

# POISS: A Large-Scale Multilingual Dataset for Point-of-Interest Search

**Nicola Maritan**  
Amazon  
nicolamaritan01@gmail.com

**Alessandro Moschitti**  
Amazon  
amosch@amazon.com

**Federico Borazio**  
Amazon  
University of Rome Tor Vergata  
rfeboraz@amazon.com

**Xiaokun Zhou**  
Amazon  
angzhou@amazon.com

**Zhengwei Bai**  
Amazon  
zwbai@amazon.com

## Abstract

Points-of-Interest (POI) search, the task of retrieving and ranking geographically situated entities in response to a query issued from a specific location, underpins major mapping and navigation services. We present POI Search Suite (POISS), the first large-scale dataset designed for this task, pairing over 300K geo-anchored natural language queries across 7 languages and 14 locales with more than 4M graded relevance labels across over 90M candidate POIs drawn from the 72M-entity Overture Maps Places index, jointly capturing semantic match, geographic proximity, and entity quality. Our experiments assess the utility of the proposed dataset for advancing research on POI search.

## 1 Introduction

A Point-of-Interest (POI) is a geographically grounded location of potential relevance to a user: a restaurant, museum, hospital, transit station, or any of the many named places that structure the physical world. POI search, the task of retrieving and ranking relevant POIs from a large collection in response to a user’s information need, powers widely used services such as Google Maps, Yelp, and TripAdvisor.

What distinguishes POI search from general-purpose document retrieval is the tight coupling between language and geography. Queries vary in their spatial specificity: some explicitly express proximity (“coffee shops near me”), others imply it (“bars”), while others refer to a specific place independent of the user’s location (“where is Central Park”). A correct system must jointly model semantic intent and geographic scope. For instance, a pizzeria in Rome should be retrieved for users in Rome searching for pizza, or for users elsewhere explicitly searching for pizza in Rome, but not for users in New York seeking nearby pizza or for users in Rome searching for fish restaurants. This raises

the challenge of how to model distances and geographic boundaries effectively, particularly when geographic intent is implicit or expressed at different levels of specificity. In addition, POIs are associated with numerical metadata such as quality score, and the system should learn to balance semantic relevance, geographic proximity, and such signals when ranking competing POIs.

Despite the practical importance and inherent challenges of this task, no public resource pairs a large-scale POI corpus with natural language queries and graded relevance labels suitable for standard retrieval evaluation. We address this gap with a dataset built on the Overture Maps Places index (Overture Maps Foundation, 2026), a collection of more than 72 million POIs that serves as the candidate pool for each query’s graded relevance list. Our queries derive from real-world POI search logs spanning multiple locales and languages, with an LLM filtering private or sensitive content and de-identifying query texts and coordinates.

We associate each query with annotated POI candidates from Overture Maps Places, sourced through two complementary pipelines: matching the results returned by Google Maps to their Overture counterparts, and running a retrieve-and-rerank system directly over the corpus to recover entities the matching step misses. An LLM-as-a-Judge then builds the ground truth by discarding irrelevant candidates and duplicates and ranking the rest by semantic relevance, geographic proximity, and quality signals.

In this work, we contribute: (i) POI Search Suite (POISS), the first dataset pairing a large-scale, openly licensed POI corpus with geo-anchored multilingual queries and graded relevance labels; (ii) an empirical assessment of the dataset’s value as a research resource for POI retrieval from a large corpus and ranking over a pool of candidates, demonstrating that fine-tuning on the training split yields strong gains in full-corpus retrieval ( $\sim 82.8$

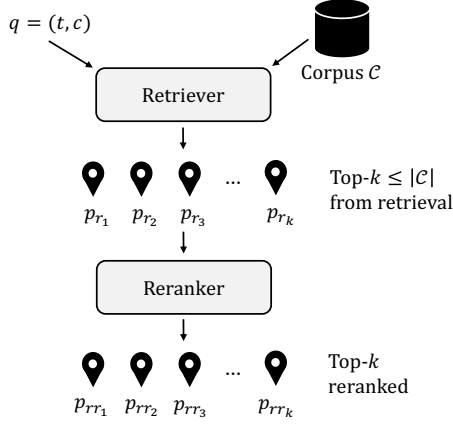


Figure 1: POI search system. Given a query  $q = (t, c)$ , the retriever returns the top- $k$  POIs from the corpus  $\mathcal{C}$  using an offline-built index, denoted  $p_{r_1}, \dots, p_{r_k}$  in retriever-rank order. The reranker jointly scores each candidate against the query and reorders them into the final ranking  $p_{rr_1}, \dots, p_{rr_k}$ .

R@100) and reranking ( $\sim 89.0$  MRR,  $\sim 66.0$  NDCG@20), which confirms both that the task challenges current methods and that the labels provide strong training supervision.

POISS is publicly available at <https://huggingface.co/datasets/amazon/poiss>.

## 2 Background and Related Work

In this section we outline how a POI search system is typically built over a corpus of POIs and survey existing tasks and datasets that motivate our contribution.

### 2.1 POI Search System

Given a query  $q = (t, c)$ , where  $t$  is a natural language text and  $c \in [-90, 90] \times [-180, 180]$  is a geographic coordinate, a POI search system maps  $q$  to a ranked list of  $k$  POIs from a corpus  $\mathcal{C}$ , with  $k \leq |\mathcal{C}|$ . Following the retrieve-and-rerank paradigm (Yates et al., 2021), the system decomposes this task into a high-recall retrieval stage over the full corpus  $\mathcal{C}$  followed by a reranking stage that improves both precision and ranking quality over the resulting candidate set. Figure 1 illustrates the architecture.

The *retriever* scores POIs against the query and returns the top- $k$  candidates from the entire corpus of POIs  $\mathcal{C}$ . Two families of models are commonly used: lexical retrievers such as BM25 (Robertson and Zaragoza, 2009), which match query and POI text via sparse term overlap, and dense bi-encoders (Reimers and Gurevych, 2019; Karpukhin

et al., 2020), which encode queries and POIs independently into a shared vector space and score candidates by inner product. To scale to a corpus of tens of millions of POIs, both variants index the corpus *offline*, so that retrieval at query time reduces to a lookup against this precomputed structure. This efficiency comes at the cost of limited query-POI interaction: because POIs must be encoded independently of the query, the retriever cannot directly model joint features that depend on both inputs, such as the geographic distance between the issuing coordinate  $c$  and a candidate POI or query-dependent combinations of POI metadata.

Operating on the much smaller retrieved set, the *reranker* can afford a *joint* encoding of the query and each candidate POI, typically realized as a cross-encoder (Nogueira and Cho, 2019; Humeau et al., 2020), whose accuracy on ranking tasks was first established by TANDA (Garg et al., 2020) via transfer-then-adapt fine-tuning, which requires in-domain labeled data of the kind POISS provides. The reranker can therefore recover fine-grained query-POI interactions that are difficult to model during efficient full-corpus retrieval, including geographic relationships and numerical metadata such as a quality score. Such non-textual evidence can be supplied in the input or fused with the [CLS] embedding as a feature vector (Lauriola and Moschitti, 2021; Han et al., 2021), an approach Bai et al. (2026) apply to POI reranking. However, this additional modeling capacity is restricted to the retrieved candidate set: POIs that are not selected by the retriever cannot be recovered by the reranker.

Since reranking cost grows with the candidate count, Matsubara et al. (2020) introduced two-stage reranking, placing a fast ranker before an accurate transformer; later work prunes candidates across encoder layers (Soldaini and Moschitti, 2020) or scores the top- $k$  jointly (Zhang et al., 2021).

### 2.2 Related Tasks and Datasets

Despite the practical significance of POI search, the academic community has long lacked the infrastructure to study it in a reproducible, standardized way. The core obstacle is data: while large-scale POI corpora are publicly available (Yang et al., 2019; Yelp Inc., 2023; Haklay and Weber, 2008; Overture Maps Foundation, 2026), none pairs geographic and categorical place information with the natural language queries and graded relevance labels that retrieval and ranking evaluation requires. This stands in stark contrast with general-purpose

information retrieval, where benchmarks such as MS MARCO (Bajaj et al., 2018), BEIR (Thakur et al., 2021), and MIRACL (Zhang et al., 2023) have driven consistent, reproducible progress over the past decade.

Adjacent efforts have addressed related but distinct problems. Geocoding maps a natural-language location description to a geographic location, e.g., a coordinate pair (Cui et al., 2026; Gritta et al., 2018), while entity linking, which resolves a mention in text to a specific knowledge-base entity (Le and Titov, 2018; Orlando et al., 2024), has been applied to POI entities in spatial-domain queries (Zhang et al., 2020). POI search, in contrast, retrieves and ranks multiple candidate POIs by relevance to user intent, proximity, and quality. The GeoCLEF shared tasks (Gey et al., 2006) evaluate geographic information retrieval over news documents but not POI entities. GeoGLUE (Li et al., 2025) introduces a Chinese geographic language understanding benchmark with tasks such as textual similarity and entity alignment, but targets NLU subtasks rather than end-to-end retrieval from a large corpus. LocalSearchBench (He et al., 2026) evaluates multi-hop reasoning over local life services but focuses on agentic question answering rather than standard retrieval and ranking. Mai et al. (2018) introduce POIReviewQA, a dataset of 20k questions paired with annotated Yelp review sentences, but their focus is on extracting answers from review text rather than on retrieving and ranking POIs themselves. POI recommendation benchmarks built on check-in data (Lian et al., 2020; Yang et al., 2019) evaluate sequential next-POI prediction, not query-driven retrieval.

**Our recent work.** POISS builds on a parallel line of research we have been conducting on POI search. Bai et al. (2026) propose a heterogeneous two-stage cross-encoder for POI reranking, and Shah et al. (2026) present the query understanding component of the same system, which prompts an LLM to predict the search area implied by a query. Finding that standard ranking metrics do not capture the specificity of POI search, we introduced Rank-Deviation Quality (RDQ) (Zhou et al., 2026), which scores a ranking against an ordinal ground truth rather than absolute relevance grades. Kambhatla et al. (2026) then made such supervision practical at scale with an LLM-based automatic annotator, which we use here to label POISS (Section 3.4).

None of these efforts, however, provides a pub-

licly available dataset that combines a large-scale POI corpus with natural language queries and graded relevance labels for standard information retrieval evaluation, which we describe and implement in this paper.

### 3 Dataset

Our dataset consists of three components, presented in the order of the subsections that follow: (i) the Overture Maps Places corpus  $\mathcal{C}$ , which serves as the shared POI index from which all relevant entities are drawn (Section 3.1); (ii) a set of queries derived from real POI search queries spanning multiple locales and languages, filtered and de-identified by an LLM to remove private or sensitive content (Section 3.2); and (iii) for each query, up to 20 POIs from  $\mathcal{C}$  annotated with graded relevance labels that reflect semantic match, geographic proximity, and quality signals, obtained by first assembling a candidate pool (Section 3.3) and then labeling it (Section 3.4). We take  $\mathcal{C}$  as given and construct the other two components.

#### 3.1 Points-of-Interest Corpus

Our dataset is built on version 2026-03-18 of the Overture Maps Places dataset (Overture Maps Foundation, 2026), an openly licensed corpus of over 72 million POIs with global coverage, released by the Overture Maps Foundation through the conflation of data from multiple commercial providers. Its scope extends well beyond businesses: alongside restaurants, shops, and service establishments, spanning multinational chains as well as long-tail independents, it covers schools, hospitals, places of worship, cultural and tourist landmarks, transportation hubs, parks, and natural features such as mountain peaks and beaches. Each POI is characterized by a name, address, coordinates, a primary category and a set of alternate ones, and may include additional attributes such as phone number, email, website, and social media links.

The corpus reaches more than 200 countries and territories, with coverage concentrated in a relatively small set of countries: the United States accounts for around 21% of all POIs, the top-10 countries together account for about 60%, and the top-20 for close to 80%, with a long tail of more than a hundred countries contributing the remainder. By continent, Europe and North America each cover around 29% of the index, followed by Asia at 26% and South America at 11%, while Africa and

Oceania each contribute a few percent. Tables 7 and 8 in Appendix A report the top-20 country breakdown and the continental distribution. The corpus is dense in the regions our queries come from, as the main source countries (Section 3.2) jointly account for about 54% of all POIs.

Categories are drawn from a shared vocabulary of over 2,000 labels, with 2.17 labels per POI on average. The most frequent category, restaurants, covers 7.2% of the corpus, the top-5 categories cover 22.4%, and the top-20 cover 57.0%. The head is dominated by service businesses (restaurants, professional services, shopping), with civic and educational entities such as schools, community services, and places of worship further down. Table 9 in Appendix A reports the full top-20.

### 3.2 Query Generation

Our queries  $q = (t, c)$  as defined in Section 2.1 are derived from raw user queries  $\tilde{q} = (\tilde{t}, \tilde{c})$  collected from real internal POI search logs spanning multiple locales and languages. To address privacy concerns, since  $\tilde{t}$  may contain sensitive textual information (e.g., addresses or residential references) and  $\tilde{c}$  may identify specific households in low-density regions, we transform each  $\tilde{q}$  into a privacy-safe  $q$  by de-identifying both text and location. A prompted LLM (Qwen3-235B-A22B (Yang et al., 2025)) first filters out queries whose  $\tilde{t}$  lacks clear POI intent or contains sensitive content (personal identifiers, full addresses, or residential references), then rewrites each remaining  $\tilde{t}$  into a safe query text  $t$  while preserving intent and language. The coordinate  $\tilde{c}$  is independently de-identified to a nearby  $c$  that preserves geographic context while obscuring the original location. For additional safety, we discard any query whose  $\tilde{q}$  yields no results when issued to Google Maps, ensuring all retained  $q$  correspond to publicly retrievable geographic context; in practice, this removes about 0.5% of queries. Full filtering and de-identification details are provided in Appendix B. Finally, the LLM-based annotation step in Section 3.4 provides an additional safeguard, as queries referencing private information produce no valid Overture POIs and are filtered out during relevance labeling.

### 3.3 Retrieval and Ranking Candidates

For each query  $q$ , we assemble a candidate pool from  $\mathcal{C}$  that allows us to capture the most relevant POIs. In order to retrieve the best possible candidates, these are sourced through two complemen-

tary paths. The first path retrieves up to 20 highly relevant candidate POIs by querying Google Maps, then matches each returned name and coordinate pair to its corresponding POI in Overture Maps Places, where one exists (Appendix C). Through this path, a first portion of the candidate set consists of the best Overture Maps Places POIs according to a well-established search engine. The second path runs our retrieve-and-rerank system of Section 2.1 directly on  $\mathcal{C}$ , retrieving the top-300 candidates and reranking them. The retriever and reranker used are BGE-M3 and BGE-reranker-v2-M3 (Chen et al., 2024). We first label the candidates retrieved by the base models, then use these labels to train the retriever and reranker of our POI search system. We then use the trained system to generate a refined version of the dataset. This process is repeated for several iterations, with each cycle improving both the models and the resulting candidates. Direct retrieval from  $\mathcal{C}$  also recovers relevant POIs that may be missed by Google Maps matching due to naming differences or coordinate inconsistencies between Google Maps and Overture Maps Places.

The up to 20 candidates from the Google Maps path and the 300 retrieval candidates are all Overture POIs and therefore lack native quality signals. We enrich them with a quality score feature in the 0–5 range generated by Qwen3-235B-A22B from the available Overture POI metadata. This score is intended to provide a numerical signal for the design of ranking experiments in POI search, rather than to precisely estimate ground-truth real-world quality. Importantly, the same score is used consistently throughout the labeling pipeline, and this imperfect approximation of real-world quality does not diminish the research utility of the resulting labels. The reranker is then applied to the 300 enriched retrieval candidates to produce a short-list of 30, after which candidates from both paths are passed to the annotation pipeline described in Section 3.4.

The retrieve-and-rerank pipeline serves only as a candidate generation mechanism. All surfaced candidates undergo independent validity and relevance assessment through the LLM-based annotation process in Section 3.4. This decouples the final graded labels from any specific retrieval system, ensuring supervision reflects the LLM’s broader knowledge rather than the behavior of a particular search model.

### 3.4 Data Labeling

We label the candidates with the automatic annotator of [Kambhatla et al. \(2026\)](#), summarized below; the dataset we build with it is new to this paper. Several stages of the annotation pipeline delegate labeling decisions to an LLM used as an automated annotator. In the LLM-as-a-Judge (LLMaaJ) paradigm ([Zheng et al., 2023](#)), a model is prompted to assess the quality or graded relevance of a given input according to criteria specified in the prompt. Given the up to 20 candidates from the Google Maps path and the 30 from our POI search system, we annotate them with graded relevance labels by relying on the LLMaaJ paradigm to perform relevance validation, deduplication, and user preference ranking.

**Graded Relevance Judgment Generation.** We produce graded relevance labels by prompting Qwen3-235B-A22B to annotate the candidate pool. The model receives the set of candidates to be annotated, each described by name, address, categories, coordinates, and quality score, alongside the original query, and is instructed to: (1) filter out candidates that do not plausibly answer the query, i.e., those that do not match the business type it specifies or, when a geographic area is mentioned, do not fall within or close to it, while ignoring subjective qualifiers such as “best” or “closest”; (2) collapse near-duplicates based on name, address, geographic proximity, phone number, and quality score, retaining the most complete entry per group; and (3) rank without ties the remaining candidates in a listwise fashion by estimated user preference, considering semantic match to the query intent, geographic proximity, and quality score. The top-20 in the final ordering constitute the ground-truth list, with rank position serving as the graded relevance labels.

**Alignment with Human Judgments.** We validated each stage of the annotation pipeline through agreement studies with trained human annotators on stratified query samples. This evaluation followed an iterative calibration process rather than a formal benchmark protocol: disagreements were manually reviewed, prompts were refined accordingly, and the evaluator was calibrated to better align with expert judgments. We employed a dataset comprising 161 queries associated with approximately 3,000 POIs ( $\mu = 18.4$  POIs,  $\sigma = 8.3$ ). The validation set was annotated by a domain ex-

| Statistic                   | Value      |
|-----------------------------|------------|
| Queries                     | 309,386    |
| Train                       | 267,368    |
| Test                        | 42,018     |
| Candidate query–POI pairs   | 94,416,318 |
| Relevant                    | 4,363,559  |
| Non-relevant                | 90,052,759 |
| Avg. candidates per query   | 305.2      |
| Avg. relevant per query     | 14.1       |
| Avg. non-relevant per query | 291.1      |

Table 1: Overall dataset scale.

pert with extensive experience in POI evaluation, following detailed guidelines covering POI validity, location, intent matching, proximity, and quality. The filtering stage achieved a macro-F1 of 92.9% and a Cohen’s  $\kappa$  of 0.86, indicating strong agreement with human judgments. The deduplication stage reached an F1 of 97.2%, demonstrating accurate identification of duplicate groups while retaining the most informative entries. For listwise ranking, agreement with human-produced rankings reached 91.8% NDCG under binary relevance and 80.2% under graded relevance. Across all stages, model performance was comparable to human judges, supporting the reliability of the pipeline. Disagreements between humans and LLM judges mainly occur in inherently ambiguous cases, such as co-located POIs, missing metadata, or distant but category-matching results. We refer the reader to [Kambhatla et al. \(2026\)](#), which also reports this study, for the detailed methodology and analysis underlying it.

## 4 Dataset Analysis

In this section, we report and analyze statistics of the final dataset, covering its overall scale, query characteristics, geographic and linguistic distribution, and category composition.

**Overall scale.** Table 1 summarizes the dataset scale, comprising over 4M query–POI pairs with graded relevance labels drawn from 90M candidates, split into 267,368 training and 42,018 test queries. The split is created via stratified sampling over POI categories and intent groups, preserving near-identical category, intent, and language distributions across train and test, with minor deviations in low-resource locales (see Appendix D for full breakdowns). Candidate retrieval and relevance annotation are performed identically for both splits. The training set supports the development of both the retriever and reranker models. Specifically, we

treat as relevant the top-20 candidate POIs that were assigned non-zero graded relevance labels by our LLMaaJ framework, while all remaining candidates are treated as irrelevant. Although this strategy yields a relatively noisy set of negative examples, it is consistent with prior work that uses top-ranked retrieved documents as hard negatives during training (Garg et al., 2020; Zhan et al., 2021; Xiong et al., 2020), starting with the ASNQ corpus, where unlabeled candidates from the retrieved document are graded by confusability and used as negatives. POI-level overlap between the two splits is expected rather than accidental: both draw their candidates from a single Overture Maps Places snapshot, and popular POIs are relevant to many distinct queries. Concretely, 56.1% of the POIs relevant to a test query also appear among the relevant POIs of some training query, while 12.1% of the training-relevant POIs recur in test; the two relevant-POI sets have a Jaccard similarity of 0.11, computed over 490,212 test and 2,279,801 training POIs. Our splits enforce disjointness at the *query* level: a test query is removed if the same query text was issued within 0.5 km of its test coordinate in training, so no test query–POI pair is ever observed during training. In addition, for 18.1% of test queries none of the ground-truth POIs appears in any training query, which isolates a subset on which retrieval and ranking are evaluated over entirely unseen entities.

For retriever training, the relevant candidate POIs serve as positive examples, whereas negatives are sampled from the non-relevant candidate POIs within each query-specific candidate set. For reranker training, the graded relevance annotations can be leveraged directly to optimize fine-grained ranking objectives, or alternatively binarized into positive–negative labels in the same manner as for the retriever. The same test set serves to evaluate both components: for the retriever, measuring how many relevant POIs are surfaced from the corpus; for the reranker, assessing the ranking quality over the retrieved relevant POIs.

**Query Length Distribution.** Queries are short, with a mean length of 3.26 tokens, a median of 3, and a maximum of 14. The distribution is heavily concentrated in the 2–4 token range: only 11.2% of queries contain five or more tokens, and queries exceeding seven tokens are exceedingly rare (<1%). The full histogram is shown in Figure 2 (Appendix D).

**Geographic and linguistic distribution.** The dataset covers 7 languages across 14 (country, language) locales, split almost evenly between the Americas and Europe, with the remaining queries spreading across over 100 additional countries. English dominates (50.2%), followed by Spanish (15.3%), French (8.9%), German (8.0%), Portuguese (7.3%), Italian (6.4%), and Dutch (2.3%). Crucially, language and geography are decoupled: four languages span multiple countries and two countries contribute queries in more than one language, preventing models from exploiting spurious locale–language correlations (full breakdown in Appendix D).

**Query Categories.** Each query is annotated with a POI category and a use-case label reflecting the user’s search intent. The dataset spans 22 POI categories, with RESTAURANTS & CAFES accounting for the largest share (27.4%), followed by SHOPPING MALLS & STORES (12.6%) and HOTELS & LODGING (6.8%) (full breakdown in Appendix D).

**Query Intents.** We group queries into four high-level intents based on whether they include an explicit *name anchor*. *Name-anchored* queries target a specific POI by name: SEARCH (50.4%) for lookup (e.g., “Starbucks Times Square”) and DETAIL (0.4%) for additional information such as hours or pricing (e.g., “IKEA opening hours”). *Open-ended* queries seek places or activities based on category, attributes, or context: RECOMMEND (47.5%) for category-based discovery (e.g., “Italian restaurants near me”) and THINGS-TO-DO (1.7%) for activity suggestions (e.g., “fun things to do in Barcelona”). Name-anchored intents provide strong lexical cues, while open-ended intents require reasoning over entity attributes and geographic context (see Appendix D for a full breakdown).

## 5 Experiments

Our experimental study aims to evaluate the utility of the proposed dataset for advancing research on POI search systems. We conduct experiments by training on the dataset’s train split and evaluating on the test split, using the graded relevance labels of valid POIs as ground truth. We fine-tune neural retrieval and reranking models on the train split and measure system performance on each of these two stages in turn. A validation set comprising 5% of the train queries, held out at random, is

used for model selection. All fine-tuned models are trained on  $8 \times$  NVIDIA A100 40GB GPUs. Full hyperparameters, optimizer settings, and early-stopping criteria are reported in Appendix F.

## 5.1 Retrieval Experiments

In the *full-corpus retrieval* task, a system retrieves relevant POIs directly from the full Overture Maps Places index of over 72 million entities given a query and its issuing coordinate, measuring how effectively it can surface relevant POIs from a large corpus.

### 5.1.1 Retrieval Setup

**Representation:** Queries are represented by concatenating the query text with a textual encoding of the issuing coordinate; POIs by concatenating name, address, coordinate, and categories (full input in Appendix E). We also evaluated two alternative approaches for encoding geographic information when fine-tuning BGE-M3: a geohash-based representation of both user and POI coordinates<sup>1</sup> and a toponym-based representation of the user’s geographic information using their city, country, region, and ZIP code. The coordinate-based representation achieved the highest overall recall and was therefore adopted for the final retrieval setup. Detailed comparisons with these alternative geographic representations are provided in Appendix H.

#### Models and Baselines:

**Distance**, a baseline that ranks candidates purely by distance to the issuing coordinate, providing a lower bound for any content-aware system.

**BM25** (Robertson and Zaragoza, 2009) used as a training-free lexical baseline. All text is tokenized with Pyserini’s (Lin et al., 2021) Lucene analyzer (Porter stemming, stopword removal) and scored with BM25 ( $k_1=0.9$ ,  $b=0.4$ ). We additionally report a **BM25 + Distance** variant that retrieves the top-1,000 candidates by BM25 and re-scores them via linear interpolation of the per-query BM25 score  $\hat{s}_{\text{BM25}}$  with a normalized inverse-distance score  $\hat{s}_{\text{dist}} = 1/(1+d)$ , where  $d$  is the haversine distance in km between the issuing coordinate and the POI:  $s_{\text{final}} = \alpha \hat{s}_{\text{BM25}} + (1-\alpha) \hat{s}_{\text{dist}}$ , with  $\alpha \in [0, 1]$  tuned on the validation set.

**Multilingual bi-encoders** of two types: BGE-M3 (Chen et al., 2024), built on the encoder-based model XLM-RoBERTa (Conneau et al., 2020),

| System         | R@10 | R@100 | R@1k | N@5  | N@20 |
|----------------|------|-------|------|------|------|
| Distance       | 0.6  | 2.4   | 10.2 | 0.5  | 0.7  |
| BM25           | 17.8 | 29.9  | 42.8 | 16.0 | 18.0 |
| BM25 + Dist.   | 18.4 | 31.2  | 42.8 | 17.1 | 19.1 |
| BGE-M3 (ZS)    | 15.2 | 26.4  | 40.0 | 13.7 | 15.3 |
| BGE-M3 (FT)    | 45.2 | 83.2  | 95.6 | 58.3 | 60.3 |
| Qwen3-Emb (ZS) | 22.2 | 35.3  | 49.7 | 22.6 | 24.1 |
| Qwen3-Emb (FT) | 45.0 | 82.5  | 94.7 | 58.7 | 60.7 |

Table 2: Retrieval results (test split).

and the 0.6B-parameter Qwen3-Embedding (Zhang et al., 2025), a decoder-based model with instruction-aware prompting. Both produce single  $\ell_2$ -normalized embeddings used for dense retrieval, and we report zero-shot (ZS) and fine-tuned (FT) variants. POI embeddings are precomputed offline and indexed with FAISS (Johnson et al., 2021) for nearest-neighbor search.

**Training details:** Fine-tuned bi-encoders are trained with InfoNCE loss (van den Oord et al., 2018) and in-batch negatives, augmented with 100 hard negatives per query mined by combining LLM-judged non-relevant candidates from the POI search system path (Section 3.3) with lower-ranked candidates from the cross-encoder shortlist, which are semantically related to the query but typically exhibit lower quality score, greater distances, or fewer reviews.

**Evaluation:** We report Recall@ $k$  ( $R@k$ ) for  $k \in \{10, 100, 1,000\}$  and NDCG@ $k$  ( $N@k$ ) for  $k \in \{5, 20\}$ . All values are reported as percentages.

### 5.1.2 Retrieval Results and Analysis

Table 2 reports Recall@ $k$  and NDCG@ $k$  for retrieval. The Distance baseline performs poorly, confirming that proximity alone is insufficient. BM25 consistently outperforms zero-shot BGE-M3, suggesting that many POI queries remain strongly lexical and benefit from exact matching. Zero-shot Qwen3-Embedding performs best among untuned models, likely due to its larger scale and more recent training. Incorporating distance into BM25 provides modest gains, indicating that geographic proximity acts as a useful secondary signal. After fine-tuning, BGE-M3 and Qwen3-Embedding substantially outperform all baselines and converge to nearly identical performance. We therefore use the smaller and faster fine-tuned BGE-M3 model to generate reranking candidates in Section 5.2.

Table 3 shows a pronounced gap between name-anchored intents (SEARCH, DETAIL) and

<sup>1</sup>Leveraging <https://pypi.org/project/geohash2/>.

| Model       | Metric | Search | Rec  | TTD  | Detail |
|-------------|--------|--------|------|------|--------|
| BM25        | R@100  | 50.5   | 9.0  | 9.5  | 58.4   |
|             | N@20   | 31.2   | 4.4  | 5.3  | 40.6   |
| BGE-M3 (ZS) | R@100  | 45.4   | 7.3  | 7.3  | 55.2   |
|             | N@20   | 27.1   | 3.4  | 3.6  | 33.4   |
| BGE-M3 (FT) | R@100  | 89.1   | 77.3 | 74.0 | 88.0   |
|             | N@20   | 69.0   | 51.6 | 48.7 | 71.7   |

Table 3: Per-intent retrieval performance (test split). Rec: Recommend, TTD: Things-to-do.

| Model       | Metric | EN   | ES   | FR   | PT   | DE   | IT   | NL   |
|-------------|--------|------|------|------|------|------|------|------|
| BM25        | R@100  | 25.0 | 32.8 | 36.4 | 32.5 | 31.4 | 43.6 | 42.6 |
|             | N@20   | 13.7 | 20.0 | 22.7 | 19.2 | 22.1 | 30.2 | 31.4 |
| BGE-M3 (ZS) | R@100  | 21.3 | 27.5 | 32.7 | 28.8 | 35.5 | 37.7 | 47.3 |
|             | N@20   | 11.9 | 15.2 | 19.3 | 16.2 | 23.3 | 23.7 | 30.9 |
| BGE-M3 (FT) | R@100  | 82.4 | 83.9 | 85.1 | 80.5 | 86.1 | 84.7 | 79.8 |
|             | N@20   | 58.8 | 61.1 | 63.9 | 57.4 | 64.2 | 64.4 | 58.5 |

Table 4: Per-language retrieval performance (test split).

open-ended ones (RECOMMEND, THINGS-TO-DO). BM25 achieves 50.5% R@100 on SEARCH but only 9.0% on RECOMMEND; BGE-M3 exhibits a similar pattern. Name-anchored queries benefit from direct lexical cues, whereas open-ended queries require matching intent to entity attributes and context. Fine-tuning narrows this gap but does not eliminate it, indicating that open-ended retrieval remains substantially harder due to the need to distinguish among many plausible candidates without access to the quality score.

Table 4 analyzes results by language. Zero-shot performance varies considerably across languages, with English and Spanish proving most difficult due to greater geographic ambiguity and a higher prevalence of RECOMMEND queries. Dutch and Italian perform better largely because of more favorable intent distributions, although the Dutch sample size is small. Fine-tuning substantially reduces cross-language variation, shrinking the R@100 gap from  $2.2\times$  to  $1.08\times$  and similarly improving NDCG. This suggests that the training signal generalizes effectively across languages despite an English-heavy dataset, and that gains cannot be explained solely by proper-name matching.

## 5.2 Reranking Experiments

In the *candidate reranking* task, a system reorders a per-query candidate set to match the ground-truth graded relevance labels, measuring its ability to discriminate fine-grained relevance within a fixed pool.

### 5.2.1 Reranking setup

For each query, the pool is the top-1,000 POIs retrieved from the full Overture Places index by the fine-tuned BGE-M3 retriever (Section 5.1), with each candidate enriched by the quality score feature described in Section 3.3; rerankers thus have access to quality signals at scoring time. Since the pool is produced by the retriever, reranking scores depend jointly on retrieval Recall@1,000 and the reranker’s ordering of the returned candidates; graded POIs missed by the retriever contribute zero to reranking precision and NDCG. Unlike the retriever, which must encode queries and POIs independently to index the 72M-entity corpus offline, the per-query pool here is small enough to afford a cross-encoder that jointly encodes each query–candidate pair and exploits *joint* features such as the geographic distance between the issuing coordinate and the POI. Concretely, query–candidate pairs are encoded jointly by concatenating the query text with the candidate’s name, categories, address, and the quality score and distance signals, the latter two passed as natural-language quantization buckets (full input in Appendix E).

**Baselines:** We evaluate the cross-encoder counterparts of the two retrieval models: BGE-Reranker-v2-m3, the reranking variant of BGE-M3; and the 0.6B-parameter Qwen3-Reranker, the reranking variant of Qwen3-Embedding. Both jointly encode each query–candidate pair. For each pair, we rely solely on the cross-encoder’s logit score for reranking, discarding the retrieval-stage ranking score. For each model, we evaluate zero-shot (ZS) and fine-tuned (FT) variants. We also report the ranking induced by our BGE-M3 retriever’s own scores.

**Training details:** Fine-tuned cross-encoders are trained with binary cross-entropy on the train split, using the same hard-negative mining strategy as the bi-encoders (Section 5.1.1).

**Metrics:** We report Precision@ $k$  ( $P@k$ ) and NDCG@ $k$  ( $N@k$ ) for  $k \in \{5, 20\}$ , and Mean Reciprocal Rank (MRR). All values are reported as percentages.

### 5.2.2 Reranking Results and Analysis

Table 5 reports Precision@ $k$ , MRR, and NDCG@ $k$  for the candidate reranking task. Our fine-tuned cross-encoders improve over the retriever’s own ranking across all metrics, confirming the dataset

| System           | P@5  | P@20 | MRR  | N@5  | N@20 |
|------------------|------|------|------|------|------|
| BGE-M3 retriever | 57.6 | 35.5 | 85.5 | 61.5 | 62.6 |
| BGE-RR (ZS)      | 24.6 | 14.3 | 50.0 | 32.1 | 34.8 |
| BGE-RR (FT)      | 63.0 | 38.3 | 89.7 | 66.8 | 66.6 |
| Qwen3-RR (ZS)    | 38.3 | 21.3 | 68.3 | 43.8 | 44.6 |
| Qwen3-RR (FT)    | 61.8 | 37.4 | 88.8 | 65.7 | 65.7 |

Table 5: Reranking results (test split). BGE-RR: BGE-Reranker-v2-m3. Qwen3-RR: Qwen3-Reranker-0.6B.

enables meaningful research on the reranking stage of POI search systems.

As on the retrieval track, fine-tuned Qwen3 and BGE reranker variants perform comparably, indicating that the supervision transfers across architectures. To assess statistical reliability, we perform paired significance tests over the 42,018 test queries with Holm–Bonferroni correction for multiple comparisons (Holm, 1979). The fine-tuned BGE-Reranker significantly outperforms the BGE-M3 retriever on all five metrics (all  $p < 10^{-199}$  after correction), with per-query win rates ranging from 59.9% to 65.3% and small effect sizes ( $d_z = 0.149\text{--}0.229$ ). Given the large test set, statistical significance is readily attained; we therefore emphasize win rates and effect sizes. Full per-metric results are reported in Appendix J.

The intent-level picture (Table 6) is shaped by the size of the ground-truth relevant set, which varies systematically across intents: name-anchored SEARCH and DETAIL queries typically carry fewer valid POIs than RECOMMEND and THINGS-TO-DO queries, which often saturate the 20-POI cap (see Appendix D). The higher  $P@k$  on RECOMMEND versus SEARCH for the fine-tuned BGE-Reranker mostly reflects this. NDCG exposes the opposite gap:  $N@5$  drops 16.4% from SEARCH to RECOMMEND and  $N@20$  drops 20.5%. The reranker reliably surfaces a relevant POI at the very top for open-ended queries, which is consistent with the strong MRR, but reproducing the ground-truth *ordering*, which requires distinguishing candidates of comparable semantic relevance by quality and proximity, remains the harder challenge.

The retrieval-side ordering carries over to reranking (Table 26 in Appendix I): German and Italian remain strongest, English and Portuguese weakest on rank-aware metrics. All seven languages improve on every metric, and the cross-language gap narrows on  $P@5$  ( $1.65\times \rightarrow 1.57\times$ ),  $P@20$  ( $2.00\times \rightarrow 1.78\times$ ), and  $N@5$  ( $1.14\times \rightarrow 1.11\times$ ), while  $N@20$  stays comparable ( $1.13\times \rightarrow 1.14\times$ ): the reranker benefits all languages roughly uni-

| Model              | Metric | Search | Rec  | TTD  | Detail |
|--------------------|--------|--------|------|------|--------|
| BGE-M3 (retriever) | P@5    | 50.8   | 64.8 | 61.1 | 42.1   |
|                    | P@20   | 28.6   | 42.7 | 38.8 | 21.1   |
|                    | N@5    | 70.0   | 53.0 | 50.4 | 73.1   |
|                    | N@20   | 72.3   | 53.0 | 49.9 | 75.5   |
| BGE-Reranker (FT)  | P@5    | 53.8   | 72.5 | 70.0 | 43.9   |
|                    | P@20   | 30.9   | 45.9 | 44.1 | 22.4   |
|                    | N@5    | 75.0   | 58.6 | 57.3 | 76.5   |
|                    | N@20   | 76.8   | 56.3 | 55.0 | 78.9   |

Table 6: Per-intent reranking performance (test split). Rec: Recommend, TTD: Things-to-do.

formly despite an English-heavy training distribution.

## 6 Conclusion

We presented POI Search Suite (POISS), a large-scale open benchmark for POI search, pairing over 300K geo-anchored multilingual queries with more than 4M graded relevance labels over 90M candidates from the Overture Maps Places corpus, and supporting both full-corpus retrieval and reranking evaluation. To our knowledge, this is the first public dataset at this scale that combines a large POI corpus with natural language queries and graded relevance labels, addressing a long-standing gap in POI search research. We release queries, graded labels, and reranking candidate sets under the Apache 2.0 license rather than a pre-built index, which can be built directly from Overture Maps Places; since reranking pools contain all annotated relevant POIs, researchers using custom indices can treat any retrieved POI absent from the labeled set as non-relevant.

Fine-tuning substantially improves full-corpus retrieval, reaching  $\sim 82.8$  R@100, while cross-encoders further improve reranking by leveraging quality and query–candidate interactions, reaching  $\sim 89.0$  MRR and  $\sim 66.0$  NDCG@20. Gains are consistent across model families and languages despite an English-heavy training distribution, showing that the graded labels provide a strong learning signal while leaving meaningful headroom. The largest remaining challenge is open-ended RECOMMEND and THINGS-TO-DO queries, where many candidates are plausible and ranking requires subtle proximity and quality signals.

## Limitations

Dataset labeling relied on an LLM judge rather than direct human annotation. The annotation pipeline, including validation, deduplication, and ranking,

was calibrated against trained human annotators using a stratified sample. The full calibration study is described in a companion paper currently in preparation. Nevertheless, known biases in LLM-based evaluation, including positional and verbosity effects (Zheng et al., 2023; Wang et al., 2023), may still propagate into the resulting labels.

Locale coverage is also uneven. The dataset includes 7 languages across 14 primary locales, with representation concentrated in the Americas (50.7%) and Europe (47.6%). There is no coverage from Asia, Africa, or the Middle East, no inclusion of low-resource languages, and English accounts for approximately half of all queries. In addition, several per-locale test splits (e.g., Dutch,  $n=193$ ) are sufficiently small that estimates at the language level may exhibit non-negligible variance. Consequently, conclusions regarding cross-lingual generalization should be interpreted with these distributional imbalances in mind.

A limitation of our evaluation is that we do not report comparisons with proprietary, real-world services such as Google Maps or Yelp. Although we conducted these evaluations internally, we are unable to report the corresponding results due to potential internal restrictions. We therefore omit these results from the paper.

## Ethical Considerations

The query set is derived from real POI search logs. As detailed in Section 3.2 and Appendix B, we apply a multi-stage privacy pipeline before any query enters the released dataset: an LLM filter discards queries containing personal identifiers, full street addresses, or residential references; remaining query texts are rewritten to remove any residual identifying information while preserving intent and language; and issuing coordinates are perturbed so that no released coordinate can be traced back to a specific household or user. Queries that yield no public results when issued to Google Maps are dropped, and the downstream LLM relevance annotation provides an additional safeguard, since queries referencing private locations produce no valid Overture POIs and are filtered out during labeling. Despite these safeguards, no de-identification process is perfect, and we encourage users of the dataset to treat queries as derived from real user behavior and to refrain from attempting to re-identify or re-link them to other sources.

## References

- Zhengwei Bai, Moreno D’Incà, Alessandro Moschitti, and Danielle Class. 2026. [H2ce: Modeling geo-semantic interactions for POI reranking with heterogeneous two-stage cross-encoders](#). In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management*, Rome, Italy.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. [Ms marco: A human generated machine reading comprehension dataset](#). Preprint, arXiv:1611.09268.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jian Cui, Zhiyuan Ren, Desheng Weng, Yongqi Zhao, Gong Wenbin, Yu Lei, and Zhenning Dong. 2026. [ReaGeo: Reasoning-enhanced end-to-end geocoding with LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 28012–28023, San Diego, California, United States. Association for Computational Linguistics.
- Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. [TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7780–7788.
- Fredric Gey, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, and Vivien Petras. 2006. Geoclef: The clef 2005 cross-language geographic information retrieval track overview. In *Accessing Multilingual Information Repositories*, pages 908–919, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Milan Gritta, Mohammad Taher Pilehvar, Nut Limsoopatham, and Nigel Collier. 2018. [What’s missing in geographical parsing?](#) *Language Resources and Evaluation*, 52(2):603–623.
- Mordechai (Muki) Haklay and Patrick Weber. 2008. [Openstreetmap: User-generated street maps](#). *IEEE Pervasive Computing*, 7(4):12–18.

- Rujun Han, Luca Soldaini, and Alessandro Moschitti. 2021. [Modeling context in answer sentence selection systems on a latency budget](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3005–3010, Online. Association for Computational Linguistics.
- Hang He, Chuhuai Yue, Chengqi Dong, Mingxue Tian, Hao Chen, Zhenfeng Liu, Jiajun Chai, Xiaohan Wang, Yufei Zhang, Qun Liao, Guojun Yin, Wei Lin, Chengcheng Wan, Haiying Sun, and Ting Su. 2026. [Localsearchbench: Benchmarking agentic search in real-world local life services](#). *Preprint*, arXiv:2512.07436.
- Sture Holm. 1979. [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.
- Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. [Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring](#). In *8th International Conference on Learning Representations (ICLR)*.
- Jeff Johnson, Matthijs Douze, and Herve Jegou. 2021. [Billion-Scale Similarity Search with GPUs](#). *IEEE Transactions on Big Data*, 7(03):535–547.
- Gauri Kambhatla, Xiaokun Zhou, and Alessandro Moschitti. 2026. Automatic construction of ranked ground truth for poi retrieval evaluation using large language models. *arXiv preprint*.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Ivano Lauriola and Alessandro Moschitti. 2021. [Answer sentence selection using local and global context in transformer models](#). In *Advances in Information Retrieval (ECIR)*, volume 12656 of *Lecture Notes in Computer Science*, pages 298–312. Springer.
- Phong Le and Ivan Titov. 2018. [Improving entity linking by modeling latent relations between mentions](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.
- Dongyang Li, Ruixue Ding, Qiang Zhang, Zheng Li, Boli Chen, Pengjun Xie, Yao Xu, Xin Li, Ning Guo, Fei Huang, and Xiaofeng He. 2025. Geoglu: A chinese geographic language understanding evaluation benchmark. In *Advanced Data Mining and Applications*, pages 198–212, Singapore. Springer Nature Singapore.
- Defu Lian, Yongji Wu, Yong Ge, Xing Xie, and En-hong Chen. 2020. [Geography-aware sequential location recommendation](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, page 2009–2019, New York, NY, USA. Association for Computing Machinery.
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.
- Ilya Loshchilov and Frank Hutter. 2017. [Fixing weight decay regularization in adam](#). *CoRR*, abs/1711.05101.
- Gengchen Mai, Krzysztof Janowicz, Cheng He, Sumang Liu, and Ni Lao. 2018. [Poireviewqa: A semantically enriched poi retrieval and question answering dataset](#). In *Proceedings of the 12th Workshop on Geographic Information Retrieval, GIR'18*, New York, NY, USA. Association for Computing Machinery.
- Yoshitomo Matsubara, Thuy Vu, and Alessandro Moschitti. 2020. [Reranking for efficient transformer-based answer selection](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1577–1580. Association for Computing Machinery.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with bert](#). *arXiv preprint arXiv:1901.04085*.
- Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. [ReLiK: Retrieve and LiK, fast and accurate entity linking and relation extraction on an academic budget](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.
- Overture Maps Foundation. 2026. Overture maps foundation places dataset. <https://overturemaps.org>. Data release 2026-03-18, accessed March 19, 2026.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Foundations and Trends in Information Retrieval*, 4(1-2):1–174.

- Avijit Shah, Srini Dwarakanathan, Yash Kamdar, Danielle Class, and Alessandro Moschitti. 2026. Omnium: Multilingual query-to-location understanding model for local search. In *Proceedings of EMNLP (Industry Track)*.
- Luca Soldaini and Alessandro Moschitti. 2020. [The cascade transformer: an application for efficient answer sentence selection](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5697–5708, Online. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. [Is ChatGPT a good NLG evaluator? a preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. [Approximate nearest neighbor negative contrastive learning for dense text retrieval](#). *Preprint*, arXiv:2007.00808.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudre-Mauroux. 2019. [Revisiting user mobility and social relationships in lbsns: A hypergraph embedding approach](#). In *The World Wide Web Conference, WWW '19*, page 2147–2157, New York, NY, USA. Association for Computing Machinery.
- Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. [Pretrained transformers for text ranking: BERT and beyond](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.
- Yelp Inc. 2023. Yelp open dataset. <https://www.yelp.com/dataset>.
- Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. [Optimizing dense retrieval model training with hard negatives](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 1503–1512, New York, NY, USA. Association for Computing Machinery.
- Lei Zhang, Runze Wang, Jingbo Zhou, Jingsong Yu, Zhenhua Ling, and Hui Xiong. 2020. [Joint intent detection and entity linking on spatial domain queries](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4937–4947, Online. Association for Computational Linguistics.
- Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. [MIRACL: A multilingual retrieval dataset covering 18 diverse languages](#). *Transactions of the Association for Computational Linguistics*, 11:1114–1131.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *ArXiv*, abs/2506.05176.
- Zeyu Zhang, Thuy Vu, and Alessandro Moschitti. 2021. [Joint models for answer verification in question answering systems](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3252–3262, Online. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Xiaokun Zhou, Alessandro Moschitti, and Danielle Class. 2026. [Rank-deviation quality: A distance-aware metric for multi-answer retrieval and ranking evaluation](#). *Preprint*, arXiv:2608.25318.

## A Overture Maps Places Detailed Statistics

This appendix reports the geographic distribution of the Overture Maps Places corpus (version 2026-03-18) summarized in Section 3.1. Table 7 lists the 20 countries with the largest POI counts, which together cover roughly 79% of the corpus. Table 8 reports the distribution by continent.

| Country        | POIs       | %    |
|----------------|------------|------|
| United States  | 15,561,168 | 21.4 |
| Brazil         | 5,192,697  | 7.1  |
| India          | 4,071,204  | 5.6  |
| United Kingdom | 3,384,346  | 4.6  |
| Germany        | 3,213,189  | 4.4  |
| Mexico         | 2,902,728  | 4.0  |
| Japan          | 2,748,605  | 3.8  |
| Italy          | 2,443,389  | 3.4  |
| France         | 2,230,836  | 3.1  |
| Thailand       | 1,921,069  | 2.6  |
| Indonesia      | 1,883,496  | 2.6  |
| Turkey         | 1,740,215  | 2.4  |
| Canada         | 1,653,878  | 2.3  |
| Spain          | 1,641,989  | 2.3  |
| Vietnam        | 1,386,755  | 1.9  |
| Australia      | 1,349,438  | 1.9  |
| Poland         | 1,105,930  | 1.5  |
| Philippines    | 1,053,310  | 1.4  |
| Taiwan         | 882,624    | 1.2  |
| Argentina      | 829,127    | 1.1  |
| Top-20 total   | 57,195,993 | 78.6 |

Table 7: Top-20 countries in the Overture Maps Places corpus by number of POIs (version 2026-03-18). The remaining countries and territories together account for the other  $\sim 21\%$  of the corpus.

| Region          | Share (%) |
|-----------------|-----------|
| Europe          | 29.4      |
| North America   | 28.6      |
| Asia            | 26.4      |
| South America   | 10.9      |
| Africa          | 2.5       |
| Oceania         | 2.2       |
| Other / Unknown | <0.1      |

Table 8: Geographic distribution of Overture Maps Places POIs by continent (version 2026-03-18). “Other / Unknown” covers Antarctic territories and entries with missing country codes.

## B Query Generation Details

This appendix provides full details of the query generation pipeline summarized in Section 3.2. Importantly, all reference search engine calls described here are performed at the *original* issuing coordinates and are used solely for privacy de-identification and filtering. The candidate retrieval described in Section 3.3 is a separate process that operates on the final, de-identified coordinates.

### B.1 Source Queries

The starting point for our query set is a collection of real user queries  $\tilde{q} = (\tilde{t}, \tilde{c})$  for POI retrieval sourced internally. The queries span multiple locales and languages, covering regions across the Americas, Europe, and Asia, reflecting the geographic breadth of the Overture Maps dataset on

| Category                         | POIs       | %     |
|----------------------------------|------------|-------|
| restaurant                       | 5,252,879  | 7.22  |
| professional_services            | 3,126,572  | 4.30  |
| shopping                         | 3,021,670  | 4.15  |
| beauty_salon                     | 2,521,898  | 3.46  |
| health_and_medical               | 2,380,383  | 3.27  |
| hotel                            | 2,006,664  | 2.76  |
| beauty_and_spa                   | 1,999,582  | 2.75  |
| accommodation                    | 1,929,412  | 2.65  |
| bar                              | 1,873,903  | 2.57  |
| cafe                             | 1,771,059  | 2.43  |
| landmark_and_historical_building | 1,764,372  | 2.42  |
| clothing_store                   | 1,725,095  | 2.37  |
| school                           | 1,720,217  | 2.36  |
| retail                           | 1,613,221  | 2.22  |
| real_estate                      | 1,572,998  | 2.16  |
| education                        | 1,492,054  | 2.05  |
| community_services_non_profits   | 1,463,368  | 2.01  |
| church_cathedral                 | 1,461,625  | 2.01  |
| grocery_store                    | 1,424,347  | 1.96  |
| fast_food_restaurant             | 1,338,200  | 1.84  |
| Top-20 total                     | 41,459,519 | 56.96 |

Table 9: Top-20 categories in the Overture Maps Places corpus (version 2026-03-18), pooled across the primary and alternate category fields and deduplicated within each POI. Percentages report the share of POIs tagged with each category. Out of more than 2,000 distinct labels, the top 20 together tag nearly 57% of all POIs.

which the benchmark is grounded.

### B.2 Textual Filtering and De-identification

The filtering step uses an LLM that classifies each query text  $\tilde{t}$  as *safe* or *unsafe*. A query is marked unsafe if it contains private personal information (full names with surname, phone numbers, email addresses, personal IDs), private or residential locations, full street addresses, or any information that could identify the query author. Queries containing well-known public entities, generic business types, or well-known geographic references are retained.

In the de-identification step, the model generates a privacy-safe query text  $t$  that removes any identifying information while preserving the original language of  $\tilde{t}$  (e.g., a query issued in Italian is rewritten in Italian), ensuring that the linguistic diversity of the source data is maintained in the final dataset.

### B.3 Geographic De-identification

The perturbation applied to the issuing coordinate  $\tilde{c}$  depends on whether the query is location-dependent. For queries whose results are independent of the precise issuing location (e.g., “where is Central Park”), we obtain  $c$  by displacing  $\tilde{c}$  along a random bearing by a distance uniformly sampled in  $[0, 5 \text{ km}]$ . For location-dependent queries (e.g.,

“coffee shops near me”), we instead set  $c$  to a random convex combination of the coordinates of the top-5 results returned by querying the reference search engine with  $\tilde{q} = (\tilde{t}, \tilde{c})$  placing the perturbed point within the convex hull of relevant results to preserve geographic coherence while obscuring  $\tilde{c}$ . When fewer than two reference results are available, we fall back to the random offset.

### C POI Matching Details

Each result returned by the reference search engine Google Maps must be matched to its corresponding POI in the Overture Maps Places corpus, as the two sources use independent entity namespaces. We solve this with a two-stage procedure.

In the first stage, we retrieve the closest 100 Overture POIs within a radius of 100 m from the Google Maps result’s coordinate. In the second stage, among the retrieved candidates we select the one whose name is most semantically similar to the Google Maps POI name, as measured by cosine similarity of sentence-transformer embeddings (Reimers and Gurevych, 2019) using the all-MiniLM-L6-v2 model. A match is accepted only if the similarity exceeds a threshold of 0.6; otherwise the Google Maps POI is marked as unmatched.

### D Detailed Dataset Statistics

For each split (train and test), we report the country–language distribution of queries (Tables 11 and 12), the full POI category distribution across all 22 categories (Tables 13 and 15), and the intent group distribution (Tables 17 and 18). We also cross-tabulate language and intent, reporting the per-language intent distribution for all queries, train, and test (Tables 19, 20, and 21). Finally, we report per-intent ground-truth relevant-set size statistics for all queries, train, and test (Tables 22, 23, and 24).

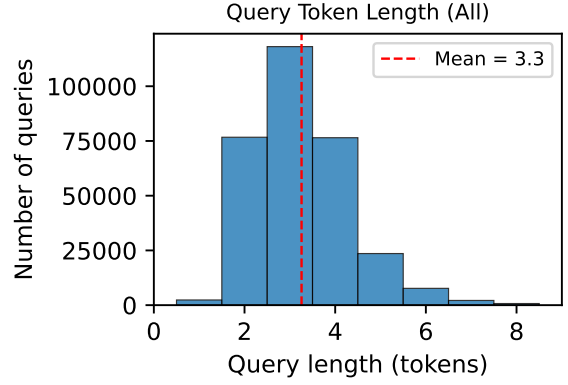


Figure 2: Distribution of query length in tokens. The dashed line indicates the mean (3.26 tokens).

| Country        | Language   | Queries | %     |
|----------------|------------|---------|-------|
| United States  | English    | 65,527  | 21.2  |
| United Kingdom | English    | 44,036  | 14.2  |
| Canada         | English    | 42,723  | 13.8  |
| France         | French     | 24,365  | 7.9   |
| Spain          | Spanish    | 24,219  | 7.8   |
| Brazil         | Portuguese | 22,530  | 7.3   |
| Germany        | German     | 22,190  | 7.2   |
| Mexico         | Spanish    | 20,750  | 6.7   |
| Italy          | Italian    | 19,684  | 6.4   |
| Netherlands    | Dutch      | 7,260   | 2.3   |
| Canada         | French     | 3,104   | 1.0   |
| Ireland        | English    | 3,080   | 1.0   |
| Austria        | German     | 2,519   | 0.8   |
| United States  | Spanish    | 2,409   | 0.8   |
| Others         |            | 4,990   | 1.6   |
| Total          |            | 309,386 | 100.0 |

Table 10: Distribution of queries by country and language (all queries).

| Country        | Language   | Queries | %     |
|----------------|------------|---------|-------|
| United States  | English    | 57,241  | 21.4  |
| United Kingdom | English    | 37,787  | 14.1  |
| Canada         | English    | 36,846  | 13.8  |
| Spain          | Spanish    | 20,848  | 7.8   |
| France         | French     | 20,774  | 7.8   |
| Germany        | German     | 19,311  | 7.2   |
| Brazil         | Portuguese | 18,835  | 7.0   |
| Mexico         | Spanish    | 17,548  | 6.6   |
| Italy          | Italian    | 16,717  | 6.3   |
| Netherlands    | Dutch      | 7,068   | 2.6   |
| Ireland        | English    | 2,648   | 1.0   |
| Canada         | French     | 2,641   | 1.0   |
| United States  | Spanish    | 2,289   | 0.9   |
| Austria        | German     | 2,177   | 0.8   |
| Others         |            | 4,638   | 1.7   |
| Total          |            | 267,368 | 100.0 |

Table 11: Distribution of queries by country and language (train).

| Country        | Language   | Queries | %     |
|----------------|------------|---------|-------|
| United States  | English    | 8,286   | 19.7  |
| United Kingdom | English    | 6,249   | 14.9  |
| Canada         | English    | 5,877   | 14.0  |
| Brazil         | Portuguese | 3,695   | 8.8   |
| France         | French     | 3,591   | 8.5   |
| Spain          | Spanish    | 3,371   | 8.0   |
| Mexico         | Spanish    | 3,202   | 7.6   |
| Italy          | Italian    | 2,967   | 7.1   |
| Germany        | German     | 2,879   | 6.9   |
| Canada         | French     | 463     | 1.1   |
| Ireland        | English    | 432     | 1.0   |
| Austria        | German     | 342     | 0.8   |
| Netherlands    | Dutch      | 192     | 0.5   |
| United States  | Spanish    | 120     | 0.3   |
| Others         |            | 352     | 0.8   |
| Total          |            | 42,018  | 100.0 |

Table 12: Distribution of queries by country and language (test).

| Category                          | Queries | %     |
|-----------------------------------|---------|-------|
| Restaurants & Cafes               | 73,898  | 27.6  |
| Shopping Malls & Stores           | 33,903  | 12.7  |
| Hotels & Lodging                  | 17,897  | 6.7   |
| Tourist Attractions               | 17,074  | 6.4   |
| Grocery & Supermarkets            | 13,973  | 5.2   |
| Services                          | 12,501  | 4.7   |
| Parks & Recreation Areas          | 11,160  | 4.2   |
| Entertainment & Event Venues      | 10,735  | 4.0   |
| Hospitals & Clinics               | 9,872   | 3.7   |
| Schools & Universities            | 9,608   | 3.6   |
| Transportation                    | 9,218   | 3.4   |
| Gyms & Fitness Centers            | 8,357   | 3.1   |
| Car Services                      | 6,754   | 2.5   |
| Hair Salons & Spas                | 6,361   | 2.4   |
| Government Offices & Services     | 5,434   | 2.0   |
| Churches & Religious Institutions | 5,022   | 1.9   |
| Banks & ATMs                      | 3,843   | 1.4   |
| Pharmacies                        | 3,275   | 1.2   |
| Real Estate Listings              | 2,924   | 1.1   |
| Gas Stations                      | 2,158   | 0.8   |
| Pet Services                      | 1,772   | 0.7   |
| Libraries                         | 1,629   | 0.6   |
| Total                             | 267,368 | 100.0 |

Table 13: Distribution of queries by POI category (train).

| Category                          | Queries | %     |
|-----------------------------------|---------|-------|
| Restaurants & Cafes               | 84,843  | 27.4  |
| Shopping Malls & Stores           | 38,941  | 12.6  |
| Hotels & Lodging                  | 21,157  | 6.8   |
| Tourist Attractions               | 20,030  | 6.5   |
| Grocery & Supermarkets            | 15,918  | 5.1   |
| Services                          | 14,391  | 4.7   |
| Parks & Recreation Areas          | 13,084  | 4.2   |
| Entertainment & Event Venues      | 12,499  | 4.0   |
| Hospitals & Clinics               | 11,450  | 3.7   |
| Schools & Universities            | 11,268  | 3.6   |
| Transportation                    | 10,679  | 3.5   |
| Gyms & Fitness Centers            | 9,712   | 3.1   |
| Car Services                      | 7,854   | 2.5   |
| Hair Salons & Spas                | 7,309   | 2.4   |
| Government Offices & Services     | 6,248   | 2.0   |
| Churches & Religious Institutions | 5,933   | 1.9   |
| Banks & ATMs                      | 4,434   | 1.4   |
| Pharmacies                        | 3,703   | 1.2   |
| Real Estate Listings              | 3,421   | 1.1   |
| Gas Stations                      | 2,537   | 0.8   |
| Pet Services                      | 2,085   | 0.7   |
| Libraries                         | 1,890   | 0.6   |
| Total                             | 309,386 | 100.0 |

Table 14: Distribution of queries by POI category (all queries).

| Category                          | Queries | %     |
|-----------------------------------|---------|-------|
| Restaurants & Cafes               | 10,945  | 26.0  |
| Shopping Malls & Stores           | 5,038   | 12.0  |
| Hotels & Lodging                  | 3,260   | 7.8   |
| Tourist Attractions               | 2,956   | 7.0   |
| Grocery & Supermarkets            | 1,945   | 4.6   |
| Parks & Recreation Areas          | 1,924   | 4.6   |
| Services                          | 1,890   | 4.5   |
| Entertainment & Event Venues      | 1,764   | 4.2   |
| Schools & Universities            | 1,660   | 4.0   |
| Hospitals & Clinics               | 1,578   | 3.8   |
| Transportation                    | 1,461   | 3.5   |
| Gyms & Fitness Centers            | 1,355   | 3.2   |
| Car Services                      | 1,100   | 2.6   |
| Hair Salons & Spas                | 948     | 2.3   |
| Churches & Religious Institutions | 911     | 2.2   |
| Government Offices & Services     | 814     | 1.9   |
| Banks & ATMs                      | 591     | 1.4   |
| Real Estate Listings              | 497     | 1.2   |
| Pharmacies                        | 428     | 1.0   |
| Gas Stations                      | 379     | 0.9   |
| Pet Services                      | 313     | 0.7   |
| Libraries                         | 261     | 0.6   |
| Total                             | 42,018  | 100.0 |

Table 15: Distribution of queries by POI category (test).

| Type          | Intent group | Queries | %     |
|---------------|--------------|---------|-------|
| Name-anchored | Search       | 155,883 | 50.4  |
|               | Detail       | 1,196   | 0.4   |
| Open-ended    | Recommend    | 146,886 | 47.5  |
|               | Things-to-do | 5,421   | 1.8   |
| Total         |              | 309,386 | 100.0 |

Table 16: Distribution of queries by intent group (all queries).

| Type          | Intent group | Queries | %     |
|---------------|--------------|---------|-------|
| Name-anchored | Search       | 134,973 | 50.5  |
|               | Detail       | 989     | 0.4   |
| Open-ended    | Recommend    | 126,793 | 47.4  |
|               | Things-to-do | 4,613   | 1.7   |
| Total         |              | 267,368 | 100.0 |

Table 17: Distribution of queries by intent group (train).

| Type          | Intent group | Queries | %     |
|---------------|--------------|---------|-------|
| Name-anchored | Search       | 20,910  | 49.8  |
|               | Detail       | 207     | 0.5   |
| Open-ended    | Recommend    | 20,093  | 47.8  |
|               | Things-to-do | 808     | 1.9   |
| Total         |              | 42,018  | 100.0 |

Table 18: Distribution of queries by intent group (test).

| Intent group | Queries | Mean GT | Median GT |
|--------------|---------|---------|-----------|
| Search       | 155,883 | 10.0    | 8         |
| Recommend    | 146,886 | 18.4    | 20        |
| Things-to-do | 5,421   | 17.8    | 20        |
| Detail       | 1,196   | 7.9     | 4         |

Table 22: Per-intent ground-truth (GT) relevant-set size statistics (all queries). Each query has between 1 and 20 relevant POIs; *Mean GT* and *Median GT* report the mean and median number of relevant POIs per query within each intent group.

| Intent group | Queries | Mean GT | Median GT |
|--------------|---------|---------|-----------|
| Search       | 134,973 | 10.1    | 8         |
| Recommend    | 126,793 | 18.5    | 20        |
| Things-to-do | 4,613   | 17.8    | 20        |
| Detail       | 989     | 8.1     | 4         |

Table 23: Per-intent ground-truth (GT) relevant-set size statistics (train). Each query has between 1 and 20 relevant POIs; *Mean GT* and *Median GT* report the mean and median number of relevant POIs per query within each intent group.

| Intent group | Queries | Mean GT | Median GT |
|--------------|---------|---------|-----------|
| Search       | 20,910  | 9.5     | 6         |
| Recommend    | 20,093  | 18.1    | 20        |
| Things-to-do | 808     | 17.8    | 20        |
| Detail       | 207     | 6.9     | 3         |

Table 24: Per-intent ground-truth (GT) relevant-set size statistics (test). Each query has between 1 and 20 relevant POIs; *Mean GT* and *Median GT* report the mean and median number of relevant POIs per query within each intent group.

## E Model-Specific Prompting and Inputs

**Bi-encoder and BM25 inputs.** The two bi-encoders and the BM25 baseline operate on the same query and POI input representation, so that

lexical and dense retrievers see identical inputs. Queries are formatted as

```
User Query: {query}; Coordinate:
{lat}{N|S} {lon}{E|W};
```

and POIs as

```
Location Name: {POI name}; Address:
{street}; Position: {city}, {state},
{country}, {postal code}; Coordinate:
{lat}{N|S} {lon}{E|W}; Categories:
{categories};
```

Coordinates are written as decimal degrees with a hemisphere suffix at four-digit precision, giving the models a textual proxy for geographic proximity. Optional POI fields like individual address components are omitted when absent rather than left as empty placeholders.

For Qwen3-Embedding, queries are additionally prefixed with the task instruction

```
Instruct: Given a local search query
with user coordinates, retrieve relevant
points of interest
```

per the model’s instruction-tuning format; POI inputs receive no prefix. POI embeddings for both bi-encoders are precomputed offline and indexed with FAISS (Johnson et al., 2021) for nearest-neighbor search on the retrieval task.

**Cross-encoder template.** All cross-encoders represent each query–candidate with query text, POI name, quality score, and distance:

```
<Query>: {query}
<Document>: Name: {POI name};
Categories: {categories}; Distance:
{distance quantization}; Score: {score
quantization}; Address is {address}
```

Distance, quality score values are passed as natural language quantization buckets (e.g., distance is extremely close) so that BGE-Reranker-v2-m3 and Qwen3-Reranker, which were not trained on numeric features, can attend to them directly.

Qwen3-Reranker additionally requires its native instruction format, so the template above is prepended with

```
<Instruct>: Given a local search query,
retrieve relevant points of interest
that match the query. The document
may include distance, and quality score
information.
```

BGE-Reranker-v2-m3 receives the same query–document body without any instruction prefix.

| Language     | Search (%) | Recommend (%) | Things-to-do (%) | Detail (%) | Queries |
|--------------|------------|---------------|------------------|------------|---------|
| English      | 41.7       | 55.6          | 2.5              | 0.3        | 156,387 |
| Spanish      | 62.1       | 36.7          | 0.7              | 0.5        | 49,469  |
| French       | 57.0       | 41.3          | 1.3              | 0.4        | 27,623  |
| German       | 53.8       | 44.7          | 1.0              | 0.5        | 24,751  |
| Italian      | 65.7       | 32.9          | 0.8              | 0.5        | 19,872  |
| Dutch        | 59.7       | 37.2          | 2.3              | 0.8        | 8,501   |
| Portuguese   | 56.3       | 42.4          | 0.8              | 0.5        | 22,783  |
| <b>Total</b> | 50.4       | 47.5          | 1.8              | 0.4        | 309,386 |

Table 19: Per-language intent distribution as percentages (all queries). Rows sum to 100%.

| Language     | Search (%) | Recommend (%) | Things-to-do (%) | Detail (%) | Queries |
|--------------|------------|---------------|------------------|------------|---------|
| English      | 41.8       | 55.5          | 2.4              | 0.3        | 135,406 |
| Spanish      | 62.1       | 36.7          | 0.7              | 0.5        | 42,677  |
| French       | 56.9       | 41.5          | 1.2              | 0.4        | 23,542  |
| German       | 54.0       | 44.5          | 1.0              | 0.5        | 21,525  |
| Italian      | 65.8       | 32.9          | 0.8              | 0.5        | 16,874  |
| Dutch        | 59.7       | 37.4          | 2.3              | 0.7        | 8,308   |
| Portuguese   | 56.8       | 41.9          | 0.8              | 0.5        | 19,036  |
| <b>Total</b> | 50.5       | 47.4          | 1.7              | 0.4        | 267,368 |

Table 20: Per-language intent distribution as percentages (train). Rows sum to 100%.

## F Training Details

This appendix reports the full hyperparameters and optimization settings used to fine-tune the supervised baselines summarized in Section 5.1.2 and Section 5.2.1.

**Common setup.** All fine-tuned models are trained on the train split using AdamW (Loshchilov and Hutter, 2017) with gradient clipping (max norm 1.0), in mixed precision on  $8 \times$  NVIDIA A100 40 GB GPUs.

**Hard negative mining.** We mine 100 negatives per query by first including all candidates from the retrieval path (Section 3.3) that the LLM-as-a-Judge relevance validation labeled as invalid, then padding to 100 by sampling from the cross-encoder reranked top-300 starting at position 30. These lower-ranked POIs, while still semantically related to the query, typically exhibit lower quality score, greater distances from the issuing coordinate, making them suitable hard negatives for fine-grained relevance discrimination.

**Bi-encoders.** Trained with the InfoNCE loss (van den Oord et al., 2018) with a learnable temperature initialized at 0.01 and in-batch negatives enabled, lr  $2e-5$  on a linear schedule, per-GPU batch size 16 across 8 GPUs, and 4 hard negatives per query sampled from the mined set of 100, for up to 20 epochs with early stopping on validation MRR.

**Cross-encoders.** Trained with binary cross-entropy, lr  $1e-6$ , per-GPU batch size 4 across 8 GPUs, and 16 hard negatives per query sampled from the mined set of 100, for up to 20 epochs with the best checkpoint selected by validation NDCG@20.

**Compute budget.** Total GPU usage was approximately 90 A100-40 GB hours: 77 h for fine-tuning, BGE-M3 (5 h), Qwen3-Embedding (12 h), BGE-Reranker (20 h), and Qwen3-Reranker (40 h), and the remaining 13 h for inference and evaluation.

## G Graded Relevance for NDCG

For each query in our benchmark, the ground truth is an ordered list of up to 20 POIs, where rank reflects the combined criterion of semantic match, distance, quality score (Section 3.3). To reflect this ordering in the evaluation, we assign rank-derived graded relevance scores rather than binary labels. For a query with  $N$  ground-truth POIs ( $1 \leq N \leq 20$ ), the POI at rank  $i \in \{1, \dots, N\}$  in the ground truth receives gain  $g_i = N - i + 1$ , so the top-ranked POI scores  $N$ , the second  $N - 1$ , and so on down to 1 for the lowest-ranked. All other POIs (including graded items missed by the retriever in the reranking task) receive gain 0.

NDCG@ $k$  is then computed in its standard linear-gain form,

$$\text{DCG}@k = \sum_{j=1}^k \frac{g_{\pi(j)}}{\log_2(j+1)}, \quad (1)$$

| Language     | Search (%)  | Recommend (%) | Things-to-do (%) | Detail (%) | Queries       |
|--------------|-------------|---------------|------------------|------------|---------------|
| English      | 40.9        | 55.9          | 2.8              | 0.3        | 20,981        |
| Spanish      | 61.7        | 36.9          | 0.8              | 0.6        | 6,792         |
| French       | 57.4        | 40.2          | 1.7              | 0.7        | 4,081         |
| German       | 52.6        | 45.8          | 0.9              | 0.6        | 3,226         |
| Italian      | 65.6        | 32.9          | 0.9              | 0.7        | 2,998         |
| Dutch        | 63.2        | 32.1          | 1.0              | 3.6        | 193           |
| Portuguese   | 53.7        | 45.0          | 1.0              | 0.4        | 3,747         |
| <b>Total</b> | <b>49.8</b> | <b>47.8</b>   | <b>1.9</b>       | <b>0.5</b> | <b>42,018</b> |

Table 21: Per-language intent distribution as percentages (test). Rows sum to 100%.

$$\text{NDCG}@k = \frac{\text{DCG}@k}{\text{IDCG}@k}, \quad (2)$$

where  $\pi(j)$  is the index of the POI a system ranks at position  $j$ , and  $\text{IDCG}@k$  is the DCG of the ideal ordering of the graded ground truth. This grading rewards systems for recovering the ground-truth ranking and not merely the unordered set of relevant items, providing a sharper signal of ranking capability than a binary scheme would.

## H Alternative Geographic Representations

Before selecting the final geographic representation, we conducted additional experiments comparing two alternatives to the coordinate-based representation used in our retrieval experiments. Specifically, when fine-tuning BGE-M3, we evaluated a geohash-based representation of both user and POI coordinates, as well as a toponym-based representation of the user’s geographic information using city, country, region, and ZIP code.

When employing geohashes, queries are formatted as

```
User Query: {query}; Geohash:
{geohash};
```

while POIs are formatted as

```
Location Name: {POI name}; Address:
{street}; Position: {city}, {state},
{country}, {postal code}; Geohash:
{geohash}; Categories: {categories};
```

For the toponym-based geographic representation, queries are formatted as

```
User Query: {query}; Position: {city},
{state}, {country}, {postal code};
```

while POIs are formatted as

```
Location Name: {POI name}; Address:
{street}; Position: {city}, {state},
{country}, {postal code}; Categories:
{categories};
```

Results are reported in Table 25. The coordinate-based representation achieved higher recall than both the geohash-based and toponym-based alternatives, and also yielded higher NDCG than the toponym-based representation. While the geohash-based representation achieved higher NDCG than the coordinate-based representation, the stronger recall of the latter motivated its selection for the final retrieval setup. Based on these results, we use query and POI coordinates as the geographic representation in all subsequent experiments.

| Representation | R@10 | R@100 | R@1k | N@5  | N@20 |
|----------------|------|-------|------|------|------|
| Coordinate     | 45.2 | 83.2  | 95.6 | 58.3 | 60.3 |
| Geohash        | 45.2 | 82.7  | 95.0 | 59.2 | 60.9 |
| Toponym-Based  | 43.8 | 79.8  | 94.0 | 55.9 | 57.6 |

Table 25: Retrieval results for alternative geographic representations (test split).

## I Per-Language Reranking Results

This appendix reports the per-language breakdown of reranking performance on the test split, complementing the aggregate reranking results of Table 5 and the per-language retrieval results of Table 4. The trends are discussed in Section 5.2.2.

| Model              | Metric | EN   | ES   | FR   | PT   | DE   | IT   | NL   |
|--------------------|--------|------|------|------|------|------|------|------|
| BGE-M3 (retriever) | P@5    | 62.5 | 53.6 | 52.5 | 53.7 | 55.9 | 48.0 | 37.8 |
|                    | P@20   | 40.0 | 31.3 | 30.4 | 32.3 | 33.6 | 27.1 | 20.0 |
|                    | N@5    | 60.5 | 61.8 | 64.7 | 57.4 | 65.5 | 64.9 | 58.0 |
|                    | N@20   | 60.9 | 63.5 | 66.5 | 59.3 | 67.0 | 67.0 | 60.3 |
| BGE-RR (FT)        | P@5    | 68.4 | 57.2 | 57.6 | 60.0 | 60.6 | 53.8 | 43.7 |
|                    | P@20   | 42.6 | 33.7 | 33.9 | 35.5 | 36.3 | 31.4 | 23.9 |
|                    | N@5    | 65.8 | 65.7 | 70.7 | 64.0 | 71.1 | 70.8 | 67.1 |
|                    | N@20   | 64.3 | 66.8 | 71.4 | 64.2 | 71.3 | 72.9 | 69.8 |

Table 26: Per-language reranking performance (test split). BGE-RR: BGE-Reranker-v2-m3.

## J Statistical Significance Tests

We assess the statistical significance of the reranking improvements using paired tests over the same 42,018 test queries. Pairing is performed at the query level, comparing the metric obtained by the

fine-tuned BGE-Reranker with that obtained by the BGE-M3 retriever for each query. We apply Holm–Bonferroni correction for multiple comparisons (Holm, 1979). In addition to statistical significance, we report 95% confidence intervals, paired standardized effect sizes ( $d_z$ ), and per-query win rates to characterize the magnitude and uncertainty of the improvements.

Table 27 reports the results. Relative to the BGE-M3 retriever, the fine-tuned BGE-Reranker improves P@5 by 5.38 points ( $d_z = 0.227$ , Win= 65.3%), P@20 by 2.83 points ( $d_z = 0.204$ , Win= 62.4%), MRR by 4.19 points ( $d_z = 0.149$ , Win= 64.6%), N@5 by 5.34 points ( $d_z = 0.229$ , Win= 61.3%), and N@20 by 3.96 points ( $d_z = 0.210$ , Win= 59.9%). All five improvements are statistically significant after Holm–Bonferroni correction. Given the large number of test queries, we report effect sizes and win rates alongside  $p$ -values to provide a more informative characterization of practical differences between the systems.

| Metric | $\Delta$ | 95% CI       | $d_z$ | Win% | Adj. $p$               |
|--------|----------|--------------|-------|------|------------------------|
| P@5    | +5.38    | [5.15, 5.61] | 0.227 | 65.3 | $< 10^{-308}$          |
| P@20   | +2.83    | [2.70, 2.96] | 0.204 | 62.4 | $< 10^{-308}$          |
| MRR    | +4.19    | [3.93, 4.46] | 0.149 | 64.6 | $5.8 \times 10^{-200}$ |
| N@5    | +5.34    | [5.12, 5.56] | 0.229 | 61.3 | $< 10^{-308}$          |
| N@20   | +3.96    | [3.78, 4.14] | 0.210 | 59.9 | $< 10^{-308}$          |

Table 27: Paired significance tests comparing the fine-tuned BGE-Reranker with the BGE-M3 retriever over the 42,018 test queries.  $\Delta$  denotes the absolute improvement in percentage points. CI: 95% confidence interval;  $d_z$ : paired standardized effect size; Win%: percentage of queries for which the reranker achieves a higher per-query metric value; Adj.  $p$ : Holm–Bonferroni-adjusted  $p$ -value.