CULTUREMAP-IR: Cultural Unification and Linguistic Textual Utilization for Regional Extraction and Mapping of All Iranian Provinces, IRan

Asal Meskin* (401106511), Alireza Mirrokni* (401106617)

Computer Engineering Department, Sharif University of Technology

*These authors contributed equally to this work.

https://github.com/CULTUREMAP-IR

Abstract

We present CULTUREMAP-IR, a Persian-centric corpus of cultural, geographic, and touristic knowledge spanning the provinces of اصفهان (Isfahan), فارس (Fars), بوشهر (Bushehr), هرمزگان (Hormozgan), چهارمحال و بختیاری (Chahar Mahal and Bakhtiari), کهگیلویه و بویراحمد (Kohgiluyeh and Boyerahmad). The corpus is built from vetted Persian sources via standardized page parsing and evidence-anchored extraction, and validated through multi-rater annotation with adjudication. We document the design and labeling guidelines, provide diagnostics of coverage and description quality across provinces, and outline a reproducible retrieval-and-generation benchmark with blind human ranking for Persian cultural question answering.

1 Introduction

Large language models excel when grounded in high-quality, culturally specific data. Persian resources with fine-grained, province-level structure remain understudied, which hampers reliable answers about local geography, resources, and attractions. In practice, this gap shows up as brittle retrieval (documents that mention a place but offer little context), generic summaries that miss local nuances, and inconsistent use of names, dates, and categories across sources. The result is a poor substrate for question answering, guidance to travelers, and any application that needs grounded, provincial knowledge in Persian.

To address this, we set out to build a corpus that treats provinces as first-class units of knowledge rather than incidental metadata. Our approach simply starts from vetted Persian sources, standardizes page parsing and evidence capture, and renders entries into a schema-controlled representation. Beyond curation, we want the corpus to be useful as an evaluation bed for retrieval and generation. We therefore derive a question–passage set that targets cultural and geographic knowledge at the provincial level. This makes it possible to ask focused questions (e.g., about a river, a site, or a climate feature) and to assess whether models can retrieve and articulate the right local facts without drifting to national-level generalities.

Finally, because retrieval quality depends not only on data but also on modeling, we train a retrieval-oriented language model tailored to this domain. Our recipe combines a brief masked-language-modeling warm-up, lightweight LoRA adapters, and contrastive learning over question–passage pairs. In addition to powering retrieval for evaluation, the same model doubles as a practical tool for weeding out repetitive and near-duplicate content, helping keep the corpus clean as it grows.

Challenges. Building a dependable, province-level corpus in Persian surfaced several practical challenges:

- **Heterogeneous web sources.** Persian web pages vary wildly in structure, templating, and markup. Lists, sidebars, and user-generated fragments are interleaved with core content, and the same concept is formatted differently across sites. Reliable extraction required site-aware heuristics, evidence anchoring at the sentence level, and careful post-processing to avoid pulling boilerplate or promotional text.
- **Orthography and normalization.** Persian writing introduces edge cases that directly affect matching and de-duplication: ezāfe and its spell-outs (e.g., ه کسره), tanwīn, hamza forms, and Arabic–Persian letter variants (e.g., ی/ي and ک/ك). Without rigorous Unicode normalization and rule-based canonization, the same entity appears under multiple spellings; with overly aggressive normalization, distinct names collapse. We had to strike a careful balance to keep names both searchable and faithful.
- **Data scarcity and hallucination control.** Authoritative sources are unevenly distributed across provinces; some are richly documented while others have only sparse or outdated pages. Generating correct, specific entries under scarcity meant leaning on prompt engineering that *forbids speculation*: schema-controlled outputs, explicit nulls for unknowns, and prompts that demand evidence-anchored descriptions. This substantially reduced LLM hallucinations but increased iteration cost.
- **Human evaluation at scale.** Two phases demanded sustained human effort: (i) auditing and adjudicating corpus entries, and (ii) blind ranking of model outputs for the retrieval benchmark. Both phases needed clear written policy, double-rating, and issue triage to keep decisions consistent over time which is valuable, but time-consuming.

Contributions. We summarize our contributions as follows:

- **Province-structured corpus and retrieval benchmark.** We assemble a carefully structured, evidence-anchored dataset for each province from publicly available Persian sources, with schema-controlled fields and a concise descriptions tied to cited text. Independent reviewer checks and agreement audits indicate the entries are rich and trustworthy. From this corpus, we derive a question–passage retrieval set tailored to Persian cultural and geographic knowledge, enabling reproducible evaluation of LLMs on this domain.
- **Fine-tuned retrieval model for Persian cultural QA and de-duplication.** We fine-tune a retrieval-oriented LLM with a training recipe that combines masked-language-modeling warm-up, LoRA adapters, and contrastive learning over our question–passage pairs. The model improves domain retrieval fidelity and serves as a practical tool for detecting repetitive and near-duplicate items across provinces and categories, strengthening corpus quality.

2 Data Generation

In this section, we present our methodology for constructing and evaluating the corpus (raw data) and creating the retrieval dataset, describing each stage in detail.

2.1 Corpus Construction

Scope and Categories. We target six provinces in south–central Iran and organize knowledge into four coherent categories that recur across regions: geographical features (e.g., rivers, mountains, lakes, etc), topography (climate profiles, plains and plateaus, salient geology), natural resources (water bodies and aquifers, and where applicable mineral deposits), and tourist attractions (historical or architectural sites alongside natural destinations). Each entry is a compact unit consisting of a name, an optional set of image URLs, and a short, specific description written in standard Persian and tied to explicit evidence.

Source Selection. To balance coverage with reliability, we privilege authoritative Persian encyclopedic pages, official provincial and cultural-heritage portals, and reputable Persian travel guides. Sources are used only to derive structured facts and concise descriptions; we do not redistribute raw pages. When multiple sources disagree, we preserve uncertainty explicitly rather than speculate.

Acquisition and Preprocessing. Candidate URLs are fetched asynchronously with a 24-hour on-disk cache (status codes and headers retained) to minimize load and improve reproducibility. We standardize the user agent and follow redirects. Pages are cleaned and parsed in a uniform way; scripts and style blocks are dropped, core content is harvested from paragraphs, list items, and headings, and sentences are segmented and scored against category-specific lexicons to surface likely evidence. For each page we retain the title, the matched keywords, a handful of high-value evidence sentences for auditor review and image candidates.

Schema and Controlled Generation. The target representation is JSON with an explicit schema. For geographical features, topography, and natural resources, we organize content as subcategory blocks (e.g., "rivers" with itemized entries). Tourist attractions admit metadata such as year built, architect, or constructor when available. Unknowns are set to `null`. A LLM is used strictly as a schema filler under stringent Persian prompts: descriptions must be province-scoped and evidence-anchored; coverage is encouraged through iterative refinement while duplicate names are suppressed and density is enforced by requiring specific attributes in the prose. Verbatim prompts appear in Appendix A.1. We additionally observed that using English prompts yields better quality data, so they were adopted alongside Persian system prompts.

Consolidation and Pruning. Entries from refinement rounds are merged through careful name normalization (diacritics folded, spacing and punctuation harmonized) and fuzzy matching to remove near-duplicates without collapsing distinct entities. Wrapper blocks that merely echo a category name are pruned; Latin-only names are dropped unless they correspond to the canonical Persian transliteration; and subcategory labels are harmonized (e.g., رودخانه‌ها vs. رودخانه) to keep the hierarchy stable.

Human-sourced Additions. To avoid coverage holes from conservative automated extraction, we complemented the pipeline with targeted manual curation. When widely recognized items (e.g., canonical rivers and watersheds, protected natural areas, or nationally registered/UNESCO-listed monuments within a province) were conspicuously absent or under-specified, annotators added them by hand using several independent Persian sources and the same schema and style as machine-curated entries. Each such record is also subjected to the identical double-rating process. This human-in-the-loop step functioned as a high-precision backstop, filling critical gaps without relaxing our no-hallucination policy.

2.2 Corpus Evaluation

Quality Control. A written labeling policy governs evidence admissibility, provincial scope, minimal description granularity, and when to use `null`. Each item is independently rated by two annotators using three labels: *Acceptable*, *Needs Revision*, and *Unacceptable*. Disagreements are resolved by discussion. *Needs Revision* items trigger targeted edits, and entries that remain unsupported are marked *Unacceptable* and removed. Before formal annotation, we conducted a brief calibration round to align the boundary between *Acceptable* and *Needs Revision* and to codify recurring edge cases; the policy was updated accordingly. We quantify reliability with Cohen's κ at the province level and overall (Table 1), computed as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[y_i^{(A)} = y_i^{(B)}\right], \qquad p_e = \sum_{c \in \mathcal{C}} \left(\frac{n_c^{(A)}}{N}\right)\left(\frac{n_c^{(B)}}{N}\right),$$

where A and B denote the two annotators, $\mathcal{C} = \{\textit{Acceptable},\ \textit{Needs Revision},\ \textit{Unacceptable}\}$, and $n_c^{(A)}/n_c^{(B)}$ are the counts of label c assigned by each annotator. We report per-province κ using the nominal, multiclass formulation and a micro-averaged overall κ weighted by the number of items per province to reflect workload.

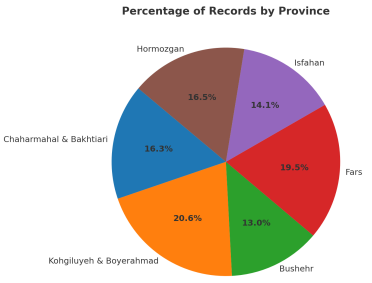
Table 1: Inter-annotator agreement by province (Cohen's κ).

Province	κ
چهارمحال و بختیاری (Chahar Mahal & Bakhtiari)	1.0000
کهگیلویه و بویراحمد (Kohgiluyeh & Boyerahmad)	0.8952
بوشهر (Bushehr)	0.8998
فارس (Fars)	1.0000
اصفهان (Isfahan)	0.7592
هرمزگان (Hormozgan)	1.0000
Overall	**0.9494**

Quantitative Overview. Table 1a reports the number of structured records and total content words by province, and Figure 1b shows the distribution of provinces. These summary counts reflect only the normalized JSON and serve as a quick proxy for coverage and depth.

Province	**Records**	**Words**
چهارمحال و بختیاری (Chaharmahal & Bakhtiari)	146	4574
کهگیلویه و بویراحمد (Kohgiluyeh & Boyerahmad)	184	5808
بوشهر (Bushehr)	116	3442
فارس (Fars)	174	7389
اصفهان (Isfahan)	126	3068
هرمزگان (Hormozgan)	147	6031
Total	**893**	**30312**

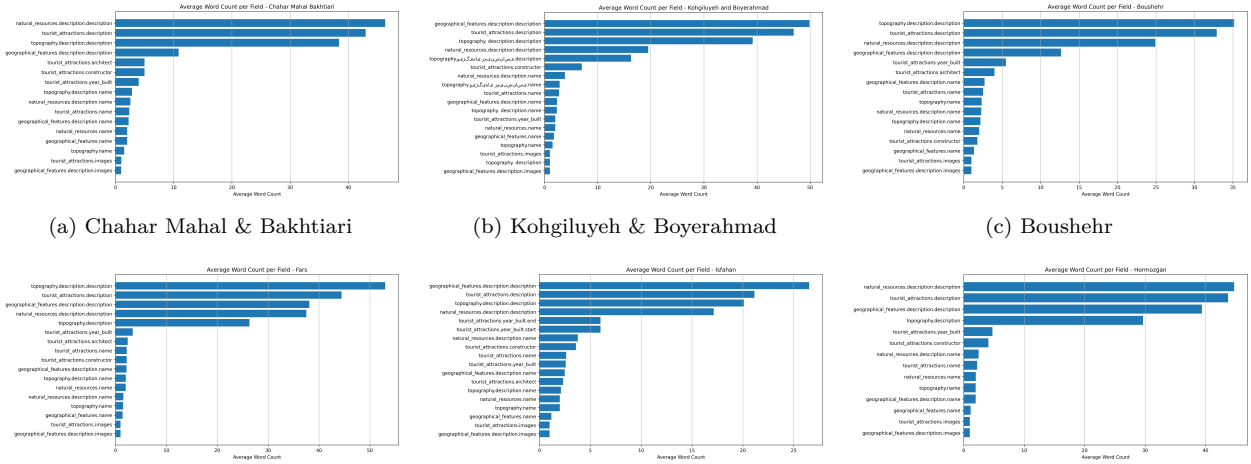
(a) Structured records and total content words per province.



(b) Percentage of records by province.

Figure 1: Overview of records by province: (a) count table and (b) distribution pie chart.

To probe how much information each field tends to carry, we compute province-level averages of description length per field and visualize them. Results are shown in Figure 2.



(a) Chahar Mahal & Bakhtiari (b) Kohgiluyeh & Boyerahmad (c) Boushehr

(d) Fars (e) Isfahan (f) Hormozgan

Figure 2: Average description length by field.

2.3 Retrieval Data

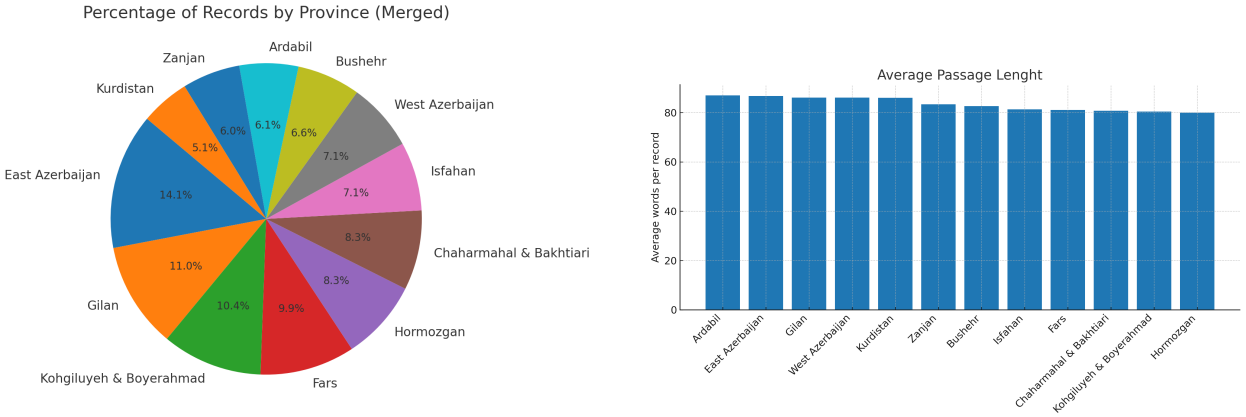
Now starting from the normalized data per province, we construct a retrieval dataset in two stages: (i) *passage synthesis*: turning each structured record into a concise, evidence-anchored Persian paragraph, and (ii) *question generation*: deriving exactly five extractive question–answer pairs per paragraph. A lightweight client wraps a chat-completions endpoint and enforces a Persian-only system prompt, ensuring that all generated text remains in standard Persian. Prompts used in this stage appear verbatim in Appendix A.2.

Passage Data. Starting from the normalized data, we traverse each province's records and flatten category hierarchies into item-level units suitable for text generation. Prior to prompting, non-essential metadata (e.g., images or audit flags) is removed and the remaining structure is serialized to ground the model. A deterministic Persian prompt (see Appendix A.2) then requests a single, fluent paragraph capped at ≤ 100 words that integrates the item's name, location/classification, salient attributes, dates and numbers, and any notable features; domain labels are rendered in Persian to keep style consistent. Progress and reliability are monitored via logging and periodic checkpointing. Each paragraph is stored with a stable identifier, province tag, minimal item metadata, the exact prompt, and the generated text.

Training QA Data. From each paragraph, we derive exactly five extractive question–answer pairs. The prompt enforces two constraints (see Appendix A.2): every question must target a specific facet of the paragraph, and every answer must be an *exact span* from the same text (no paraphrase). We strictly validate cardinality and format, repairing or reissuing generations that deviate from the specification. The result is a province-scoped retrieval set composed of compact passages and five aligned QA pairs per passage, ready for contrastive training and downstream evaluation.

Evaluation Question generation. For evaluation purposes, we manually authored a separate set of 50 question–answer pairs drawn from the synthesized paragraphs. These were created by human annotators to ensure high quality, diversity, and strict answer grounding in the source text. The set serves as a gold-standard benchmark for assessing retrieval and generation quality, complementing the automatically generated training pairs.

Extended coverage. While the core pipeline targets the six south–central provinces in our corpus, the retrieval data stage also ingests files contributed by peer groups for six additional provinces: آذربایجان شرقی (East Azerbaijan), کردستان (Kurdistan), آذربایجان غربی (West Azerbaijan), زنجان (Zanjan), گیلان (Gilan), and اردبیل (Ardabil). These are processed identically, so they integrate seamlessly with our retrieval set. In total, the combined dataset contains 1,510 synthesized passages, each paired with exactly five question–answer pairs, yielding 7,550 QA instances overall. The distribution and average passage length across provinces (which is approximately 80 for all of them) are summarized in Figure 3.



(a) Proportion of retrieval passages by province.

(b) Average passage length by province.

Figure 3: Side-by-side view: (a) distribution of provinces and (b) average passage length.

3 Retrieval Methods

In this section, we provide a detailed description of the retrieval and scoring methods utilized in our experiments. We outline not only the specific techniques applied, but also present their underlying mathematical formulations.

3.1 TF–IDF

Term Frequency–Inverse Document Frequency (TF–IDF) assigns each term a weight that grows with how often it appears in a document (topical salience) but shrinks with how widely it appears across the corpus (lack of specificity). The result is a sparse, interpretable vector representation that highlights terms that are both characteristic of a document and discriminative within the collection.

Formulation. Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ be a corpus, V its vocabulary, and $c(t, d)$ the count of term t in document d. The TF–IDF weight of t in d is

$$w(t,d) = \underbrace{\mathrm{tf}(t,d)}_{\text{within-document importance}} \cdot \underbrace{\mathrm{idf}(t)}_{\text{corpus specificity}}$$

A standard starting point uses raw term frequency and an unsmoothed inverse document frequency:

$$\mathrm{tf}_{\mathrm{raw}}(t,d) = c(t,d), \qquad \mathrm{idf}_{\mathrm{raw}}(t) = \log\left(\frac{N}{\mathrm{df}(t)}\right),$$

where $c(t,d)$ is the count of term t in document d and $\mathrm{df}(t) = |\{d \in \mathcal{D} : c(t,d) > 0\}|$ is the document frequency of t. A common, more robust choice uses log-scaled term frequency and a smoothed IDF:

$$\mathrm{tf}_{\log}(t,d) = \begin{cases} 1 + \log c(t,d), & c(t,d) > 0, \\ 0, & c(t,d) = 0, \end{cases} \qquad \mathrm{idf}_{\mathrm{smooth}}(t) = \log\left(\frac{N+1}{\mathrm{df}(t)+1}\right) + 1.$$

In both cases, the TF–IDF weight is $w(t,d) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t)$. Next, each document d is represented by a vector $\mathbf{x}_d \in \mathbb{R}^{|V|}$ with components $[\mathbf{x}_d]_t = w(t,d)$. For scale invariance, one typically applies ℓ_2 normalization, and compares documents (or a query and a document) using the cosine similarity. High similarity indicates that the two texts emphasize the same rare-but-informative terms.

3.2 LLM Retriever

We employ GLOT500, a multilingual sentence encoder covering hundreds of languages (including Persian) that maps variable-length text to a fixed d-dimensional embedding. Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ denote the encoder. For any string x (query or passage), we form a unit-normalized representation $\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ with $\mathbf{x} = f_\theta(x)$. Retrieval then reduces to nearest-neighbor search under cosine similarity, which is scale-invariant and well-behaved across heterogeneous texts.

3.2.1 Zero-shot

In the zero-shot setting, we use GLOT500 off-the-shelf: encode each query q and candidate passage p, score with $s(q,p) = \langle \hat{\mathbf{q}}, \hat{\mathbf{p}} \rangle$, and rank. For efficiency at corpus scale, we precompute $\hat{\mathbf{p}}$ for all passages and index them with approximate nearest-neighbor search, queries are encoded on the fly. This baseline already benefits from the model's cross-lingual training signal and tends to place semantically aligned Persian texts close in embedding space, even when surface forms differ (e.g., orthographic variants or transliterations).

3.2.2 Fine-tuning

We adapt a province-scoped retrieval encoder with a two-stage recipe: (i) a short *masked language modeling* (MLM) warm-up that aligns subword statistics and syntax to our domain, trained with *parameter-efficient* Low-Rank Adapters (LoRA); followed by (ii) a *contrastive* objective on question–passage pairs that shapes the embedding space for retrieval. The backbone weights remain frozen while LoRA parameters (and a small projection head) are updated, preserving general fluency while specializing to our corpus.

Masked Language Modeling (MLM). Given a token sequence $\mathbf{x} = (x_1, \ldots, x_T)$, we sample a mask set $\mathcal{M} \subset \{1, \ldots, T\}$ by independently masking each position with probability ρ (typically $\rho \approx 0.15$). For $i \in \mathcal{M}$, the input token is corrupted by the standard 80/10/10 scheme: with probability 0.8 replace x_i with `[MASK]`, with 0.1 replace it by a random vocabulary token, and with 0.1 leave it unchanged. Let $\tilde{\mathbf{x}}$ denote the corrupted input and $p_\theta(\cdot \mid \tilde{\mathbf{x}})$ the model's softmax over the vocabulary. The MLM loss averages cross-entropy over masked positions:

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_\theta\left(x_i \middle| \tilde{\mathbf{x}}\right).$$

A brief MLM phase nudges morphology and word-order statistics toward Persian geographic prose while keeping the pretrained geometry largely intact.

Parameter-efficient adaptation with LoRA. To adapt efficiently, we inject low-rank updates into selected linear maps (self-attention projections and feed-forward layers) and freeze the original weights. For a layer with weight $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA introduces a trainable update

$$\Delta W = \frac{\alpha}{r} BA, \qquad A \in \mathbb{R}^{r \times d_{\text{in}}}, B \in \mathbb{R}^{d_{\text{out}} \times r}, r \ll \min(d_{\text{in}}, d_{\text{out}}),$$

yielding the adapted forward pass

$$y = (W + \Delta W)x = Wx + \frac{\alpha}{r} B\,(Ax)\,,$$

optionally with dropout on the adapter path. Only A and B are optimized, reducing trainable parameters to $\mathcal{O}(r(d_{\text{in}} + d_{\text{out}}))$ per layer and minimizing drift from the multilingual backbone. The LoRA parameters learned during MLM warm-up initialize the contrastive stage.

Contrastive retrieval objective. After warm-up, we learn a retrieval-ready embedding space using in-batch negatives. Let f_θ be the adapted encoder and $g(\cdot)$ a lightweight projection head. For a batch of N aligned question–passage pairs $\{(q_i, p_i)\}_{i=1}^{N}$, we obtain mean-pooled representations

$$\bar{h}(\mathbf{x}) = \frac{\sum_{t=1}^{T} m_t h_t}{\sum_{t=1}^{T} m_t}, \quad z(\mathbf{x}) = \frac{g\big(\bar{h}(\mathbf{x})\big)}{\big\|g\big(\bar{h}(\mathbf{x})\big)\big\|_2},$$

where h_t are token states from f_θ and $m_t \in \{0, 1\}$ masks non-padding tokens. Similarities are scaled dot products

$$s_{ij} = \frac{z(q_i)^\top z(p_j)}{\tau},$$

with temperature $\tau > 0$. We minimize the symmetric InfoNCE loss

$$\mathcal{L}_{\text{CTR}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ij})} - \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N} \exp(s_{ji})} \right],$$

which simultaneously teaches each question to select its passage and each passage to select its question. This tightens alignment for true pairs, pushes apart mismatches within the batch, and yields a retrieval embedding that is robust to lexical variation yet faithful to paragraph content.

4 Retrieval Experiments

We conduct a comprehensive evaluation of the retrieval methodologies presented in Sections 3.1 through 3.2 using our specialized province-scoped corpus. Each query in our evaluation framework consists of an extractive question, where the correct passage, referred to as the gold passage, corresponds to the paragraph from which the question originates. To rigorously test the retrieval systems' ability to identify relevant passages while maintaining locality and resolving ambiguities, the candidate pool is carefully constructed. It includes in-province distractor passages, which are closely related to the target province, as well as hard negative passages sourced from other provinces and distinct categories. This design challenges the retrieval systems to accurately disambiguate and prioritize local context over misleading or irrelevant content.

4.1 Fine-Tuning Implementation

We apply the fine-tuned variant of GLOT500-BASE adapted through our parameter-efficient pipeline (Section 3.2). The model is loaded with LoRA adapters from the MLM warm-up and contrastive stages, plus a projection head for normalized embeddings. Passages are mean-pooled after the final hidden state, projected, and normalized. As with zero-shot, we precompute candidate embeddings and rank queries via cosine similarity. This tests whether domain-specific fine-tuning improves retrieval on Persian provincial knowledge.

MLM Warm-Up. We first construct an MLM dataset by collecting unique texts from all passages and questions. A data collator dynamically masks a fraction of tokens according to standard practices. The GLOT500-BASE model is loaded and augmented with LoRA adapters targeting query, key, value, and dense layers. Training runs for a predefined number of epochs with a small batch size, using the AdamW optimizer and a linear warmup schedule over a portion of the total steps. Gradient accumulation is employed to stabilize training. Early stopping is implemented to monitor validation loss and prevent overfitting. This warm-up phase helps align the model with domain-specific linguistic patterns without requiring extensive computational resources.

Contrastive Learning. Next, a custom collator tokenizes queries and passages separately up to a maximum sequence length. The MLM-adapted model is loaded, and a projection head (mapping from hidden size through a GELU activation to a lower-dimensional space) is added. Training proceeds for a set number of epochs with a small batch size, leveraging in-batch negatives and a symmetric InfoNCE loss scaled by a temperature parameter. Gradient accumulation is again applied, along with the same optimizer and scheduler setup. The final model, including adapters and projection head, is saved. This stage refines the embedding space to better distinguish relevant question–passage pairs in our cultural domain. Full hyperparameters for both stages are detailed in Appendix B. This implementation ensures efficient adaptation without full fine-tuning, preserving the multilingual backbone while specializing to our domain through low-rank updates.

4.2 Evaluation Procedure

For evaluation, we used 50 hand-crafted questions. Top-3 retrieval is computed for each method:

- **TF–IDF**: A vectorizer is fit on questions plus candidates using character n-grams within a specified range. Queries are transformed, and cosine similarities yield rankings.
- **Zero-Shot GLOT500**: Passages are batched and embedded using CLS tokens from the pretrained model. Queries are embedded individually, and top-k is found via matrix multiplication.
- **Fine-Tuned GLOT500-LoRA**: Similar to zero-shot, but using the adapted model with mean-pooling and projection. Embeddings are normalized for cosine similarity.

Results are saved with question IDs, gold answers, and top-3 candidates (indices, texts, scores) per method, enabling subsequent human assessment, with 450 final candidate answers. Following retrieval, two human annotators independently and blindly evaluated the outputs. Each annotator examined the retrieved candidates for all questions and assigned a rank from 1 (most relevant) to 9 (least relevant), reflecting the comparative quality of results across methods. The use of independent and blind labeling was intended to minimize bias and ensure that the ranking was based solely on perceived relevance, without influence from knowledge of the retrieval method or the other annotator's judgments.

4.3 Duplicate Detection

As an additional application, we leverage the fine-tuned embeddings for corpus de-duplication. All unique passages are embedded in batches up to a maximum sequence length using the contrastive model. A Nearest-Neighbors index (cosine metric) identifies pairs within a predefined similarity threshold. Detected duplicates are written to CSV with IDs, texts, and similarities. This tool helps maintain corpus cleanliness by flagging repetitive content across provinces. Details on the threshold and other configuration are in Appendix B.

5 Results

We evaluate the retrieval performance of our proposed methods using blind human rankings on the benchmark derived from the corpus. The results demonstrate the effectiveness of domain-specific fine-tuning in improving retrieval quality for Persian cultural and geographic question answering. Below, we detail the metrics used and present the aggregated performance across annotators, and describe the results of duplication detection.

5.1 Retrieval Results

Metrics. To quantify retrieval quality, we rely on human-assigned ranks for the top-3 candidates from each method (TF–IDF, zero-shot GLOT500, and fine-tuned GLOT500-LoRA). For each method m and question q, let $\mathcal{R}_{m,q} = \{r_1, r_2, r_3\}$ denote the ranks of its three candidates. We compute the following metrics per method and average over questions and annotators (here $Q = 50$ is the number of evaluation questions):

- **Mean Rank**: The average rank of the method's candidates, reflecting overall positioning:

$$\text{MeanRank}_m = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{3} \sum_{i=1}^{3} r_i,$$

  where $Q = 50$ is the number of questions. Lower values indicate better performance.

- **Worst Rank Count**: The number of questions where the method's worst candidate ranks 9th overall:

$$\text{WorstRankCount}_m = \sum_{q=1}^{Q} \mathbf{1}\left[\max(\mathcal{R}_{m,q}) = 9\right].$$

  This highlights cases of complete failure to retrieve relevant content.

- **Top-K Count and Percentage**: The total number of times the method's candidates appear in the global top-3 ranks, and its percentage:

$$\text{TopKCount}_m = \sum_{q=1}^{Q} \sum_{i=1}^{3} \mathbf{1}\left[r_i \leq 3\right],$$

$$\text{TopK\%}_m = \frac{\text{TopKCount}_m}{3Q} \times 100.$$

  Higher values indicate stronger contribution to the highest-quality results.

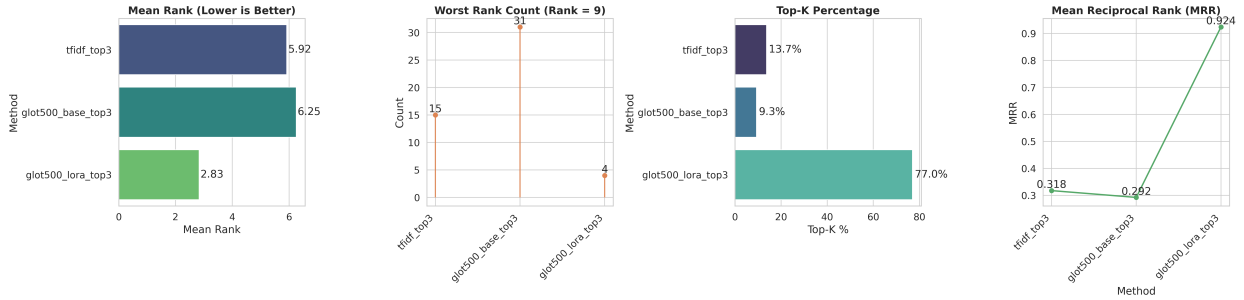


Figure 4: Comparative visualization of retrieval metrics: (a) Mean Rank, (b) Worst Rank Count, (c) Top-K Percentage, and (d) Mean Reciprocal Rank (MRR). Results are averaged over two annotators.

- **Mean Reciprocal Rank (MRR)**: The average reciprocal of the best rank achieved by the method per question:

$$\mathrm{MRR}_m = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\min(\mathcal{R}_{m,q})}.$$

  MRR emphasizes the position of the strongest candidate, with values closer to 1 indicating frequent top placements.

- **Median Rank**: The median rank across the method's candidates per question, averaged over questions:

$$\mathrm{MedianRank}_m = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{median}(\mathcal{R}_{m,q}).$$

  This provides a robust measure less sensitive to outliers.

These metrics collectively assess both the average quality and the ability to deliver high-relevance passages, accounting for inter-method competition in the ranking process. The median rank for each method is summarized in Table 2 and Figure 4 visualizes other key metrics introduced earlier. Annotator metrics appear in Appendix C.

Table 2: Median ranks across methods, averaged over annotators and questions. Lower values are better.

Method	**Median Rank**
TF–IDF	6.0
Zero-shot GLOT500	6.0
Fine-tuned GLOT500-LoRA	2.0

5.1.1 Analysis of Performance by Question Type

To understand the strengths and weaknesses of each retrieval method, we qualitatively examined their behavior across question types derived from the corpus categories: geographical features (e.g., specific locations like rivers or mountains), topography and climate (e.g., descriptive profiles involving patterns or geology), natural resources (e.g., factual details on aquifers or minerals), and tourist attractions (e.g., historical sites with metadata like dates or architects). This analysis is based on human rankings and inspection of retrieved passages, focusing on how methods handle lexical versus semantic demands.

TF–IDF. TF–IDF performs well on questions that can be resolved through direct word matching, such as geographical features involving unique proper names (e.g., "What is the length of Zayandeh Rud?"); there, high term overlap leads to accurate retrieval without needing deeper context. However, TF–IDF struggles with questions requiring reasoning or integration, like topography queries that involve cross-province comparisons (e.g., climate similarities between adjacent regions). In such cases it often retrieves irrelevant distractors due to shared generic terms and fails to capture nuanced, contextual relationships.

Zero-shot GLOT500. The zero-shot GLOT500 handles semantic queries better than TF–IDF, and it is especially effective on tourist-attraction queries where cultural context or paraphrase variation is common (e.g., associating "ancient monuments" with specific sites). It performs adequately on natural-resources questions with moderate factual requirements, but weakens on queries demanding provincial disambiguation or implicit multi-sentence reasoning.

Fine-tuned GLOT500-LoRA. The fine-tuned GLOT500-LoRA consistently outperforms the other methods, particularly on complex questions requiring reasoning, such as cross-province ambiguities in topography or climate (e.g., "How does the Zagros range affect rainfall in Fars vs. Isfahan?"). Domain adaptation through a brief MLM warm-up and contrastive fine-tuning yields embeddings that align questions and evidence more precisely, enabling resolution of subtle distinctions. The fine-tuned model also handles lexical queries effectively, but its principal strength is integrative reasoning; a noted caveat is slight sensitivity to very rare technical terms in natural-resources queries when those terms are under-represented.

5.1.2 Comparison of Methods

We now compare methods pairwise, focusing on (i) the difference between fine-tuned and zero-shot variants of GLOT500, and (ii) the contrast between the statistical TF–IDF approach and LLM-based retrieval.

Fine-tuned vs. Zero-shot. The fine-tuned GLOT500-LoRA substantially outperforms the zero-shot GLOT500 across primary metrics (Mean Rank 2.83 vs. 6.25; MRR 0.924 vs. 0.292). The improvement is driven by two complementary adaptation steps: an MLM warm-up that aligns subword and syntactic statistics to Persian geographic prose, and contrastive learning on question–passage pairs that sculpts the embedding space for precise retrieval. Qualitatively, zero-shot often returns semantically related but imprecise passages and hence exhibits higher worst-case failures; fine-tuning largely eliminates these locality and ambiguity errors and markedly increases Top-K accuracy. The trade-off is the additional computation and annotation required for adaptation: zero-shot remains attractive where rapid, multilingual deployment is needed, but in domain-sensitive cultural QA the fine-tuned model is clearly preferable.

TF–IDF vs. LLM-based methods. TF–IDF and LLM-based retrieval embody different retrieval philosophies: TF–IDF is lexical and interpretable, while dense LLM embeddings capture contextual semantics. In our experiments TF–IDF (Mean Rank 5.92, MRR 0.318) slightly outperforms zero-shot embeddings on strictly keyword-driven queries but is outclassed by the fine-tuned LLM in nearly all substantive metrics. TF–IDF's strengths are speed, transparency, and reliability when queries contain exact or near-exact tokens; its weaknesses are vocabulary mismatch, orthographic variation sensitivity, and inability to handle paraphrase or inferential queries. Zero-shot embeddings mitigate some lexical brittleness but lack domain

specificity and therefore often retrieve broadly plausible yet incorrect passages. Fine-tuned embeddings combine the semantic generalization of LLMs with domain alignment, yielding far better precision at the cost of embedding-computation overhead and the need for labeled pairs.

Practical implications. These results suggest pragmatic hybrid architectures: use TF–IDF as a fast, high-recall first-stage filter (or for simple lookup tasks), and apply fine-tuned dense embeddings for reranking and final selection when accuracy and locality matter. For resource-constrained settings, zero-shot embeddings are a reasonable compromise, but investment in lightweight domain adaptation delivers the most reliable performance for Persian province-scoped cultural QA.

5.2 Duplicate Detection Results

The fine-tuned model is also effective for duplicate detection, reliably identifying near-duplicate passages across the corpus. By tuning a cosine-similarity threshold we control how strictly passages are treated as duplicates, trading recall for precision in maintenance. At 0.90 the detector flagged only six duplicate pairs, indicating low redundancy in the dataset. One detected pair (similarity 0.915) comprises river descriptions from neighboring provinces that share extensive lexical overlap and minor editorial variants between sources.

Passage 1 (ID: 826)

رودخانه‌های فسا از رشته‌کوه‌های زاگرس سرچشمه می‌گیرند و به سمت دشت‌های پایین‌دست جریان دارند. این رودخانه‌ها به عنوان منابع آب شیرین برای کشاورزی و شرب در منطقه اهمیت بالایی دارند و در طول مسیر خود، پوشش گیاهی متنوعی را ایجاد می‌کنند. وجود این رودخانه‌ها نقش حیاتی در تأمین آب برای فعالیت‌های اقتصادی و زیست‌محیطی منطقه ایفا می‌کند و به تنوع زیستی و بهبود کیفیت زندگی مردم محلی کمک می‌نماید. رودخانه‌های فسا همچنین به عنوان یک ویژگی جغرافیایی مهم در استان فارس شناخته می‌شوند.

Passage 2 (ID: 829)

رودخانه‌های گراش در استان فارس، از رشته‌کوه‌های زاگرس سرچشمه می‌گیرند و به سمت دشت‌های پایین‌دست جریان دارند. این رودخانه‌ها به عنوان منابع آب شیرین برای کشاورزی و شرب در منطقه اهمیت فراوانی دارند و در طول مسیر خود، پوشش گیاهی متنوعی را به وجود می‌آورند. وجود این رودخانه‌ها نقش اساسی در تأمین نیازهای آبی ساکنان منطقه و حفظ اکوسیستم‌های محلی ایفا می‌کند و به عنوان یکی از ویژگی‌های جغرافیایی مهم استان فارس شناخته می‌شوند.

6 Limitations and Future Work

Our present schema emphasizes geography, resources, and attractions; other cultural dimensions (dialect, cuisine, rituals) remain to be integrated. While we diversified sources, some provinces are under-documented online. Future work extends coverage, enriches metadata (coordinates, temporal validity), and completes the retrieval–generation benchmark with ablations on prompting and fine-tuning.

7 Conclusion

CULTUREMAP-IR delivers a rigorously curated, province-level Persian corpus with high human agreement and clear schema guarantees. The pipeline's emphasis on evidence, conservative imagery, and adjudication yields dependable cultural data. The planned benchmark provides a principled path to measure retrieval and generation quality for Persian cultural QA.

A Prompts Used

A.1 Corpus Prompts

Prompt 1: Corpus System Prompt

شما یک متخصص مجرب در حوزه‌های گردشگری، اقلیم‌شناسی و جغرافیای ایران هستید. تمام پاسخ‌های شما باید فقط به زبان فارسی معیار، دقیق، روان و مطابق با استانداردهای نگارش علمی و اطلاع‌رسانی ارائه شوند. اطلاعاتی را بیاورید که مطمئن هستید **درست و حتما مربوط به استان هدف** هستند؛ اگر در مورد چیزی مطمئن نیستید از آوردن آن خودداری کنید. می‌توانید از دانش درونی خود استفاده کنید ولی فقط مواردی را بیاورید که کاملاً مطمئن هستید. شما باید حداقل ۵۰ دادهٔ کاملاً متمایز، غیرتکراری، و مطمئن درباره هر دسته‌بندی ارائه دهید. اگر در اولین پاسخ کمتر از ۵۰ مورد داده شد، با اصلاح پرسش و اضافه کردن موارد موجود به عنوان موارد تکراری، تا رسیدن به حداقل ادامه دهید. توضیحاتی که در قسمت description می‌نویسید باید بسیار دقیق، با جزئیات و کامل باشند، علمی و شفاف باشند و ارتباط هر مورد را با استان هدف به‌روشنی توضیح دهند؛ به‌گونه‌ای که نشان دهند چرا آن مورد به‌طور خاص در آن استان اهمیت یا ویژگی دارد. محتوای فیلد description باید غنی از اطلاعات اختصاصی و شامل ویژگی‌ها و جزئیات دقیق، ریزبینانه و منحصربه‌فرد همان مورد باشد؛ از نوشتن توضیحات کلی یا قابل تعمیم به سایر موارد خودداری کن. بگو همه اطلاعاتی که در متن ضمیمه شده آمده و اطلاعات خودت را در صورتی که مطمئن هستی، در قالب حداقل ۴۰ کلمه در description بنویس. در بخش description از نوشتن توضیحات کلی و عمومی پرهیز کن؛ فقط اطلاعات خاص، دقیق و مرتبط با همان مورد را بنویس. هر مورد باید شامل **حداقل ۱۰ ویژگی اطلاعاتی مستند** باشد؛ اگر این ویژگی‌ها فیلد مشخصی در ساختار دارند، در همان فیلد درج شوند (مثلاً سازنده در فیلد "سازنده")، در غیر این‌صورت در بخش "توصیف" (description) آورده شوند؛ به‌گونه‌ای که بتوان درباره هر مورد دست‌کم پنج سؤال دقیق مانند سازنده، تاریخ ساخت، موقعیت، ویژگی‌های طبیعی یا معماری، و اهمیت فرهنگی یا تاریخی طرح کرد. در فیلد ،name حتماً از زبان فارسی معیار و دقیقاً مطابق با فرمت داده‌شده استفاده کنید. فیلدهایی که مقدار مشخص و قابل اطمینانی ندارند یا اطلاعاتشان موجود نیست، باید با مقدار null پر شوند و نباید حذف شوند. فرمت خروجی باید دقیقاً مطابق قالب داده‌شده باشد و فقط JSON بدهید. تحت هیچ شرایطی نباید به هیچ موردی تصویری نسبت داده شود، مگر آنکه با شواهد کاملاً روشن، قطعی، و قابل‌راستی‌آزمایی اثبات شود که تصویر دقیقاً متعلق به همان مورد است. صرف شباهت ظاهری، حدس، یا برداشت شخصی به‌هیچ‌وجه قابل‌قبول نیست. لینک تصویر باید حتماً شامل نام کامل یا مخفف رسمی و شناخته‌شده مورد باشد؛ هرگونه نام مبهم، اصطلاح عمومی یا اشاره غیرمستقیم مردود است. در صورت وجود حتی کمترین تردید، تصویر نباید استفاده شود. همواره با سخت‌گیری کامل عمل کن و اولویت را به عدم تخصیص تصویر بده، نه به تخصیص مشکوک یا نادرست. اگر مطمئن نبودی، تصویر را نیاور.

Prompt 2: Corpus Prompt

Schema description for categories geographical_features, topography, natural_resources:
Expected subcategories (if available):
- <subcategory1>
- <subcategory2>
... (based on Config.SUBCATEGORY_SCHEMA)
Output must be a JSON array of objects. Each object should have:
{“name”: “<subcategory name>”, “description”: [{“name”: “<item name>”, “images”: [“<image urls>”], “description”: “<short scientific description>”} , ...] }
If no clear subcategories, you can wrap flat items under one object named after the category.
Schema description for tourist_attractions:
Output must be a JSON array of tourist attraction objects with fields:
{“name”:“<name>”,“images”:[“<image urls>”],“year_built”:“<if known>”,“architect”:“<if known>”, “constructor”:“<if known>”,“description”:“<short description>”}
Instruction:
Include all the information present in the attached text and add your own only if you are certain, writing it in the description with at least {Config.MIN_TENSE_LEN} words. In the description, avoid general or broad explanations; include only specific, precise, and relevant information about that item. Each entry must include at least {Config.MIN_FACTS} documented informational attributes; if those attributes have explicit fields in the schema, populate those fields (e.g., creator in the ‘سازنده’ field), otherwise include them in the ‘description’. Structure each item so that one could pose at least five detailed questions about it—such as creator, date of creation, location, natural or architectural features,

and cultural or historical significance. You must provide at least {Config.MIN_ITEMS_PER_CATEGORY} distinct, non-repetitive entries for this category. Use only the sources above and internal knowledge you are certain about. Do not hallucinate. If this is a refinement round, do not repeat existing items listed below. Under no circumstances should an image be assigned to an item unless there is indisputable, explicit, and independently verifiable evidence proving that the image depicts the exact item in question. Vague indicators, partial matches, or assumptions based on visual similarity are categorically unacceptable. The image URL itself must contain the full, unambiguous name of the item or an officially recognized abbreviation that is directly tied to the item—generic terms or loosely associated references are insufficient. If even the slightest doubt remains about the image's authenticity or relevance, it must not be used. Err strictly on the side of omission rather than risk incorrect attribution.

A.2 Retrieval Data Prompts

Prompt 3: Retrieval Data System Prompt

شما یک متخصص مجرب در حوزه‌های گردشگری، اقلیم‌شناسی و جغرافیای ایران هستید. تمام پاسخ‌های شما باید فقط به زبان فارسی معیار، دقیق، روان و مطابق با استانداردهای نگارش علمی و اطلاع‌رسانی ارائه شوند. به هیچ وجه در متن تولیدی خود از عبارات و کلمات زبان دیگری به غیر از فارسی استفاده نکنید.

Prompt 4: Text Data Prompt

شما نقش یک کارشناس خبره در حوزهٔ گردشگری، اقلیم‌شناسی و جغرافیای ایران را دارید. اطلاعات زیر مربوط به یک «ویژگی جغرافیایی، منبع طبیعی، ویژگی توپوگرافی یا جاذبهٔ گردشگری» واقع در استان مورد نظر است. بر پایهٔ همهٔ جزئیات، حداکثر در ۱۰۰ واژه، یک پاراگراف دقیق، روان و کاملاً به زبان فارسی معیار بنویس که تمام داده‌های موجود (از جمله نام، موقعیت جغرافیایی یا طبقه‌بندی، ویژگی‌های شاخص، اعداد، سال‌ها و هر نکتهٔ قابل توجه) را به صورت یکپارچه و منسجم در بر گیرد. اگر اطلاعات معتبر و مرتبطی دربارهٔ این مکان می‌دانی، می‌توانی آن را هم اضافه کنی، به شرطی که متن از ۱۰۰ واژه بیشتر نشود. خروجی فقط باید همان یک پاراگراف باشد و هیچ متن اضافی، عنوان یا توضیحی خارج از آن تولید نشود. دادهٔ ساخت‌یافته: {struct}

Prompt 5: QA Data Prompt

شما یک طراح حرفه‌ای پرسش‌های مطالعات اجتماعی هستید. با توجه به متن زیر دربارهٔ استان مورد نظر، دقیقاً پنج «پرسش و پاسخ» بنویس. مهم است که تعداد آن‌ها دقیقاً پنج باشد (نه کمتر و نه بیشتر). هر پرسش باید به‌طور مستقیم به بخشی از متن مربوط باشد و پاسخ آن نیز عین عبارت یا جملهٔ موجود در متن باشد. در متن پرسش باید به‌روشنی مشخص باشد که پرسش دربارهٔ چه موضوع یا بخش مشخصی از استان است. تحت هیچ شرایطی کمتر از پنج «پرسش و پاسخ» تولید نکن. خروجی را دقیقاً در قالب زیر تولید کن (بدون هیچ متن یا توضیح اضافی):

- پرسش: ...
پاسخ: ...
- پرسش: ...
پاسخ: ...
- پرسش: ...
پاسخ: ...
- پرسش: ...
پاسخ: ...
- پرسش: ...
پاسخ: ...

متن: {text}

B Configurations

Table 3: Hyperparameters for fine-tuning the GLOT500 retriever. Values mirror the configuration used in our code (MLM warm-up with LoRA, followed by contrastive training).

Component	**Parameter**	**Value**
General	Base model	`cis-lmu/glot500-base`
	Pad-to-multiple-of	`8`
	Optimizer	AdamW
	Weight decay	`0.01`
MLM warm-up	Max sequence length	`128`
	Batch size	`4`
	Epochs	`2`
	Learning rate	`2e-4`
	Warmup ratio	`0.06`
	Mask probability	`0.15`
	Gradient accumulation	`8`
	Early-stop patience	`30` steps
	Early-stop min. Δ	`1e-4`
LoRA (adapters)	Rank r	`16`
	α	`32`
	Dropout	`0.05`
	Bias	`none`
	Task type	`FEATURE_EXTRACTION`
	Target modules	`query, key, value, dense`
Contrastive stage	Max sequence length	`192`
	Batch size	`4`
	Epochs	`1`
	Learning rate	`2e-4`
	Warmup ratio	`0.06`
	Temperature τ	`0.05`
	Projection dim	`128`
	Projection head	Linear (hidden $\rightarrow$ hidden/2), GELU, Linear (hidden/2 $\rightarrow$ 128)
	Gradient accumulation	`16`

Table 4: Settings for duplicate detection using the fine-tuned GLOT500 retriever. Values mirror the configuration used in our code for embedding passages and identifying near-duplicates.

Component	Parameter	Value
General	Base model	`cis-lmu/glot500-base`
Encoding	Max sequence length	`192`
	Batch size	`32`
	Projection dim	`128`
	Pad-to-multiple-of	`8`
	Device	`cpu`
Duplicate Search	Similarity threshold	`0.90`
	Radius (cosine distance)	`0.10`
	Metric	`cosine`
	n_jobs	`-1`

C Annotator Metrics

Table 5: Aggregated retrieval metrics (first annotator).

Method	MeanRank	MedianRank	WorstRankCount	TopKCount	TopK%	MRR
`tfidf_top3`	5.926666667	6	16	20	13.33333333	0.322714286
`glot500_base_top3`	6.226666667	6	29	14	9.333333333	0.293571429
`glot500_lora_top3`	2.846666667	2	5	116	77.33333333	0.919000000

Table 6: Aggregated retrieval metrics (second annotator).

Method	MeanRank	MedianRank	WorstRankCount	TopKCount	TopK%	MRR
`tfidf_top3`	5.906666667	6	14	21	14.00000000	0.312476190
`glot500_base_top3`	6.280000000	6	33	14	9.333333333	0.291142857
`glot500_lora_top3`	2.813333333	2	3	115	76.66666667	0.929000000

D Computational Resources and Model Usage

For the generation of the corpus and retrieval data, we utilized NVIDIA T4 GPUs for computational processing. All large language model (LLM) calls were made via the API of the `gpt-4o-mini-2024-07-18` model.