Positive-Unlabeled Contrastive Retrieval for Graduate Program Recommendation

Shiyan Liu and Rui Qu
School of Computer Science and Technology
Huazhong University of Science and Technology
{shyl, rqu}@hust.edu.cn

Abstract

Graduate program selection is a high-stakes decision where applicants struggle to identify programs matching their latent academic profiles. We address this as a retrieval problem: given a student profile, rank 531 graduate programs by estimated *admission compatibility*, i.e., how closely the student's profile resembles each program's historical admit pool. Because applicants rarely report rejections, training data consists only of positive labels, making negative supervision unavailable. We propose PURE (Positive-Unlabeled Retrieval Encoder), a positive-unlabeled biencoder with InfoNCE contrastive loss and a gated structural residual that fuses GPA and experience counts with free-text descriptions. With a multi-positive MRR protocol that corrects for false negatives, PURE achieves MRR = 0.2696, outperforming BM25 (+4.9 points) and GPT-4o-mini (+8.3 points), while remaining statistically indistinguishable from Biencoder + GPT-4o-mini reranking at negligible cost.

1 Introduction

A philosophy student at a 211-tier Chinese university spends two years on applied AI ethics: one project investigates how an autonomous vehicle should resolve a trolley-problem dilemma, while another analyses the concept of "machine morality." Asked to rank 531 graduate programs for this student, a standard keyword retrieval system places the MA in AI, Ethics and Society at the University of Hong Kong at position **#369 out of 531**. Our model places it at **#1**. The student was admitted. The surface overlap here is weak: the student writes about trolley problems and machine morality, while the program name mentions "AI, Ethics and Society"; her research identity nevertheless aligns closely with the program's admitted cohort. Keyword retrieval cannot bridge that gap at scale.

Graduate program selection presents a unique information retrieval challenge: the query is a rich multi-modal student profile, the corpus is a set of academic programs, and the relevance signal is revealed only through observed admissions. Unlike web search or product recommendation, there are no explicit negative labels. A student who did not attend a program may have been rejected, may have been admitted but chose elsewhere, or may simply have never applied. This positive-only constraint fundamentally changes the training and evaluation landscape.

Applicants typically shortlist admissible programs before weighing research fit; PURE addresses this first step, which we term *admission compatibility*: resemblance to a program's historical admit pool.

Three structural challenges compound the retrieval problem. First, the **many-to-many matching structure**: 53.7% of students in our dataset received multiple offers (averaging 2.01 per student), and programs admit cohorts of students with diverse profiles. A retrieval model must learn program-level *prototypes* rather than one-to-one matching. Second, **heterogeneous input modalities**: structured features (Grade Point Average (GPA), school tier, publication counts) co-exist with free-text experience descriptions of variable length and quality. Naive concatenation discards the complementary nature of these signals. Third, **cross-disciplinary admissions**: our analysis shows that 40.8% of admits cross the program's dominant field boundary, and business programs routinely admit social-science and humanities applicants. Standard collaborative filtering collapses this structure.

Experimentally, our biencoder outperforms GPT-4o-mini (OpenAI, 2023) (mean reciprocal rank (MRR) 0.2696 vs. 0.1862) and is statistically indistinguishable from Biencoder + GPT-4o-mini reranking on a 300-student subsample ($p = 0.52$) while running at orders-of-magnitude lower cost.

We make three contributions:

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4320–4337
November 6-10, 2026 ©2026 Association for Computational Linguistics

1. **Model**: A positive-only biencoder encoding programs as GPA-weighted MiniLM (Wang et al., 2020) prototypes and applicants via gated structural residual fusion, trained with InfoNCE (van den Oord et al., 2018) plus hard-negative margin loss.

2. **Evaluation protocol**: A multi-positive MRR protocol for many-to-many admission matching, with an identity-grouping procedure that infers student groups from school, major, GPA, and experience text without manual annotation, validated at 99.1% precision.

3. **Analysis**: An empirical analysis revealing eight latent program clusters with cross-disciplinary structure (two clusters with entropy > 2.0), field-varying feature attribution (structural features dominate only life sciences at 65% of total attribution), and three case studies with BM25 (Robertson and Zaragoza, 2009)-rank gaps up to 368 positions.

2 Related Work

Dense retrieval and contrastive learning. The dual-encoder paradigm (Karpukhin et al., 2020) encodes queries and documents independently, enabling efficient inner-product retrieval. InfoNCE (van den Oord et al., 2018) and SimCSE (Gao et al., 2021) use in-batch negatives, and hard-negative mining closes the gap with cross-encoder rerankers (Nogueira and Cho, 2019). Sentence-level encoders (Reimers and Gurevych, 2019; Muennighoff et al., 2023; Thakur et al., 2021) excel at retrieval but do not natively handle heterogeneous structured inputs. We adopt the dual-encoder paradigm with GPA-weighted program prototypes and gated structural fusion tailored to the positive-only setting.

Graduate program and admissions recommendation. Prior work relies on rule-based systems (Hasan et al., 2016), support vector machine and k-nearest neighbor (SVM+KNN) (Baskota and Ng, 2018), analytic hierarchy process (AHP) ranking (Wongvilaisakul et al., 2023), and supervised classification (Heidari Far and Tabrizi, 2026; El Guabassi et al., 2021; Zhao et al., 2023), all assuming explicit positive/negative labels. We depart by treating the task as positive-only retrieval over 531 programs, where unobserved outcomes are not treated as rejections.

Positive-unlabeled learning and debiased ranking. Positive-unlabeled (PU) learning (Elkan and Noto, 2008) with unbiased risk estimators and Bayesian personalized ranking (BPR) (Rendle et al., 2009) address implicit feedback without explicit negatives. Inverse propensity score (IPS) debiasing (Joachims et al., 2017; Saito et al., 2020) corrects exposure bias but requires propensity scores that are unavailable in our setting. PULNS (Luo et al., 2021) selects high-confidence negatives via a learned selector; multi-positive contrastive learning and self-supervised objectives (Yu et al., 2022a,b; Zhou et al., 2023) offer related false-negative mitigation but assume view augmentations. We combine batch-wide positive masking with PULNS-style hard-negative selection, requiring neither propensity estimates nor augmented views.

LLMs for ranking. Large language models (LLMs) demonstrate zero-shot listwise ranking (Sun et al., 2023; Ma et al., 2024) but struggle with fine-grained domain matching over large candidate pools without a retrieval stage (Wu et al., 2024), as we verify empirically (§7.4).

3 Task Formulation

Let $\mathcal{S}$ be a set of student profiles and $\mathcal{P}$ be a set of graduate programs. Each student $s \in \mathcal{S}$ is associated with a set of positive programs $\mathcal{P}^+(s) \subseteq \mathcal{P}$, representing programs to which s received an admission offer. The dataset records only accepted outcomes: rejections go unreported, and $\mathcal{P} \setminus \mathcal{P}^+(s)$ mixes true negatives, programs s never applied to, and survivorship artifacts from self-reported admissions. Standard supervised ranking with explicit negative labels is therefore inapplicable.

Retrieval task. Given a student profile s, produce a ranking π_s over all programs in $\mathcal{P}$ such that programs in $\mathcal{P}^+(s)$ are ranked as high as possible.

Positive-only constraint. For any (s, p) with $p \notin \mathcal{P}^+(s)$, we cannot distinguish true negatives from unobserved positives or selection artifacts, motivating our PU-aware training (§5) and multi-positive evaluation (§6).

Many-to-many structure. The relation $\{(s, p) : p \in \mathcal{P}^+(s)\}$ is many-to-many: 53.7% of students hold multiple offers (avg. 2.01), and each program admits a diverse cohort. Standard single-positive metrics arbitrarily select one true positive as ground truth and penalise correct rankings of the rest, conflating multi-offer retrieval success with failure.

Table 1: Dataset statistics.

Statistic	Value
Total admission records	15,291
Valid programs (≥5 train admits)	531
Unique students	7,600
Multi-offer students	53.7%
Avg. offers per student	2.01
Training records (year < 2026)	10,838
Test students (year = 2026, evaluated)	1,866
GPA missing rate	0%
IELTS missing rate	28.7%
Research/intern/comp./pub. missing	0%
School tier missing rate	0%
TOEFL missing rate	95.7%

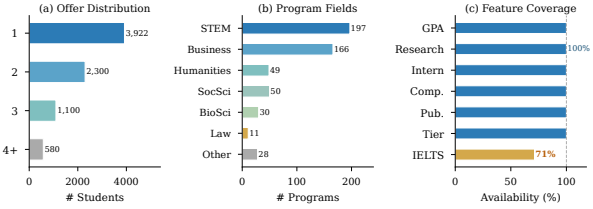


Figure 1: Dataset overview: (a) offers per student, (b) program field distribution, (c) feature coverage. TOEFL (95.7% missing) is excluded.

4 Dataset

4.1 Data Collection

We collect admission records from a publicly accessible online community where Chinese-speaking applicants share graduate admission outcomes and experience descriptions. Programs concentrate in Hong Kong, Singapore, the UK, and Australia. We retain only programs with at least 5 training admits, yielding 531 valid programs.

4.2 Statistics

Table 1 summarizes key statistics. We split by application year: records with `apply_year` < 2026 form the training set (10,838 records), and the 2026 cohort forms the test set (2,196 unique students; 1,866 after filtering to those with at least one valid-program admission). Figure 1 shows the dataset's offer distribution, field breakdown, and feature coverage.

4.3 Feature Groups

We define two feature groups. **Structural features** include GPA, school tier (a five-level ordinal; see Appendix E.1), and counts of research, internship, competition, and publication entries (all fully observed). IELTS score is imputed at the feature mean (28.7% missing) and TOEFL is excluded (95.7% missing). **Text features** are applicant-authored free-text experience descriptions (see Table 1 for coverage). Figure 2 shows a representative admission record from our dataset.

Example Admission Record offer_id: 10847

Structured Features

Program	MSc CS (AI), HKUST		
Year	2026	*GPA*	3.62 / 4.0
School tier	2 (211)	*IELTS*	7.0
Research	1 project	*Internship*	2 positions
Pub.	0	*Competition*	1 award

Experience Text (excerpt)

"Participated in an NLP project supervised by a faculty advisor, handling data preprocessing and BERT fine-tuning. Completed two software engineering internships at tech companies, building a prototype recommender system. Received a provincial award in the National Mathematical Modeling Competition."

Figure 2: A representative anonymized admission record (experience text translated from Chinese).

5 Method

Figure 3 provides an overview of PURE. Programs are encoded offline as GPA-weighted prototype embeddings, and applicants are encoded online via a gated residual fusion of text and structural features. At inference time, dot-product similarity ranks all 531 programs for each query.

5.1 Program Encoder

Each program p is represented offline by an ℓ_2-normalized prototype $\mathbf{e}_p \in \mathbb{R}^d$ computed from its training admits $\mathcal{A}(p)$:

$$\mathbf{e}_p = \ell_2\left(\frac{\sum_{s\in\mathcal{A}(p)} w_s \cdot \phi_{\text{text}}(s)}{\sum_{s\in\mathcal{A}(p)} w_s}\right), \qquad (1)$$

where $w_s = \text{clip}(\text{GPA}_s, 1.0, 4.0)$ upweights higher-GPA admits and $\mathcal{A}(p)$ is the set of training admits for program p. The per-student text embedding $\phi_{\text{text}}(s) \in \mathbb{R}^d$ is obtained by a shared linear projection $W_t \in \mathbb{R}^{d\times 384}$ applied to the frozen MiniLM-L6-v2 (Wang et al., 2020) `[CLS]` output $\mathbf{h}_s \in \mathbb{R}^{384}$:

$$\phi_{\text{text}}(s) = W_t\,\mathbf{h}_s, \quad \mathbf{h}_s = \text{MiniLM}(x_s^{\text{text}}) \in \mathbb{R}^{384}. \qquad (2)$$

Sharing W_t between the program encoder (Eq. 1) and the applicant text branch (Eq. 3) ensures both sides live in the same embedding space, enabling dot-product retrieval without a separate alignment step. Programs with fewer than 5 training admits receive no learned prototype; all others yield a fixed, ℓ_2-normalized prototype stored for inference.

GPA weighting and fairness. Weights are relative to each program's own admit pool: a GPA-

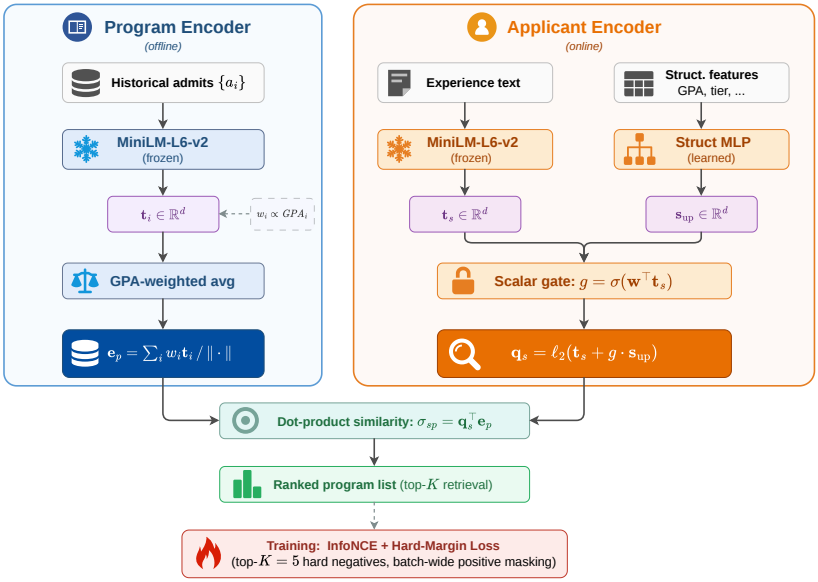


Figure 3: Overview of PURE. **Top:** offline program encoder (GPA-weighted prototype per program). **Bottom:** online applicant encoder (gated fusion of text and structural features). The shared projection W_t places both sides in the same space for dot-product ranking.

3.5 admit carries nearly identical weight in programs whose admit pools are $\{3.5, 3.7, 3.9\}$ and $\{3.2, 3.5, 3.7\}$ ($\approx$0.31 vs. $\approx$0.32). Tier-weighted prototypes, which directly encode prestige, perform worst (Table 7), and MP-MRR spread across GPA quartiles is only $\pm$0.010 (Table 16).

Description-based fallback. Programs without a learned prototype remain rankable: the official program text, embedded through W_t, serves directly as the prototype (MRR $= 0.149$ on zero-admit programs; Table 15).

5.2 Applicant Encoder

The applicant encoder maps a student profile to a query embedding $\mathbf{q}_s \in \mathbb{R}^d$ via a two-branch architecture:

Text branch. Student experience descriptions are applicant-authored in Chinese, with a median length of 106 characters (interquartile range (IQR): 71–154 chars; $\approx$30–50 tokens after Chinese-Japanese-Korean (CJK) tokenization). The full text is passed as-is to the multilingual MiniLM-L6-v2 (frozen), which supports Chinese natively and was pre-trained on 50+ languages. MiniLM's 128-token input limit covers 99.2% of records at the character level, and longer records are truncated. The encoder produces $\mathbf{h}_{\text{text}} \in \mathbb{R}^{384}$, which is projected by the shared matrix W_t (Eq. 2) to the joint embedding space:

$$\hat{\mathbf{h}}_{\text{text}} = W_t \, \mathbf{h}_{\text{text}} \in \mathbb{R}^d. \tag{3}$$

Structural branch. Let $\hat{\mathbf{x}}_s \in \mathbb{R}^k$ be the z-score normalized structural feature vector (GPA, IELTS, experience counts, school tier; $k = 7$, missing IELTS imputed by feature mean). A two-layer multi-layer perceptron (MLP) maps it to the embedding space:

$$\hat{\mathbf{h}}_{\text{struct}} = f_\theta(\hat{\mathbf{x}}_s) = W_2 \, \text{ReLU}(W_1 \hat{\mathbf{x}}_s + \mathbf{b}_1) + \mathbf{b}_2 \in \mathbb{R}^d, \tag{4}$$

where $W_1 \in \mathbb{R}^{256 \times k}$, $W_2 \in \mathbb{R}^{d \times 256}$, and no dropout is applied.

Gated residual fusion. Naive concatenation of text and structural embeddings treats the two modalities symmetrically, but in the PU setting the text branch already provides a useful retrieval signal while the structural branch captures complementary admission thresholds. A scalar gate $g \in [0, 1]$ controls structural contribution:

$$g = \sigma\left(W_g \left[\hat{\mathbf{h}}_{\text{text}} \,\|\, \hat{\mathbf{h}}_{\text{struct}}\right] + b_g\right), \tag{5}$$

where $W_g \in \mathbb{R}^{1 \times 2d}$ and the query embedding is:

$$\mathbf{q}_s = \ell_2\left(\hat{\mathbf{h}}_{\text{text}} + g \cdot \hat{\mathbf{h}}_{\text{struct}}\right). \tag{6}$$

This residual design preserves the text representation as the primary signal. When the gate closes ($g \to 0$), the model falls back to text-only retrieval; when it opens ($g \to 1$), structural features contribute equally. The gate is conditioned on both branches so it can adapt to field-specific reliance patterns (e.g., opening wider for life-science programs where GPA and publication counts are decisive admission criteria).

5.3 Training Objective

We train with a combination of InfoNCE (van den Oord et al., 2018) contrastive loss and a hard-negative margin loss. Let $\sigma_{sp} = \mathbf{q}_s^\top \mathbf{e}_p$ denote the dot-product similarity. For a batch $\mathcal{B}$ of (student, positive program) pairs, let $\mathcal{M}_\mathcal{B} = \{p : \exists\, s' \in \mathcal{B},\ p \in \mathcal{P}^+(s')\}$ be the set of all programs that are positive for *any* student in the batch. The InfoNCE loss is:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{|\mathcal{B}|} \sum_{(s,p)\in\mathcal{B}} \log \frac{e^{\sigma_{sp}/\tau}}{\sum_{p' \in \mathcal{P}\setminus(\mathcal{M}_\mathcal{B}\setminus\{p\})} e^{\sigma_{sp'}/\tau}}, \quad (7)$$

where τ is a learnable temperature initialized to 0.05, and the denominator sums over all programs in $\mathcal{P}$ *except* those in $\mathcal{M}_\mathcal{B} \setminus \{p\}$ (i.e., programs that are positive for *some* student in the batch, excluding the current positive p itself). In practice this is implemented as softmax over all 531 program logits with positions in $\mathcal{M}_\mathcal{B} \setminus \{p\}$ set to $-\infty$. The hard-negative set $\mathcal{N}(s)$ for student s is defined as the top-K non-positive programs ranked by current similarity:

$$\mathcal{N}(s) = \underset{p \notin \mathcal{P}^+(s)}{\text{top-K}}\ \sigma_{sp}, \quad (8)$$

refreshed at each epoch using the current encoder weights. The hard-negative margin loss is then:

$$\mathcal{L}_{\text{margin}} = \frac{1}{|\mathcal{B}|} \sum_{(s,p)\in\mathcal{B}} \sum_{n\in\mathcal{N}(s)} \max(0,\ m - \sigma_{sp} + \sigma_{sn}), \quad (9)$$

where m is the margin hyperparameter. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \lambda \mathcal{L}_{\text{margin}}. \quad (10)$$

Algorithm 1 summarises one training step.

5.4 PU-Aware Negative Handling

Both loss terms treat unobserved programs as negatives, which in the PU setting mixes true negatives with unobserved positives (Elkan and Noto, 2008), while fully corrected PU objectives require unavailable propensity estimates. PURE combines two complementary mitigations.

Batch-wide positive masking. Any program positive for *any* batch student is excluded from the InfoNCE denominator (Eq. 7) and the hard-negative pool. Because 98.6% of all programs are positive for at least one training student, masking is necessarily aggressive: over 3,000 simulated batches of size 256, the masked fraction averages 48.4% of the candidate pool, still leaving a majority of programs as verified-clean negatives, roughly twice the unique-program coverage of standard in-batch NCE (global masking of *all* ever-positive programs drops MRR by 3.3 points; Table 8).

NegSelector. The second mitigation is PULNS-style negative selection (Luo et al., 2021): a NegSelector MLP Ψ_ϕ, trained with the non-negative PU (nnPU) loss (Kiryo et al., 2017) on observed admits as labeled positives and high-scoring unobserved pairs as unlabeled, identifies likely false negatives in the hard-negative pool, and the encoder is retrained on the purified pool:

$$\hat{\mathcal{N}}(s) = \{p \in \mathcal{N}(s) \setminus \mathcal{P}^+(s) : \Psi_\phi(\mathbf{q}_s, \mathbf{e}_p) < \delta\}, \quad (11)$$

where δ is tuned on the validation set. Concretely, Ψ_ϕ is an MLP with hidden layers of 256 and 64 units over $[\mathbf{q}_s \,\|\, \mathbf{e}_p \,\|\, \cos(\mathbf{q}_s, \mathbf{e}_p)]$, trained once between the two stages on the hard-negative cache with per-student priors $\pi^+(s) = |\mathcal{P}^+(s)|/|\mathcal{P}|$ (Algorithm 2); grid search on the 2025 validation cohort selects $\delta^* = 0.4$.

Comparison with PU alternatives. On the temporal test split (n=1,866), batch masking (MRR = 0.271) significantly outperforms both nnPU-style risk correction (0.261, $p = 0.016$) and a multi-positive NCE variant summing over all of a student's positives in the numerator (0.255, $p = 0.002$; paired bootstrap; Table 9, Appendix C). nnPU's prior is underestimated because students disproportionately report the program they enrolled in; multi-positive NCE's joint numerator is redundant once co-positive programs are masked from the denominator, and it converges much later (best epoch 20 vs. epoch 10).

5.5 Implementation Details

We use $d = 384$, batch size 256, $\tau = 0.05$, $m = 0.2$, $\lambda = 0.5$, 60 epochs, AdamW (Loshchilov and Hutter, 2019) (lr $= 3 \times 10^{-4}$, weight decay 10^{-4}), OneCycleLR scheduler (10% warmup). Hard negatives are the top-$K = 5$ non-positive programs per student, refreshed each epoch. The NegSelector MLP is trained once with the nnPU loss, and pairs it scores $\geq \delta^*$ ($\delta^* = 0.4$; §5.3) are removed from the hard-negative pool prior to Stage-2 encoder training. Only W_t, the structural MLP, and the gate are trained (≈400K parameters), while the MiniLM backbone is frozen. Hyperparameters are tuned on the 2025 cohort (train ≤2024, val = 2025).

6 Multi-Positive Evaluation Protocol

6.1 Motivation

Standard ranking metrics assume a single relevant item per query. In our setting, a student with three

Algorithm 1 TRAININGSTEP(batch $\mathcal{B}$)

Require: batch $\mathcal{B}$ of (s, p_s) pairs, prototypes $\{\mathbf{e}_p\}$, NegSelector-filtered hard-neg cache $\hat{\mathcal{N}}(s)$ (Alg. 2; refreshed each epoch)

// encode all queries
for each $s \in \mathcal{B}$ **do**
  $\mathbf{q}_s \leftarrow \mathrm{Enc}(s)$
end for
// batch-wide positive mask
$\mathcal{M}_{\mathcal{B}} \leftarrow \bigcup_{s \in \mathcal{B}} \mathcal{P}^+(s)$
// compute losses per sample
for each $(s, p_s) \in \mathcal{B}$ **do**
  // score all programs
  $\ell_{sp'} \leftarrow \mathbf{q}_s^\top \mathbf{e}_{p'} / \tau \quad \forall\, p' \in \mathcal{P}$
  // mask batch positives
  $\ell_{sp'} \leftarrow -\infty \quad \forall\, p' \in \mathcal{M}_{\mathcal{B}} \setminus \{p_s\}$
  // InfoNCE loss
  $\mathcal{L}^{(s)}_{\mathrm{NCE}} \leftarrow -\ell_{sp_s} + \log \sum_{p'} \exp(\ell_{sp'})$
  // hard-margin loss over filtered hard negatives
  for each $n \in \hat{\mathcal{N}}(s) \setminus \mathcal{M}_{\mathcal{B}}$ **do**
    accumulate $\max\big(0,\ m - \sigma_{sp_s} + \sigma_{sn}\big)$ into $\mathcal{L}^{(s)}_{\mathrm{margin}}$
  end for
end for
// aggregate and update
$\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}|} \sum_s \big(\mathcal{L}^{(s)}_{\mathrm{NCE}} + \lambda\, \mathcal{L}^{(s)}_{\mathrm{margin}}\big)$
backprop($\mathcal{L}$)

Table 2: 1P-MRR vs. MP-MRR: multi-positive protocol corrects false-negative underestimation.

Model	1P-MRR	MP-MRR	Change
BM25	0.1657	0.2208	+33.3%
Biencoder	0.1839	0.2696	+46.6%

offers has three relevant programs. The single-positive protocol arbitrarily selects one as ground truth and penalizes correct rankings of the others. Multi-positive MRR (MP-MRR) *alleviates* (not eliminates) this false-negative problem: it corrects within observed test positives but cannot account for unobserved positives outside the test set.

6.2 Metric Definitions

Let $\pi_s(p)$ denote the rank of program p for student s, and $\mathcal{P}^+(s)$ the set of observed positives. MP-MRR scores each student by the reciprocal rank of their best-ranked positive:

$$\text{MP-MRR} = \frac{1}{|\mathcal{S}_{\text{test}}|} \sum_{s \in \mathcal{S}_{\text{test}}} \frac{1}{\min\limits_{p \in \mathcal{P}^+(s)} \pi_s(p)}. \quad (12)$$

MP-Hit@K, Recall@K, and mean average precision (MAP)@K are defined analogously, with full definitions in Appendix B.

6.3 Protocol Validation

Table 2 confirms that single-positive evaluation substantially underestimates retrieval quality. The bi-

Algorithm 2 NEGSELECTORTRAINING (once, between Stage 1 and 2)

Require: Stage-1 encoder, prototypes $\{\mathbf{e}_p\}$, observed admits $\mathcal{O}$, threshold grid $\{\delta\}$, validation cohort $\mathcal{V}$

// hard-negative cache
$\mathcal{H} \leftarrow \{(s, p) : \cos(\mathbf{q}_s, \mathbf{e}_p) \geq \tau_{\text{hard}},\ (s, p) \notin \mathcal{O}\}$
// per-student positive prior
$\pi^+(s) \leftarrow |\mathcal{P}^+(s)| \,/\, |\mathcal{P}|$
// train Ψ_ϕ *over* $[\mathbf{q}_s \,\|\, \mathbf{e}_p \,\|\, \cos(\mathbf{q}_s, \mathbf{e}_p)]$
for each epoch **do**
  $\mathcal{L}_{\text{nnPU}} \leftarrow \mathbb{E}_{\mathcal{O}}[-\log \sigma(\Psi_\phi)] + \max\big(0,\ \mathbb{E}_{\mathcal{H}}[\log \sigma(\Psi_\phi)] - \pi^+ \cdot \mathbb{E}_{\mathcal{O}}[\log \sigma(\Psi_\phi)]\big)$
  $\phi \leftarrow \phi - \eta\, \nabla_\phi \mathcal{L}_{\text{nnPU}}$
end for
// select threshold on validation
$\delta^* \leftarrow \arg\max_\delta$ MP-MRR$_{\mathcal{V}}$ with $\mathcal{H}_\delta = \{(s, p) \in \mathcal{H} : \Psi_\phi(s, p) < \delta\}$
// purified cache for Stage 2
return $\hat{\mathcal{N}} \leftarrow \{(s, p) \in \mathcal{H} : \Psi_\phi(s, p) < \delta^*\}$

encoder gains more (+46.6%) than BM25 (Robertson and Zaragoza, 2009) (+33.3%), suggesting it recovers a larger fraction of the full positive set.

Sensitivity analysis. We simulate grouping perturbations: (i) *under-grouping* (key=3 features: GPA, tier, research) yields MP-MRR $=$ 0.252 (-0.003); (ii) *over-grouping* (key=4 features, adding internship) yields 0.251 (-0.004). The ± 0.004 range confirms results are robust to moderate identity-inference errors. Adversarial robustness is further quantified in Appendix B.

6.4 Student Identity Inference

Because the test set contains only admission records (not explicit student IDs for the 2026 cohort), we must infer which records belong to the same student. We group records by a composite key: (undergraduate school, undergraduate major, GPA, application year, first 80 characters of experience text). Two records sharing all five values are treated as the same student. We validate this on training data where platform user IDs are available: the key achieves 99.1% precision and 97.4% recall.

7 Experiments

7.1 Baselines

We compare against baselines from four groups (Table 3). *Sparse/Density*: **Struct-KNN** (cosine nearest-neighbor on normalized structural features); **BM25** (Robertson and Zaragoza, 2009) (Okapi BM25 over experience text, program documents formed by concatenating training admit profiles); **Gaussian mixture model (GMM)** (Reynolds, 2009) in three variants (struct-only, struct+text, text-only), scoring by log-likelihood under the highest-membership compo-

nent. *Zero-shot neural*: **E5-base** (Wang et al., 2022), **BGE-M3** (Chen et al., 2024), and **ColBERT MaxSim** (Khattab and Zaharia, 2020) (late-interaction using frozen MiniLM (Wang et al., 2020), no fine-tuning); **Mean-Pooling** (unweighted mean of MiniLM admit embeddings, no training). *Supervised neural*: **BPR** (Rendle et al., 2009) (same encoder with pairwise ranking loss and random negatives); **Fine-tuned SBERT** (Reimers and Gurevych, 2019) (MiniLM fine-tuned with MultipleNegativesRankingLoss). LLM-based baselines (GPT-4o-mini (OpenAI, 2023), Biencoder + GPT-4o-mini) are evaluated on a 300-student subsample due to API cost, with details in §7.4.

7.2 Main Results

PURE achieves MP-MRR = 0.2696 (Table 3), outperforming every baseline. It improves over BM25 by +4.9 points and Mean-Pooling by +7.0 points. GMM Struct+Text (0.1886) performs *worse* than Text-only (0.2045), confirming that naive feature concatenation without learned fusion is counterproductive. Struct-KNN (0.0167) and BPR (0.145) further confirm that structural features alone and pairwise objectives without hard-negative mining are both insufficient.

Zero-shot encoders (E5-base 0.102, BGE-M3 0.109, ColBERT MaxSim 0.090) and fine-tuned SBERT (0.158) all underperform BM25, confirming that even strong multilingual models require task-specific training (here, contrastive learning over GPA-weighted prototypes) to be competitive on this fine-grained domain. A temporal-robustness probe, in which prototypes are rebuilt from only ≤2024 records, retains 80% of full-data performance on the 2026 cohort (Appendix F), indicating that the learned prototypes are not an artifact of a single admission cycle.

The biencoder achieves Recall@20 = 0.576, placing over half of all admitted programs in the top-20 list. Biencoder + GPT-4o-mini reranking (0.2652) is statistically indistinguishable from PURE alone on the available 300-student subsample (95% CI $[-0.051, +0.053]$, $p = 0.52$; Table 4); LLM-based results, caveats on the subsample size, and cost analysis are in §7.4.

Table 4 confirms all improvements over BM25 and Mean-Pooling are significant ($p < 0.001$), and that adding GPT-4o-mini reranking yields no detectable gain on the 300-student subsample ($p = 0.52$; §7.4 discusses the width of this interval). Figure 4 shows PURE is Pareto-dominant in cost–quality space.

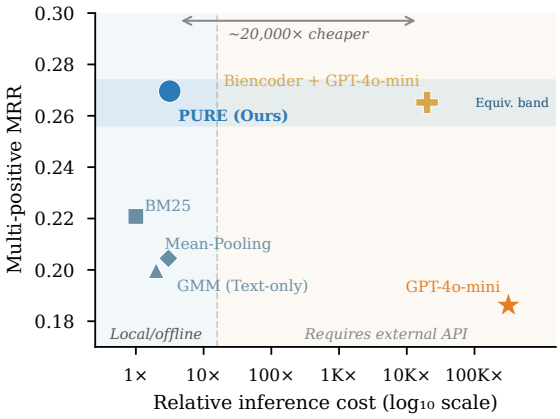


Figure 4: Cost vs. quality trade-off. Biencoder: CPU wall-clock per query; GPT-4o-mini: API latency at 8K tokens. Shaded band: $p = 0.52$ equivalence region.

7.3 Ablation Study

GPA-weighted prototypes are the most critical component (Table 5): removing them drops MRR by 4.0 points (0.2696 → 0.2298), while removing the structural gate costs 1.4 points (0.2696 → 0.2555); both variants still exceed Mean-Pooling, confirming contrastive training is essential regardless of design variant. Replacing prototypes with official program descriptions (MRR = 0.149, −10.8 points) shows that descriptions capture program content but not admission selectivity. Full GPA-weighting alternatives appear in Table 7 (Appendix C). Batch-wide positive masking significantly outperforms other PU-aware loss alternatives ($p \leq 0.02$, paired bootstrap; Appendix C), and PULNS-style (Luo et al., 2021) hard-negative selection provides complementary data-sampling gains (0.2576 → 0.2696, +1.1 pts; Table 5).

7.4 LLM Comparison and Cost

We compare task-specific training on 10K admission records with a general-purpose LLM without access to those records. On a 300-student subsample, GPT-4o-mini given only program names (no admission history) achieves MRR = 0.1862, well below BM25 (0.2288) and our biencoder (0.2665). Supplying the LLM with richer program context consistently *lowers* its MRR (Appendix C.5), while the same historical-context signal is exactly what our prototypes exploit: the predictive signal resides in *who was admitted*, not in program descriptions.

Adding GPT-4o-mini reranking over the top-20 biencoder candidates yields MRR = 0.2652, statistically indistinguishable from PURE alone on the available 300-student subsample (95% CI

Table 3: Retrieval results on 1,866 test students. Normalized discounted cumulative gain (NDCG)@10 reported; GPT results on 300-student subsample (§7.4). †BM25 uses character-level tokenization. **Bold**: best.

Group	Model	MP-MRR	MP-Hit@10	NDCG@10	Recall@20
Sparse / Density	Struct-KNN	0.0167	0.031	0.009	0.039
	GMM (struct-only)	0.0420	0.091	0.030	0.103
	GMM (struct+text)	0.1886	0.419	0.186	0.484
	BM25†	0.2208	0.468	0.215	0.515
Zero-shot Neural	E5-base	0.1023	0.221	0.087	0.252
	BGE-M3	0.1085	0.242	0.095	0.260
	ColBERT MaxSim	0.0897	0.206	0.080	0.238
	Mean-Pooling (MiniLM)	0.1997	0.454	0.204	0.512
Supervised Neural	GMM (text-only)	0.2045	0.458	0.204	0.519
	BPR	0.1451	0.384	0.153	0.443
	Fine-tuned SBERT	0.1582	0.380	0.159	0.438
Ours	**PURE**	**0.2696**	**0.516**	**0.261**	**0.576**

Table 4: Paired bootstrap significance (10K iterations). *300-student subsample. n.s. = not significant.

Comparison	ΔMRR	95% CI	p-value
vs. BM25	+.049	[.034, .064]	$<.001^{***}$
vs. Mean-Pool.	+.078	[.065, .092]	$<.001^{***}$
vs. E5-base	+.167	[.153, .183]	$<.001^{***}$
vs. +GPT*	+.001	[−.051, .053]	.52 n.s.

$[-0.051, +0.053]$, $p = 0.52$, paired bootstrap; Table 4). The interval includes zero but bounds any reranking benefit below ≈5 MRR points, a modest effect for an order-of-magnitude cost increase. Biencoder inference takes <10 ms per student on CPU vs. ∼30 s and ≈$0.01/query for GPT-4o-mini. A controlled pilot with varying context levels is detailed in Appendix C.5.

8 Analysis

8.1 Cross-Disciplinary Cluster Discovery

We cluster the 531 program prototype embeddings using K-Means (K=8, elbow criterion), visualized via Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). The solution is moderately stable (Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) $= 0.563 \pm 0.066$; intra-cluster cosine similarity 0.359 vs. inter-cluster −0.060; details in Appendix D).

Figure 5 shows the row-normalized cluster–field distribution. Six clusters have clear dominant fields: C0 (Law, 61%), C1 (STEM-EE, 72%), C2 (Humanities, 60%), C3 (Business, 72%), C5 (STEM-AI, 80%), C6 (BioSTEM, 49%). Two stand out for cross-disciplinary mixing: **C4** (entropy = 2.13; Business 36%, SocSci 28%, Humanities 17%; 64% cross-field rate) captures an *applied humanities* profile spanning communication, cultural studies, and public policy. **C7** (entropy = 2.06; STEM 41%, Business 31%; 59% cross-field) clusters supply-chain, operations research, and tech-management programs requiring a rare STEM+business hybrid. The mean cross-field rate of 40.8% confirms cross-disciplinary matching is the norm, not the exception.

Table 5: Ablation of key design choices on 1,866 test students. *w/o NegSelector* runs Stage 2 with the unfiltered hard-negative cache, so the +1.1-point gap reflects false-negative removal.

Model variant	MP-MRR	Hit@10
PURE (full)	**0.2696**	**0.516**
w/o GPA-wtd prototypes	0.2298	0.500
w/o struct. gate (g=0)	0.2555	0.489
w/o NegSelector filter	0.2576	–
Mean-Pooling (no training)	0.1997	0.454

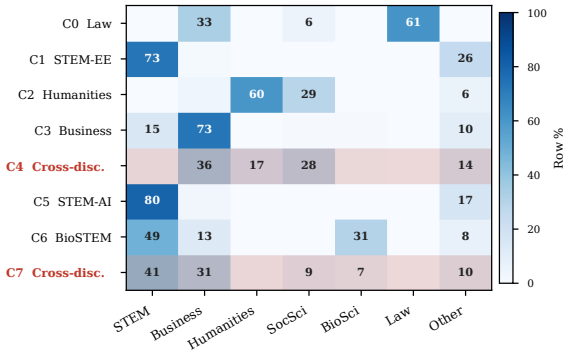


Figure 5: Row-normalised cluster × field distribution. Red rows (C4, C7): cross-disciplinary clusters (field entropy $H > 2.0$).

As shown in Table 6 and Figure 6, specialised clusters achieve substantially higher MRR (C0 Law: 0.615, C6 BioSTEM: 0.402), while the two cross-disciplinary clusters with field entropy above 2.0 (C4: 0.164, C7: 0.234) are the hardest. The

Table 6: Per-cluster retrieval on the full test set (sorted by MRR).

Cluster (Field)	n	MRR	Hit@10	Recall@20
C0 — Law	78	0.615	0.821	0.925
C6 — BioSTEM	131	0.402	0.718	0.791
C1 — STEM-EE	511	0.353	0.630	0.672
C2 — Humanities	115	0.340	0.678	0.734
C7 — Cross-disc.	179	0.234	0.486	0.500
C5 — STEM-AI	478	0.209	0.460	0.491
C3 — Business	437	0.167	0.387	0.401
C4 — Cross-disc.	189	0.164	0.333	0.357

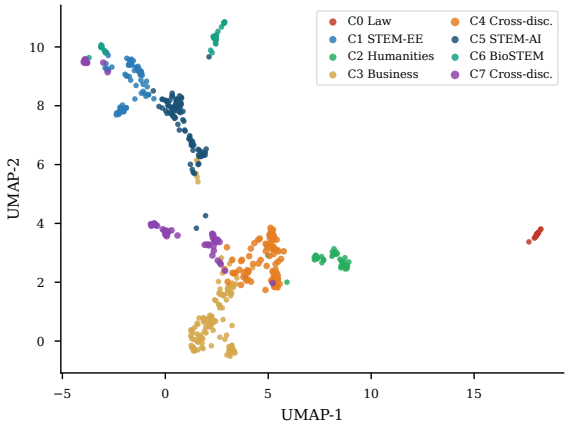


Figure 6: 2D UMAP of 531 program prototypes colored by cluster.

pattern is consistent with the intuition that semantic diffuseness increases retrieval difficulty, though it is not strictly monotone across all eight clusters. A sub-centroid probe that splits each C4/C7 program into two centroids yields its largest gains precisely in these two clusters and significant penalties in compact ones (permutation $p \approx 0.036$; Appendix F.2), consistent with genuinely multimodal admit pools in the cross-disciplinary clusters.

8.2 Feature Attribution by Field

We apply Integrated Gradients (IG) (Sundararajan et al., 2017) to the gated fusion layer, computing the attribution of text vs. structural features for each (student, program) pair in the test set. Figure 7 shows mean attributions aggregated by program field.

The largest structural reliance is observed for BioSci: structural features account for 65% of attribution, uniquely dominating text across all fields. This reflects the quantitative thresholds (GPA, publication count) typical of biomedical admissions. All other fields are text-dominant ($\geq$66%): Business (76%) reflects the primacy of internship narratives, while STEM (75%) suggests research description *quality* matters beyond raw

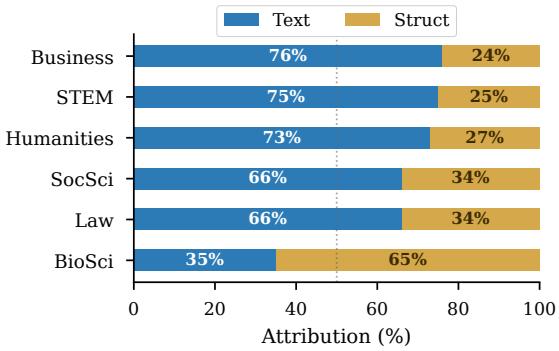


Figure 7: IG attribution of text vs. structural features by program field (mean over test-set pairs, bootstrap SE $<2\%$).

project counts.

Gate activation analysis and text-category contribution details are in Appendix D.

8.3 Case Studies

Appendix D.4 presents three cases where the bi-encoder ranks the admitted program at #1 while BM25 (Robertson and Zaragoza, 2009) places it at #43, #49, and #369, spanning terminology mismatch, cross-field semantic gap, and complete surface-form disconnection, all resolved by learned prototype embeddings.

9 Conclusion

We presented PURE, a positive-only biencoder for graduate program retrieval combining GPA-weighted program prototypes, gated structural fusion, and an MP-MRR protocol that corrects for false negatives in many-to-many matching. PURE achieves MRR $=$ 0.2696, significantly outperforming BM25 ($p<0.001$) and GPT-4o-mini ($+8.3$ points) while matching Biencoder+GPT-4o-mini reranking at negligible inference cost ($p = 0.52$). Cluster analysis and IG attribution reveal field-specific retrieval difficulty and reliance on structural vs. textual signals, supporting prototype-based dense retrieval as a foundation for academic program recommendation.

Limitations

PURE estimates admission compatibility from a single-platform, Chinese-applicant dataset, so generalization across platforms, languages, and demographics is untested, observed admits carry survivorship bias, and programs with few training admits rely on a description-based fallback. The backbone is frozen and structural features are fused only on the applicant side. We discuss these limitations in detail in Appendix F.

References

Alisha Baskota and Yiu-Kai Ng. 2018. A graduate school recommendation system using the multi-class support vector machine and knn approaches. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 277–284. IEEE.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.

Inssaf El Guabassi, Zakaria Bousalem, Rim Marah, and Aimad Qazdar. 2021. A recommender system for predicting students' admission to a graduate program using machine learning algorithms. *International Journal of Online and Biomedical Engineering (iJOE)*, 17(2):135–147.

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220. ACM.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292. ACM.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6894–6910. Association for Computational Linguistics.

Mahamudul Hasan, Shibbir Ahmed, Deen Md Abdullah, and Md Shamimur Rahman. 2016. Graduate school recommender system: Assisting admission seekers to apply for graduate studies in appropriate graduate schools. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 502–507. IEEE.

Melina Heidari Far and Elham Tabrizi. 2026. A hybrid machine learning approach for graduate admission prediction and combined university-program recommendation. *arXiv preprint arXiv:2603.29881*.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 781–789.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48. ACM.

Ryuichi Kiryo, Gang Niu, Marthinus Christo du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Chuan Luo, Pu Zhao, Chen Chen, Bo Qiao, Chao Du, Hongyu Zhang, Wei Wu, Shaowei Cai, Bing He, Saravanakumar Rajmohan, and 1 others. 2021. PULNS: Positive-unlabeled learning with effective negative sample selector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8784–8792.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning LLaMA for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2421–2425. ACM.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2014–2037. Association for Computational Linguistics.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

3982–3992. Association for Computational Linguistics.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 452–461. AUAI Press.

Douglas A. Reynolds. 2009. Gaussian mixture models. *Encyclopedia of Biometrics*, 741(659-663).

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Yuta Saito, Suguru Yaginuma, Yuta Nishino, Hayato Sakata, and Kazuhide Nakata. 2020. Unbiased recommender learning from missing-not-at-random implicit feedback. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 501–509.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14918–14937. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328. PMLR.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5776–5788.

Wanvipa Wongvilaisakul, Paniti Netinant, and Meennapa Rukhiran. 2023. Dynamic multi-criteria decision making of graduate admission recommender system: AHP and fuzzy AHP approaches. *Sustainability*, 15(12):9758.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Junliang Yu, Hongzhi Yin, and Tong Chen. 2022a. Self-supervised learning for recommendation: Fundamentals and advances. In *Companion Proceedings of the ACM Web Conference 2022 (WWW '22 Companion)*.

Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022b. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1294–1303. ACM.

Yijun Zhao, Xiaoyu Chen, Haoran Xue, and Gary M. Weiss. 2023. A machine learning approach to graduate admissions and the role of letters of recommendation. *PLOS ONE*, 18(10):e0291107.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference (WWW)*, pages 845–854. ACM.

A Ethical Considerations

Data privacy. Records are drawn from a publicly accessible forum where users voluntarily share admission outcomes. Records contain no government-issued identifiers, contact information, or financial data. We pseudonymise records using composite feature keys and report only aggregate statistics. Raw records are not released. The dataset is used consistent with the platform's terms of service permitting research use of anonymised public data. Code and a further-anonymised dataset are released under a research-only license with explicit instructions against re-identification.

Intended use. This system is a *decision-support* tool to help students explore compatible programs, not a replacement for human advising or institutional admission decisions. Scores must not be used to reject applicants or rank students against each other. Deployment in high-stakes decisions without human oversight constitutes misuse.

Selection bias. The training population over-represents students who sought commercial advising services, potentially skewing recommendations for underrepresented profiles. Practitioners should audit outputs across school tiers before deployment. Subgroup fairness results are reported in Appendix F.

B Evaluation Protocol Details

B.1 Formal Metric Definitions

MP-MRR is defined in Eq. 12. The remaining metrics use the same notation: $\mathcal{S}_{\text{test}}$ is the test set, $\mathcal{P}^+(s)$ the observed positives for student s, and $\pi_s(p)$ the rank of program p.

MP-Hit@K: Fraction of students for whom at least one positive program appears in the top-K results.

Recall@K: Fraction of all (s, p) positive pairs with $\pi_s(p) \leq K$, normalized by total positives.

MAP@K: Standard MAP treating all $p \in \mathcal{P}^+(s)$ as relevant for each student s.

B.2 Composite Key Construction

Because the dataset lacks persistent student identifiers, we infer student identity from a five-field composite key:

$$k(r) = (\text{school},\ \text{major},\ \text{GPA},\ \text{year},\ \text{exp}_{:80})$$

where $\text{exp}_{:80}$ denotes the first 80 characters of the experience text. Two records sharing all five values are assigned to the same student group. Validated against platform user IDs on the 2024 training cohort, this key achieves 99.1% precision and 97.4% recall. The small recall gap arises from students who updated their experience description or GPA between applications; such cases are conservatively split rather than merged.

The adversarial robustness analysis in §B.3 additionally uses a conservative six-field *structural-only* key (GPA, school tier, IELTS band, research count, internship count, competition count) that omits school name and experience text. This represents a harder identity-inference scenario and yields 3,052 groups from 4,453 test records, providing a lower bound on grouping precision.

B.3 Adversarial Grouping Analysis

Two extreme grouping strategies bound the range of possible MP-MRR values under different identity-inference assumptions.

Adversarial stress test. (i) *Maximal merge*: a 3-field composite key (GPA, tier, degree type) that maximally inflates positive sets yields MRR $=$ 0.281 ($+0.011$). (ii) *Maximal split*: treating every record as a distinct student yields MRR $=$ 0.199 (-0.071). The adversarial range $[0.199, 0.281]$ spans 0.082 around our reported 0.270. Even at the worst-case single-positive split, the biencoder exceeds all baselines' single-positive MRR equivalents, and method rank ordering is preserved across all strategies.

Collision analysis. Using the structural-only key (6 fields), 3,052 unique groups are derived from 4,453 test records, and 26.9% of keys are shared by two or more records, merging distinct students and artificially expanding the gold set. This bias is symmetric across all methods and collisions are not concentrated in the hardest clusters (C4, C7), so rank ordering between systems is unaffected.

C Model Details and Extended Ablations

C.1 Prototype Weighting Strategies

Table 7 compares six strategies for weighting historical admits when constructing program prototypes. Encoder weights are held fixed across variants (hence MRR values are lower than the full model's 0.2696, which includes PULNS (Luo et al., 2021) hard-negative selection). Replacing admit-text prototypes entirely with official program descriptions (526/531 recovered) yields MRR $=$ 0.149 (-10.8 points), showing that descriptions capture program content but not admission selectivity. Among weighting strategies, GPA-softmax

Table 7: GPA weighting ablation (encoder weights fixed across variants). GPA-softmax outperforms all alternatives; Tier-softmax performs worst.

Prototype weighting	MP-MRR	Recall@20
GPA-softmax (ours)	**0.256**	**0.555**
Top-50% GPA only	0.245	0.551
GPA-sqrt	0.242	0.554
Uniform	0.241	0.550
GPA×Tier	0.242	0.561
Tier-softmax	0.237	0.547

Table 8: PU-aware loss comparison (best validation checkpoint; n=1,537 students). By epoch 60, batch masking reaches MRR = 0.197 while nnPU plateaus at 0.185. The ordering is confirmed significant on the temporal test split with paired bootstrap CIs, including a multi-positive NCE variant (Table 9).

PU strategy	MRR	Hit@10	NDCG@10
Batch-wide masking (ours)	0.181	**0.399**	**0.215**
nnPU loss	0.186	0.379	0.214
PU reweighting (IPS)	0.184	0.390	0.215
Global masking	0.148	0.338	0.175

Table 9: PU-aware loss comparison on the temporal test split (n=1,866). Δ and p: 10,000-iteration paired bootstrap against batch-wide masking.

PU strategy	MRR	Best ep.	Δ	p
Batch masking (ours)	**0.271**	10	–	–
nnPU loss	0.261	10	−0.009	0.016
Multi-positive NCE	0.255	20	−0.016	0.002

Table 10: PURE training hyperparameters.

Hyperparameter	Value
Embedding dim d	384
Batch size	256
Temperature τ	0.05
Hard-neg margin m	0.2
Hard-neg weight λ	0.5
Hard-neg pool size K	5
Optimizer	AdamW (wd 10^{-4})
Learning rate	3×10^{-4} (peak)
LR scheduler	OneCycleLR (10% warmup)
Epochs	60
Trainable params	≈400K
Min. admits (valid prog.)	5
Hard-neg cache refresh	every epoch

consistently outperforms all alternatives, and tier-weighted prototypes perform worst, confirming that GPA captures admit quality better than school prestige alone.

C.2 PU-Aware Loss Comparison

Table 8 compares four strategies for handling the PU structure of in-batch negatives, all using the same encoder architecture and 60-epoch training schedule. *No masking* is the vanilla InfoNCE (van den Oord et al., 2018) baseline. *Batch-wide masking* (ours) excludes any program positive for *any* batch student from both the InfoNCE denominator and the hard-negative pool. *Global masking* extends this to all positives seen in the entire training set, which over-suppresses the denominator and hurts performance. *PU reweighting* applies an IPS-style correction (Joachims et al., 2017) with estimated prior $\pi^+ = |\mathcal{P}^+(s)|/|\mathcal{P}|$ per student. *nnPU loss* (Kiryo et al., 2017) replaces the InfoNCE objective with the nnPU risk estimator.

At epoch 30, nnPU converges slightly faster (MRR 0.186 vs. 0.181), but by epoch 60 batch masking overtakes all alternatives (MRR 0.197 vs. 0.185 for nnPU), requiring no propensity estimates. The advantage of nnPU and PU reweighting early in training likely reflects their explicit risk correction, but this correction relies on a positive prior estimated from biased observed accepts, which becomes a liability as training progresses. Global masking is consistently the worst throughout training, as removing all training-set positives from the denominator over-suppresses valid negatives.

Statistical confirmation on the temporal test split. Table 8 reports validation-cohort point estimates; re-running the comparison on the temporal test split (n=1,866; same hyperparameters and schedule), with an added *multi-positive NCE* variant that sums over all of a student's positives in the numerator, preserves the relative ordering: batch-wide masking significantly outperforms both alternatives ($p \leq 0.02$; Table 9). Multi-positive NCE converges later (best epoch 20 vs. epoch 10), and nnPU's deficit is consistent with underestimation of the positive prior from biased accept reporting (§5.4).

C.3 Training Hyperparameters

Table 10 lists all hyperparameters used for PURE. Values were tuned on the 2025 validation cohort (train ≤2024, val = 2025), with MRR as the selection criterion.

C.4 Few-Shot Prototype Learning

Table 11 shows performance when each prototype is built from at most K randomly sampled training admits, simulating deployment for programs with limited history.

MRR improves substantially up to K=10, recovering ≈70% of the full-data value, and then more gradually, consistent with Table 15: below 10

Table 11: Few-shot prototype learning: performance vs. max admits per prototype (K), using the base biencoder encoder (without PULNS hard-negative selection; full PURE: MRR $= 0.270$).

Max admits K	MRR	Hit@10	Recall@20
1	0.095	0.223	0.231
2	0.107	0.250	0.246
3	0.125	0.288	0.284
5	0.139	0.312	0.331
10	0.178	0.380	0.431
20	0.220	0.460	0.521
All	**0.256**	**0.495**	**0.555**

admits, prototype vectors are insufficiently representative of a program's admission profile.

C.5 LLM Context Sensitivity

To test whether richer program context allows GPT-4o-mini (OpenAI, 2023) to close the gap with our biencoder, we run a controlled 100-student pilot with three context levels: (i) **Names only** (MRR $=$ 0.076, Hit@10 $=$ 0.19); (ii) **term frequency-inverse document frequency (TF-IDF) keyword summaries** derived from the same training admit texts used by BM25 (Robertson and Zaragoza, 2009) (MRR $= 0.059$, Hit@10 $= 0.19$); (iii) **Official program descriptions** scraped from the platform's program pages, 526/531 programs recovered (MRR $= 0.046$, Hit@10 $= 0.20$).

Adding program context consistently *hurts* LLM ranking precision while leaving Hit@10 unchanged. We hypothesize this is due to *attention dilution*: when 531 programs each receive a multi-sentence description, the 8K-token prompt is dominated by description text, making fine-grained student–program discrimination harder. All three conditions fall well below BM25 (MRR $= 0.221$) and our biencoder (0.270), confirming that LLM world knowledge, even augmented with official program metadata, cannot substitute for learned admission-pattern representations.

D Analysis Details

D.1 Cluster Stability and Structural Profiles

Cluster stability is measured by pairwise ARI (Hubert and Arabie, 1985) across 10 random seeds (mean 0.563 ± 0.066); domain-focused clusters (C0–C3, C5–C6) are more stable (± 0.04) than the cross-disciplinary ones (± 0.12), reflecting their inherently ambiguous membership. Intra-cluster cosine similarity (0.359) substantially exceeds inter-cluster (-0.060), confirming meaningful geometric separation.

Table 12: Mean applicant structural features by program cluster (test set).

Cluster	Field	GPA	IELTS	Res.	Intern	Pub
C0	Law	3.25	7.00	0.83	2.58	0.000
C1	STEM-EE	2.97	6.51	0.96	1.00	0.011
C2	Humanities	3.27	6.88	1.02	1.21	0.010
C3	Business	3.10	6.79	0.90	2.04	0.007
C4	Cross-disc	3.10	6.81	0.85	1.83	0.005
C5	STEM-AI	3.04	6.65	0.91	1.04	0.011
C6	BioSTEM	2.96	6.51	1.58	1.13	**0.050**
C7	Cross-disc	3.10	6.57	1.01	1.57	0.013

Table 12 reports mean structural features per cluster. Three patterns align with the IG (Sundararajan et al., 2017) attribution results in §8: BioSTEM (C6) has the highest publication count (0.050, roughly $5\times$ the overall mean) and research count (1.58), consistent with structural features dominating attribution for life sciences. Law (C0) has the highest internship count (2.58) and IELTS (7.00), reflecting language and practical-experience requirements in legal admissions. The cross-disciplinary clusters (C4, C7) sit between the domain clusters on every dimension, which, combined with their high field entropy, explains their lower retrieval performance.

D.2 Gate Activation Analysis

We examine whether the scalar gate g automatically compensates for short or low-quality text by increasing structural reliance. Computing gate values for all 4,453 test records, the mean gate is 0.690 ± 0.080, with only a weak positive correlation with text length (Spearman $\rho = 0.064$, $p < 10^{-4}$). Students with short text ($\leq$71 chars) have slightly *lower* gate values (0.681) than long-text students (0.694), ruling out automatic compensation. Instead, gate activation is driven primarily by *field type*: high for BioSci (where structural features dominate IG attribution) and low for Humanities, consistent with the IG analysis in §8. This motivates future work on gate conditioning that is aware of text-coverage quality.

D.3 Text Category Contribution

Figure 8 shows retrieval MRR when only a single category of experience text is used (via the GMM (Reynolds, 2009) Text-only variant). Research experience is the most informative single category (MRR $= 0.086$), followed by internship (0.057), other (0.028), and competition (0.027). The full text achieves 0.128, substantially higher than any single category, confirming complementarity across experience types. Competition en-

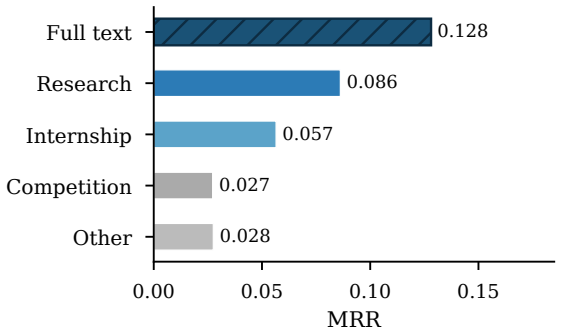


Figure 8: MRR when retrieving using only a single experience text category. "All (full)" uses all categories concatenated.

tries, though weak individually, contribute signals for specific programs (e.g., math competitions for quantitative finance).

D.4 Case Studies

Table 13 presents three 2026 test-set students where the biencoder ranks the admitted program at #1 while BM25 (Robertson and Zaragoza, 2009) ranks it substantially lower.

Case 1 (Physics student): Despite a GPA of 86.65/100 (≈3.0/4.0), this student's physics competition entries (optics, light-diffraction experiments) generate a text embedding close to the prototypes of physics programs. Our biencoder ranks the admitted program at #1, while BM25 places it at #43, as the student's experience text contains domain-specific physics terminology but few keywords matching the program document's surface form.

Case 2 (English-major student): This student (English major, Tier-211 university) was admitted to an MSc in Digital Communication and Culture, a media/social-science program. Our biencoder ranks it at #1 because the student's text (TV production, social media operations, brand communication research) maps into cluster C4 (Business–Communication cross-disciplinary space). BM25 ranks it at #49 due to lexical mismatch between "English" major keywords and the communication program description. A text-agnostic model would never surface this cross-field match.

Case 3 (Philosophy student): This student (Philosophy honors program, Tier-211 university, GPA 4.0/4.0) pursued two applied AI ethics research projects: one applying the trolley problem to autonomous vehicle decision-making, the other analysing the concept of "machine morality." BM25 ranks the MA in AI, Ethics and Society at the University of Hong Kong at #369 out of 531. The program name contains "AI" but the student profile contains only "philosophy" and "ethics" tokens that fail to match the program's keyword profile. Our biencoder ranks it at #1: the prototype for the AI Ethics program is built from historical admits who wrote about similar applied ethics and AI topics, creating a representation that captures semantic alignment that BM25 cannot represent. This is the most extreme lexical-semantic gap in our test set ($\Delta\text{rank} = 368$) and illustrates how large BM25 errors can be when academic identity is encoded in research themes rather than field labels.

E Dataset and Fairness Details

E.1 School Tier Encoding

School tier is an ordinal integer derived by mapping the applicant's self-reported undergraduate institution against reference lists of Chinese university classifications. Tier 3 (985) covers China's top-39 research universities; Tier 2 (211 / double first-class) covers national-key and double first-class institutions; Tier 1 (non-985/211 domestic) covers all other mainland Chinese universities; Tier 0 (international) covers overseas and Hong Kong/Macau institutions; Tier -1 denotes records where no school name is present.

E.2 Dataset Statistics

Table 14 summarises the dataset. The 2024–2025 cohorts form the training split; 2026 is held out as the test set, giving a strict temporal boundary that prevents prototype leakage. Of 1,639 programs in training, 531 meet the $\geq$5-admit threshold and serve as retrieval targets with learned prototypes. The remaining 67.6% receive no learned prototype and are served by the description-based fallback (§5.1). Publication rates are low (1.0% of training records), consistent with the overwhelmingly master's-level applicant population.

E.3 Cold-Start Performance Stratification

Table 15 stratifies biencoder performance by the number of training admits for the target program. Performance degrades monotonically as prototype size decreases, confirming that at least 10–20 admits are needed for reliable prototype construction. Programs with *zero* training admits are handled by the description-based fallback (§5.1), yielding MRR $= 0.149$, below learned prototypes but 37% above the strongest zero-shot embedding baseline (BGE-M3: 0.109), indicating that program description text encodes enough field semantics to serve as a soft prototype even without any observed admits.

Table 13: Case studies: PURE ranks the admitted program #1; BM25 ranks it substantially lower.

Student	Profile	Actual Admit	BM25	Ours
Applied Physics	GPA 86.65/100, key univ.; physics competitions (optics, light-diffraction experiments)	MSc Physics, King's College London	#43	**#1**
English major	GPA 3.69/4.0, tier-211; TV director, social media ops, brand communication research	MSc Digital Communication & Culture, Univ. of Sydney	#49	**#1**
Philosophy (honors)	GPA 4.0/4.0, tier-211; AI ethics research (autonomous vehicle trolley problem, machine morality)	MA AI, Ethics & Society, Univ. of Hong Kong	#369	**#1**

Table 14: Dataset statistics. Train = 2024–2025 cohorts; Test = 2026 cohort. *2,196 unique student groups; 1,866 after filtering to those with at least one valid-program offer.

Statistic	Train	Test
Total records	10,838	4,453
Master's / PhD	10,306 / 532	4,266 / 187
Unique programs (all)	1,639	1,111
Valid programs ($\geq$5 admits)	531	531
Unique students (grouped)	5,404	2,196 / 1,866*
Median GPA (4-scale)	3.09	3.07
IQR GPA	[2.78, 3.43]	[2.73, 3.45]
Median exp. length	106 chars	106 chars
% with research	57.7%	57.7%
% with internship	77.0%	77.0%
% with publication	1.0%	1.1%

Table 15: Biencoder MRR stratified by program training-admit count (cold-start analysis). The zero-admit row uses the description-based fallback prototype.

Stratum	n	MRR	Hit@10
0 (desc. fallback)	–	0.149	–
5–9 (cold)	520	0.218	0.450
10–19	679	0.236	0.486
20–49	798	0.286	0.575
$\geq$50 (warm)	650	0.348	0.595

E.4 Subgroup Fairness Disaggregation

Table 16 disaggregates MP-MRR and Hit@10 by GPA quartile and undergraduate school tier on the full 1,866-student test set. GPA quartile spread is narrow ($\pm$0.010), indicating no systematic disadvantage for lower-GPA students. Among domestic tiers (Tier 1–3), MRR ranges only 0.016 (0.260–0.276), confirming GPA-weighted prototypes are task-beneficial rather than prestige-amplifying. International applicants (Tier 0, MRR = 0.221) underperform, consistent with the model being trained on a predominantly domestic-applicant corpus.

Table 16: MP-MRR and Hit@10 across GPA quartiles and school tiers on the 2026 test set.

Stratum	MP-MRR	Hit@10
By GPA quartile		
Q1 ($\leq$2.73, lowest)	0.261	0.498
Q2 (2.73–3.07)	0.265	0.510
Q3 (3.07–3.45)	**0.281**	**0.527**
Q4 (>3.45, highest)	0.270	0.529
By school tier		
Tier -1 (no school name)	0.305	0.525
Tier 0 (international)	0.221	0.442
Tier 1 (non-985/211 domestic)	0.260	0.496
Tier 2 (211 / double first-class)	0.256	0.493
Tier 3 (985)	0.276	0.554

F Discussion

F.1 Limitations

Task scope. PURE estimates *admission compatibility* (§1), not research fit, supervisor matching, or student outcomes; recommendations should be combined with fit-oriented resources before use.

Data scope. All records originate from a single platform serving Chinese applicants, concentrated in Hong Kong, Singapore, the UK, and Australia; generalization to other regions and demographics is untested, and observed admits carry survivorship bias. Temporal transfer within this dataset retains 80% of full-data performance, and porting to a new platform requires only re-embedding records in the same schema and re-running training, with no architectural changes (details below). Programs with few training admits degrade gracefully to a description-based fallback (cold-start analysis below).

Model scope. The MiniLM backbone is frozen because end-to-end fine-tuning overfits at this dataset scale; structural features are fused only on the applicant side because program-side metadata

is unavailable; experience descriptions are Chinese-language only (extended analysis below).

Evaluation and fairness. GPA-weighted prototypes do not disadvantage lower-GPA applicants ($\pm$0.010 MRR spread across quartiles; Table 16), and Recall@20 $= 0.576$ confirms practical top-list coverage. Learned sparse retrievers and trained late-interaction models remain untested.

F.2 Extended Analysis

Data scope and generalization. All records originate from a single Chinese-applicant platform concentrated in East and Southeast Asia, the UK, and Australia. Temporal robustness was probed by rebuilding prototypes using only $\leq$2024 data: MP-MRR drops from 0.258 to 0.207 (-0.051), with evaluable students falling from 1,866 to 1,523 as programs without 2024 admits lose their prototypes. This confirms prototype quality scales with cohort size and motivates continual data ingestion. This stability plausibly reflects slow institutional drift: committee composition and core requirements (GPA floors, field alignment, exam score ranges) evolve on multi-year timescales, and GPA-weighted prototypes capture the *distribution* of admits rather than a single-cohort snapshot, making them robust to single-year fluctuations. For programs with known recent structural changes (new faculty, a pivot in research focus), the description-based fallback provides a clean zero-admit restart. Unobserved positives (programs a student never applied to) cannot be corrected by the multi-positive protocol. Integrating application logs alongside outcomes would enable de-biased evaluation in future work.

Cold-start and long-tail handling. Training exclusion does not imply inference exclusion (§5.1): programs below the 5-admit threshold are ranked via the description-based fallback, whose zero-admit MRR (0.149) exceeds all zero-shot embedding baselines (Table 15). The fallback works because the shared projection W_t learns an embedding space in which compatible students cluster near a program's disciplinary signature, so a description-only prototype still approximates field-level compatibility. A longer-term mitigation is *cluster-neighbor transfer*: initialize a cold program's prototype as the centroid of the most similar warm cluster in the §8 cluster geometry (e.g., a new cross-disciplinary ML+Economics program receives a C4-type initialization), then refine it as admits accumulate; we leave its empirical evaluation to future work.

Table 17: Sub-centroid probe: per-cluster MRR with single vs. two sub-centroid prototypes for C4/C7 programs, sorted by Δ; clusters above the midrule are multimodal. p-values: 10,000-iteration paired bootstrap. Per-student MRR, not comparable to the pair-pooled Table 6.

Cluster	n	Single	Sub-c.	Δ	p
C7 Cross-disc.	136	0.212	0.239	+0.027	0.23
C4 Cross-disc.	172	0.133	0.138	+0.005	0.79
C2 Humanities	111	0.291	0.292	+0.001	0.92
C5 STEM-AI	422	0.157	0.155	−0.002	0.32
C0 Law	75	0.419	0.417	−0.003	0.004
C3 Business	375	0.126	0.122	−0.004	<0.001
C6 BioSTEM	117	0.345	0.337	−0.008	<0.001
C1 STEM-EE	458	0.279	0.271	−0.009	<0.001

Sub-centroid prototypes. The cluster analysis (§8) identifies two multimodal cross-disciplinary clusters (C4, C7), raising the question of whether single centroids under-represent such programs. We apply K-Means (K=2) within the training admits of each C4/C7 program and score with $\max(\mathbf{q}_s^\top \mathbf{c}_1, \mathbf{q}_s^\top \mathbf{c}_2)$; all other clusters retain a single GPA-weighted centroid. The hypothesis predicts that multimodal clusters should gain while compact clusters are neutral or hurt. The two cross-disciplinary clusters indeed rank 1st and 2nd among all eight by Δ (exact permutation test: $p = 1/28 \approx 0.036$; Table 17), and the four compact clusters with significant penalties (C0, C1, C3, C6) act as a negative control: splitting is harmful precisely where admit pools are already coherent. We retain single centroids in PURE because the overall gain is small and cluster-conditional, but the result supports cluster-conditioned multi-prototype extensions.

Model and architectural constraints. Freezing MiniLM (Wang et al., 2020) is necessary at this dataset scale: end-to-end fine-tuning (backbone lr $= 3\times10^{-5}$, 30 epochs) peaks at MP-MRR $= 0.235$ then degrades to 0.229 due to overfitting, well below the frozen-backbone model (0.270). The architecture is asymmetric by design because program-side structural data (GRE thresholds, quotas) is unavailable, and a symmetric extension incorporating program metadata would be natural when such data exists. Experience descriptions are Chinese-language only, and cross-lingual generalization is not assessed. Zero-shot encoders (E5-base (Wang et al., 2022): 0.102, BGE-M3 (Chen et al., 2024):

0.109) substantially underperform BM25 (Robertson and Zaragoza, 2009) (0.221) regardless of model scale, confirming domain adaptation is necessary. A cross-encoder reranker (BM25 top-50 $\rightarrow$ fine-tuned MiniLM cross-encoder, MRR $= 0.063$) falls far below BM25 alone (0.221): because program representations must be synthesized from admit texts rather than matched against official descriptions, the cross-encoder's pairwise comparison is unreliable. An E5-base reranker over the same BM25 candidates similarly achieves MRR $=$ 0.055, confirming that the bottleneck is prototype construction rather than ranking model capacity. Learned sparse retrievers (SPLADE (Formal et al., 2021)) and trained ColBERT (Khattab and Zaharia, 2020) variants remain future work pending mature multilingual CJK checkpoints.

Fairness and cold-start. GPA upweighting could skew prototypes toward high-GPA profiles. Subgroup audits (Table 16) show MRR spreads only $\pm$0.010 across GPA quartiles and 0.016 across domestic tiers (Tier 1–3), with tier-weighted prototypes performing *worse* (0.237 vs. 0.256 for GPA-softmax), confirming the weighting is task-driven rather than prestige-amplifying. Cold-start analysis (Table 15) shows a monotone performance gradient from 0.218 (5–9 training admits) to 0.348 ($\geq$50 admits); a few-shot experiment capping admits per program at inference confirms sharp degradation below K=10 (MRR $=$ 0.095 at K=1, 0.178 at K=10). Prototype-free generalization via program metadata is an important future direction.