Low-Rank Embedding Adaptation for Models with Expanding Vocabularies

Rishabh Agrawal [1]

Abstract

Pretrained models are routinely extended with new vocabulary entries, for example new items in recommender systems and new tokens in large language models (LLMs). We show that jointly training old and new embeddings has a hidden failure mode: old-entry quality degrades while new entries are still learning. On sequential recommendation, old items overfit while new items improve, forcing premature early stopping; on LLMs, base-token perplexity rises within 1–2 epochs of joint training. We propose *population-specific low-rank subspaces* that extend the low-rank adaptation principle to embedding tables: each entry is parameterized as $e_i = \bar{e}_i + u_i V_k^\top$, where V_k is shared within a population and separate across populations with different data regimes. The shared V_k provides implicit regularization for sparse new entries while the separation prevents gradient interference between populations. Three instantiations (Freeze-SV, Freeze1-SV, Dual-SV) cover different regimes. On sequential recommendation (SASRec and GRU4Rec architectures; Taobao and MerRec datasets), our methods Pareto-dominate joint fine-tuning and continual-learning baselines. On LLM vocabulary expansion (GPT-2, Pythia-410M, Pythia-1.4B on BillSum and PubMed), freezing base embeddings (the core of our approach) improves overall perplexity over joint training in 8 of 9 model-scale cells, with the low-rank parameterization providing 2–4$\times$ compression at matched quality.

1. Introduction

Pretrained models are routinely extended with new vocabulary entries: recommender systems add new items as catalogs grow, and language models acquire domain-specific tokens for specialized corpora. When these new entries are trained jointly with existing ones, a hidden failure mode emerges. On sequential recommendation, base-item quality degrades over training while new items are still improving, forcing premature early stopping. On LLMs, base-token perplexity rises within 1–2 epochs of joint training. This failure mode is related to catastrophic forgetting in continual learning (Kirkpatrick et al., 2017): adapting the model to serve new entries degrades its performance on existing ones.

The dominant approach for parameter-efficient adaptation, LoRA (Hu et al., 2022), targets transformer weight matrices via low-rank updates $\Delta W = BA$. However, in sequential recommenders like SASRec (Kang & McAuley, 2018), 99.9% of parameters reside in the embedding table, not the transformer. LoRA adapts the 0.1% and leaves the dominant block untouched. Similarly, new vocabulary entries added to any model must learn their embeddings from scratch; adapting transformer layers alone does not provide per-entry representations.

Background and vocabulary. We study two settings where embedding tables expand incrementally. In *sequential recommendation* ("seqrec" hereafter), a model predicts the next item a user will interact with, given their interaction history. We evaluate using Normalized Discounted Cumulative Gain at rank 10 (NDCG@10), a standard ranking metric. In *LLM vocabulary expansion*, a pretrained language model is extended with new domain-specific tokens (e.g., legal or medical terminology) and further pretrained on domain text; we evaluate using perplexity (PPL), the exponentiated average cross-entropy loss. In both settings, the vocabulary is partitioned into *base entries* (present during pretraining) and *new entries* (added at adaptation time). We use fine-tuning (FT) to refer to joint training of all parameters, and early stopping (ES) to refer to halting training when a validation metric degrades.

Low-rank embedding subspaces. We propose *population-specific low-rank embedding subspaces*, extending the low-rank adaptation principle from weight matrices to embedding tables. Our factorization parameterizes each

[1]AWS AI, Santa Clara, CA, USA. Correspondence to: Rishabh Agrawal <riagrawa@amazon.com>.

ICML 2026 Workshop on Connecting Low-rank Representations in AI (CoLorAI), Seoul, South Korea.

entry as:

$$e_i = \bar{e}_i + u_i \cdot V_k^\top, \quad u_i \in \mathbb{R}^r,\ V_k \in \mathbb{R}^{d \times r} \tag{1}$$

where $\bar{e}_i$ is a frozen anchor (the pretrained embedding for base entries, or an initialization for new entries), V_k is a projection matrix shared within a population, and u_i provides per-entry coordinates in the shared subspace. The shared V_k acts as a regularizer for sparse new entries and enables information pooling across entries within a population. Combined with LoRA on the transformer layers, this gives a complete low-rank adaptation covering both weight matrices and embedding tables.

Population-specific structure. A key challenge is that vocabularies expand incrementally: new entries arrive with sparse interaction data alongside well-established entries. Naïve low-rank adaptation of the full table with a single shared V fails because established-entry gradients dominate the shared subspace, preventing new entries from learning useful directions. We show that population-specific V_k matrices (separate projections for base and new entries, where "populations" denote groups of entries with different data regimes) are essential. This separation resolves an *anti-correlation* phenomenon where base-entry quality degrades while new entries are still improving (Figure 1), eliminating the need for premature early stopping.

We propose three instantiations that differ in how they handle the base population:

1. **Freeze-SV**: base embeddings are frozen throughout training; only new-entry parameters are learned. This is the simplest variant and is most competitive on well-pretrained LLMs where base representations are already strong.
2. **Freeze1-SV**: all parameters are trainable for one warmup epoch, then base embeddings are frozen. This variant is epoch-robust (it trains for 20 epochs without base degradation) and dominates on seqrec tasks and smaller LLMs.
3. **Dual-SV**: base and new entries each have their own subspace (V_{base}, V_{new}) with the base subspace trained at a $10\times$ lower learning rate. This achieves the highest new-entry quality under early stopping on seqrec.

Contributions.

1. We extend low-rank adaptation from weight matrices to embedding tables, providing parameter efficiency ($4\times$ compression at matched quality) in embedding-dominated models and preventing base-entry overfitting during LLM vocabulary expansion.
2. We show that population-specific structure is necessary when vocabularies expand incrementally, supported by an ablation (single shared V: 0.229 new NDCG vs. population-specific: 0.877) and a convergence analysis predicting $O(d/r)$ variance reduction for sparse entries.
3. We demonstrate Pareto-dominance over fine-tuning and continual-learning baselines on two seqrec datasets (Taobao, MerRec) and two architectures (SASRec, GRU4Rec). On LLM vocabulary expansion, base-freezing combined with our low-rank parameterization improves overall PPL over joint training in 8 of 9 model-scale cells across two domains (BillSum, PubMed), with the low-rank structure providing 2–4$\times$ parameter efficiency at matched quality relative to full-rank frozen baselines.

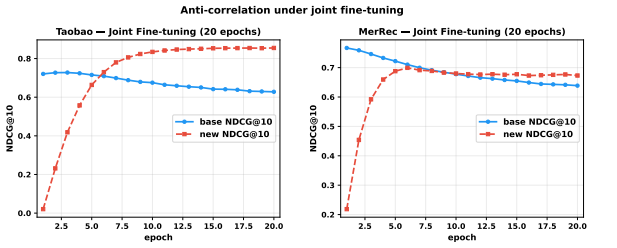

Figure 1. Anti-correlation under joint fine-tuning (W0, 20 epochs without early stopping). Blue: base NDCG@10 (drops). Red: new NDCG@10 (rises). Left: Taobao. Right: MerRec.

2. Related Work

Low-rank adaptation of weight matrices. LoRA (Hu et al., 2022) decomposes weight updates as $\Delta W = BA$, training only low-rank factors. GaLore (Zhao et al., 2024) projects gradients onto a low-rank subspace; QLoRA (Dettmers et al., 2023) combines quantization with LoRA. All target weight matrices; none address the embedding table. SD-LoRA (Wu et al., 2025) uses separate LoRA branches for old and new tasks, making it the closest structural precedent to our approach, though it operates on weight matrices for task-level separation rather than on embedding tables for population-level separation.

Incremental recommendation. EWC (Kirkpatrick et al., 2017) adds Fisher-weighted regularization; ADER (Mi et al., 2020) uses replay with distillation. Both focus on preventing forgetting but do not examine the base-vs-new interaction during joint training. Fatkulin et al. (Fatkulin et al., 2025) propose bounded-delta embeddings but do not prevent base degradation (our reimplementation shows -5.8pp base NDCG over 20 epochs).

LLM vocabulary expansion. Cui et al. (Cui et al., 2023b) and Yamaguchi et al. (Yamaguchi et al., 2024) study token initialization but do not identify that joint training degrades base-token quality during vocabulary expansion.

Embedding compression. ALBERT (Lan et al., 2020)

factorizes the entire embedding layer as $E = E_{\text{small}}P$ for uniform parameter reduction during pretraining. Mixed Dimension Embeddings (Ginart et al., 2021) assign different dimensions by frequency; FIITED (Lian et al., 2021) prunes per-embedding dimensions. All three are static techniques applied uniformly. Our method is selective (new entries only), incremental (post-deployment), and population-specific (separate projections for base vs. new).

3. The Anti-Correlation Problem

We formalize the setting as follows. A pretrained model with embedding table $\bar{E}$ is extended with new entries and fine-tuned on adaptation data containing both populations. The joint objective $\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{new}}$ shares encoder parameters θ, creating a coupling through which improving one population's loss can degrade the other's representations.

Sequential recommendation. We train SASRec (Kang & McAuley, 2018) on Taobao and MerRec with 20 epochs of joint fine-tuning (no early stopping). Base-only NDCG@10 drops by 13% on Taobao and 16% on MerRec over these 20 epochs, while new-item NDCG rises continuously (Figure 1). Our methods hold base quality within $\pm 1\%$ of the pretrained level.

Table 1. Training dynamics (W0, 20 epochs, no early stopping, r=128). Δ: change in NDCG@10 from ep1→20; @20: absolute value at epoch 20.

Dataset	Method	base (Δ / @20)	new (Δ / @20)	peak ovrl	ep
Taobao	FT	−0.093 / 0.628	+0.834 / 0.855	0.729	3
	Freeze1-SV	+0.001 / 0.720	+0.538 / 0.871	0.730	19
	Dual-SV	−0.013 / 0.699	+0.538 / 0.875	0.733	7
MerRec	FT	−0.128 / 0.639	+0.455 / 0.673	0.751	3
	Freeze1-SV	−0.009 / 0.755	+0.080 / 0.672	0.768	18
	Dual-SV	−0.021 / 0.743	+0.140 / 0.722	0.767	2

Table 1 quantifies the tradeoff: FT peaks on overall NDCG at epoch 3 because base quality (which dominates the overall metric) is already declining, while new-item quality is still rising. Our methods achieve higher peak overall NDCG at later epochs because base quality is preserved, allowing the overall metric to benefit from continued new-item improvement.

LLM vocabulary expansion. We observe the same phenomenon on GPT-2 and Pythia (410M, 1.4B) when expanded with 2000 domain tokens (BillSum legal text, PubMed medical abstracts). Under joint training, base-token perplexity rises within 1–2 epochs even as new-token perplexity improves. Freezing base embeddings eliminates this degradation and wins on overall perplexity in 8 of 9 model-scale cells (Figure 4).

4. Method

4.1. Framework

The general formulation (Eq. 1) has three desirable properties. First, gradient decoupling: separate V_k matrices prevent gradient interference between populations. Second, implicit regularization: sharing V_k within a population pools information across entries, acting as a data-dependent regularizer for sparse items. Third, parameter efficiency: each new entry learns only r coordinates rather than d independent parameters, yielding $4\times$ compression at $r/d = 25\%$. Figure 2 illustrates the end-to-end pipelines for both sequential recommendation and LLM vocabulary expansion.

While structurally analogous to LoRA (which decomposes $\Delta W = BA$ shared across all inputs), our factorization operates at the level of individual entries. The full embedding update can be written as $\Delta E = UV^\top$ where $V \in \mathbb{R}^{d\times r}$ is shared across entries and $U \in \mathbb{R}^{n\times r}$ stores per-entry coordinates. Unlike LoRA, where the correction ΔW is identical for every input, our correction $\Delta e_i = u_i V^\top$ differs per entry.

4.2. Freeze-SV (No Warmup)

In this simplest variant, base embeddings are frozen from epoch 0 and never updated. New entries are parameterized as $e_j = \bar{e}_j + u_j V^\top$ with a shared projection V. Freeze-SV is most effective for LLM vocabulary expansion, where the pretrained base representations are already strong and even one epoch of joint training begins overfitting base tokens.

4.3. Freeze1-SV (Warmup + Freeze)

This variant trains all parameters for one warmup epoch, allowing the encoder to observe the joint distribution of base and new entries. From epoch 2 onward, base embeddings are frozen. New entries use the same low-rank parameterization: $e_j = \bar{e}_j + u_j V^\top$. Freeze1-SV is epoch-robust, training for 20 epochs without base degradation (less than 0.01 NDCG drift). It dominates on sequential recommendation where the encoder is small relative to the embedding table and benefits from one epoch of joint adaptation.

Algorithm 1 Freeze1-SV Training (per window)

1: **Input:** Pretrained $\bar{E}$, new entries $\{j\}$, rank r
2: Initialize anchors: $\bar{e}_j \leftarrow$ task-dependent (mean of base embeddings, content embedding, or multivariate normal)
3: Init: $u_j \leftarrow \mathbf{0}$, $V \leftarrow \text{Xavier}(r, d)$; embedding: $e_j = \bar{e}_j + u_j V^\top$
4: **Epoch 1 (warmup):** Train all ($\bar{E}$, $\{u_j\}$, V, encoder + LoRA)
5: **Epoch 2+:** Freeze $\bar{E}$ and anchors $\{\bar{e}_j\}$; train $\{u_j\}$, V, encoder
6: **Return:** Updated model for next window

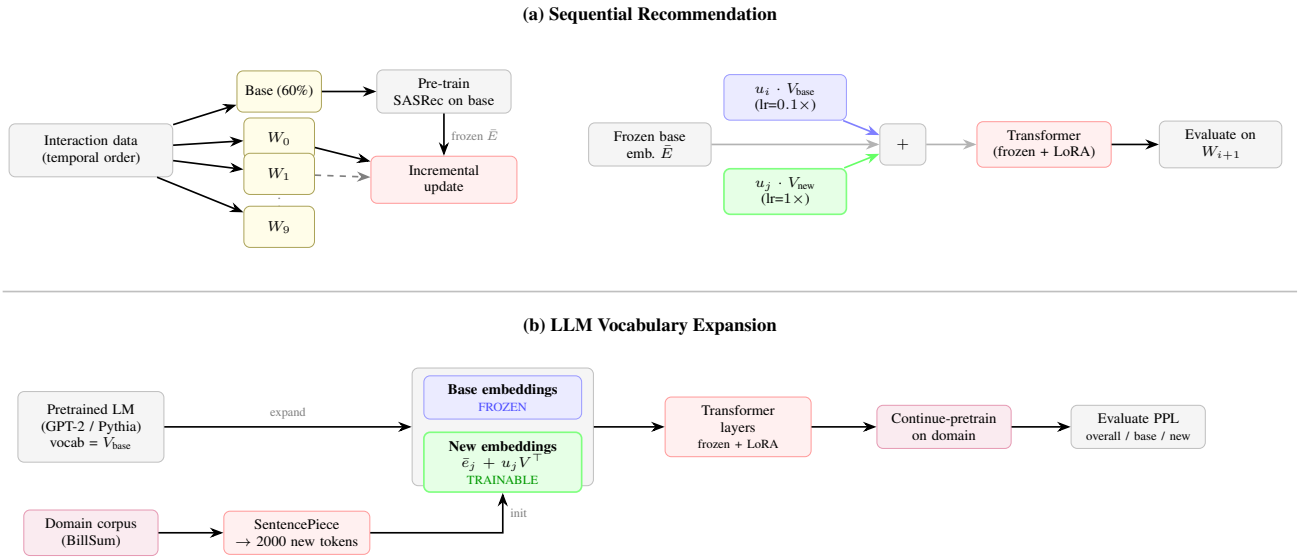

Figure 2. End-to-end pipelines. **(a)** Sequential recommendation: frozen base embeddings with separate low-rank corrections (V_{base}, V_{new}), LoRA on attention. **(b)** LLM vocabulary expansion: new rows parameterized as $\bar{e}_j + u_j V^\top$, base frozen, LoRA on all layers.

4.4. Dual-SV (Separate Subspaces)

The most expressive variant assigns separate subspaces to base and new entries, trained at different learning rates:

$$e_i^{\text{base}} = \bar{e}_i + u_i^{\text{base}} (V^{\text{base}})^\top, \ \text{lr} = 0.1 \times \text{lr}_{\text{new}} \quad (2)$$

$$e_j^{\text{new}} = \bar{e}_j + u_j^{\text{new}} (V^{\text{new}})^\top, \ \text{lr} = \text{lr}_{\text{new}} \quad (3)$$

Base entries receive a controlled low-rank delta at $10\times$ lower learning rate, allowing mild co-adaptation without catastrophic drift. This achieves the best new-entry quality at the best-overall epoch under standard early stopping, because the encoder can co-adapt to serve both populations simultaneously.

Algorithm 2 Dual-SV Training (per window)

Input: Pretrained $\bar{E}$, new entries $\{j\}$, rank r
Initialize anchors: $\bar{e}_j \leftarrow$ task-dependent (same as Freeze1-SV)
Base anchors: $\bar{e}_i \leftarrow$ pretrained embeddings from $\bar{E}$
Init: $u_j^{\text{new}} \leftarrow \mathbf{0}$, $V^{\text{new}} \leftarrow \text{Xavier}(r, d)$
Init: $u_i^{\text{base}} \leftarrow \mathbf{0}$, $V^{\text{base}} \leftarrow \text{Xavier}(r, d)$
Set $\text{lr}_{\text{base}} = 0.1 \times \text{lr}_{\text{new}}$
for each epoch **do**
  Train $\{u_j^{\text{new}}\}$, V^{new} at lr_{new}
  Train $\{u_i^{\text{base}}\}$, V^{base} at lr_{base}
  Train encoder (+ LoRA)
  Stop if val-overall metric degrades
end for
Return: Updated model for next window

4.5. Trainable Parameters

For a window with n_b active base items and n_n new items at rank r, Dual-SV trains:

$$(n_b + n_n) \cdot r + 2rd + \text{LoRA params} \quad (4)$$

On MerRec (n_b=1.5M, n_n=1.5M, r=32, d=150): 106M embedding params vs. 678M for full FT (6.4× reduction). Each item learns only r parameters regardless of d. Freeze1-SV is even cheaper: only new items have trainable u_j ($n_n \cdot r + rd$).

4.6. Theoretical Motivation

We define the *shared-V estimator* as the estimator $\hat{e}_j = \hat{u}_j \hat{V}^\top$ that estimates each entry's coordinate u_j individually while sharing V across all entries, in contrast to estimating each $e_j \in \mathbb{R}^d$ independently.

Proposition 1 (Shared-V error bound)**.** *Consider n entries, each observed k_j times with noise variance σ^2. Under approximate low-rank structure ($\|e_j^* - u_j^* V^*\| \le \tau$), the shared-$V$ estimator achieves per-entry error $O(r\sigma^2/k_j)$ vs. $O(d\sigma^2/k_j)$ for independent estimation – an $O(d/r)$ improvement.*

Proof sketch. The shared V is estimated from all $N = \sum_j k_j$ observations jointly. Once V is known, each entry estimates only its r-dimensional coordinate u_j from k_j samples, reducing variance by factor d/r compared to estimating the full d-dimensional embedding independently. The condition for this to help is $N \gg rd/(d-r)$, which is satisfied by orders of magnitude on both datasets (Taobao: $N \approx 10^7$, threshold ≈ 873).

The bound predicts: (P1) advantage concentrated in sparse entries; (P2) diminishing returns as per-entry data grows; (P3) a data-scale crossover where full-rank becomes preferable. P1 follows from the bound structure (r/k_j dominates when k_j is small) and is consistent with our new-item results (median $k_j = 1$). P2 is confirmed by rank ablations. P3 is confirmed by the GPT-2 15K exception.

5. Experiments

5.1. Sequential Recommendation

Setup. We use SASRec (Kang & McAuley, 2018) with embedding dimension d=150 and 2 transformer layers. We evaluate on two datasets that represent different item-turnover regimes: Taobao (Zhu et al., 2018) (91M interactions, 1.2M items, 2-hour windows with 10% new items per window) and MerRec (Sato et al., 2025) (280M interactions, 4.5M items, 1-day windows with 61% new items per window). We perform 10-window continual evaluation (W0 through W9, where each window introduces new items) with 20 epochs per window. We use the sampled-100 evaluation protocol, which ranks the ground-truth next item among 100 randomly sampled negatives (Figure 2a).

Baselines. (1) Stale: no incremental training. (2) FT: all parameters trainable. (3) EWC (Kirkpatrick et al., 2017): Fisher-weighted regularization. (4) ADER (Mi et al., 2020): experience replay with distillation.

Our methods. Freeze1-SV and Dual-SV (r=128) with LoRA (rank 8) on Q/K/V attention projections, giving a complete low-rank adaptation of both the embedding table and the transformer encoder.

Table 2. 10-window continual: mean NDCG@10 (val-overall ES).

	Taobao		MerRec	
Method	new	base	new	base
FT	0.067	0.717	0.661	0.742
EWC	0.063	0.717	0.671	0.748
ADER	0.147	0.712	0.660	0.753
Freeze1-SV	0.071	0.717	0.659	0.754
Dual-SV	**0.624**	0.709	**0.711**	0.754

Under val-overall early stopping (Table 2), Dual-SV achieves 0.624 new NDCG on Taobao, a 4× improvement over ADER (0.147). This is because its decoupled optimization allows the overall metric to keep improving without early stopping prematurely from base degradation. Freeze1-SV's advantage is epoch-robustness: without early stopping it reaches 0.770 new NDCG while maintaining stable base quality (0.728). In the 10-window continual evaluation, Freeze1-SV converges in 3.3 epochs per window (under early stopping) versus 12.1 for FT, at 27s versus 46s per epoch (a 6× wall-clock speedup).

Pareto dominance (Figure 3). On window 0 without early stopping, Freeze1-SV and Dual-SV occupy the top-right quadrant of the base-vs-new NDCG plane (high base *and* high new), dominating every point on FT's trajectory. Without early stopping, FT eventually reaches 0.855 new NDCG, but base quality collapses to 0.632 (a 12% drop from its epoch-1 peak).

GRU4Rec (architecture generality). To verify that the phenomenon is not specific to transformers, we evaluate on GRU4Rec (Hidasi et al., 2016), an RNN-based model without attention. The same pattern holds: Freeze1-SV achieves higher overall NDCG than FT on both datasets (Taobao: 0.833 vs. 0.820; MerRec: 0.920 vs. 0.890).

Rank ablation (Table 3). At r=64 (43% of d), Dual-SV matches full FT on new-item quality (0.863 vs. 0.867) while preserving base quality. At r=128 (85%), it exceeds FT (0.872 vs. 0.867), suggesting that the rank constraint provides beneficial regularization even at high capacity.

Table 3. Rank ablation (Taobao W0, no ES). Higher rank improves new quality with diminishing returns; base is stable.

Variant	r	r/d	new	base
Dual-SV	16	11%	0.604	0.721
Dual-SV	32	21%	0.794	0.718
Dual-SV	64	43%	0.863	0.718
Dual-SV	128	85%	0.872	0.722
Full FT	d	100%	0.867	0.690

Content-initialized embeddings. When new items have content features (E5 (Wang et al., 2024) text embeddings), the anchor $\bar{e}_j$ can be set to the content embedding rather than the base mean. On MerRec with content initialization (Table 4), lower rank suffices: Dual-SV r=32 achieves 0.849 new NDCG, outperforming r=128 (0.832), because the content anchor already provides a good starting point. The Fatkulin (Fatkulin et al., 2025) bounded-delta baseline does not prevent base degradation (−5.8pp).

Table 4. Content-initialized embeddings on MerRec (W0, best-overall-NDCG epoch). Anchor is the E5 content embedding. Δbo denotes the base-only NDCG change from epoch 1 to epoch 20.

Method	rank	ep	overall	base	new	Δbo
FT (content)	full	1	0.804	0.807	0.784	−0.074
Fatkulin (δ)	full	2	0.803	0.801	0.813	−0.058
Freeze1-SV	128	9	0.817	0.819	0.800	+0.016
Freeze1-SV	32	6	0.810	0.804	0.861	−0.002
Dual-SV	128	6	**0.822**	0.820	0.832	+0.006
Dual-SV	32	7	0.820	0.816	**0.849**	−0.000

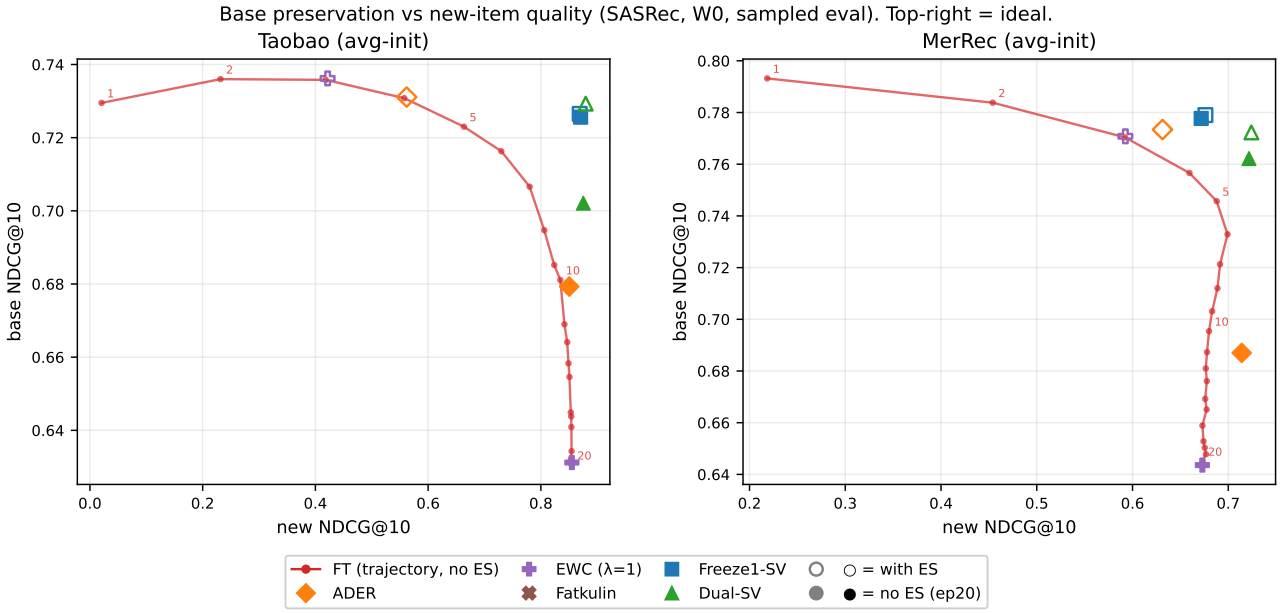

Figure 3. Pareto comparison (W0, 20 epochs without early stopping). Gray trajectory: FT epochs 1–20. Our methods (Freeze1-SV = blue, Dual-SV = green) Pareto-dominate every point on FT's trajectory. Left: Taobao 2h. Right: MerRec 1d.

5.2. LLM Vocabulary Expansion

Setup. GPT-2 (Radford et al., 2019) (124M, d=768), Pythia-410M and Pythia-1.4B (Biderman et al., 2023) (d=1024, d=2048). We add 2000 new tokens via SentencePiece trained on BillSum (Kornilova & Eidelman, 2019), a corpus of US Congressional bills whose specialized legal vocabulary is poorly served by general-purpose tokenizers, making it a natural test case for vocabulary expansion. We further pretrain on 1K/5K/15K documents with LoRA rank 8 on all linear layers (Figure 2b). New-token embeddings are initialized via multivariate normal matching the base embedding distribution.

Baselines. (1) Basetrn: base embeddings trainable, full-dim new rows (Cui et al. (2023b) Stage-1 recipe). (2) Full-rank frozen: base frozen, full-dim new rows.

Our methods. Freeze-SV, Freeze1-SV, and Dual-SV at $r/d \approx 50\%$.

Table 5. BillSum: best-epoch overall PPL (↓). Bold indicates the best result in each row.

Model	Scale	Basetrn (full)	Frozen (full)	Freeze-SV	Freeze1-SV	Dual-SV
GPT-2	1K	6.40	6.42	6.39 / 6.49	**6.30** / 6.40	6.55 / 6.65
	5K	5.95	5.98	5.96 / 6.04	**5.84** / 5.90	5.94 / 6.02
	15K	**4.79**	5.27	5.26 / 5.29	5.02 / 5.05	5.11 / 5.18
P-410M	1K	4.85	4.80	4.80 / 4.81	4.79 / 4.81	**4.78** / 4.80
	5K	5.06	**4.69**	**4.69** / 4.71	4.83 / 4.84	**4.69** / 4.70
	15K	4.13	4.06	4.06 / 4.06	4.08 / 4.09	**4.04** / 4.05
P-1.4B	1K	4.14	4.09	4.09 / 4.10	**4.07** / **4.07**	**4.07** / 4.09
	5K	4.40	4.04	**4.03** / 4.04	4.09 / 4.11	**4.03** / 4.04
	15K	3.67	3.54	3.54 / 3.54	3.56 / 3.56	**3.52** / 3.53

r/d: 50% / 25%. GPT-2 r=384/192; P-410M r=512/256; P-1.4B r=1024/512.

Table 5: Base-freezing wins in 8/9 cells (Figure 4). The exception (GPT-2 15K) occurs when the small model has sufficient data to tolerate joint training. At both $r/d = 50\%$ and 25%, our low-rank methods match full-rank frozen within 0.01–0.02 PPL, confirming the rank constraint does not hurt while providing 2–4× parameter efficiency. Multi-seed stability is confirmed in Appendix A: across 3 seeds, standard deviation is below 0.03 PPL for all cells and method ordering is preserved.

At $r/d = 50\%$, our low-rank methods match or slightly beat full-rank frozen (e.g., Dual-SV 4.03 vs. 4.04 on Pythia-1.4B 5K). At $r/d = 25\%$, quality remains within 0.01–0.02 PPL at 4× fewer parameters. The improvement over joint training comes from base preservation: basetrn degrades base-token PPL by 7–18% while our frozen methods prevent

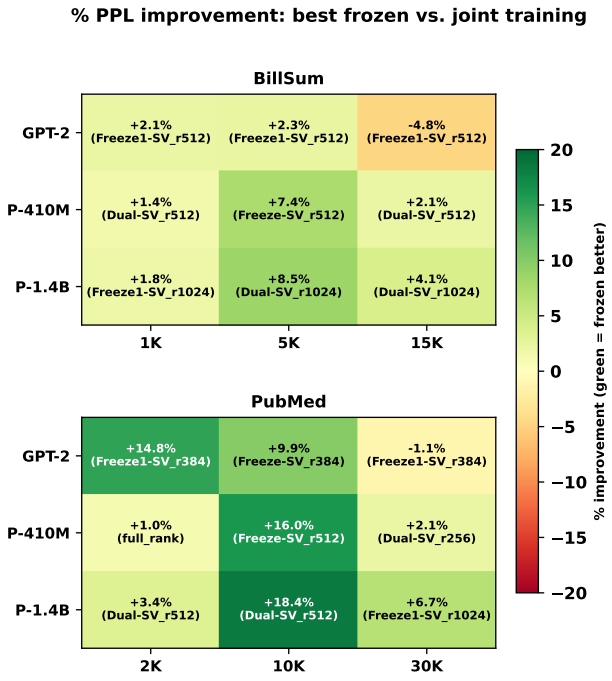

Figure 4. Best frozen method vs. joint training on BillSum (top) and PubMed (bottom). Green = frozen better. Base-freezing wins in 8/9 cells on both domains.

this, winning on overall PPL (Figure 5).

PubMed replication. We replicate on PubMed medical abstracts (U.S. National Library of Medicine, 2024) (token-matched scales: 2K/10K/30K documents) and observe the same 8/9 pattern (Table 6).

Table 6. PubMed: best-epoch overall PPL (↓). Bold indicates the best result in each row.

Model	Scale	Basetrn (full)	Frozen (full)	Freeze-SV	Freeze1-SV	Dual-SV
GPT-2	2K	51.6	45.6	**44.0** / 44.2	**44.0** / 44.1	48.0 / 46.1
	10K	41.4	37.7	**37.3** / 37.6	37.5 / 37.7	41.7 / 40.1
	30K	**29.8**	31.3	31.1 / 31.5	30.1 / 30.6	32.0 / 32.3
P-410M	2K	23.8	**23.5**	23.6 / 23.8	23.6 / 23.8	23.7 / 23.7
	10K	26.1	**21.9**	**21.9** / 22.1	22.1 / 22.4	22.3 / 22.0
	30K	14.0	**13.8**	14.1 / 13.9	14.0 / 13.8	13.9 / **13.7**
P-1.4B	2K	17.5	17.4	**17.0** / 17.1	**17.0** / 17.1	17.1 / **16.9**
	10K	22.6	19.3	19.2 / 18.8	19.4 / 19.2	**19.0** / **18.4**
	30K	10.7	10.1	10.1 / 10.1	**10.0** / 10.1	10.3 / 10.1

r/d: 50% / 25%. GPT-2 r=384/192; P-410M r=512/256; P-1.4B r=1024/512.

6. Analysis

Why separate V matrices? When a single shared V is used for both populations on Taobao (Table 7), new-item NDCG reaches only 0.229, compared to 0.877 with population-specific projections (Freeze-SV). Base gradients dominate the shared subspace because base items are more numerous and have denser interactions, preventing new items from learning useful directions. The population-specific structure is therefore essential, not merely beneficial.

Table 7. Ablations (Taobao W0, no ES, best epoch). All use LoRA on attention.

Variant	new NDCG	base NDCG
LoRA only (frozen emb.)	0.000	0.710
LoRA + full emb. update	0.867	0.690
Shared-V (single V, all items)	0.229	0.600
Freeze-SV (r=128)	0.877	0.720
Freeze1-SV (r=128)	0.872	0.722
Dual-SV (r=128)	0.872	0.718

LoRA alone is insufficient. Frozen embeddings with LoRA applied only to the attention layers yields zero new-item quality (Table 7). The embedding delta is the critical component; LoRA on the encoder provides complementary adaptation but cannot substitute for embedding-level learning when new entries have no pretrained representation.

Architecture generality. As shown in Section 5.1, the anti-correlation and the benefit of population-specific factorization also hold on GRU4Rec (Hidasi et al., 2016), an RNN without attention. The phenomenon is not specific to the transformer architecture.

V-initialization sensitivity. The shared projection V can be initialized from the top-r PCA directions of the base embedding table or via Xavier random initialization. On Taobao, PCA outperforms Xavier by $+0.8$pp NDCG; on MerRec, Xavier outperforms PCA by $+1.9$pp. Zero initialization fails completely, as V cannot discover useful directions from zero within the training budget. The sensitivity is mild: both non-zero initializations converge to similar quality by epoch 10, suggesting that V learns task-appropriate directions regardless of starting point when given sufficient training signal.

Rank selection for LLM vocabulary expansion. At $r/d \geq 50\%$, low-rank methods match or slightly beat full-rank frozen (the rank constraint provides mild regularization). At $r/d \approx 25\%$ with multivariate anchors, quality remains within 0.01 PPL. Below $r/d = 15\%$, the low-rank constraint becomes too restrictive. The relevant quantity is the ratio r/d, not the absolute value of r.

The parameter inversion across architectures. In LLMs (GPT-2: 124M parameters, 31% in embeddings), most parameters reside in transformer layers, which is where LoRA operates. In sequential recommendation (SASRec with 4.5M items: 99.9% in embeddings), the distribution inverts. Here our method adapts the dominant parameter block while providing $4\times$ compression at $r/d = 25\%$. For LLM vocabulary expansion, new embedding rows constitute a small

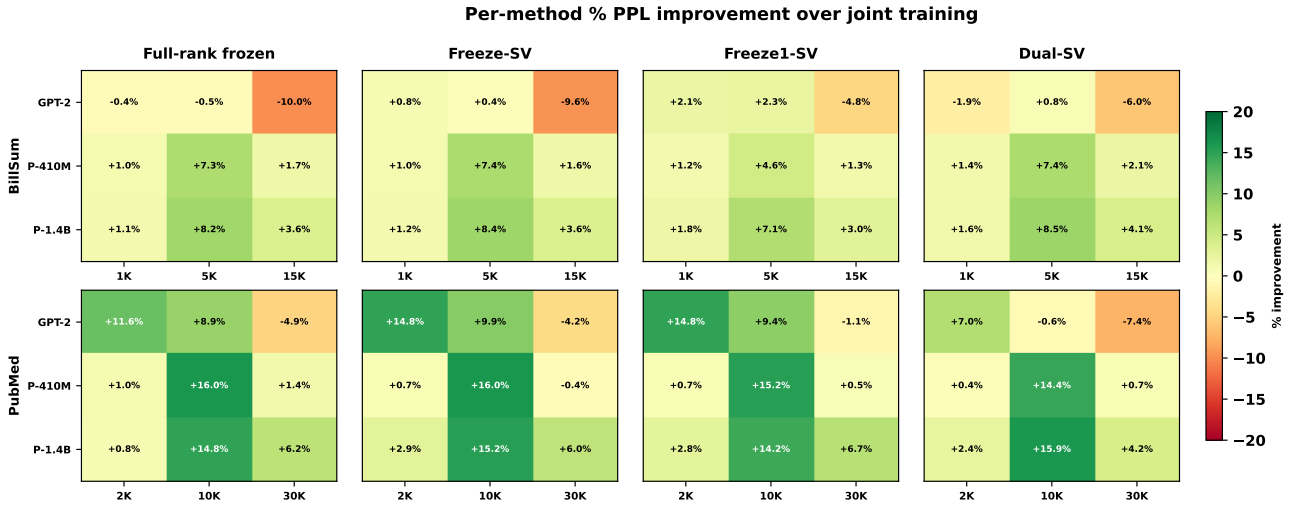

Figure 5. Per-method PPL improvement over joint training at $r/d \approx 50\%$. Top row: BillSum. Bottom row: PubMed. Ranks: GPT-2 r=384; P-410M r=512; P-1.4B r=1024.

fraction of total model parameters, so parameter efficiency is a secondary benefit; the primary value is preventing the base-token overfitting that joint training causes within 1–2 epochs.

Limitations. Our LLM experiments reach 1.4B parameters; verification at 7B+ is future work. We evaluate on two LLM domains (BillSum legal text, PubMed medical abstracts) and two seqrec datasets (Taobao 91M interactions, MerRec 280M interactions). The anti-correlation persists across LLM scales (124M to 1.4B) and seqrec architectures (SASRec, GRU4Rec), but our seqrec encoder is small (2 layers, d=150); scaling to larger seqrec encoders is left to future work. On GPT-2 at 15K documents, joint training wins because small models with abundant data tolerate the anti-correlation. On knowledge graph completion (TransE (Bordes et al., 2013) on ENTITY (Cui et al., 2023a)), the anti-correlation does not manifest and low-rank hurts because TransE requires precise pairwise distances without a shared encoder. Our seqrec evaluation uses sampled-100 protocol; base-only evaluation confirms method ordering.

7. Discussion

LoRA for embeddings vs. LoRA for weights. The structural parallel between our method and LoRA is precise but the design considerations differ. In LoRA, the shared $\Delta W = BA$ applies identically to every input token. In our factorization, the shared V defines the subspace but each entry has its own coordinate u_i – making it closer to matrix completion than to adapter-based PEFT. This per-entry structure is necessary because embedding tables store per-entity information, not shared transformations.

When does population-specific structure help? The separation into population-specific V_k is essential when: (1) populations have different data densities (sparse new vs. abundant base), causing gradient imbalance in a shared subspace; (2) the encoder creates gradient coupling between populations through shared computation layers. On knowledge graph completion (TransE (Bordes et al., 2013) on the ENTITY benchmark (Cui et al., 2023a)), where each embedding is trained independently without a shared encoder, the anti-correlation does not manifest.

Connection to continual learning. EWC (Kirkpatrick et al., 2017) implicitly implements a similar base/new separation: new entries have zero Fisher importance (unconstrained) while base entries have high importance (penalized). Our method achieves the same effect structurally: the rank constraint limits base drift while separate V_{new} allows unconstrained new-entry learning. The structural approach avoids EWC's computational cost (Fisher computation scales poorly to million-item catalogs) and its sensitivity to the regularization strength λ.

Practical deployment considerations. Freeze1-SV is the most deployment-friendly variant: it requires no early stopping criterion, no population-specific validation set, and trains for a fixed number of epochs with guaranteed base stability. Dual-SV achieves higher new-entry quality but benefits from monitoring overall validation quality to select the best epoch. In production systems where the validation set composition changes daily, Freeze1-SV's epoch-robustness is a significant practical advantage.

Multi-population extensions. The framework generalizes to $K > 2$ populations. For example, items could be

stratified by interaction frequency into cold ($k < 5$), warm ($5 \leq k < 50$), and hot ($k \geq 50$) cohorts, each with its own V_k and learning rate. The theoretical bound suggests that sparser cohorts benefit more from the shared subspace, so rank could also be adapted per cohort ($r_{\text{cold}} < r_{\text{warm}} < r_{\text{hot}}$). We leave this exploration to future work.

8. Conclusion and Future Work

We extend low-rank adaptation from weight matrices to embedding tables for models with expanding vocabularies. The key insight is that the parameter distribution inverts across architectures: in LLMs, LoRA adapts the dominant transformer weights; in recommendation, our method adapts the dominant embedding table. Population-specific subspaces resolve the anti-correlation between base and new entries by decoupling their optimization while providing implicit regularization for sparse entries.

The framework Pareto-dominates baselines on sequential recommendation and wins in 8/9 LLM cells, with $4\times$ parameter efficiency at matched quality. Future directions include: (1) extending to $K > 2$ populations (e.g., frequency-based cohorts), (2) scaling to 7B+ LLMs where the anti-correlation effect may grow with model size, and (3) applying to DLRM-style models where embedding tables dominate even more than in sequential recommendation.

References

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *NeurIPS*, 2013.

Cui, Y., Wang, Y., Sun, Z., Liu, W., Jiang, Y., Han, K., and Hu, W. Lifelong embedding learning and transfer for growing knowledge graphs. In *AAAI*, 2023a. ENTITY benchmark dataset derived from Freebase (CC BY 2.5).

Cui, Y., Yang, Z., and Yao, X. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*, 2023b.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized language models. In *NeurIPS*, 2023.

Fatkulin, A., Zhukova, A., and Makarov, I. Let it go? not quite: Addressing item cold start in sequential recommendations with content-based initialization. In *Proceedings of the Web Conference*, 2025.

Ginart, A., Neel, S., Roth, A., and Waldon, B. Mixed dimension embeddings with application to memory-efficient recommendation systems. In *IEEE International Symposium on Information Theory*, 2021.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kang, W.-C. and McAuley, J. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining*, 2018.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Vinyals, O., Mohamed, S., et al. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences*, 2017.

Kornilova, A. and Eidelman, V. Billsum: A corpus for automatic summarization of us legislation. In *Workshop on New Frontiers in Summarization*, 2019.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*, 2020.

Lian, D., Wang, H., Liu, Z., Lian, J., Chen, E., and Xie, X. Personalized embedding size search under memory constraint. In *SIGIR*, 2021.

Mi, F., Lin, B., and Luo, X. Ader: Adaptively distilled exemplar replay towards continual learning for session-based recommendation. In *RecSys*, 2020.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Sato, L. et al. Merrec: A large-scale multipurpose mercari dataset for consumer-to-consumer recommendation systems. In *KDD*, 2025.

U.S. National Library of Medicine. Pubmed/medline baseline, 2024. Courtesy of the U.S. National Library of Medicine.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2024.

Wu, J. et al. SD-LoRA: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv preprint arXiv:2501.13198*, 2025.

Yamaguchi, A., Chrysostomou, G., and Aletras, N. An empirical study on vocabulary expansion for language models. *arXiv preprint arXiv:2404.03416*, 2024.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-efficient LLM training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

Zhu, H., Li, X., Zhang, P., Li, G., He, J., Li, H., and Gai, K. Learning tree-based deep model for recommender systems. *KDD*, 2018. Dataset: https://tianchi.aliyun.com/dataset/649, CC BY-NC-SA 4.0.

A. Multi-Seed Stability

Table 8. SASRec multi-seed stability (W0, sampled eval). Mean ± std across 3 seeds (42, 123, 456). Method ordering is stable.

Dataset	Method	new NDCG	base NDCG
Taobao	FT	0.806 ± 0.000	0.696 ± 0.000
	FT + freeze@ep1	0.865 ± 0.001	0.727 ± 0.000
	Freeze1-SV (r=64)	0.871 ± 0.001	0.723 ± 0.000
	Freeze1-SV (r=128)	0.874 ± 0.001	0.725 ± 0.001
MerRec	FT	0.674 ± 0.014	0.751 ± 0.006
	FT + freeze@ep1	0.698 ± 0.000	0.767 ± 0.000
	Freeze1-SV (r=64)	0.721 ± 0.003	0.762 ± 0.003
	Freeze1-SV (r=128)	0.734 ± 0.001	0.766 ± 0.001

Standard deviation is below 0.003 NDCG on Taobao and below 0.014 on MerRec, confirming that method ordering is stable across seeds for the sequential recommendation experiments.

Table 9. LLM multi-seed stability (BillSum, $r/d \approx 50\%$). Mean ± std of best-epoch overall PPL across 3 seeds (42, 123, 456). Standard deviation is below 0.03 PPL across all cells, confirming that method ordering is stable.

Model	Scale	Basetrn	Frozen	Freeze-SV	Freeze1-SV	Dual-SV
GPT-2	1K	6.40 ± 0.01	6.41 ± 0.01	6.37 ± 0.01	6.29 ± 0.01	6.55 ± 0.01
	5K	5.96 ± 0.01	5.98 ± 0.00	5.96 ± 0.00	5.84 ± 0.01	5.94 ± 0.00
	15K	4.79 ± 0.00	5.27 ± 0.00	5.26 ± 0.01	5.03 ± 0.01	5.11 ± 0.00
P-410M	1K	4.83 ± 0.03	4.83 ± 0.02	4.83 ± 0.02	4.81 ± 0.01	4.81 ± 0.02
	5K	5.08 ± 0.02	4.69 ± 0.01	4.68 ± 0.01	4.80 ± 0.02	4.68 ± 0.01
	15K	4.13 ± 0.01	4.08 ± 0.04	4.05 ± 0.01	4.08 ± 0.00	4.03 ± 0.01
P-1.4B	1K	4.12 ± 0.01	4.09 ± 0.00	4.09 ± 0.00	4.07 ± 0.01	4.08 ± 0.01
	5K	4.41 ± 0.01	4.04 ± 0.00	4.03 ± 0.00	4.10 ± 0.01	4.03 ± 0.00
	15K	3.67 ± 0.01	3.54 ± 0.01	3.54 ± 0.01	3.56 ± 0.01	3.52 ± 0.01

B. Empirical Validation of Theoretical Bound

We estimate the low-rank approximation quality from trained models on Taobao (W0). Computing the SVD of centered new-item embeddings from a full-rank frozen model, the top-128 singular vectors capture 95% of variance ($\tau^2 = 0.006$ residual per item). Even at r=32, 54% of variance is captured ($\tau^2 = 0.056$), validating the approximate low-rank structure assumed by the proposition.

The bound predicts shared-V outperforms independent estimation when k_j is small relative to d/r. On both datasets, new items have median $k_j = 1$. At our primary rank r=128 ($d/r \approx 1.2$), the estimation-error improvement is modest; the empirical gains at high rank come primarily from the regularization effect of the shared subspace (preventing overfitting on sparse items) rather than pure variance reduction. At lower ranks (r=32, $d/r \approx 5$), the variance reduction is larger but approximation error increases. Prediction (P3) – data-scale crossover – is confirmed by the LLM experiments: at 15K documents on GPT-2, full-rank becomes preferable as per-token data grows large relative to model capacity. Prediction (P2) – diminishing returns – is confirmed by the rank ablation (Table 3): gains from r=64 to r=128 are smaller than from r=32 to r=64. Prediction (P1) – advantage concentrated in sparse entries – follows from the bound structure and is consistent with our new-item evaluation (all new items have $k_j \leq 5$, well below the crossover threshold).