Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multidimensional semantic retrieval of classical Chinese poetry for literary heritage discovery

Zhaojiang Wang
Shanghai Normal University
Chen Ma
Shanghai Normal University
Wei Duan
Shanghai Normal University
Zongzhe Wang
Shanghai Normal University
Aihua Sun
Shanghai Normal University
Wei Dai
Shanghai Normal University
Xiaping Wu
Shanghai Normal University
Hengnian Qi
qihengnian@foxmail.com

Huzhou Normal University

Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11510169/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multidimensional semantic retrieval of classical Chinese poetry for literary heritage discovery

Zhaojiang Wang[1,2], Chen Ma[1], Wei Duan[1], Zongzhe Wang[1], Aihua Sun[1], Wei Dai[1], Xiaping Wu[1] and Hengnian Qi[2]*

[1] School of Humanities, Shanghai Normal University, Shanghai 200234, China
[2] School of Artificial Intelligence, Huzhou Normal University, Huzhou 313000, China

Correspondence: Hengnian Qi (qihengnian@foxmail.com)

Abstract

Classical Chinese poetry expresses meanings that keyword searches often miss. We construct a retrieval dataset covering emotion, authorial intent, theme and thought, and develop an embedding optimization method for literary heritage discovery. Large language model analyses, semantic sampling, multi-model candidate pooling and four-level relevance annotation produce 29,080 queries. Human annotation provides a gold test set of 8,400 query–poem pairs. Gain-aware Listwise Ordinal Loss (GLO) combines graded-gain listwise supervision with constraints on similarity-score gaps between relevance levels. Using Qwen3-Embedding-0.6B, default GLO achieves the highest macro-averaged normalized discounted cumulative gain at 10 (nDCG@10) at relevance thresholds 1 and 2 across three evaluation resources. When only highly relevant, grade-3 poems qualify, the human gold score increases from 0.174 to 0.391; LambdaLoss@10 scores slightly higher. The dataset and method support the discovery and comparison of poems according to different aspects and degrees of semantic relevance.

Introduction

Classical Chinese poetry records emotional experiences, values and aesthetic expression across historical periods and forms an important body of literary heritage. Digitized collections give researchers access to large numbers of poems and create a need for retrieval that responds to specific semantic questions. Researchers may seek poems expressing different emotional attitudes toward a shared theme or compare authorial intent and thought behind similar imagery. Searches by author, title and keyword help locate known works, while predefined labels support browsing within established categories. Detailed semantic needs require retrieval beyond these access points. Fig.1 illustrates how dimension-specific retrieval connects semantic research questions with poems for comparative reading and cultural interpretation.

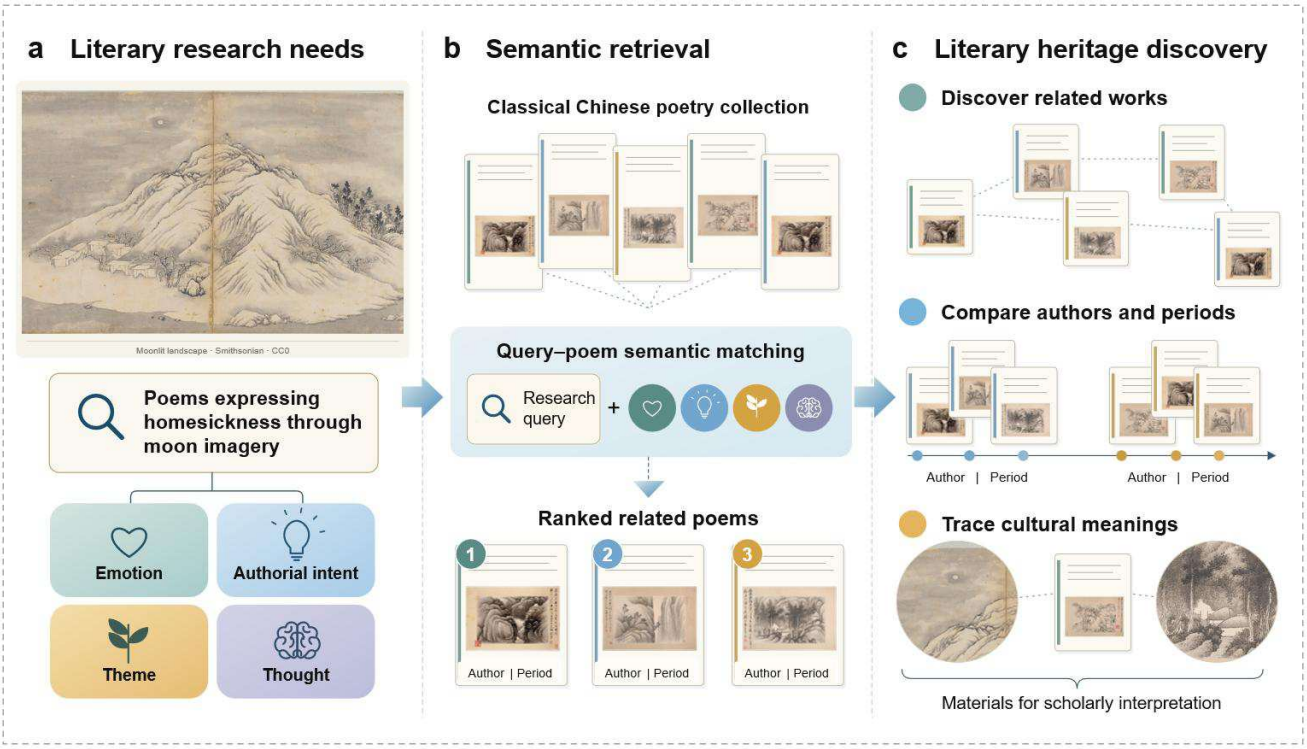


Fig. 1 | Multidimensional semantic retrieval for literary heritage discovery. a Literary research needs expressed through emotion, authorial intent, theme and thought; **b** semantic matching and ranking of related poems from a digitized collection; **c** use of retrieved poems for discovering related works, comparing authors and periods, and interpreting cultural meanings. Historical image details were obtained from the Smithsonian's National Museum of Asian Art (https://asia-archive.si.edu/object/F1911.166j/) and The Metropolitan Museum of Art (https://www.metmuseum.org/art/collection/search/36131). All source images are in the public domain and are provided under CC0; details were cropped and arranged for this figure.

Recent heritage research has begun to use the layered semantics of poetry. Poe2CLP combines phrase-level attention and cross-modal semantic alignment to generate landscape paintings[1]. Xu and Zhou integrate emotional, imagery and rhetorical features with consistency evaluation for poetry visualization[2]. Yan et al. extract cultural-space semantics from a large poetry corpus and demonstrate their interpretive use through Yellow Crane Tower materials[3]. These studies emphasize visual transformation or the interpretation of particular cultural spaces. We address the preceding problem of material discovery: retrieving original texts for an explicit semantic need and distinguishing degrees of relevance, thereby providing traceable candidates for close reading, comparison and heritage interpretation.

Poetry retrieval faces vocabulary mismatch, where a query and a relevant poem express related meanings using different words, as well as relevance that varies with the requested semantic dimension. Two poems may share a theme but differ emotionally, or share imagery while expressing different ideas. Foley identifies topic, event, mood, metadata and other queries, including combinations of these needs[4]. ACP-RAG constructs 180,000 positive and 500,000 negative pairs and fine-tunes BERT-Guwen models for coarse semantic retrieval and fine-grained reranking using the binary cross-entropy implementations BCELoss and BCEWithLogitsLoss, respectively[5]. Its retrieval-augmented question-answering setting differs from our separate evaluation of four semantic dimensions and relevance grades. Dense retrieval represents queries and documents as numerical vectors, or embeddings, in a shared space[6] and supplies materials for retrieval-augmented generation[7]. Domain-specific data and human judgments are needed to evaluate these semantic distinctions in poetry retrieval.

Relevant experience exists in scientific document representation and retrieval[8,9], cultural-memory representation[10], and legal[11], financial[12], commercial[13,14], code[15] and biomedical[16,17] applications. Retrieval resources combine logged, human-written and synthetic queries, as illustrated by MS MARCO[18], DuReader-retrieval[19], T2Ranking[20], Multi-CPR[21], TyDi QA[22], MIRACL[23] and CIRAL[24]. Large language model (LLM) query construction and rewriting[25,26,27], multi-retriever pooling[28,29], hard-example refreshing and relabeling[30,31,32], and shallow pooling[33] provide foundations for domain resource construction. Because poetic meaning admits

interpretive variation, automatic judgments also require human assessment of their grading scales and biases[34,35,36,37,38].

Training objectives specify how a model adjusts similarity scores to match relevance judgments. InfoNCE learns through positive–negative comparisons[39], but its standard binary setting does not explicitly distinguish degrees of relevance among positive examples. Learning-to-rank methods[40] model relevance at three levels: pointwise methods assess individual query–document pairs, pairwise methods compare two candidates, and listwise methods consider a candidate list jointly. Soft-label contrastive learning[41] and MarginMSE[42] transfer teacher-model judgments through candidate distributions and positive–negative score differences, respectively.

RankNet[43] and PairDistill[44] learn pairwise relations, while LambdaLoss[45] introduces weights tied to ranking metrics. ListNet[46] constructs a loss from ranking probabilities; RocketQAv2[47] matches list distributions to jointly train retrieval and reranking models; SyCL[34] combines multigrade synthetic documents with a Wasserstein-based objective. BiXSE[48] fits continuous relevance targets using pointwise binary cross-entropy with a bias term in the scoring function. Mine and Refine[31] combines label-aware supervised contrastive learning and multiclass Circle Loss with grade-dependent similarity boundaries, while WeMM-Embedding[49] weights CoSENT-style ranking terms by relevance-score differences. These approaches motivate combining listwise supervision with explicit score gaps between relevance levels.

We construct a retrieval resource covering emotion, authorial intent, theme and thought. LLM analyses provide proxy queries, which undergo semantic cluster sampling, multi-model candidate mining and relevance annotation from 0 to 3. Independent assessments by three annotators and adjudication establish a human gold test set. We then introduce Gain-aware Listwise Ordinal Loss (GLO), combining graded-gain listwise supervision with similarity-margin constraints between grades. Comparisons across general-purpose models and training objectives assess the retrieval performance of an adapted lightweight model. Fig.2 summarizes the workflow. The experiments ask whether general-purpose models distinguish these semantic dimensions, whether GLO improves retrieval over alternative objectives, and how its two components and the fine-tuning strategy affect performance.

Methods

Source corpus and four retrieval perspectives

The dataset covers four semantic perspectives. Emotion concerns the feelings expressed in a poem; authorial intent concerns its expressed communicative or creative purpose; theme concerns its subject; and thought concerns the ideas or values it articulates. Each query addresses one perspective. A candidate pool is a set of poems selected for relevance annotation for that query. The training set supports model fine-tuning, the silver test set uses automatic relevance labels, and the gold test set uses final human judgments. A pool-completion subset extends the judgments for selected silver queries by adding retrieved poems outside their original pools (Fig.2).

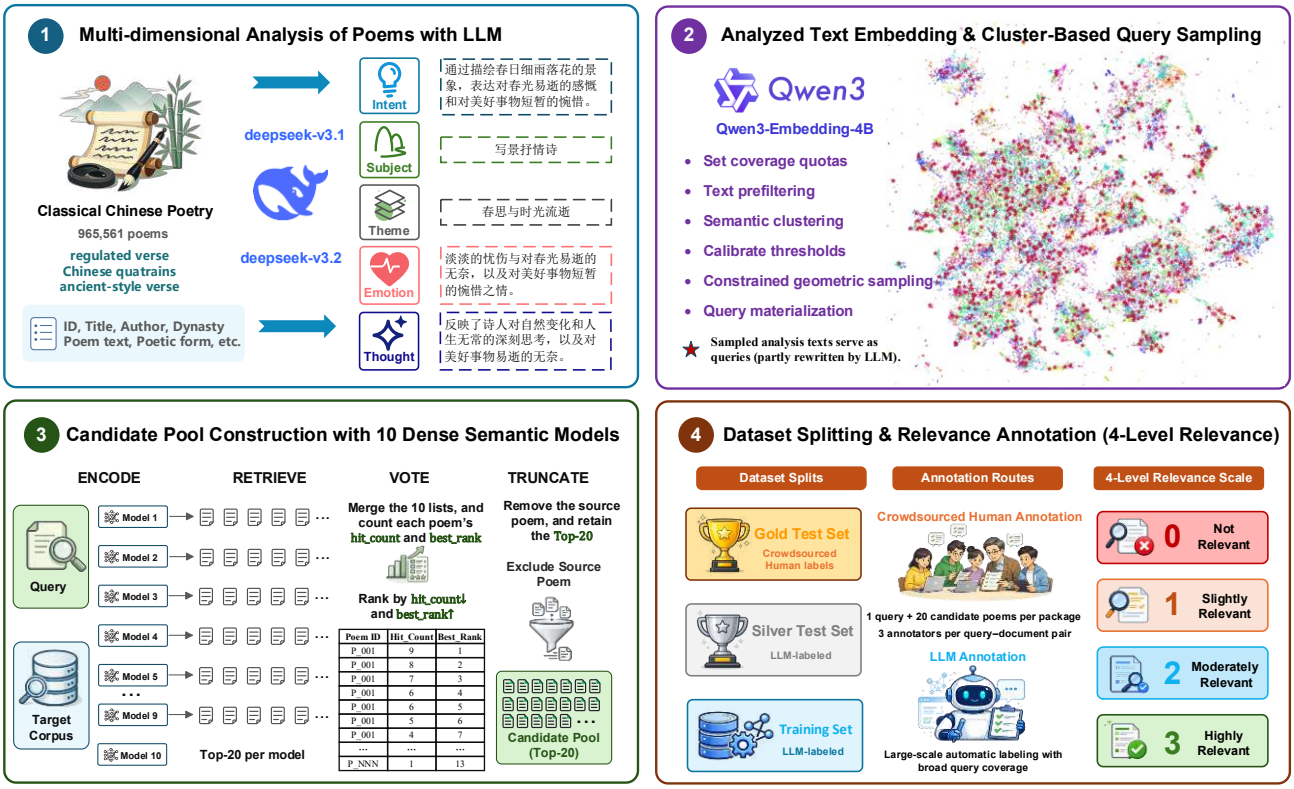


Fig. 2 | Construction of the four-dimensional poetry retrieval dataset. 1 LLM-based analysis of source poems; **2** analysis encoding, semantic clustering and query sampling; **3** candidate pooling from ten retrieval models; **4** dataset splitting and four-level relevance annotation. Candidate poems are merged and ranked by cross-model retrieval frequency and best rank, with the source poem excluded before retaining 20 candidates per query.

The source is the open poetry collection compiled by Chen Yiyun and distributed through the Chinese Ancient Books Knowledge Graph website (https://cnkgraph.com/Home/OpenResources). Filtering the genre field T for regulated verse (*lüshi*), Chinese quatrains (*jueju*), and ancient-style verse (*gufeng*) yields 965,563 poems. Each record retains a unique identifier, title, author, dynasty, text and genre. These records support query generation, sampling and candidate discovery. The source document collection is defined as follows:

$$\mathcal{D} = \{d_i\}_{i=1}^{N} \tag{1}$$

Here, $\mathcal{D}$ denotes the source poetry collection, $N = 965{,}563$ is the number of poems, and d_i is its ith poem. A poem record is represented as:

$$d_i = (t_i, a_i, x_i, m_i) \tag{2}$$

Here, t_i, a_i, x_i and m_i denote the title, author, poem text and other metadata, respectively. Source poems are used to generate proxy queries and identify neighboring poems in the analysis representation space. Dimension-specific retrieval corpora are then constructed for candidate mining and evaluation.

LLM-generated queries, documents and hard negatives provide training data for embedding models[50,51]. We use dimension-specific prompts to analyze source poems for authorial intent, theme, emotion and thought, and treat each analysis as a proxy query for the corresponding poem. Some analyses are rewritten and shortened by an LLM to resemble a retrieval query. Here, a proxy query is a textual statement of a poem's meaning in one dimension, used as a search request; the poem from which it is generated is called its source poem. The resulting poem–query record is:

$$\mathcal{P}_i = (d_i, Q_i), \qquad Q_i = \{I_i, T_i, E_i, C_i\} \tag{3}$$

Here, $\mathcal{P}_i$ pairs poem d_i with its proxy-query set Q_i. The texts I_i, T_i, E_i and C_i represent authorial intent, theme, emotion and thought, respectively.

Query encoding and semantic clustering

Qwen3-Embedding-8B[51] separately encodes each dimension's analyses. Vectors are L2-normalized. A Faiss `IndexHNSWFlat` index is built for each dimension using inner product similarity. Spherical k-means clusters the normalized analysis vectors; cluster counts are selected independently for each dimension. Clustering groups analyses with similar meanings, allowing subsequent sampling to cover varied semantic needs while limiting repeated formulations.

Cluster selection has two stages. Candidate counts are screened on random subsamples using objective curves, cosine-distance silhouette scores, and stability across runs measured by adjusted Rand index (ARI) and normalized mutual information (NMI). On a fixed subsample, semantic separation, silhouette score, objective value and minimum cluster size receive weights of 40%, 25%, 20% and 15%. Indicator directions are aligned and values min–max normalized before aggregation; manual checks inform the final selection. Counts are 800 for intent, 1,500 for emotion, 900 for theme and 800 for thought.

Stratified sampling and granularity control

Stratified sampling balances query quality, semantic coverage and granularity. Text prefiltering, offline threshold calibration and online sampling are combined with cluster quotas and semantic deduplication to reduce repeated high-frequency meanings and similar expressions.

Prefiltering removes empty, punctuation-only, unusually short or long, failed, template-like and excessively repetitive analyses. Length and character-diversity thresholds vary by dimension. Pilot samples proportional to cluster size provide neighbor-structure measurements after self-matches are removed. Empirical quantiles and preset boundaries define screening intervals; simulated filtering checks the acceptance rate.

Online sampling draws unprocessed candidates in stratified batches. Hard filters constrain neighbor counts at cosine similarity thresholds of 0.90, 0.92 and 0.95, the proportion of the nearest 50 neighbors in the same cluster, and the count in the effective similarity domain. That domain contains neighbors with similarity at least 0.85, whose count must reach the pilot-calibrated minimum. Table 1 gives the high-similarity count and same-cluster intervals.

Table 1 | Filtering intervals for neighbor counts and cluster proportions

Dimension	Neighbors ≥0.90	Neighbors ≥0.92	Neighbors ≥0.95	Same-cluster fraction in top 50
Intent	3–800	1–500	0–118	0.18–0.98
Emotion	32–800	3–500	0–200	0.14–0.98
Theme	3–800	1–500	0–200	0.18–0.98
Thought	7–800	1–500	0–200	0.22–0.98

Soft filters constrain the similarity difference between the 10th and 50th neighbors, the coefficient of variation (CV) within the effective domain, and the minimum 10th-neighbor similarity (Table 2). The 50th-neighbor upper bound is a calibration record rather than an active filter. Together, hard and soft filtering reduce overly broad queries and queries with insufficient neighborhood support.

Table 2 | Calibration parameters for neighbor similarity distributions

Dimension	10th–50th similarity gap	Effective-domain CV	10th-neighbor minimum	50th-neighbor upper bound
Intent	0.0080–0.0276	0.0097–0.0238	0.8736	0.9591
Emotion	0.0064–0.0208	0.0088–0.0200	0.9109	0.9777
Theme	0.0050–0.0352	0.0125–0.0348	0.8838	0.9940
Thought	0.0050–0.0262	0.0102–0.0247	0.8973	0.9901

Surviving candidates undergo quota and deduplication checks. A cluster quota is the ceiling of target query count multiplied by cluster size proportion and a buffer coefficient. A query is accepted only if its quota is unfilled and its maximum cosine similarity to previously accepted queries in the dimension does not exceed the deduplication threshold. Thresholds are 0.87, 0.90, 0.89 and 0.89 for intent, emotion, theme and thought. Sampling stops at the target or when candidate batches are exhausted. Query text, source and cluster identifiers, screening metrics and rejection reasons are retained. The sample comprises 29,080 queries covering 3,773 clusters; coverage ranges from 90.9% to 97.6% (Table 3).

Table 3 | Query sampling and semantic cluster coverage

Dimension	Clusters	Sampled queries	Covered clusters	Coverage (%)	Deduplication threshold
Intent	800	7,120	781	97.6	0.87
Emotion	1,500	7,491	1,390	92.7	0.90
Theme	900	7,914	875	97.2	0.89
Thought	800	6,555	727	90.9	0.89
Total	4,000	29,080	3,773	—	—

Cluster coverage is the percentage of clusters represented by at least one sampled query.

Dimension corpora and candidate mining

Dimension-specific corpora control mining and evaluation costs. Each sampled query searches the full analysis index of its dimension. After removing its source-poem self-match, the top 10 neighbors identify additional poems. The union of these poems and all query source poems, deduplicated by identifier, forms the shared corpus for that dimension. Analysis similarity only identifies potential candidates; relevance is judged from the query and poem text.

Emotion, intent, theme and thought corpora contain approximately 79,000, 76,000, 83,000 and 69,000 poems, respectively; overlap across corpora is allowed. These document collections remain fixed for subsequent mining and retrieval evaluation. Only selected query–poem pairs receive four-level judgments. The design emphasizes fine semantic discrimination among plausible neighbors while controlling computation and annotation costs.

Ten embedding models reduce dependence on one retriever: EmbeddingGemma[52], KaLM-Embedding-V2.5[53,54], multilingual-e5-large-instruct[55], jina-embeddings-v5-text-small[56], Qwen3-Embedding-0.6B, -4B and -8B[51], Linq-Embed-Mistral[57], e5-mistral-7b-instruct[58], and KaLM-Embedding-Gemma3-12B-2511[53].

Each model independently retrieves the top 20 poems for each query from the corresponding dimension-specific corpus. Results are merged and deduplicated by poem identifier. Candidates are ordered by the number of models retrieving them, in descending order, with ties resolved by their best rank across models, in ascending order. After excluding the source poem, the first 20 candidates are selected for labeling. Existing positive or negative judgments for candidates in the pool are retained. This strategy combines cross-model retrieval consensus with ranking information.

The 20 candidates and one source poem form a core annotation package. Four levels distinguish highly relevant, strongly relevant, weakly relevant and irrelevant matches. Irrelevant candidates supply explicit hard negatives.

These negatives are considered hard because retrieval models rank them highly even though relevance annotation finds them unsuitable for the query.

For source poem d_i with proxy-query set $Q_i = \{I_i, T_i, E_i, C_i\}$, the annotation package associated with query $q \in Q_i$ is defined as:

$$\begin{aligned} \mathcal{B}_{i,q} &= \left(q, \{d_i\} \cup \mathcal{K}_{i,q}\right), \\ \left|\mathcal{K}_{i,q}\right| &= 20, \qquad d_i \notin \mathcal{K}_{i,q} \end{aligned} \tag{4}$$

Here, $\mathcal{B}_{i,q}$ denotes the annotation package, and $\mathcal{K}_{i,q}$ denotes the corresponding set of candidate poems after deduplication and exclusion of the source poem. Each package contains one query and 21 poems. Query–poem pairs receive ordinal relevance labels from 0 to 3 according to the query's semantic dimension.

Dimension-specific annotation prompts share the grade structure in Table 4 while adapting the criteria to each dimension. Before full annotation, 100 queries and their 20 candidates per dimension are scored by DeepSeek-V3.2, GPT-5.4 mini and Doubao Seed 2.1 Pro, selected for comparable annotation costs. Their median serves as a proxy reference on pairs whose score range does not exceed one grade. Spearman correlation, mean absolute error, severe-disagreement rate, positive-class F1 and leave-one-model-out bias inform the choice of DeepSeek-V3.2 for automatic annotation. Human gold assessment is conducted separately as described below.

Table 4 | Four-level relevance criteria for query–poem pairs

Score	Grade	Criterion
3	Highly relevant	Accurately and fully captures the requested dimension, with clear textual evidence and no subst antial deviation.
2	Strongly relevant	Broad agreement in the requested dimension; slightly broader or narrower scope or minor devia tions are allowed.
1	Weakly relevant	Only partial overlap in meaning, imagery or another dimension, with weak evidence for the req uested dimension.
0	Irrelevant	Clearly mismatched in the requested dimension, or insufficient textual evidence for relevance.

Human annotation and quality control

A smaller human gold set provides the primary evaluation alongside larger automatically labeled resources. We randomly sample 100 queries per dimension, totaling 400 queries. Each includes 20 mined candidates and one source poem, giving 8,400 pairs. The human annotation and quality-control workflow is summarized in Fig.3.

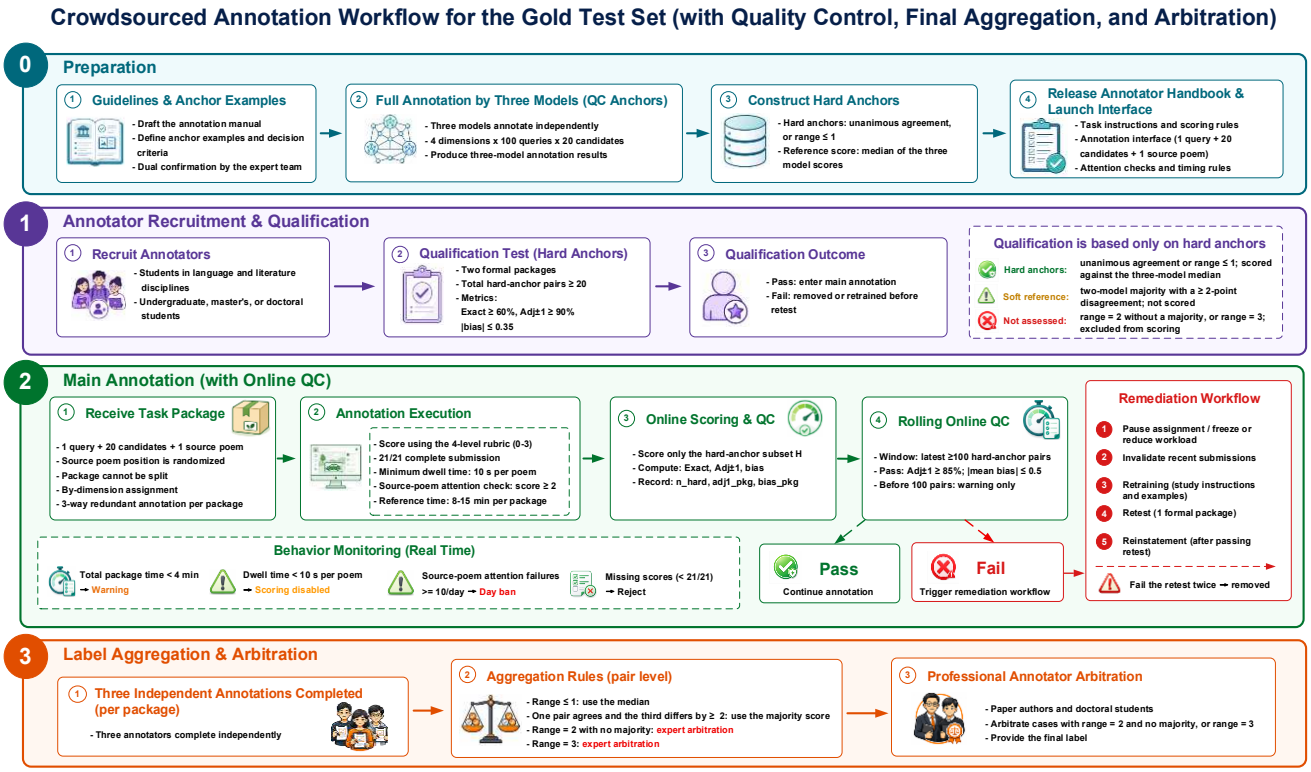


Fig. 3 | Human annotation and quality control for the gold test set. 0 Preparation of the annotation rubric and model-consensus anchors; **1** annotator recruitment and qualification; **2** independent annotation with rolling quality control; **3** label aggregation and expert adjudication. Each package contains one query, 20 candidate poems and one source poem and is scored independently by three annotators.

Packages are assigned by dimension, with three trained annotators independently scoring all 21 poems. Source-poem positions are randomized. Two experts confirm anchor examples and grade boundaries in the shared annotation handbook. Valid submissions come from 46 students in linguistics, literature and related programs: 20 undergraduates, 16 master's students and 10 doctoral students. Main annotation ran from 18 July to 1 August 2026; additional labeling and adjudication ended on 3 August. The median package time was 11.75 minutes, with 278.4 recorded person-hours excluding training, qualification tests and expert adjudication. Fig.4 shows the annotation interface.



Fig. 4 | Crowdsourcing interface for query–poem relevance annotation. The original Chinese interface displays the query and its semantic dimension, candidate poem texts and the four-level relevance scale. Annotators score each poem independently and submit a complete query package.

Poetry relevance allows interpretive variation. Testing annotators directly on highly disputed pairs risks treating reasonable disagreement as poor quality. We therefore use pairs with relatively consistent model scores as hard anchors for qualification and online monitoring, to identify clear departures from the scoring rules.

Before human annotation, Doubao-Seed-2.1-pro, DeepSeek-V3.2 and Qwen3.7-plus independently assign preliminary relevance labels to the 8,000 query–candidate pairs. Let $\mathcal{A}$ be the pair set, and let $s_g(q,d) \in \{0,1,2,3\}$ denote model g's score. The hard-anchor set and its reference scores are defined as:

$$\begin{aligned} \mathcal{H} &= \{(q,d) \in \mathcal{A} : \max_g s_g(q,d) - \min_g s_g(q,d) \le 1\}, \\ r(q,d) &= \text{median}_{g\in\{1,2,3\}} s_g(q,d), \qquad (q,d) \in \mathcal{H} \end{aligned} \tag{5}$$

Reference scores are hidden from annotators and used only for quality control, not as final human labels. Pairs with substantial model disagreement are excluded from qualification scoring.

Qualification requires two annotation packages containing at least 20 hard-anchor pairs. Relative to their reference scores, exact agreement must be at least 60%, agreement within one grade at least 90%, and the absolute mean signed bias no more than 0.35 grades. Only qualified annotators proceed to formal annotation (Fig.3, stage 1).

During annotation, a rolling window with at least 100 anchors requires within-one-grade agreement of at least 85% and absolute mean signed bias no more than 0.5 grades. Smaller windows generate warnings only. The system checks completeness and viewing time; source-poem scores below 2 generate attention flags. Failed quality control pauses assignments; affected submissions are invalidated and reassigned, with retraining and requalification before resumption.

When the three scores span at most one grade, their median becomes the final label. Remaining pairs meeting the majority rule take the majority score. Larger disagreements are reviewed by authors and doctoral researchers against the query, full poem and annotation rules; adjudication reasons are recorded. Raw scores, aggregation routes and anonymous annotator identifiers are retained.

The three ratings for each of the 8,400 query–poem pairs yield 25,200 pairwise annotator comparisons. Exact agreement is 57.9%, agreement within one grade is 96.7%, and quadratically weighted kappa is 0.663. Most differences occur at adjacent grades. Median aggregation resolves 7,778 pairs (92.60%), majority voting 181 (2.15%), and adjudication 441 (5.25%).

Dataset splitting and expanded judgments

After automatic annotation, queries are stratified by semantic cluster and divided approximately 8:2 into training and test sides within each cluster. Gold queries are removed from both sides. Training, silver and gold queries are disjoint; all candidates and labels move with their query. Gold labels are the final human judgments.

The pool-completion subset contains 100 queries per dimension, 400 in total, sampled from the silver set by semantic cluster. For each sampled query, the top-10 poems retrieved by the ten models from the complete dimension-specific corpus are merged and deduplicated by poem identifier. The original 20 candidates and source poem are excluded, and all remaining poems receive additional DeepSeek four-level relevance judgments for that same query. The 20,889 additional query–poem judgments are merged with the original silver labels. Pool completion expands judgment coverage for selected silver queries rather than creating an independent query split (Table 5).

Table 5 | Label sources and uses of dataset subsets

Subset	Query allocation	Relevance labels	Use
Training	Training side within clusters, excluding gold queries	Automatic grades, mapped earlier positives and default source labels	Domain fine-tuning
Silver	Test side within clusters, excluding gold queries	Same labeling rules as training	Auxiliary evaluation and model selection
Pool completion	100 silver queries per dimension, 400 in total	Silver labels merged with additional automatic judgments	Auxiliary evaluation with expanded judgments
Gold	100 per dimension, 400 in total	Three independent ratings, aggregation and adjudication	Primary evaluation

GLO formulation

A loss function guides training by penalizing similarity scores that disagree with the relevance labels. GLO combines two such signals. The listwise term allocates greater target weight, or gain, to more relevant poems within a query's candidate list. The ordinal term penalizes score gaps that are smaller than the prescribed margins between higher and lower relevance levels. Together, they train the model to order candidates and distinguish the strength of their relevance.

For query q in semantic dimension v, let $\mathcal{D}_q = \{d_i\}_{i=1}^{n_q}$ be its labeled candidate-poem set. Grades $y_i \in \{0,1,2,3\}$ indicate irrelevant, weakly relevant, strongly relevant and highly relevant poems, respectively. The index sets $P_q = \{i: y_i > 0\}$ and $H_q = \{i: y_i = 0\}$ identify relevant candidates and explicit hard negatives. The encoder produces L2-normalized query and poem vectors $\mathbf{e}_q$ and $\mathbf{e}_i$. The model uses cosine similarity, so vectors pointing in similar directions receive higher matching scores:

$$s_i = \mathbf{e}_q^\top \mathbf{e}_i, \qquad \| \mathbf{e}_q \|_2 = \| \mathbf{e}_i \|_2 = 1 \tag{6}$$

Each training unit consists of a query and its complete labeled candidate list, deduplicated by poem identifier. The same poem can have different relevance grades for different queries or semantic dimensions. Let B_q index candidates supplied by other queries in the batch after excluding poems already in $\mathcal{D}_q$. These candidates have no relevance judgment for the current query q. A batch is a group of queries processed together during training. Candidates associated with the other queries broaden the alternatives against which the current query's labeled poems compete.

Gain-aware listwise supervision

The exponential gain function $g(y) = 2^y - 1$ maps grades 0–3 to gains 0, 1, 3 and 7. For relevant candidates $i \in P_q$, normalized gains define the target distribution:

$$t_i = \frac{g(y_i)}{\sum_{j \in P_q} g\left(y_j\right)}, \qquad i \in P_q \tag{7}$$

Let $\tau_l > 0$ be the listwise temperature and $0 \le \rho \le 1$ the competition weight for unjudged batch candidates. The relevant-candidate probability and query-level listwise loss are defined as:

$$\begin{aligned} p_i &= \frac{\exp(s_i/\tau_l)}{\sum_{j \in P_q \cup H_q} \exp\left(s_j/\tau_l\right) + \rho \sum_{k \in B_q} \exp\left(s_k/\tau_l\right)}, \quad i \in P_q, \\ \mathcal{L}_{\text{list}}(q) &= -\sum_{i \in P_q} t_i \log p_i \end{aligned} \tag{8}$$

Higher relevance grades receive greater target mass, while grade-0 candidates participate at full weight in the denominator. The parameter ρ adjusts the contribution of unjudged batch candidates to reduce interference from

potential false negatives. When $\rho = 0$, only the current query's labeled candidates are used. The target distribution uses the same gain function as graded DCG, without a rank-position discount. The temperature controls how concentrated the predicted distribution is: a lower value places more weight on the highest-scoring candidates. The batch-competition weight controls how strongly candidates from other queries enter the denominator.

Ordinal constraints with nonuniform margins

Nonuniform grade positions define target score margins:

$$\begin{aligned} u_0 &= 0, \quad u_1 = 0.20, \quad u_2 = 0.75, \quad u_3 = 1.00, \\ m_{r,s} &= \alpha(u_r - u_s), \qquad r > s, \quad \alpha > 0 \end{aligned} \tag{9}$$

The scale α controls the target margins. The grade-2 versus grade-1 distinction receives the largest adjacent margin; nonadjacent margins follow the corresponding position differences. At $\alpha = 0.16$, the adjacent margins are $m_{1,0} = 0.032$, $m_{2,1} = 0.088$ and $m_{3,2} = 0.040$.

For a pair of labeled candidates with $y_a > y_b$, define the margin violation as $x_{ab} = m_{y_a,y_b} - (s_a - s_b)$. A temperature-smoothed Softplus penalty is then applied:

$$\ell_{ab} = \tau_o \log\left(1 + \exp\left(\frac{x_{ab}}{\tau_o}\right)\right), \qquad \tau_o > 0 \tag{10}$$

When $x_{ab} > 0$, the predicted score difference falls short of its target margin. As $\tau_o \to 0^+$, the penalty tends to $\max(0, x_{ab})$. The temperature τ_o controls smoothness near the margin boundary.

To reduce the influence of unequal grade frequencies, pair penalties are averaged first within each relevance-grade combination and then equally across the nonempty combinations. For each query, define:

$$\begin{aligned} \mathcal{O}_{r,s}(q) &= \{(a,b): a,b \in P_q \cup H_q,\ y_a = r,\ y_b = s\}, \\ \mathcal{R}_q &= \{(r,s): r > s,\ |\mathcal{O}_{r,s}(q)| > 0\} \end{aligned} \tag{11}$$

The grade-combination and query-level ordinal losses are:

$$\begin{aligned} \mathcal{L}_{r,s}(q) &= \frac{1}{|\mathcal{O}_{r,s}(q)|} \sum_{(a,b)\in\mathcal{O}_{r,s}(q)} \ell_{ab}, \quad (r,s) \in \mathcal{R}_q, \\ \mathcal{L}_{\text{ord}}(q) &= \frac{1}{|\mathcal{R}_q|} \sum_{(r,s)\in\mathcal{R}_q} \mathcal{L}_{r,s}(q) \end{aligned} \tag{12}$$

Every observed grade combination receives equal query-level weight, avoiding domination by combinations with more pairs. Ordinal supervision uses only judged candidates; it does not assign grade relations to unjudged batch candidates.

Joint training objective

Let $\lambda_{\text{ord}} \geq 0$ control the ordinal component. The query-level and batch-level objectives are:

$$\begin{aligned} \mathcal{L}_{\text{GLO}}(q) &= \mathcal{L}_{\text{list}}(q) + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}}(q), \\ \mathcal{L}_{\text{batch}} &= \frac{1}{|Q_B|} \sum_{q\in Q_B} \mathcal{L}_{\text{GLO}}(q) \end{aligned} \tag{13}$$

Here, Q_B is the set of valid queries in the batch. Equation (7) requires $P_q \neq \varnothing$, so queries without relevant candidates are excluded from the batch loss. If $\mathcal{R}_q = \varnothing$, the query's ordinal loss is set to zero. The listwise

component allocates supervision using graded gains, and the ordinal component penalizes insufficient score differences between grades. At inference, poems are ranked using the cosine similarity in Equation (6).

Comparator objectives

The experiments first evaluate the ten models used for candidate pooling without domain fine-tuning. They then compare training objectives using the same Qwen3-Embedding-0.6B starting model and the same underlying graded training resource. Each objective organizes samples according to its training requirements, as described below. This comparison examines how the objectives use relevance supervision to improve the same model architecture. Separate experiments compare the graded training pipeline with binary hard-negative training after one epoch and assess LoRA adaptation on lower-memory hardware.

RankNet[43] uses pairwise logistic ranking to score a higher-grade poem above a lower-grade poem for the same query. Pair losses are averaged within grade combinations and then equally across nonempty combinations.

$\mathbf{XE_{NDCG}}$ [59] uses listwise cross-entropy between a relevance target proportional to $2^{y_i} - \gamma_i$ and a Softmax distribution of temperature-scaled cosine scores. Independent values $\gamma_i \sim U[0,1]$ are drawn for each candidate on every forward pass. Each query's complete labeled candidate pool forms the training list.

LambdaLoss@K[60] adjusts pair weights using graded gain differences, predicted ranks and the cutoff, retaining supervision involving candidates outside the top K. We use $K = 10$ and cross-grade pairs within each pool. Deduplicated candidates from other queries that have no judgment for the current query are treated as grade-0 batch negatives.

SyCL-WS2[34] uses batch-level distribution matching based on the squared 2-Wasserstein distance. Only queries with candidates at all four relevance grades are used. Six training groups are sampled per query, each containing four poems ordered by grades 3, 2, 1 and 0. Each batch contains queries from the same semantic dimension. Softmax distributions are computed from temperature-scaled similarities and relevance labels, and their squared 2-Wasserstein distance defines the loss.

Score-gap-weighted CoSENT (SGW-CoSENT)[49] weights ranking terms by relevance-grade differences. Comparisons are restricted to one query's full judged pool to avoid direct comparisons of similarity scales across queries. Within-query cross-grade errors form a query loss, then losses are averaged over queries.

Training and evaluation settings

Main experiments use one NVIDIA RTX PRO 6000 GPU with 96 GB GPU memory and 110 GB system memory. All fine-tuning starts from Qwen3-Embedding-0.6B and the 22,828 training queries. The stored resource retains full grade-0–3 pools without truncation or additional random negatives from the corpus. Sample use follows the objective-specific rules above. Full model parameters are updated. Queries use dimension-specific instructions; poem inputs concatenate dynasty, author, title and text. The exact query instructions and input formats are provided in Supplementary Note 1 and Supplementary Table S1. The query instruction specifies the requested semantic perspective, so the same poem can be ranked differently for emotion, intent, theme or thought. The model encodes the query and poem independently; it does not require an LLM to generate relevance judgments at retrieval time.

Main runs last three epochs, with 952 optimization steps per epoch and 2,856 steps in total. Query batch size is 24, gradient accumulation 1, maximum sequence length 512 and encoding micro-batch size 32. AdamW uses learning rate 2×10^{-5}, linear scheduling with 0.1 warmup proportion, weight decay 0.01, gradient clipping 1.0 and seed 42. Loss-specific settings are listed in Table 6.

Table 6 | Objective-specific training parameters

Method	Objective-specific settings	Batch negatives
RankNet[43]	Pairwise temperature $\tau_r = 0.05$	No

Method	Objective-specific settings	Batch negatives
XE_{NDCG}[59]	Score temperature $\tau = 0.06$	No
SGW-CoSENT[49]	Ranking coefficient $\gamma_c = 20$; numerical stabilizer $\varepsilon = 10^{-6}$	No
LambdaLoss@10[60]	Cutoff $K = 10$; pairwise score coefficient $\sigma = 16.67$	Yes
SyCL-WS2[34]	Prediction Softmax temperature $\tau = 0.05$; batch size 144	Yes
GLO	Listwise temperature $\tau_l = 0.06$; batch-candidate weight $\rho = 0.50$; ordinal weight $\lambda_{\text{ord}} = 0.25$; margin scale $\alpha = 0.16$; ordinal temperature $\tau_o = 0.05$	Yes

After each epoch, the same evaluation pipeline retrieves poems from each fixed dimension-specific corpus and reports nDCG@10 at minimum relevance grades 1, 2 and 3. Main checkpoints are selected by silver threshold-3 macro-averaged nDCG@10, and the same checkpoint is evaluated on all three resources. The pool-completion subset shares its queries with the silver set.

Corpus-wide retrieval searches every poem in the fixed corpus for the query's dimension, rather than only its annotated candidate pool. The primary metric, nDCG@10, measures how well the first ten retrieved poems are ordered relative to their relevance judgments; it compares the retrieved ordering with an ideal ordering based on the available judgments, with higher scores indicating that relevant poems appear closer to the top. Thresholds 1, 2 and 3 require a relevance grade of at least 1, 2 or 3, respectively. Threshold 1 includes weak matches, threshold 2 retains strongly and highly relevant poems, and threshold 3 retains only highly relevant poems. Scores are computed for each dimension and averaged equally across the four dimensions. Full-list candidate-pool nDCG instead evaluates the ordering of all judged poems in a query's pool, without the top-ten cutoff.

Results

Resource scale and label quality

The resource contains 29,080 distinct queries: 22,828 for training, 5,852 for silver evaluation and 400 for human gold evaluation (Table 7). The 400 pool-completion queries remain part of the silver set and are not counted again.

Table 7 | Query counts and splits by dimension

Dimension	Training	Silver	Gold	Total	Pool completion
Emotion	5,881	1,510	100	7,491	100
Intent	5,578	1,442	100	7,120	100
Theme	6,239	1,575	100	7,914	100
Thought	5,130	1,325	100	6,555	100
Total	22,828	5,852	400	29,080	400

Pool-completion queries are included in silver and do not add independent queries.

Each dimension contains 100 gold queries and 2,100 final judgments (Table 8). Grades 0–3 comprise 1,359, 3,365, 2,759 and 917 pairs, respectively, or 16.18%, 40.06%, 32.85% and 10.92%. Grades 1 and 2 dominate, retaining partial and stronger correspondences rather than collapsing them into one positive class. Grades 1–3 total 7,041 pairs; grade-0 pairs remain available as judged hard negatives. Fig.3 and Fig.4 document the human annotation workflow and interface.

Table 8 | Gold judgments and aggregation routes

Dimension	Queries	Judged pairs	Grades 1–3	Grade 0	Median	Majority	Adjudication
Intent	100	2,100	1,653	447	1,944	37	119
Emotion	100	2,100	1,855	245	2,003	35	62
Theme	100	2,100	1,788	312	1,916	52	132
Thought	100	2,100	1,745	355	1,915	57	128
Total	400	8,400	7,041	1,359	7,778	181	441

Final human labels provide a reference for estimating the agreement and bias of automatic relevance judgments. Candidate-poem scores and the default source-poem grade are assessed separately.

The 8,000 mined candidate pairs from 400 gold queries are aligned with DeepSeek-V3.2 prelabels, excluding source poems. Exact agreement is 50.4%, agreement within one grade 94.2%, quadratically weighted kappa 0.610, and mean absolute error 0.558 grades (Table 9). The mean automatic-minus-human difference is +0.168 grades, indicating mild relevance overestimation. Exact agreement requires identical grades; agreement within one grade allows a one-level difference. Weighted kappa discounts chance agreement and penalizes larger grading differences more strongly, while mean absolute error reports the average size of the grade difference.

Table 9 | Agreement between DeepSeek and final human labels

Measure	Estimate	95% CI
Exact agreement	50.4%	49.3%–51.5%
Within-one-grade agreement	94.2%	93.7%–94.7%
Quadratically weighted kappa	0.610	0.596–0.623
Mean absolute error (grades)	0.558	0.545–0.572
Mean signed bias (grades)	+0.168	+0.151–+0.185

With human judgments as reference, precision, recall and F1 are 0.920, 0.882 and 0.900 at relevance threshold 1, and 0.680, 0.828 and 0.747 at threshold 2. Automatic grade-3 labels account for 18.8% of candidate pairs, compared with 7.6% under human assessment. Thus, broad relevance is recognized more reliably than stronger relevance. At threshold 1, DeepSeek-V3.2 has the highest F1 among the compared models; Doubao-Seed-2.1-pro and Qwen3.7-plus score 0.803 and 0.874.

Training and silver labels assign source poems a default grade of 3. Human grades 0–3 for the 400 source pairs number 5, 21, 67 and 307, with a mean of 2.69. Default assignment achieves 76.8% exact and 93.5% within-one-grade agreement, with mean absolute error 0.31 grades. Analysis errors and changes in query scope during rewriting can weaken correspondence to the source. Gold evaluation therefore uses the actual human source-poem grade.

Automatic labels provide scalable supervision but overestimate some high-relevance relations and source matches. Human gold evaluation supports the principal conclusions; silver and pool-completion resources provide auxiliary comparisons. Label provenance and human assessments are retained for verification. The following experiments therefore report all three evaluation resources, with the human gold set providing the primary reference for relevance.

Dimension differences in general models

The ten general-purpose embedding models differ substantially (Table 10). Qwen3-Embedding-4B leads at all thresholds on gold and silver, whereas Qwen3-Embedding-8B leads on pool completion. The 12B KaLM model does not surpass them, showing that parameter count alone does not determine domain retrieval performance. All ten models score lower as the minimum relevance threshold rises from 1 to 3, indicating a remaining challenge in retrieving highly relevant poems. Lightweight models generally perform less well.

Table 10 | Macro-averaged nDCG@10 of ten general-purpose models

Model	Gold			Silver			Pool completion		
	1	2	3	1	2	3	1	2	3
EmbeddingGemma-300M	0.290	0.255	0.188	0.261	0.239	0.194	0.366	0.323	0.224
KaLM-embedding-multilingual-mini-instruct-v2.5	0.089	0.067	0.035	0.084	0.070	0.047	0.178	0.139	0.077
multilingual-e5-large-instruct	0.201	0.160	0.112	0.187	0.166	0.125	0.325	0.276	0.173
Qwen3-Embedding-0.6B	0.279	0.247	0.174	0.257	0.236	0.189	0.359	0.320	0.218
jina-embeddings-v5-text-small	0.359	0.323	0.243	0.319	0.296	0.239	0.413	0.377	0.259
Qwen3-Embedding-4B	**0.457**	**0.427**	**0.313**	**0.431**	**0.413**	**0.349**	0.510	0.485	0.364
e5-mistral-7b-instruct	0.266	0.213	0.119	0.243	0.216	0.151	0.322	0.284	0.174
Linq-Embed-Mistral	0.278	0.235	0.149	0.258	0.236	0.178	0.356	0.322	0.213
Qwen3-Embedding-8B	0.438	0.413	0.293	0.409	0.392	0.332	**0.534**	**0.511**	**0.391**
KaLM-Embedding-Gemma3-12B-2511	0.400	0.378	0.294	0.367	0.350	0.296	0.424	0.402	0.312

1, 2 and 3 are minimum relevance-grade thresholds. Values are nDCG@10 macro-averaged across four dimensions. Bold denotes the largest reported value per column.

Fig.5 shows dimension differences at relevance threshold 2. Most models score higher on authorial intent in the gold and silver sets and on theme in the pool-completion set; emotion and thought are relatively weaker. These results identify emotion and thought as priorities for domain adaptation. Further work should strengthen the modeling of emotional and intellectual expression through imagery, metaphor and cultural context, and expand fine-grained judgments and difficult examples for these dimensions.

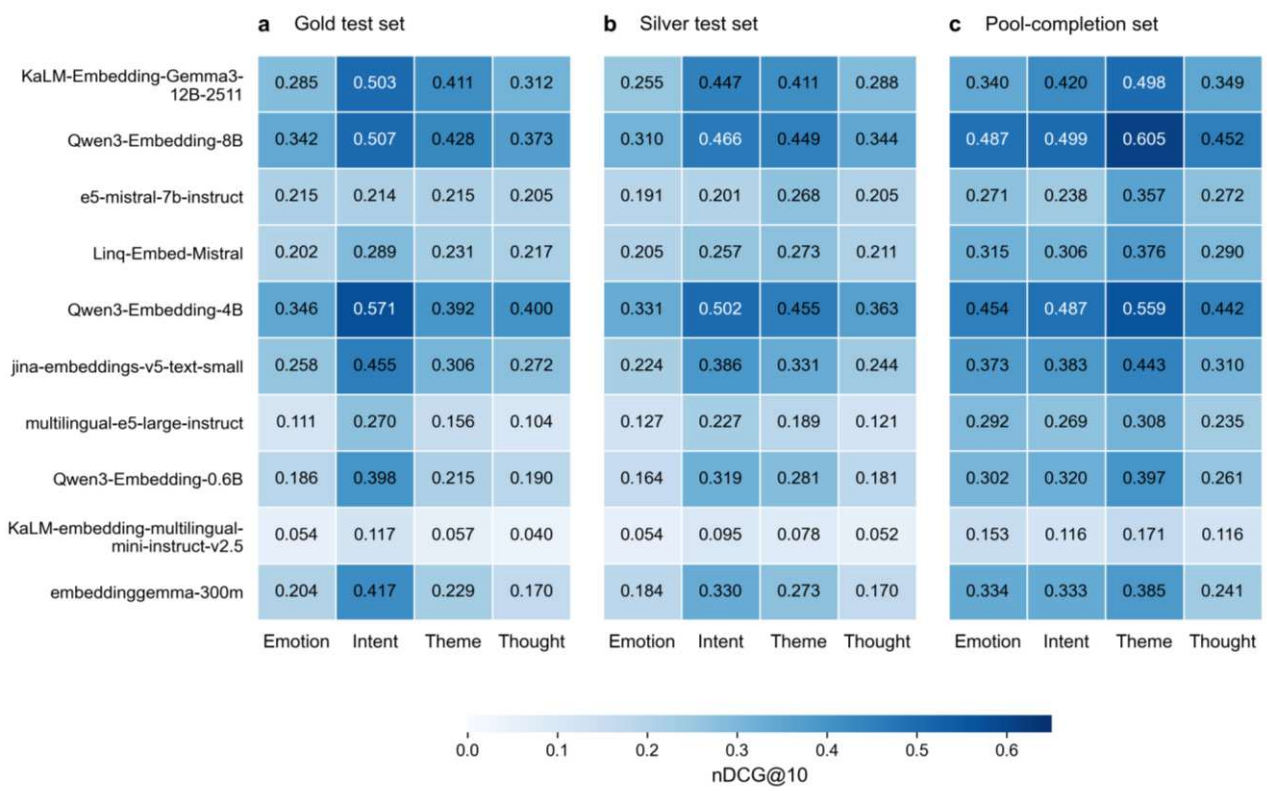


Fig. 5 | Retrieval performance of general-purpose embedding models across four semantic dimensions. a Gold test set; **b** silver test set; **c** pool-completion set. Results use a minimum relevance grade of 2. Cells report dimension-level nDCG@10 rounded to three decimal places; darker blue indicates higher performance.

Retrieval under different training objectives

Default GLO improves over the unadapted base model in all nine conditions (Table 11). At threshold 3, absolute gains are 0.217, 0.186 and 0.143 on gold, silver and pool completion. The adapted 0.6B model also exceeds all ten unadapted models in Table 10 on gold and silver at threshold 3.

Default GLO achieves the highest scores in all six conditions at thresholds 1 and 2, and outperforms RankNet, SGW-CoSENT, XE_{NDCG} and SyCL-WS2 in all nine conditions. LambdaLoss@10 scores slightly higher at threshold 3 on each test set. Parameter ablations examine the effects of adjusting the ordinal and batch-competition weights.

Table 11 | Macro-averaged nDCG@10 across training objectives

Method	Selected Epoch	Gold			Silver			Pool completion		
		1	2	3	1	2	3	1	2	3
Qwen3-Embedding-0.6B	—	0.279	0.247	0.174	0.257	0.236	0.189	0.359	0.320	0.218
RankNet	2	0.311	0.325	0.325	0.298	0.299	0.310	0.334	0.329	0.304
SGW-CoSENT	2	0.298	0.312	0.319	0.284	0.286	0.296	0.317	0.313	0.291
XE_{NDCG}	2	0.255	0.271	0.290	0.247	0.250	0.267	0.284	0.281	0.269
SyCL-WS2	3	0.351	0.359	0.331	0.346	0.344	0.338	0.384	0.377	0.331
LambdaLoss@10	2	0.388	0.403	**0.395**	0.378	0.379	**0.385**	0.404	0.398	**0.363**
GLO	3	**0.393**	**0.405**	0.391	**0.381**	**0.380**	0.375	**0.411**	**0.404**	0.361

1, 2 and 3 are minimum relevance-grade thresholds. Values are nDCG@10 macro-averaged across four dimensions. Bold denotes the largest reported value per column.

Main checkpoints are selected by silver threshold-3 macro-averaged nDCG@10 and evaluated on all three resources.

At threshold 2, GLO leads on thought across all three resources and on theme in silver and pool completion (Fig.6). LambdaLoss@10 performs better on intent and on emotion in gold and silver. The thought results show a consistent advantage for GLO across evaluation resources.

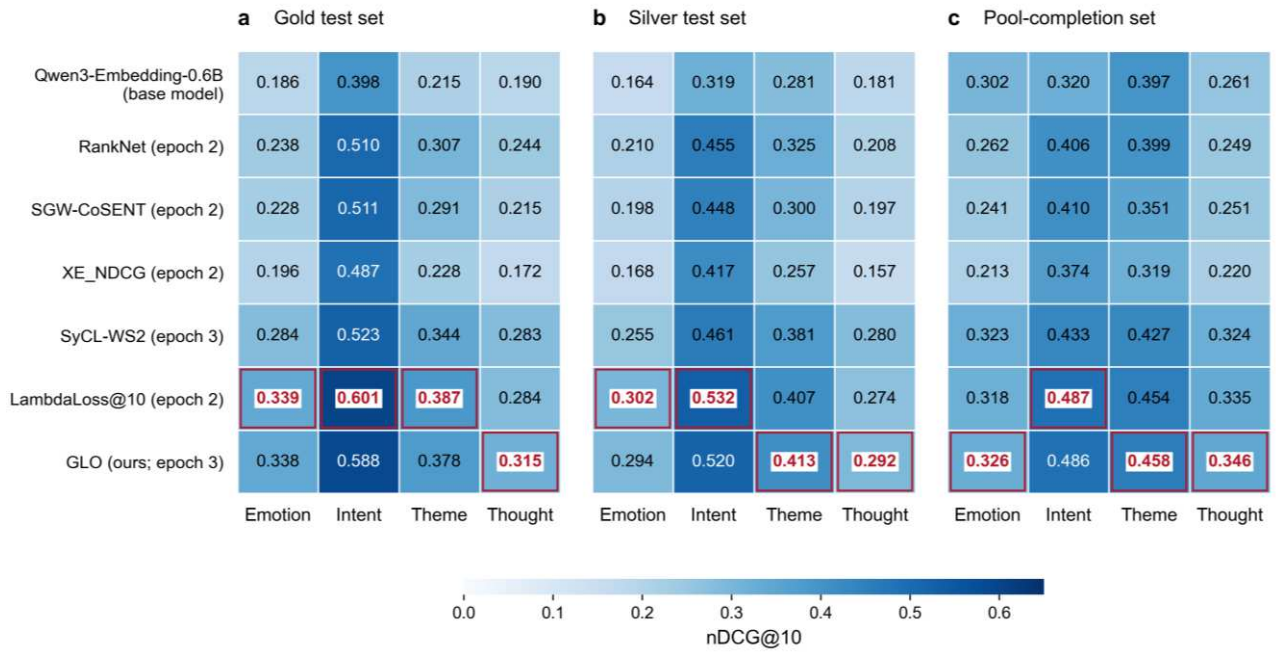


Fig. 6 | Retrieval performance of different training objectives across four semantic dimensions. a Gold test set; **b** silver test set; **c** pool-completion set. Results use a minimum relevance grade of 2. Checkpoints are selected by macro-averaged nDCG@10 at relevance threshold 3 on the silver test set. Red values and outlines identify the highest score for each dimension within each test set.

Ranking within annotated pools

All training objectives improve full-list nDCG within the annotated candidate pools (Table 12). XE_{NDCG} achieves the highest values, 0.899, 0.899 and 0.904 on the gold, silver and pool-completion sets, respectively, although its corpus-wide retrieval performance is lower than that of GLO, LambdaLoss@10 and SyCL-WS2. Better ordering within a judged pool therefore does not necessarily translate into better corpus retrieval. GLO scores 0.892, 0.885 and 0.894, exceeding the base model by 0.144, 0.137 and 0.122. These results show that its corpus-wide retrieval gains coexist with competitive candidate-pool ranking.

Table 12 | Full-list nDCG within annotated candidate pools

Method	Epoch	Gold	Silver	Pool completion
Qwen3-Embedding-0.6B	—	0.748	0.748	0.772
GLO	2	0.892	0.885	0.894
SGW-CoSENT	2	0.891	0.896	0.901
RankNet	3	0.890	0.896	0.902
SyCL-WS2	3	0.865	0.869	0.877
LambdaLoss@10	2	0.897	0.892	0.895
XE_{NDCG}	2	**0.899**	**0.899**	**0.904**

Epochs refer to the candidate-pool ranking evaluation and differ from those used in the main corpus-retrieval comparison.

Composite thematic retrieval case study

To examine retrieval of a composite theme, we analyze query 815996, which combines the grandeur of the Yangtze, reflection on dynastic change and farewell to a friend (Fig.7). The comparison traces how the leading results address these components and how related works move in the full-corpus ranking.

Query 815996

通过对长江雄浑气势和历史沧桑的描写，展现英雄气概与王朝兴替，最终回归到对友人的深情送别。

Depictions of the Yangtze's grandeur and historical vicissitudes evoke heroic spirit and dynastic change, ultimately leading to a heartfelt farewell to a friend.

a Base model

	Retrieved poem	River imagery	Dynastic change	Farewell
#1	Matching Tang Youzhi's Rhyme on His Departure for Hejin	扶藜重上渡江船 Leaning on my staff, I board the river boat once more.	Personal ageing; no dynastic reflection.	臨歧好寄兒曹話 At parting, send my words to the younger generation.
#2	Reflecting on the Past at Jingzhou	屈原祠庙枕江波 Qu Yuan's shrine rests beside the river's waves.	怀襄旧恨余三户 Old grief for Huai and Xiang lingers in Chu's last homes.	Personal lament; no farewell scene.

b Fine-tuned model (GLO)

	Retrieved poem	River imagery	Dynastic change	Farewell
#1	On the Great River: Seeing Prefect Zhou Off	大江茫茫割吳楚 The vast river divides the lands of Wu and Chu.	六朝霸業何瑣瑣 How paltry the imperial glory of the Six Dynasties!	我將江水送君去 With the river's flow, I bid you farewell.
#2	A Yangtze Journey: Seeing Du Zishi of the Revenue Ministry Off	浦口千重浪頭惡 At Pukou, countless fierce waves surge.	吴宫晉殿江聲愁 Wu's palaces, Jin's halls: the river sounds with sorrow.	别後相思雙鯉魚 After parting, longing speaks through twin-carp letters.

c Full-corpus rank improvements

Base model → GLO

On the Great River: Seeing Prefect Zhou Off	A Yangtze Journey: Seeing Du Zishi of the Revenue Ministry Off	Seeing Prefect Yang Daxun Off to Jiangnan
23 → 1	92 → 2	476 → 8

a,b: Top two results per model. c: Selected changes in full-corpus rank.

Fig. 7 | Retrieval of a composite thematic query linking river imagery, dynastic change and farewell. a The two highest-ranked poems from the base model; **b** the two highest-ranked poems from the GLO-fine-tuned model. Quotations illustrate correspondence with the query's thematic components. **c** Full-corpus rank changes for the source poem and two additional related poems, shown as base model → GLO. Only the query and quoted verses are presented in both Chinese and English.

The base model's leading results match only parts of the query (Fig.7a). *Matching Tang Youzhi's Rhyme on His Departure for Hejin* emphasizes river crossing and leave-taking, while *Reflecting on the Past at Jingzhou* emphasizes historical remembrance. Neither result brings together the query's river imagery, dynastic reflection and farewell. In contrast, the first result from the GLO-fine-tuned model, *On the Great River: Seeing Prefect Zhou Off*, follows the requested thematic progression. "The vast river divides the lands of Wu and Chu" establishes the river's grandeur; "How paltry the imperial glory of the Six Dynasties!" introduces reflection on historical change; and "With the river's flow, I bid you farewell" returns to leave-taking (Fig.7b).

The improvement also extends to poems other than the query's source. *A Yangtze Journey: Seeing Du Zishi of the Revenue Ministry Off* and *Seeing Prefect Yang Daxun Off to Jiangnan* combine river scenes, historical reflection and parting sentiment; their ranks rise from 92 to 2 and from 476 to 8, respectively (Fig.7c). Together with the source poem's movement from rank 23 to 1, these changes extend the improvement from source-poem recovery to the discovery of other works with a similar organization of themes, illustrating the model's handling of this composite semantic query.

Objective parameter ablations

Table 13 examines the ordinal loss weight λ_{ord}. Among the tested values, $\lambda_{\text{ord}} = 1$ achieves the highest score in all nine conditions. Relative to removing the ordinal term, gold-set gains are 0.008, 0.009 and 0.009 at thresholds 1, 2 and 3, with gains on the silver and pool-completion sets as well. The default value of 0.25 performs similarly to 0, indicating a limited additional benefit from the default ordinal weight. Increasing the weight to 10 and then 100 lowers performance relative to 1. Within the tested range, a weight of 1 provides the best balance between listwise supervision and inter-grade margin constraints.

Table 13 | Ablation of ordinal loss weight

λ_{ord}	Gold			Silver			Pool completion		
	1	2	3	1	2	3	1	2	3
0	0.394	0.406	0.390	0.382	0.380	0.375	0.413	0.406	0.364
0.25	0.393	0.405	0.391	0.381	0.380	0.375	0.411	0.404	0.361
1	**0.402**	**0.415**	**0.399**	**0.383**	**0.381**	**0.376**	**0.416**	**0.409**	**0.367**
10	0.389	0.401	0.389	0.373	0.373	0.370	0.412	0.405	0.366
100	0.328	0.340	0.334	0.326	0.327	0.331	0.369	0.365	0.334

1, 2 and 3 are minimum relevance-grade thresholds. Values are nDCG@10 macro-averaged across four dimensions. Bold denotes the largest reported value per column.

Table 14 examines the competition weight ρ for unjudged candidates from other queries in the batch. Raising ρ from 0 to 0.25 improves every condition; $\rho = 1$ achieves the highest score in all nine. Gold-set scores at thresholds 1 and 3 rise from 0.257 and 0.292 at $\rho = 0$ to 0.407 and 0.393 at $\rho = 1$. Cross-query candidate competition is thus an important signal for corpus-wide retrieval under the current batch construction. Full-weight competition is more effective than the default $\rho = 0.5$. Joint tuning remains to be examined for λ_{ord} and ρ.

Table 14 | Ablation of batch-candidate competition weight

ρ	Gold			Silver			Pool completion		
	1	2	3	1	2	3	1	2	3
0	0.257	0.274	0.292	0.252	0.254	0.268	0.282	0.279	0.265
0.25	0.387	0.400	0.392	0.369	0.368	0.366	0.401	0.394	0.356
0.50	0.393	0.405	0.391	0.381	0.380	0.375	0.411	0.404	0.361
1.00	**0.407**	**0.418**	**0.393**	**0.389**	**0.387**	**0.378**	**0.422**	**0.414**	**0.369**

1, 2 and 3 are minimum relevance-grade thresholds. Values are nDCG@10 macro-averaged across four dimensions. Bold denotes the largest reported value per column.

Binary and graded training pipelines

Standard InfoNCE combined with hard-negative mining is a common approach to training retrieval embeddings. We use positive-aware hard-negative mining[61] with Qwen3-Embedding-8B to obtain 20 hard negatives per training query, forming a binary relevance dataset. Standard InfoNCE[39], Masked InfoNCE and Focal InfoNCE, the latter using focal weighting[62], are each trained for one epoch. The masked variant excludes negatives whose

current cosine similarity exceeds the positive similarity by more than 0.1. Table 15 compares their first-epoch results with GLO trained for one epoch (952 optimization steps) using its default configuration (ρ=0.50 and λ_{ord}=0.25). GLO scores higher in all nine conditions. On the gold set, its macro-averaged nDCG@10 is 0.354, 0.368 and 0.366 at thresholds 1, 2 and 3, respectively.

Table 15 | Retrieval performance of binary and graded training pipelines after one epoch

Method	Gold			Silver			Pool completion		
	1	2	3	1	2	3	1	2	3
Qwen3-Embedding-0.6B	0.279	0.247	0.174	0.257	0.236	0.189	0.359	0.320	0.218
Standard InfoNCE	0.292	0.309	0.359	0.269	0.270	0.285	0.279	0.273	0.249
Masked InfoNCE	0.287	0.303	0.353	0.264	0.265	0.280	0.282	0.276	0.252
Focal InfoNCE	0.291	0.307	0.362	0.267	0.267	0.282	0.278	0.272	0.249
GLO (ρ=0.50, λ_{ord}=0.25)	**0.354**	**0.368**	**0.366**	**0.346**	**0.345**	**0.345**	**0.384**	**0.378**	**0.336**

1, 2 and 3 are minimum relevance-grade thresholds. Values are nDCG@10 macro-averaged across four dimensions. Bold denotes the largest reported value per column.

All four training objectives are evaluated after one epoch. GLO uses ρ=0.50 and λ_{ord}=0.25. Values are rounded to three decimals.

LoRA adaptation with limited GPU memory

Low-rank adaptation (LoRA) freezes the base-model weights and inserts rank-16 adapters into q_proj, k_proj, v_proj and o_proj, with scaling $\alpha = 32$ and dropout 0.05. It uses the same training resource and GLO objective on an NVIDIA GeForce RTX 4070 Ti SUPER with 16 GB GPU memory and 32 GB system memory. Query batch size is 8 with four accumulation steps, and training lasts four epochs. Other main optimization settings follow the full fine-tuning experiment. Table 16 reports LoRA epoch 3, selected using silver threshold-3 performance, against the same silver-selected epoch-3 full fine-tuning checkpoint used in the main comparison.

LoRA improves eight of nine conditions relative to the base, including threshold-3 retrieval on all resources (Table 16). Pool-completion threshold 1 decreases from 0.359 to 0.335. LoRA remains below the full fine-tuning baseline, with threshold-3 gaps of 0.065, 0.071 and 0.060. Under these configurations, LoRA enables domain adaptation on a GPU with 16 GB of memory, while full fine-tuning achieves higher retrieval scores. A separate one-epoch LoRA experiment examines the contributions of GLO components and parameter sensitivity (Supplementary Note 2, Supplementary Table S2 and Supplementary Figs. S1 and S2).

Table 16 | Retrieval performance of full and LoRA fine-tuning

Method	Gold			Silver			Pool completion		
	1	2	3	1	2	3	1	2	3
Qwen3-Embedding-0.6B	0.279	0.247	0.174	0.257	0.236	0.189	0.359	0.320	0.218
GLO full fine-tuning	**0.393**	**0.405**	**0.391**	**0.381**	**0.380**	**0.375**	**0.411**	**0.404**	**0.361**
GLO LoRA	0.307	0.318	0.326	0.300	0.300	0.304	0.335	0.331	0.301

1, 2 and 3 are minimum relevance-grade thresholds. Values are nDCG@10 macro-averaged across four dimensions. Bold denotes the largest reported value per column.

Both rows report the checkpoints selected at epoch 3 by silver threshold-3 macro-averaged nDCG@10. Full fine-tuning reproduces the main comparison. Batch configurations and total training durations differ between the adaptation settings.

Discussion

Graded relevance provides a way to organize literary heritage resources around specific semantic questions. Separating emotion, authorial intent, theme and thought allows the same poem to be retrieved for different

research needs. The predominance of grades 1 and 2 in human judgments shows that many candidate poems share only part of the requested meaning or correspond to it with varying strength. Preserving these distinctions supports the selection of related poems for close reading and comparison.

Default GLO records the highest observed scores in the six conditions at thresholds 1 and 2 and substantially raises the lightweight baseline scores at threshold 3. This supports graded training for discovering poems with different degrees of correspondence. LambdaLoss@10 performs better at threshold 3, whereas XE_{NDCG} ranks the annotated candidate pools more accurately. The difference between these results highlights the value of evaluating both corpus retrieval and ordering within judged candidate pools. The ablations identify batch-candidate competition as a major source of retrieval improvement, with additional gains from increasing the ordinal weight to 1.

Emotion and thought are priorities for improving domain retrieval. GLO's consistent advantage on thought across the three evaluation resources supports the use of domain-specific graded supervision. Further work should enrich training examples involving imagery, metaphor and historical context, and use close analysis of difficult queries and retrieved poems to identify the semantic distinctions that models fail to capture.

The present dataset primarily uses queries derived from LLM analyses, and human relevance judgments cover a subset of query–poem pairs. Evaluation is conducted within fixed, dimension-specific corpora. Future work will incorporate queries from literary researchers, expand human annotation, and evaluate retrieval over larger poetry collections to improve the dataset's representativeness and assessment coverage.

This study provides a four-dimensional retrieval dataset with 29,080 queries and a training objective that combines graded listwise supervision with relevance-level score margins. The results show how domain supervision improves a lightweight model's ability to retrieve related poems and distinguish degrees of semantic correspondence. The resource and method provide a basis for organizing classical Chinese poetry by research needs and selecting materials for comparative literary and cultural-heritage study.

Data availability

The dataset is available at https://huggingface.co/datasets/PoetryMTEB/ClassicalPoetryRetrieval. Source poetry resources are provided at https://cnkgraph.com/Home/OpenResources.

Code availability

The implementation used in this study is not publicly released at this stage, but may be made available from the corresponding author upon reasonable request, subject to institutional and project-related restrictions.

Acknowledgements

This work was supported by Shanghai Normal University through the Doctoral Program for Outstanding Innovative Talent Development under the High-Quality Graduate Education Development Project, for the project "Multidimensional Representation Learning for Classical Chinese Poetry through Distillation of LLM-Generated Analyses." We thank the relevance annotators.

Author contributions

Z.W. formulated the research questions, designed the methods, constructed the dataset, implemented models and experiments, and drafted and revised the manuscript. C.M. contributed to dataset construction, annotation and result analysis, and revised the manuscript. W.D. (Wei Duan) and Zongzhe Wang contributed to dataset construction, annotation and quality assessment, and revised the manuscript. A.S. and W.D. (Wei Dai) contributed to dataset construction and evaluation, and revised the manuscript. X.W. supervised the research direction, developed the poetry-semantic and relevance-annotation guidelines, and revised the manuscript. H.Q. supervised model optimization and experimental design, and reviewed and revised the manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

References

1. Peng, X. et al. Poe2CLP: phrase-level attention and cross-modal semantic alignment for poem generate chinese landscape paintings. *npj Herit. Sci.* **13,** 656 (2025). https://doi.org/10.1038/s40494-025-02238-0

2. Xu, C. & Zhou, S. Visualizing poetry with deep semantic understanding and consistency evaluation. *npj Herit. Sci.* **13,** 686 (2025). https://doi.org/10.1038/s40494-025-02263-z

3. Yan, S. et al. Heritage space prototypes in classical Chinese poetry. *npj Herit. Sci.* (2026). https://doi.org/10.1038/s40494-026-03018-0

4. Foley, J. J. IV. Poetry: Identification, Entity Recognition, and Retrieval. PhD thesis, University of Massachusetts Amherst (2019).

5. Liu, Y. et al. Large-scale corpus construction and retrieval-augmented generation for ancient Chinese poetry: new method and data insights. In *Findings of the Association for Computational Linguistics: NAACL 2025* 779–817 (2025). https://doi.org/10.18653/v1/2025.findings-naacl.46

6. Karpukhin, V. et al. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* 6769–6781 (2020). https://doi.org/10.18653/v1/2020.emnlp-main.550

7. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* **33,** 9459–9474 (2020).

8. Singh, A., D'Arcy, M., Cohan, A., Downey, D. & Feldman, S. SciRepEval: a multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 5548–5566 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.338

9. Men, Y., Jiao, X. & Wei, Q. Personalized academic paper retrieval integrating word embedding models and frequent sequence mining. *Libr. Inf. Serv.* **69,** 124–133 (2025). https://doi.org/10.13266/j.issn.0252-3116.2025.19.011 (In Chinese).

10. Xia, C. Knowledge representation of multimodal cultural memory resources in the GenAI environment: a case study of guqin abbreviated notation. *J. Libr. Sci. China* **52,** 126–154 (2026). https://doi.org/10.13530/j.cnki.jlis.2026015 (In Chinese).

11. Mori, L., Sousa de Oliveira, C., Yih, Y. & Ventresca, M. Assessing the performance gap between lexical and semantic models for information retrieval with formulaic legal language. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law* 114–128 (2025). https://doi.org/10.1145/3769126.3769205

12. Choi, C. et al. FinDER: financial dataset for question answering and evaluating retrieval-augmented generation. In *Proceedings of the 6th ACM International Conference on AI in Finance* 638–646 (2025). https://doi.org/10.1145/3768292.3770361

13. Fang, Z. et al. ADORE: autonomous domain-oriented relevance engine for E-commerce. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* 4259–4263 (2025). https://doi.org/10.1145/3726302.3731944

14. Soviero, B., Kuhn, D., Salle, A. & Moreira, V. P. ChatGPT goes shopping: LLMs can predict relevance in eCommerce search. In *Advances in Information Retrieval: ECIR 2024* 3–11 (2024). https://doi.org/10.1007/978-3-031-56066-8_1

15. Suresh, T. et al. CoRNStack: high-quality contrastive data for better code retrieval and reranking. In *Proceedings of the Thirteenth International Conference on Learning Representations* (2025).

16. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3,** 2 (2021). https://doi.org/10.1145/3458754

17. Han, P. & Liu, S. Research on medical query intent recognition based on LLM retrieval re-ranking and query optimization. *Libr. Inf. Serv.* https://doi.org/10.13266/j.issn.0252-3116.2026.21.005 (2026). (In Chinese).

18. Bajaj, P. et al. MS MARCO: a human generated MAchine Reading COmprehension dataset. Preprint at https://arxiv.org/abs/1611.09268 (2016).

19. Qiu, Y. et al. DuReader-Retrieval: a large-scale Chinese benchmark for passage retrieval from web search engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* 5326–5338 (2022). https://doi.org/10.18653/v1/2022.emnlp-main.357

20. Xie, X. et al. T2Ranking: a large-scale Chinese benchmark for passage ranking. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2681–2690 (2023). https://doi.org/10.1145/3539618.3591874

21. Long, D. et al. Multi-CPR: a multi domain Chinese dataset for passage retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* 3046–3056 (2022). https://doi.org/10.1145/3477495.3531736

22. Clark, J. H. et al. TyDi QA: a benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguist.* **8,** 454–470 (2020). https://doi.org/10.1162/tacl_a_00317

23. Zhang, X. et al. MIRACL: a multilingual retrieval dataset covering 18 diverse languages. *Trans. Assoc. Comput. Linguist.* **11,** 1114–1131 (2023). https://doi.org/10.1162/tacl_a_00595

24. Adeyemi, M. et al. CIRAL: a test collection for CLIR evaluations in African languages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* 293–302 (2024). https://doi.org/10.1145/3626772.3657884

25. Kuo, T.-L., Chiu, W.-N., Ma, W.-Y. & Cheng, P.-J. Doc2Query++: topic-coverage based document expansion and its application to dense retrieval via dual-index fusion. Preprint at https://arxiv.org/abs/2510.09557 (2025).

26. Chaudhary, A., Raman, K. & Bendersky, M. It's all relative! A synthetic query generation approach for improving zero-shot relevance prediction. In *Findings of the Association for Computational Linguistics: NAACL 2024* 1645–1664 (2024). https://doi.org/10.18653/v1/2024.findings-naacl.107

27. Campese, S., Moschitti, A. & Lauriola, I. Improving document retrieval coherence for semantically equivalent queries. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* 3425–3441 (2025). https://doi.org/10.18653/v1/2025.ijcnlp-long.182

28. Su, H. et al. BRIGHT: a realistic and challenging benchmark for reasoning-intensive retrieval. In *Proceedings of the Thirteenth International Conference on Learning Representations* (2025).

29. Zhang, X., Li, L., Zhou, X. & Liu, Z. R2MED: a benchmark for reasoning-driven medical retrieval. Preprint at https://arxiv.org/abs/2505.14558 (2025).

30. Yuksel, G., Rau, D. & Kamps, J. Remining hard negatives for generative pseudo labeled domain adaptation. Preprint at https://arxiv.org/abs/2501.14434 (2025).

31. Xi, J. et al. Mine and refine: optimizing graded relevance in E-commerce search retrieval. Preprint at https://arxiv.org/abs/2602.17654 (2026).

32. Tamber, M. S., Kazi, S., Sourabh, V. & Lin, J. Contrastive learning falls short: improving dense retrieval with cross-encoder listwise distillation and synthetic data. In *Advances in Information Retrieval: ECIR 2026* 104–121 (2026). https://doi.org/10.1007/978-3-032-21300-6_7

33. Arabzadeh, N., Vtyurina, A., Yan, X. & Clarke, C. L. A. Shallow pooling for sparse labels. *Inf. Retr. J.* **25,** 365–385 (2022). https://doi.org/10.1007/s10791-022-09411-0

34. Esfandiarpoor, R. et al. Beyond contrastive learning: synthetic data enables list-wise training with multiple levels of relevance. In *Findings of the Association for Computational Linguistics: EMNLP 2025* 22860–22882 (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.1245

35. Wullach, T., Shapira, O. & Cohen, A. D. N. The overlooked role of graded relevance thresholds in multilingual dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2026* 7738–7750 (2026). https://doi.org/10.18653/v1/2026.findings-acl.382

36. Li, H. et al. Semi-supervised synthetic data generation with fine-grained relevance control for short video search relevance modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence* **40,** 31591–31599 (2026). https://doi.org/10.1609/aaai.v40i37.40425

37. Thomas, P., Spielman, S., Craswell, N. & Mitra, B. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* 1930–1940 (2024). https://doi.org/10.1145/3626772.3657707

38. Fröbe, M. et al. Large language model relevance assessors agree with one another more than with human assessors. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2858–2863 (2025). https://doi.org/10.1145/3726302.3730218

39. van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. Preprint at https://arxiv.org/abs/1807.03748 (2018).

40. Liu, T.-Y. Learning to rank for information retrieval. *Found. Trends Inf. Retr.* **3,** 225–331 (2009). https://doi.org/10.1561/1500000016

41. Cheng, Q., Yang, X., Sun, T., Li, L. & Qiu, X. Improving contrastive learning of sentence embeddings from AI feedback. In *Findings of the Association for Computational Linguistics: ACL 2023* 11122–11138 (2023). https://doi.org/10.18653/v1/2023.findings-acl.707

42. Hofstätter, S., Althammer, S., Schröder, M., Sertkan, M. & Hanbury, A. Improving efficient neural ranking models with cross-architecture knowledge distillation. Preprint at https://arxiv.org/abs/2010.02666 (2020).

43. Burges, C. et al. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning* 89–96 (2005). https://doi.org/10.1145/1102351.1102363

44. Huang, C.-W. & Chen, Y.-N. PairDistill: pairwise relevance distillation for dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* 18225–18237 (2024). https://doi.org/10.18653/v1/2024.emnlp-main.1013

45. Wang, X., Li, C., Golbandi, N., Bendersky, M. & Najork, M. The LambdaLoss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* 1313–1322 (2018). https://doi.org/10.1145/3269206.3271784

46. Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F. & Li, H. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning* 129–136 (2007). https://doi.org/10.1145/1273496.1273513

47. Ren, R. et al. RocketQAv2: a joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 2825–2835 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.224

48. Tsirigotis, C., Adlakha, V., Monteiro, J., Courville, A. & Taslakian, P. BiXSE: improving dense retrieval via probabilistic graded relevance distillation. Preprint at https://arxiv.org/abs/2508.06781 (2025).

49. Zhou, J. et al. WeMM-Embedding: WeChat multi-modal embedding technical report. Preprint at https://arxiv.org/abs/2608.24053 (2026).

50. Babakhin, Y. et al. Llama-Embed-Nemotron-8B: a universal text embedding model for multilingual and cross-lingual tasks. Preprint at https://arxiv.org/abs/2511.07025 (2025).

51. Zhang, Y. et al. Qwen3 embedding: advancing text embedding and reranking through foundation models. Preprint at https://arxiv.org/abs/2506.05176 (2025).

52. Vera, H. S. et al. EmbeddingGemma: powerful and lightweight text representations. Preprint at https://arxiv.org/abs/2509.20354 (2025).

53. Zhao, X. et al. KaLM-Embedding-V2: superior training techniques and data inspire a versatile embedding model. In *Proceedings of the Fourteenth International Conference on Learning Representations* (2026).

54. Hu, X. et al. KaLM-Embedding: superior training data brings a stronger embedding model. Preprint at https://arxiv.org/abs/2501.01028 (2025).

55. Wang, L. et al. Multilingual E5 text embeddings: a technical report. Preprint at https://arxiv.org/abs/2402.05672 (2024).

56. Akram, M. K. et al. jina-embeddings-v5-text: task-targeted embedding distillation. Preprint at https://arxiv.org/abs/2602.15547 (2026).

57. Choi, C. et al. Linq-Embed-Mistral technical report. Preprint at https://arxiv.org/abs/2412.03223 (2024).

58. Wang, L. et al. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 11897–11916 (2024). https://doi.org/10.18653/v1/2024.acl-long.642

59. Bruch, S. An alternative cross entropy loss for learning-to-rank. In *Proceedings of the Web Conference 2021* 118–126 (2021). https://doi.org/10.1145/3442381.3449794

60. Jagerman, R., Qin, Z., Wang, X., Bendersky, M. & Najork, M. On optimizing top-K metrics for neural ranking models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2303–2307 (2022). https://doi.org/10.1145/3477495.3531849

61. Moreira, G. de S. P. et al. NV-Retriever: improving text embedding models with effective hard-negative mining. Preprint at https://arxiv.org/abs/2407.15831 (2024).

62. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42,** 318–327 (2020). https://doi.org/10.1109/tpami.2018.2858826

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx