] + \log\left(\frac{1}{\epsilon}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Step 3: Symmetrization and min-operator decomposition.
Our goal now is to bound the expected supremum $\mathbb{E}_S\left[\sup_{f \in \mathcal{F}^M}\left(\mathcal{L}^{\text{WTA}} - \widehat{\mathcal{L}}^{\text{WTA}}\right)\right]$ in terms of the Rademacher complexity of the hypothesis class $\mathcal{F}^M$. For doing so, we can redefine $\mathbb{E}_S[\Phi(S)]$ as follows:

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_{f \in \mathcal{F}^M}\left(\mathbb{E}_{S'}[\widehat{\mathcal{L}}_{S'}^{\text{WTA}}(f)] - \widehat{\mathcal{L}}_S^{\text{WTA}}(f)\right)\right],$$

where S' is an independent ghost sample drawn from the same distribution. Then by gathering the terms, and using Jensen's inequality, we can pull the expectation over S' outside of the supremum:

$$\mathbb{E}_S[\Phi(S)] \leq \mathbb{E}_{S,S'}\left[\sup_{f \in \mathcal{F}^M} \frac{1}{n}\sum_{i=1}^n \left(\ell(f; \mathbf{x}'_i, \mathbf{y}'_i) - \ell(f; \mathbf{x}_i, \mathbf{y}_i)\right)\right].$$

Now, since $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}'_i, \mathbf{y}'_i)$ are identically distributed, introducing independent random sign variables $\sigma_i \in \{-1, +1\}$ with $\mathbb{P}(\sigma_i = 1) = 1/2$ does not change the distribution of their difference:

$$\mathbb{E}_S[\Phi(S)] \leq \mathbb{E}_{S,S'}\left[\sup_f \frac{1}{n}\sum_{i=1}^n \sigma_i \left(\ell(f; \mathbf{x}'_i, \mathbf{y}'_i) - \ell(f; \mathbf{x}_i, \mathbf{y}_i)\right)\right],$$

and using the triangle inequality leads to the following decomposition:

$$\begin{aligned}
\mathbb{E}_S[\Phi(S)] &\leq \mathbb{E}_{S',\boldsymbol{\sigma}}\left[\sup_f \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(f; \mathbf{x}'_i, \mathbf{y}'_i)\right] \\
&\quad + \mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_f \frac{1}{n}\sum_{i=1}^n (-\sigma_i)\ell(f; \mathbf{x}_i, \mathbf{y}_i)\right] \\
&= 2\,\mathbb{E}_S\left[\widehat{\mathcal{R}}_S(\mathcal{G}_{\text{emax}})\right] = 2\,\mathcal{R}_n(\mathcal{G}_{\text{emax}}),
\end{aligned}$$

where $\widehat{\mathcal{R}}_S(\mathcal{G}_{\text{emax}})$ is the empirical Rademacher complexity of the composite pointwise minimum class:

$$\mathcal{G}_{\text{emax}} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto \min_{m \in [M]} (-\log f_{m,\mathbf{y}}(\mathbf{x})) : f_{m,\cdot} \in \mathcal{F} \right\},$$

and $\mathcal{R}_n(\mathcal{G}_{\text{emax}}) = \mathbb{E}_S[\widehat{\mathcal{R}}_S(\mathcal{G}_{\text{emax}})]$.

For a fixed sample S, using the structural property of Rademacher complexity under pointwise minimum operators over M hypothesis classes (decomposition of the pointwise minimum as a weighted sum of indicator functions and using results from Bartlett and Mendelson (2002)), we bound $\widehat{\mathcal{R}}_S(\mathcal{G}_{\text{emax}})$ as:

$$\widehat{\mathcal{R}}_S(\mathcal{G}_{\text{emax}}) \leq \sum_{m=1}^{M} \widehat{\mathcal{R}}_S(-\log \circ \mathcal{F}_{m,y}), \tag{25}$$

where $\mathcal{F}_{m,y} = \{\mathbf{x} \mapsto f_{m,y}(\mathbf{x}) : f \in \mathcal{F}\}$. Since $t \mapsto -\log t$ is $(1/\epsilon)$-Lipschitz on $[\epsilon, 1]$, applying Talagrand's Contraction Lemma (Mohri and Medina, 2014; Ledoux and Talagrand, 1991) yields:

$$\widehat{\mathcal{R}}_S(-\log \circ \mathcal{F}_{m,y}) \leq \frac{1}{\epsilon} \widehat{\mathcal{R}}_S(\mathcal{F}_{m,y}).$$

Taking expectations $\mathbb{E}_S$ on both sides gives $\mathcal{R}_n(\mathcal{G}_{\text{emax}}) \leq \frac{1}{\epsilon} \sum_{m=1}^{M} \mathcal{R}_n(\mathcal{F}_{m,y})$.

Final combination. Substituting into $\mathbb{E}_S[\Phi(S)] \leq 2\mathcal{R}_n(\mathcal{G}_{\text{emax}})$ and applying concentration yields the stated bound. □

D.2 Proof of Theorem 4

Theorem 4 (refined log-loss generalization bound)**.** *Let $\mathcal{F}$ be a hypothesis class of probability-generating heads $f_m : \mathcal{X} \to \Delta^{N-1}$. Assume there exists a constant $\epsilon > 0$ such that $f_{m,i}(\mathbf{x}) \geq \epsilon$ for all $m \in [M]$, $\mathbf{x} \in \mathcal{X}$, and $i \in [N]$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the sample $S \sim P^n_{\mathbf{x},\mathbf{y}}$, every ensemble $f = (f_1, \ldots, f_M) \in \mathcal{F}^M$ satisfies:*

$$\mathcal{L}^{\text{WTA}}(f, \mathbb{P}_S) \leq \widehat{\mathcal{L}}^{\text{WTA}}_S(f) + \frac{2}{\epsilon} \max_{m \in [M]} \mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}}) + 2 \log\left(\frac{1}{\epsilon}\right) \sqrt{\frac{2 \ln M}{n}} + \log\left(\frac{1}{\epsilon}\right) \sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{7}$$

where $\mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}})$ is the expected Rademacher complexity of $\mathcal{F}_{m,\mathbf{y}} = \{\mathbf{x} \mapsto f_{m,\mathbf{y}}(\mathbf{x}) : f_m \in \mathcal{F}\}$.

Proof of Theorem 4. To prove the theorem, we need to show that the expected Rademacher complexity of the minimum operator over M heads satisfies:

$$\mathbb{E}_{S,\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}^M} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \min_{m \in [M]} (-\log f_{m,\mathbf{y}}(\mathbf{x}_i)) \right] \leq \frac{1}{\epsilon} \max_{m \in [M]} \mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}}) + B \sqrt{\frac{2 \ln M}{n}}. \tag{26}$$

and plugging this bound into the generalization bound of Theorem 9 gives the desired result. Basically, we replace the crude sum over heads in Equation (25) by this Massart-type control of the minimum operator, which yields a sharper dependence on M. Hence, we only need to prove Equation (26). That is done in the following steps.

Let $\mathcal{G}_m = \{(\mathbf{x}, \mathbf{y}) \mapsto -\log f_{m,\mathbf{y}}(\mathbf{x}) : f_m \in \mathcal{F}\}$ be the loss function class for head m, where each function $g_m \in \mathcal{G}_m$ takes values in $[0, B]$ with $B = \log(1/\epsilon)$.

Let $g_m(\mathbf{z}_i) = -\log f_{m,\mathbf{y}}(\mathbf{x}_i)$, where $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$. We analyze the empirical Rademacher process for a fixed sample S:

$$\widehat{\mathcal{R}}_S(\mathcal{G}_{\text{emax}}) = \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{g_1 \in \mathcal{G}_1, \ldots, g_M \in \mathcal{G}_M} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \min_{m \in [M]} g_m(\mathbf{z}_i) \right].$$

Pointwise minimum upper bound. For any given input $\mathbf{z}_i$ and any choice of functions $(g_1, \ldots, g_M)$, the minimum can be written as:

$$\min_{m \in [M]} g_m(\mathbf{z}_i) = g_{m^*}(\mathbf{z}_i) \quad \text{where } m^* = \arg\min_{m \in [M]} g_m(\mathbf{z}_i).$$

Instead of summing over all m, we decouple the choice of the minimizing head index m^* from the individual head function space. For any fixed sequence of functions $(g_1, \ldots, g_M)$, we have:

$$\frac{1}{n}\sum_{i=1}^{n}\sigma_i \min_{m\in[M]} g_m(\mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{m_i^*}(\mathbf{z}_i).$$

α-cover discretization of the minima vector. To apply Massart's Lemma over the finite M choices of heads, we construct an empirical L_2-net $\mathcal{N}(\alpha, \mathcal{G}_m, S)$ for each function class $\mathcal{G}_m$ at scale $\alpha > 0$. Hence, for every $g_m \in \mathcal{G}_m$, there exists a net element $\bar{g}_m \in \mathcal{N}(\alpha, \mathcal{G}_m, S)$ such that:

$$\frac{1}{n}\sum_{i=1}^{n}|g_m(\mathbf{z}_i) - \bar{g}_m(\mathbf{z}_i)|^2 \le \alpha^2.$$

Then, using Cauchy–Schwarz, the discretization error introduced by replacing $g_{m_i^*}$ with its closest net element $\bar{g}_{m_i^*}$ is bounded by:

$$\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_{m_i^*}(\mathbf{z}_i) - \bar{g}_{m_i^*}(\mathbf{z}_i)\right) \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}|g_{m_i^*}(\mathbf{z}_i) - \bar{g}_{m_i^*}(\mathbf{z}_i)|^2} \le \alpha.$$

Decomposing into single head and Massart's finite maximum. For any tuple of true functions $(g_1, \ldots, g_M) \in \mathcal{G}_1 \times \cdots \times \mathcal{G}_M$, we find their corresponding net approximations $(\bar{g}_1, \ldots, \bar{g}_M) \in \mathcal{N}^M$. The empirical sum splits into:

$$\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{m_i^*}(\mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i \bar{g}_{m_i^*}(\mathbf{z}_i) + \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_{m_i^*}(\mathbf{z}_i) - \bar{g}_{m_i^*}(\mathbf{z}_i)\right).$$

Taking the supremum over all valid choices of tuples:

$$\begin{aligned}
&\sup_{(g_1,\ldots,g_M)\in\prod_m \mathcal{G}_m} \frac{1}{n}\sum_{i=1}^{n}\sigma_i g_{m_i^*}(\mathbf{z}_i) \\
&\quad\le \sup_{(\bar{g}_1,\ldots,\bar{g}_M)\in\mathcal{N}^M} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \bar{g}_{m_i^*}(\mathbf{z}_i) \\
&\qquad + \sup_{\substack{g_m\in\mathcal{G}_m \\ \bar{g}_m\in\mathcal{N}}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g_{m_i^*}(\mathbf{z}_i) - \bar{g}_{m_i^*}(\mathbf{z}_i)\right),
\end{aligned}$$

where the final term is at most α, which gives us the following upper bound on the empirical Rademacher process:

$$R(S) \le \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{(\bar{g}_1,\ldots,\bar{g}_M)\in\mathcal{N}^M} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \bar{g}_{m_i^*}(\mathbf{z}_i)\right] + \alpha.$$

Since $m_i^* \in [M]$, we decompose $\bar{g}_{m_i^*}(\mathbf{z}_i)$ into an arbitrary reference head (say $m = 1$) plus a residual vector in $\mathbb{R}^n$ representing the candidate selection among M heads:

$$\bar{g}_{m_i^*}(\mathbf{z}_i) = \bar{g}_1(\mathbf{z}_i) + \left(\bar{g}_{m_i^*}(\mathbf{z}_i) - \bar{g}_1(\mathbf{z}_i)\right).$$

Then, taking expectation gives the following decomposition of the empirical Rademacher process:

$$R(S) \le \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g_1\in\mathcal{G}_1}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g_1(\mathbf{z}_i)\right]}_{\text{Empirical Rademacher of single head } \widehat{\mathcal{R}}_S(\mathcal{G}_1)} + \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{m\in[M]}\frac{1}{n}\sum_{i=1}^{n}\sigma_i \mathbf{v}_m(i)\right]}_{\text{Finite maximum over } M \text{ candidate vectors}} + \alpha,$$

where $\mathbf{v}_m(i) = \bar{g}_m(\mathbf{z}_i) - \bar{g}_1(\mathbf{z}_i) \in [-B, B]$ as $\bar{g}_m$ and $\bar{g}_1$ are both in the loss function class $\mathcal{G}_m$ and hence bounded by $B = \log(1/\epsilon)$.

Table 5: Statistics of the preprocessed datasets (Zheng et al., 2025). Avg. len is the average length of the user interaction sequences.

Dataset	#Users	#Items	#Interactions	Sparsity	Avg. len
Instrument	57,439	24,587	511,836	99.964%	8.91
Scientific	50,985	25,848	412,947	99.969%	8.10
Game	94,762	25,612	814,586	99.966%	8.60

For M vectors $\mathbf{v}_1, \ldots, \mathbf{v}_M \in \mathbb{R}^n$ with bounded norm $\|\mathbf{v}_m\|_2 \leq \sqrt{n}B$, Massart's Finite Class Lemma (Massart, 2000, Lemma 5.2) states that:

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{m\in[M]} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \mathbf{v}_m(i)\right] \leq \max_{m\in[M]} \frac{\|\mathbf{v}_m\|_2}{n}\sqrt{2\ln M} \leq B\sqrt{\frac{2\ln M}{n}}.$$

We apply Talagrand's Contraction Lemma on the single-head term and use the fact that the loss function is $1/\epsilon$-Lipschitz. In addition, by taking the infimum over all valid cover radii $\alpha > 0$ (or letting $\alpha = 1/\sqrt{n}$), the residual discretization error α is absorbed into the standard covering-number bound of the individual head complexity $\mathcal{R}_n(\mathcal{F}_{1,\mathbf{y}})$. This gives us the following upper bound on the empirical Rademacher process:

$$\mathbb{E}_S[\widehat{\mathcal{R}}_S(\mathcal{G}_1)] \leq \frac{1}{\epsilon}\mathcal{R}_n(\mathcal{F}_{1,\mathbf{y}}).$$

Finally, since the choice of head $m = 1$ as reference was arbitrary, we take the maximum over all $m \in [M]$:

$$\mathbb{E}_{S,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}^M} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \min_{m\in[M]} (-\log f_m(\mathbf{x}_i)_{y_i})\right] \leq \frac{1}{\epsilon}\max_{m\in[M]} \mathcal{R}_n(\mathcal{F}_{m,\mathbf{y}}) + B\sqrt{\frac{2\ln M}{n}}.$$

Substituting this refined bound for $\mathcal{R}_n(\mathcal{G}_{\text{emax}})$ into the symmetrization argument of Appendix D.1 and combining with McDiarmid's inequality yields the corollary. □

E Dataset Description

We use three category subsets of the Amazon 2023 review corpus (Hou et al., 2024) Musical Instruments, Industrial and Scientific, and Video Games, abbreviated Instrument, Scientific, and Game which span May 1996 to September 2023. Following prior work (Rajput et al., 2023; Zhou et al., 2020; Zheng et al., 2025), we apply the benchmark *5-core* filtering (Amazon Reviews'23 Benchmark, 2023), which iteratively discards users and items with fewer than five interactions. We then group the interactions by user, order them chronologically, and truncate the histories to 50 items at both training and test time. The resulting statistics are reported in Table 5.

F Compared Models

We detail here the baselines summarized in Section 4. Numbers for the sequential and generative recommenders come from Zheng et al. (2025), who train them with RecBole (Zhao et al., 2021) under a common protocol and tune them by grid search. The multi-interest models are trained by us.

ID-based sequential recommenders.

1. **Caser** (Tang and Wang, 2018) applies horizontal and vertical convolutions to the embedded interaction sequence, capturing both positional and skip-behaviour patterns.
2. **HGN** (Ma et al., 2019) combines feature-level and instance-level gating to modulate item embeddings according to the user representation.
3. **GRU4Rec** (Hidasi et al., 2016) encodes the interaction sequence with gated recurrent units trained for session-based next-item prediction.

4. **BERT4Rec** (Sun et al., 2019) uses a bidirectional transformer encoder trained with a masked-item objective, so that each position is predicted from both past and future context.
5. **SASRec** (Kang and McAuley, 2018) uses a unidirectional (causal) self-attention encoder and predicts the next item by inner product with the item embeddings; it is the closest single-hypothesis analogue of our backbone.
6. **HSTU** (Zhai et al., 2024) introduces hierarchical sequential transducers that consume actions and timestamps jointly, and is designed for scalability while remaining ID-based.
7. **S3-Rec** (Zhou et al., 2020) pre-trains the sequence encoder with self-supervised objectives built from item–attribute correlations, then fine-tunes it for next-item prediction.

Generative recommender. **TIGER** (Rajput et al., 2023) quantizes item text embeddings into semantic identifiers with an RQ-VAE and casts recommendation as autoregressive generation of the target identifier, retrieving items by beam search over the identifier space.

Multi-interest recommenders. These models represent a user by M vectors instead of one, and are the closest existing analogue of our multi-hypothesis parameterization. We train them with the sampled-softmax objective and the readout rule of their reference implementations, and rank the catalogue by the maximum inner product over the M interest vectors, which is the retrieval rule all of them adopt.

1. **MIND** (Li et al., 2019) extracts interests with a dynamic-routing capsule network over the history and trains the interest whose inner product with the target is largest.
2. **ComiRec** (Cen et al., 2020) comes in two variants: ComiRec-DR keeps the capsule routing of MIND with a bilinear mapping, while ComiRec-SA replaces it by a multi-head self-attentive pooling of the history.
3. **SINE** (Tan et al., 2021) selects a sparse subset of concepts from a large learned concept pool, forms one interest per selected concept, and aggregates them back into a single user vector used for retrieval.
4. **Re4** (Zhang et al., 2022) augments ComiRec-SA with three backward-flow regularizers—contrasting interests against each other, realigning the attention with the retrieval criterion, and reconstructing the assigned items from each interest.
5. **REMI** (Xie et al., 2023) augments ComiRec-SA with interest-aware hard-negative mining in the sampled softmax and a variance-based regularizer on the item-to-interest routing matrix that prevents routing collapse.

G Ablation on the Score-Loss Weight

This appendix measures how much of the performance of COMPOTE comes from the score loss of Equation (31), whose weight in the training objective is α (Equation (33)).

Protocol. We train COMPOTE on Scientific with $M = 3$ for $\alpha \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5, 0.7\}$, all other hyperparameters being fixed to the values of Section 4. Rankings use max-similarity aggregation (Equation (1)), which does not read $\hat{\gamma}$ at inference. The score loss can therefore only change the ranking through its gradient on the shared encoder; the oracle assignments it regresses are detached from the computation graph. We report N@5 at the best checkpoint.

The score loss helps. The best weight, $\alpha = 0.1$, reaches 1.91 N@5 against 1.82 without the score loss (Figure 4), and four of the six positive weights improve over $\alpha = 0$. We use $\alpha = 0.1$ in all other experiments.

It is not the main contributor. With $\alpha = 0$, COMPOTE already outperforms every single-hypothesis variant of Table 3 on Scientific: 1.82 N@5 against 1.71 for MCT-SASRec, 1.64 for MCT-Router and 1.60 for MCT-EnsCE. The gain over single-hypothesis training is therefore obtained by the multiple-hypothesis objective alone. Moreover, the curve is not monotonic: $\alpha = 0.2$ and $\alpha = 0.5$ fall below $\alpha = 0$, so the effect of the score loss is of the same order as the variation across weights. The score loss acts as an auxiliary regularizer of the shared encoder rather than as the source of the improvement.

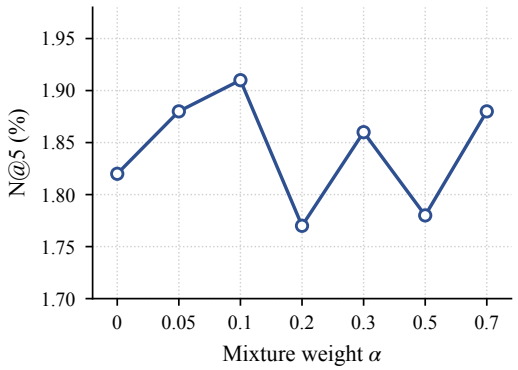


Figure 4: **Effect of the score-loss weight α.** N@5 (%) on SCIENTIFIC with $M = 3$ under max-similarity aggregation, as a function of the weight α of the score loss in Equation (33). Ticks are evenly spaced and do not follow the scale of α. Without the score loss ($\alpha = 0$), COMPOTE already reaches 1.82, above every single-hypothesis variant of Table 3; the best weight $\alpha = 0.1$ adds 0.09 points.

H THE CONFIDENCE-SCORE STABILIZATION

This appendix formalizes the regularization term mentioned at the end of Section 3. It is an auxiliary supervised objective attached to the confidence-score head $\hat{\gamma}$ of Figure 1, which is trained to predict which hypothesis the target item belongs to.

Confidence-score head. Recall the architecture of Figure 1: the shared causal encoder maps the history to the hidden state $\mathbf{z} = \mathrm{Enc}(\mathbf{x}) \in \mathbb{R}^d$ at the most recent position, the M hypothesis heads $h_1, \dots, h_M$ map $\mathbf{z}$ to query representations $h_m(\mathbf{x}) \in \mathbb{R}^d$, and the catalogue scores $\mathbf{s}_{m,i}(\mathbf{x})$ of Equation (10) are turned into the per-hypothesis distributions $f_{m,\cdot}(\mathbf{x}) = \mathrm{softmax}_i\left(\mathbf{s}_{m,\cdot}(\mathbf{x})\right) \in \Delta^{N\text{-}1}$. In parallel, an MLP taking the same $\mathbf{z}$ as input, followed by a coordinate-wise sigmoid, defines the *confidence-score head*

$$\hat{\gamma} : \mathcal{X} \to (0,1)^M , \tag{27}$$

where $\hat{\gamma}_m(\mathbf{x})$ is read as the predicted probability that hypothesis m is the one containing the target of query $\mathbf{x}$. Note that the M outputs are independent sigmoids and not a softmax: $\hat{\gamma}(\mathbf{x})$ need not sum to one.

Supervision target: belonging to $\mathcal{Y}_m$. Following the tessellation of Theorem 7, the empirical cell of hypothesis m on query $\mathbf{x}$ is the set of items for which m is the most confident hypothesis,

$$\mathcal{Y}_m(\mathbf{x}) = \left\{ y \in \mathcal{Y} \;:\; \forall m' \neq m,\; -\log f_{m,y}(\mathbf{x}) < -\log f_{m',y}(\mathbf{x}) \right\}, \tag{28}$$

so that $(\mathcal{Y}_m(\mathbf{x}))_{m \le M}$ is a partition of $\mathcal{Y}$ up to ties. The *winning hypothesis* of a training pair $(\mathbf{x}, \mathbf{y})$ is the index of the cell that contains the target,

$$m^\star(\mathbf{x}, \mathbf{y}) = \operatorname*{argmin}_{m \le M} -\log f_{m,\mathbf{y}}(\mathbf{x}) \quad \text{i.e.} \quad \mathbf{y} \in \mathcal{Y}_{m^\star(\mathbf{x},\mathbf{y})}(\mathbf{x}) , \tag{29}$$

which is exactly the hypothesis realizing the $\min$ in $\ell^{\mathrm{WTA}}(\mathbf{y}, f(\mathbf{x}))$. The regression target of $\hat{\gamma}$ is the corresponding membership vector

$$\gamma^\star(\mathbf{x}, \mathbf{y}) = \left(\mathbb{1}_{[\mathbf{y} \in \mathcal{Y}_m(\mathbf{x})]} \right)_{m=1}^{M} \in \{0,1\}^M , \tag{30}$$

which is a one-hot vector, with its non-zero coordinate at $m^\star(\mathbf{x}, \mathbf{y})$.

Binary cross-entropy on $\hat{\gamma}$. The regularizer is the binary cross-entropy between the membership indicators equation 30 and the predicted confidences, summed over the M coordinates,

$$\ell^{\mathrm{score}}(\mathbf{x}, \mathbf{y}, \hat{\gamma}) = -\sum_{m=1}^{M} \left[\mathbb{1}_{[\mathbf{y} \in \mathcal{Y}_m(\mathbf{x})]} \log \hat{\gamma}_m(\mathbf{x}) + \mathbb{1}_{[\mathbf{y} \notin \mathcal{Y}_m(\mathbf{x})]} \log\left(1 - \hat{\gamma}_m(\mathbf{x})\right) \right], \tag{31}$$

which, the target being one-hot, can be written equivalently as

$$\ell^{\text{score}}(\mathbf{x}, \mathbf{y}, \hat{\gamma}) = -\log \hat{\gamma}_{m^\star(\mathbf{x},\mathbf{y})}(\mathbf{x}) - \sum_{m \neq m^\star(\mathbf{x},\mathbf{y})} \log\big(1 - \hat{\gamma}_m(\mathbf{x})\big)\,. \tag{32}$$

The cells equation 28, and hence the labels equation 30, depend on f: they are recomputed at every training step from the current hypotheses, and are detached from the computation graph, so that no gradient flows from ℓ^{score} back to the heads h_m through $m^\star$. The only gradient path of ℓ^{score} is through $\hat{\gamma}$, i.e. through the confidence MLP and the shared encoder producing $\mathbf{z}$.

Full training objective. Weighting this term by $\alpha > 0$, the objective minimized by COMPOTE is

$$\mathcal{L}_{\tau,\alpha}(f, \hat{\gamma}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}\Big[\ell_\tau\big(\mathbf{y}, f(\mathbf{x})\big)\Big] + \alpha\, \mathbb{E}_{\mathbf{x},\mathbf{y}}\Big[\ell^{\text{score}}(\mathbf{x}, \mathbf{y}, \hat{\gamma})\Big]\,, \tag{33}$$

with ℓ_τ the annealed loss of Equation (3) – equivalently ℓ_τ^{aMCL} of Equation (4) by Proposition 2 and Corollary 3. We use $\alpha = 0.1$ in all experiments; the sensitivity to α is reported in Appendix G.

I Impact of the number of hypotheses M on learning

In Theorem 4 in Section 3, we have established that the generalization gap between the empirical WTA loss and the expected WTA loss has a sublinear dependence on the number of hypotheses M. We present here an empirical validation of this theoretical insight.

Keeping the experimental setup used for performance evaluation described in Section 4, we vary the number of hypotheses M from 1 to 12 and measure the ranking quality on SCIENTIFIC using N@5. Figure 5 shows that increasing M from 1 to 3 improves ranking quality, which suggests that having multiple hypotheses is beneficial for capturing the ambiguity of the next-item distribution. However, beyond $M = 3$, the performance plateaus, indicating that adding more hypotheses does not necessarily lead to better ranking quality. This empirical finding is consistent with the sublinear dependence on M in Theorem 4, and provides practical guidance for selecting an appropriate number of hypotheses in real-world applications.

For those larger value of M, a phenomenom we have described in Proposition 6 in Section 4 can also occur, which is the saturation of the max-similarity aggregation ω. Section 4. We have stated that this saturation can happen when at least two items have large probabilities $\mathbf{q}_a^\star, \mathbf{q}_b^\star \geq 1/M$. This means that as M increases, the probability of this saturation event increases, which can lead to a degradation of the ranking quality.

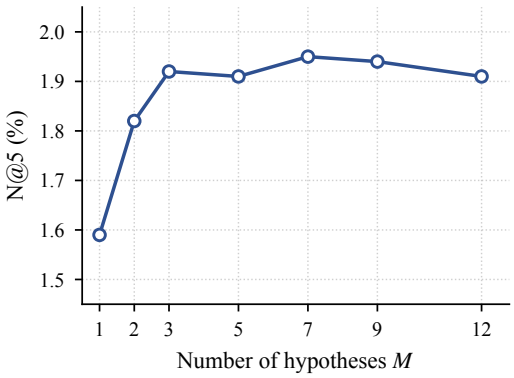


Figure 5: **Effect of the number of hypotheses.** N@5 on SCIENTIFIC under max-similarity aggregation, as a function of M. Performance improves sharply from $M = 1$ to $M = 3$, then plateaus: increasing M from 3 to 12 does not degrade ranking quality, in line with the sublinear dependence on M in Theorem 4.

Table 6: Mean pairwise cosine similarity between hypothesis heads on the test set. A value of 1 means the heads are indistinguishable, lower is more specialized.

Method	Instrument	Scientific	Game
MCT-EnsCE	1.00	1.00	1.00
MCT-Router	1.00	1.00	1.00
COMPOTE (ours)	**0.64**	**0.61**	**0.58**

Table 7: NDCG@5 (%) under max-similarity aggregation (Equation (10)) and under oracle head selection (Equation (11)). Equal columns indicate that all heads induce the same ranking, i.e. complete collapse.

	Instrument		Scientific		Game	
Method	Aggregated	Oracle	Aggregated	Oracle	Aggregated	Oracle
MCT-EnsCE	2.28	2.28	1.60	1.60	3.78	3.78
MCT-Router	2.28	2.28	1.64	1.64	3.78	3.78
COMPOTE (ours)	**2.45**	**3.75**	**1.92**	**3.52**	**4.01**	**7.78**

J Collapse and Oracle Evidence

K Proofs of Section 5

K.1 Proof of Theorem 7

Theorem 7 (Extension of Th. 1 – Rupprecht et al. (2017))**.** *Let* $\mathbf{q}^\star \in \Delta^{N-1}$*, and denote the optimal prediction* $\mathbf{p}^\star \in \operatorname{argmin}_{\mathbf{p}\in(\Delta^{N-1})^M} \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y},\mathbf{p})]$*. Then for every* $m \leq M$ *and* $y \in \mathcal{Y}$*,*

$$\mathbf{p}^\star_{m,y} = \frac{\mathbf{q}^\star_y}{Q_m}\mathbb{1}_{[y\in\mathcal{Y}_m]}\,, \quad \textit{with} \quad Q_m = \sum_{y\in\mathcal{Y}_m}\mathbf{q}^\star_y\,. \tag{12}$$

where the tessalation is $\mathcal{Y}_m = \left\{y \in \mathcal{Y} :\ \forall m' \neq m,\ -\log(\mathbf{p}^\star_{m,y}) < -\log(\mathbf{p}^\star_{m',y})\right\}$*. Moreover the tessalation* $(\mathcal{Y}_m)_{m=1}^M$ *is entropy-maximizing.*

Proof of Theorem 7. The characterization of $\mathbf{p}^\star$ using Q_m and the tessalation $\mathcal{Y}_m$ are a direct application of Theorem 7.

The only part remaining is to prove the tessalation $\mathcal{Y}_m$ is entropy-maximizing. Given a tessalation $C = \{C_1, \dots, C_m\}$ of $\mathcal{Y}$, define

$$\mathbf{p}^C_{m,y} = \frac{\mathbf{q}^\star_y}{Q^C_m}\mathbb{1}_{[y\in C_m]} \quad \text{where} \quad Q^C_m = \sum_{y\in C_m}\mathbf{q}^\star_y \tag{34}$$

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y},\mathbf{p}^C)] = -\sum_{m=1}^{M}\sum_{y\in C_m}\mathbf{q}^\star_y\log(\mathbf{p}^C_{m,y}) \tag{35}$$

$$= -\sum_{m=1}^{M}\sum_{y\in C_m}\mathbf{q}^\star_y\log\left(\frac{\mathbf{q}^\star_y}{Q^C_m}\right) \tag{36}$$

$$= H(\mathbf{q}^\star) - H(Q^C) \tag{37}$$

The choice $C^\star = (\mathcal{Y}_m)_{m=1}^M$ coincides with the optimal predictor – i.e. $\mathbf{p}^{C^\star} = \mathbf{p}^\star$, leading to

$$\mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p}^{C^\star})] = \min_{C:\text{M-tessalation of}\mathcal{Y}} \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p}^C)] \tag{38}$$

$$\Leftrightarrow H(\mathbf{q}^\star) - H(Q^{C^\star}) = \min_{C:\text{M-tessalation of}\mathcal{Y}} H(\mathbf{q}^\star) - H(Q^C) \tag{39}$$

$$\Leftrightarrow H(Q^{C^\star}) = \max_{C:\text{M-tessalation of}\mathcal{Y}} H(Q^C) \tag{40}$$

□

K.2 Optimal Risk

The following corollary makes explicit the value of the optimal winner-takes-all risk.

Corollary 10 (Optimal winner-takes-all risk)**.** *Under the setting of Theorem 7, with $Q = (Q_m)_{m=1}^M$ the cell masses of the tessalation $(\mathcal{Y}_m)_{m=1}^M$,*

$$\min_{\mathbf{p}\in(\Delta^{N-1})^M} \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p})] = H(\mathbf{q}^\star) - H(Q) \geq H(\mathbf{q}^\star) - \log M\,, \tag{41}$$

with equality in the inequality if and only if $Q_m = 1/M$ for every $m \leq M$.

Proof of Corollary 10. By Theorem 7, $\mathbf{p}^\star = \mathbf{p}^{C^\star}$ with $C^\star = (\mathcal{Y}_m)_{m=1}^M$, and the computation in the proof of Theorem 7 gives

$$\min_{\mathbf{p}\in(\Delta^{N-1})^M} \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p})] = \mathbb{E}_{\mathbf{y}\sim\mathbf{q}^\star}[\ell^{\text{WTA}}(\mathbf{y}, \mathbf{p}^{C^\star})] = H(\mathbf{q}^\star) - H(Q)\,. \tag{42}$$

Since $Q \in \Delta^{M-1}$, $H(Q) \leq \log M$, with equality if and only if Q is uniform. □

K.3 The Loss Follows the $\log M$ Law

Theorem 7 also predicts how the loss should change with M. For a given query, the optimal winner-takes-all loss is the Bayes loss minus the entropy $H(Q(\mathbf{x}))$ of the cell masses (Corollary 10). Exact querywise balance would therefore yield a $\log M$ reduction relative to the Bayes loss. Although the win frequencies of Section 5.2 establish only marginal balance, empirically $\hat{\mathcal{L}}^{\text{WTA}} + \log M$ varies by less than 0.1 across $M \in \{3, 4, 5\}$ (Table 4), showing that the learned models approximately follow this scaling law, where $\hat{\mathcal{L}}^{\text{WTA}} = \mathbb{E}[\min_m - \log f_{m,\mathbf{y}}(\mathbf{x})]$ is the test loss. Equivalently, going from $M = 3$ to $M = 5$ multiplies the geometric-mean probability that the winning head assigns to the target by $e^{0.61} \approx 1.83$, close to and slightly above the ratio $5/3 \approx 1.67$ predicted under exact balance.

L Latent Tiling Across M

This appendix details the protocol behind Table 4, reports additional diagnostics on the three COMPOTE checkpoints trained on GAME with $M \in \{3, 4, 5\}$, and checks that the partition also appears on SCIENTIFIC.

Protocol. Win frequencies, static sectors, and sector agreement use the full test set of $94{,}762$ query–target pairs. In-cell mass and the winner-takes-all loss use the softmax $f_m(\mathbf{x})$ over the whole catalogue, with $\beta = 0.07$, on the same full test set. The in-cell mass is $\sum_{y\in\mathcal{Y}_m(\mathbf{x})} f_{m,y}(\mathbf{x})$, averaged over heads and test pairs. The static direction $\widetilde{\mathbf{h}}_m$ of head m is the renormalized average of its normalized outputs over the test set. Writing $\bar{\mathbf{e}}_i = \mathbf{e}_i/\|\mathbf{e}_i\|_2$ for the normalized embedding of item i, its sector is $\operatorname{argmax}_m \langle \bar{\mathbf{e}}_i, \widetilde{\mathbf{h}}_m \rangle$. Sector agreement is the fraction of test targets whose sector is their winning head. Cross-head Jaccard, region loyalty, and single-head reach are computed from the top-10 lists on $4{,}000$ test queries sampled uniformly. Among catalogue items retrieved at least once, an item is reached by a single head if it appears in the top-10 of exactly one head. Within-head overlap compares the top-10 list of a head between randomly paired queries from this sample.

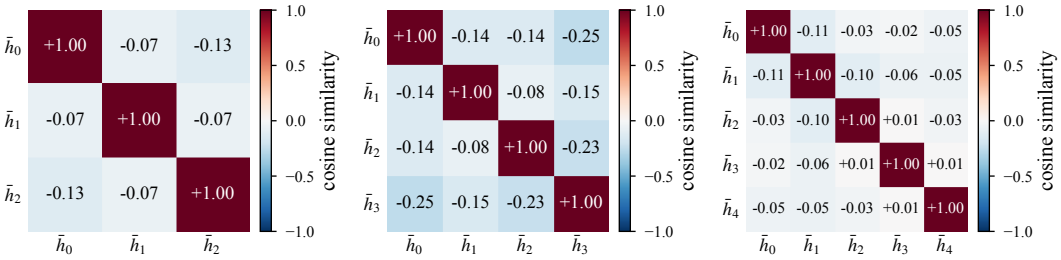
Table 8: Latent-space tiling on GAME as the number of hypotheses M varies. Mean off-diagonal cosine similarity is computed between the static head directions $\widetilde{\mathbf{h}}_m$ (the dataset-wide average of normalized head outputs). The top-10 diagnostics use 4,000 test queries sampled uniformly without replacement. Cross-head Jaccard is the mean pairwise overlap of these lists across heads. Region loyalty is the fraction of a head's top-10 items that fall in its catalogue sector.

M	Mean off-diag. $\cos(\widetilde{\mathbf{h}})$	Cross-head top-10 Jaccard	Region loyalty
3	-0.09	0.008	98.9%
4	-0.17	0.003	99.4%
5	-0.04	0.001	99.8%

Cosine and softmax winners agree. The oracle of Equation (11) selects heads by cosine similarity, whereas the cells of Theorem 7 are defined by probabilities, and each head carries its own softmax normalizer. The two definitions of the winning head agree on 98.9%, 99.4%, and 99.7% of test pairs for $M = 3, 4, 5$.

The tiling sharpens with M. The static directions are well separated for every M (Figure 6). Cross-head top-10 Jaccard falls from 0.008 at $M = 3$ to 0.001 at $M = 5$, while region loyalty rises from 98.9% to 99.8% (Table 8). Theorem 7 predicts an exact partition for every M, so this trend describes how close training gets to the optimum rather than a property of the optimum itself.

The same partition appears on SCIENTIFIC. We repeat the overlap diagnostics on 4,000 sampled test queries for the COMPOTE checkpoint trained on SCIENTIFIC with $M = 3$. The top-10 lists of two heads on the same query have a mean Jaccard overlap of 0.001, and 94.5% of the catalogue items retrieved at test time are reached by a single head only. The top-10 of a given head overlaps by 0.025 to 0.044 Jaccard between two different histories, so each head still adapts to the query. Figure 7 shows three test queries. A plain PCA of the normalized catalogue already reveals three arms, and the heads sit along distinct arms, with the target close to one of them.



(a) $M = 3$ (b) $M = 4$ (c) $M = 5$

Figure 6: **Latent-space tiling on GAME.** Cosine similarity between the static head directions $\bar{\mathbf{h}}_m$ (the dataset-wide average of normalized head outputs) for $M \in \{3, 4, 5\}$. Off-diagonal values remain near zero or negative, indicating that the heads occupy separated directions rather than collapsing onto one another.

M BEYOND THE THEOREM: A STATIC TILING

Theorem 7 allows the cells to depend on the query. To test whether they do, we average the normalized outputs of each head over the test set and assign every catalogue item to the static sector of the direction it is most aligned with (protocol in Appendix L). The sector of the target predicts its winning head on 98.5% to 99.4% of test pairs (last column of Table 4), and more than 98.9% of each head's top-10 items lie in its sector. Cell membership is thus almost a function of the item alone. COMPOTE learns a fixed tiling of the catalogue, and only the ranking inside each tile adapts to the history, as (P2) requires. A static tiling cannot balance every query: for a user whose likely next items all lie in one tile, $Q(\mathbf{x})$ is concentrated on a single head, and that head gets no boost from

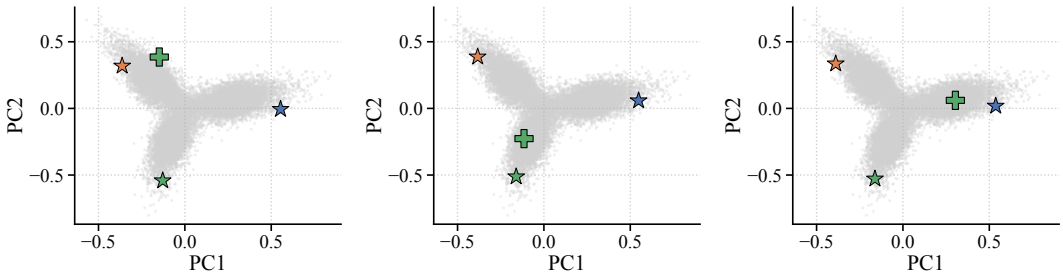


Figure 7: **Catalogue partition on SCIENTIFIC ($M = 3$).** PCA projections on three test sequences (queries 4934, 23025, and 26314) under COMPOTE. Each panel shows the L2-normalized item catalogue in two dimensions (grey), the target next item (green cross), and the three head query embeddings (stars). The heads occupy distinct regions of the catalogue rather than collapsing onto a single direction.

the partition. The balance reported in Section 5.2 holds on average over queries, which is what win frequencies measure.