Much of Geospatial Web Search Is Beyond Traditional GIS

Ilya Ilyankou[1]
SpaceTimeLab, Department of Civil, Environmental, and Geomatic Engineering, University College London, UK

Stefano Cavazzi
Ordnance Survey, Southampton, UK

James Haworth
SpaceTimeLab, Department of Civil, Environmental, and Geomatic Engineering, University College London, UK

Abstract

Web search queries concern place far more often than existing labelling schemes suggest, yet the landscape of geospatial web search queries – what people ask of place, and how often – remains poorly characterised at scale. We apply dense sentence embeddings, a lightweight SetFit classifier, and density-based clustering to the full MS MARCO corpus of 1.01 million real Bing queries without prior filtering for toponyms or spatial keywords, identifying 181,827 geospatial queries (18.0%), nearly threefold the 6.17% labelled as *Location* in the original annotations. The resulting taxonomy of 88 query categories reveals that geospatial web search is dominated by transactional and practical lookups: costs and prices alone account for 15.3% of geospatial queries, nearly twice the size of the entire physical geography theme. Much of this activity – costs, opening hours, contact details, weather, travel recommendations – falls outside the scope of what traditional GIS and knowledge graphs are built to serve. The categories vary substantially in the kind of answer they admit, from deterministic lookups answerable from spatial databases or knowledge graphs to evaluative or temporally volatile queries that require generative or real-time systems. We discuss implications for hybrid retrieval architectures and for benchmarks of geographic reasoning in large language models. We openly release the labelled dataset, classifier, and taxonomy.

2012 ACM Subject Classification Information systems → Web searching and information discovery; Information systems → Geographic information systems; Information systems → Web log analysis; Computing methodologies → Natural language processing; Computing methodologies → Cluster analysis; Human-centered computing → Human computer interaction (HCI)

Keywords and phrases Web search queries, geographic information retrieval, query classification, geospatial query taxonomy, MS MARCO, sentence embeddings, density-based clustering, GeoAI, large language models, place theory

Digital Object Identifier 10.4230/LIPIcs.COSIT.2026.10

Related Version *Full Version*: `https://arxiv.org/abs/2605.11336`

Supplementary Material *Software (Code)*: `https://github.com/ilyankou/ms-marco-geospatial` archived at `swh:1:dir:67b243446916b179b6f0a39c61afe70a2fa5f4f0`
Model (Geospatial query classifier): `https://huggingface.co/ilyankou/is-geospatial-query`

Funding *Ilya Ilyankou*: This work was supported by Ordnance Survey & UKRI Engineering and Physical Sciences Research Council [grant no. EP/Y528651/1].

[1] Corresponding author

17th International Conference on Spatial Information Theory (COSIT 2026).
Editors: Sabine Timpf, Gabriele Filomena, Armand Kapaj, Rui Zhu, Nicholas Giudice, and Ed Manley;
Article No. 10; pp. 10:1–10:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Place is central to how people make sense of the world, yet existing information systems represent it in an impoverished way, often reducing it to named points or coordinate-bound objects rather than the vague, relational, and socially constructed phenomenon it actually is [34, 10, 18]. Large-scale web search query logs offer an empirical window into what people want from place: whether a town falls in a particular county, what the weather or cost of living is somewhere, where the nearest airport is, or what language is spoken in a country. The MS MARCO corpus of 1.01 million real Bing queries [5] remains the largest publicly available record of such behaviour, despite being collected prior to 2018 and skewed toward Anglophone North American users. It captures web search rather than conversational, voice, or professional GIS use, but it is the most suitable corpus currently available for characterising geospatial web search at scale.

Prior work has identified geospatial queries primarily by syntactic form or the presence of toponyms [17, 18, 25], producing prevalence estimates that are typically too low and taxonomies too narrow to capture the full range of what people want from place. We argue that the right starting point is the query corpus itself, without prior filtering. By applying Transformer sentence embeddings [36] and density-based clustering to the full MS MARCO dataset, we build a data-driven taxonomy of geospatial web search queries that quantifies not just how many queries are geospatial, but what categories they fall into and how those categories are distributed. Much of this activity, such as costs, opening hours, travel recommendations, or weather, falls outside the scope of what traditional GIS is built to serve, with direct implications for hybrid retrieval and generative system design.

We address three research questions:

1. What proportion of web search queries are geospatial, under a definition broader than "contains toponyms"?
2. What are the dominant themes of geospatial web search, and how are they distributed?
3. How do the resulting categories vary in the kind of answer they admit, and what does this imply for the design of systems that serve such queries?

Our work makes three practical contributions. First, we release a gold human-annotated dataset of 1,200 web search queries, labelled as geospatial or non-geospatial, constructed to cover the full embedding space of MS MARCO rather than its high-density regions. Second, we train and release a lightweight binary classifier that identifies geospatial queries with $F_1 = 0.930$ (95% bootstrap CI $0.909 - 0.947$) and can be applied to web search query corpora beyond MS MARCO. Third, we derive and release a taxonomy of 88 geospatial query categories, illustrated in Figure 3, that quantifies the relative prevalence of distinct query types and provides a concrete empirical characterisation of geospatial web search at scale.

We make the code available on GitHub[2] under the MIT license for reproducibility.

2 Related work

2.1 Taxonomies of geographic questions

Hamzei et al. mined MS MARCO [5] to characterise place questions by place type and scale [17], and later by a broader semantic schema covering place names, types, activities, spatial relationships, and qualities [16]. Xu et al. showed that professional geographic questions drawn from GIScience literature are syntactically richer and semantically distinct

[2] `https://github.com/ilyankou/ms-marco-geospatial`

from web queries, centring on analytical entities such as distribution, pattern, and density [42]. Another top-down approach by Kuhn et al. derived question templates from a spatial ontology and a taxonomy of place facets, arguing that corpus-driven methods inevitably under-represent emotive and physical facets due to collection bias [25].

On the system side, Punjani et al. addressed answering geographic natural language questions over structured linked data, translating questions into GeoSPARQL queries via handcrafted templates. Their GeoQuestions201 benchmark defines seven question categories of increasing structural complexity, covering topological, cardinal direction, and distance relations [33]. Kefalidis et al. extended this line of work with GeoQuestions1089, a larger benchmark, and showed that even improved template-based systems leave substantial headroom on geographic question answering [24].

2.2 Geospatial content in search queries and on the web

Several studies have attempted to quantify how much of online information and search activity is geographic in nature, using methods that range from keyword matching to network analysis to LLM-based classification.

Sanderson and Kohler found that roughly 18.6% of 1 million Excite[3] search queries studied contained a geographic term, with spatial relationship terms appearing rarely, suggesting users at the time did not expect search engines to handle relational spatial queries [38]. Jones et al. estimated place-name prevalence of Yahoo queries at 12.7%; they found that the vast majority target cities, and demonstrated that acceptable query-to-target distance is strongly topic-dependent, with restaurant queries clustering within tens of kilometres while accommodation queries tolerate far greater distance [23]. Gan et al. analysed the AOL dataset of 36 million records from 2006[4]; the authors built a geo/non-geo classifier and identified 13.39% of geographic-intent queries, and suggested a taxonomy of 23 broad and high-level categories that "combine aspects of topicality and desired type of interaction", such as "Tourism/Travel", "Medical", "Entertainment", or "Navigational" [12]. Henrich and Lüdecke also classified AOL user intent, and found habitation, accommodation, spare time, and information as dominant concepts; 65% of queries implied physical travel to the target; and the majority were selective rather than covering, seeking a specific point rather than documents spanning an area [19].

Beyond queries, Hahmann and Burghardt tested the much-cited claim that 80% of all information is geospatially referenced, finding via network and cognitive analyses of the German Wikipedia that the defensible figure is closer to 56–59% [15]. A more recent analysis of Common Crawl found that 18.7% of web documents contain explicit geospatial information such as coordinates or addresses [22], remarkably close to Sanderson and Kohler's query-side estimate two decades earlier.

2.3 Place theory and information needs

Edwardes and Purves showed that spatial vocabularies are strongly locale-sensitive; British contributors preferred *hill* to *mountain*, and human settlements dominated over physical geography, implying that any fixed keyword list will misclassify a substantial portion of implicitly spatial queries [11]. Purves et al. argued more fundamentally that existing information systems represent place in an impoverished way, reducing it to named points or

[3] Major web portal and search engine of the late 1990s
[4] The dataset was withdrawn shortly after release following privacy concerns

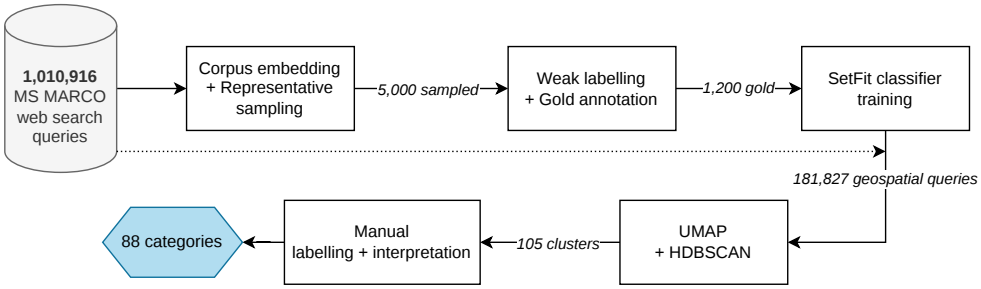

Figure 1 Overview of the methodology. Starting from the full MS MARCO corpus of 1.01M queries, we sample 5,000 representative queries for gold annotation, train a SetFit classifier to identify 181,827 geospatial queries, and apply UMAP dimensionality reduction and HDBSCAN clustering on geospatial query embeddings to derive a taxonomy of 88 geospatial query categories, grouped into 9 themes.

coordinate-bound objects, and that place is instead vague, relational, and socially constructed [34]. Therefore, it would follow that a query about living costs is just as place-anchored as any other query containing a toponym. Shanon [39] demonstrated that *where-question* responses are governed not only by containment hierarchies and physical distance but by the cognitive salience of referents and shared epistemic context, dimensions that template-driven geographic information retrieval systems do not fully address to this day.

2.4 Limitations of prior work

Previous query log studies and taxonomy building attempts share a critical limitation: they characterise geospatial queries by their syntactic form or semantic encoding, not by what people are actually trying to find out. Even sophisticated approaches to detecting geospatial language in text, such as spatial role labelling of geospatial prepositions [35], rely on surface-form signals and thus miss implicitly geospatial queries. Most rely on pre-filtered subsets: explicitly location-labelled records, toponym-containing queries, or purpose-built corpora that by design exclude such queries, producing prevalence estimates that are systematically too low when filtering on toponyms or place names (12.7% [23], 13.39% [12]), or that conflate surface form with intent when matching geographic terms more broadly. What remains unknown is the topical landscape of geospatial web search: not whether a query contains a place name or a spatial predicate, or how people phrase their queries, but whether people are searching for nearby restaurants, flight times between cities, the best places to move to, local weather, or regional costs of living. We address this gap by characterising geospatial queries by topical intent rather than syntactic form, applied without pre-filtering to a corpus of 1.01 million web search queries.

3 Methodology

Figure 1 summarises the full methodology, from corpus assembly through to taxonomy derivation.

3.1 Defining *geospatial*

No consensus definition of a spatial, geospatial, or geographic question or query exists. In the field of geographical question answering, Mai et al. broadly define *geographic* questions as those "involving geographic entities, geographic concepts, or spatial relations as parts of

the natural language questions" [27], and Kefalidis et al. define *geospatial* questions as those "requiring qualitative or quantitative geographic knowledge to be answered" [24]. We thus define a *geospatial query* as follows:

> A query is geospatial if it requires qualitative or quantitative geographic knowledge of Earth-bound features to be answered. This is usually the case if the query involves a geographic entity, a geographic concept, or a spatial relation.

In our definition, we exclude questions that are anatomical, microscopic, or astronomical in scale, and fictional or abstract "where" questions.

3.2 Identifying geospatial queries

We combined all available MS MARCO search queries[5] from the train, validation, and test subsets to produce a single corpus of 1,010,916 unique web search queries, each represented as a short[6], typically lowercased natural-language string, ranging from clearly not geospatial (e.g., "what is the human main muscles", *query ID 825350*), to clearly geospatial (e.g., "what county is badger mn in?", *602750*), to many contestable, such as "most western point in portugal" (*459663*), "how many square feet in an acre/" (*296354*), and "what all places do i need to change my address when i move" (*553148*).

To support sampling and downstream clustering, we pre-compute 384-dimensional embeddings for all 1.01M search queries using a compact, general-purpose text embedding model *BAAI/bge-small-en-v1.5*[7] via Sentence Transformers [36], which have been shown to encode some quasi-geospatial structure [21]. The chosen model offers strong semantic similarity performance for short English texts and is computationally feasible on a consumer laptop; the BGE family generally performs competitively on the Massive Text Embedding Benchmark (MTEB)[8] [32].

3.2.1 Constructing a gold dataset

We sample 5,000 queries from the entire MS MARCO corpus for labelling; exploratory manual verification suggested this would yield at least 500 positive (i.e., geospatial) samples. We use *k-means++*[9] [3] to generate 5,000 centroids in the embedding space and snap each to its nearest query. Compared to random sampling, we expect this to produce broader coverage of rare query types rather than over-representing dense clusters.

For each sampled query, we first obtain a *weak* label using a locally hosted Llama-3.1 [13] via Ollama[10] using the few-shot prompt shown in the Appendix A. The model is computationally efficient to run locally, and its instruction-following capability is sufficient for binary classification of short natural-language strings. To reduce the impact of stochastic variation in LLM outputs, we run the model 5 times per query at a low sampling temperature of 0.3 and assign the majority vote as the weak label. The lead author then manually verifies and corrects the weak labels to produce the gold dataset of $1,200$ queries, which we release

[5] Available to download at `https://msmarco.z22.web.core.windows.net/msmarcoranking/queries.tar.gz`
[6] Mean query length is 35 characters, and 99.9% of queries are under 131 characters
[7] Available on HuggingFace: `https://huggingface.co/BAAI/bge-small-en-v1.5`
[8] The leaderboard is available at `https://huggingface.co/spaces/mteb/leaderboard`
[9] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.kmeans_plusplus.html`
[10] `https://ollama.com/library/llama3.1`

on GitHub[11]. To assess the reliability of the lead author's gold labels, the remaining two authors independently of each other annotate a 200-query sample drawn from the gold set; we report pairwise Cohen's κ [9] for inter-annotator agreement across three annotator pairs.

3.2.2 Training a SetFit binary classifier

We use the gold dataset to contrastively train a lightweight geospatial/non-geospatial binary classifier using SetFit [41], a few-shot learning framework that fine-tunes a sentence embedding model via contrastive learning before fitting a classification head for strong performance with limited labelled data. We use *BAAI/bge-small-en-v1.5* as the base sentence-transformer. SetFit is designed for few-shot learning and requires few labelled examples to perform well [41]; we therefore split the dataset with stratification into train (200 samples), validation (200), and hold-out test (800) partitions for a robust initial evaluation. We train for five epochs with a batch size of 64 and a learning rate of 2×10^{-5}, keeping the SetFit default of 20 contrastive pair-sampling iterations per epoch. We evaluate every epoch and apply early stopping after 2 rounds with no improvements in embedding loss. The model is evaluated on the hold-out test set; we report 95% confidence intervals computed via 1,000 bootstrap resamples of the test set. For production inference (i.e., to label the 1.01M queries as geospatial or not), we retrain the same configuration on the full gold dataset for 3 epochs, corresponding to the best validation checkpoint observed during initial evaluation. We release the trained classifier on HuggingFace[12].

3.3 Building taxonomy

Once all geospatial queries are identified, we use clustering on their embeddings and manual interpretation to categorise geospatial web search and identify key themes.

3.3.1 Dimensionality reduction and density-based clustering

We take the subset of queries predicted as geospatial and retrieve their corresponding 384-d embeddings. Because density-based clustering in high dimensions is difficult (the problem often referred to as the *curse of dimensionality* [4]), we first reduce dimensionality using the Uniform Manifold Approximation and Projection algorithm, or UMAP[13] [29], which projects high-dimensional embeddings into a lower-dimensional space while preserving local neighbourhood structure, and then cluster the reduced representations using Hierarchical Density-Based Spatial Clustering of Applications with Noise, or HDBSCAN[14] [8]. This combination is widely used for topic modelling over transformer embeddings [2, 40, 14, 1].

3.3.2 Grid search over UMAP and HDBSCAN parameters

Density-based clustering relies on density thresholds whose appropriate values depend on the underlying data distribution; selecting a suitable density level is inherently difficult and data-dependent [7]. We therefore perform a grid search over an interpretable parameter space for both UMAP and HDBSCAN to identify a set of parameters that produces a reasonable number of high-quality clusters. For UMAP, we vary (i) output dimensionality $\in \{5, 10, 15\}$

[11] `https://github.com/ilyankou/ms-marco-geospatial`/tree/main/gold-dataset
[12] `https://huggingface.co/ilyankou/is-geospatial-query`
[13] `https://umap-learn.readthedocs.io/en/latest/`
[14] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html`

and (ii) neighbourhood size $\in \{10, 25, 50\}$. For HDBSCAN, we vary (iii) minimum cluster size $\in \{25, 50, 100, 200\}$ and (iv) minimum number of samples, set to $\{0.2, 0.5, 1.0\}\times$ *minimum cluster size*, using the default Excess of Mass (*"eom"*) cluster selection method. This gives us $3 \times 3 \times 4 \times 3 = 108$ configurations, each evaluated with a fixed random seed (42).

For each configuration, we compute four quality metrics: (a) the Density-Based Clustering Validation (DBCV) score [31], a density-aware measure of within-cluster cohesion and between-cluster separation suited to the arbitrary-shape clusters produced by methods such as HDBSCAN (range -1 to 1, higher is better), (b) the noise fraction, i.e. the proportion of queries assigned to cluster "-1"), (c) the number of non-noise clusters, and (d) the median non-noise cluster size. We use DBCV rather than silhouette scores, as the latter assumes convex clusters and has no principled treatment of noise points, both of which make it unsuitable for evaluating HDBSCAN output.

3.3.3 Consistency checks

UMAP is stochastic, and its randomness can affect downstream clustering. To assess how stable our clustering is, we pick the top-10 configurations by DBCV that produced at least 10 clusters (to exclude degenerate solutions with too few clusters to be informative), and re-run each across six random seeds $\in \{0, 1, 2, 3, 4, 42\}$. For each configuration, we report the mean and standard deviation of DBCV, noise fraction, and cluster count across seeds, and select the configuration with the strongest combination of high mean DBCV and low cross-seed variance. For final taxonomy extraction, we run the full pipeline once on all 181,827 queries using the selected configuration.

3.3.4 Cluster interpretation

To support manual naming of resulting clusters, for each cluster we (i) identify a representative query, defined as the query whose embedding is closest to the cluster centroid, (ii) extract salient unigram and bigram terms using Term Frequency-Inverse Document Frequency (TF-IDF) over concatenated cluster texts, with English stop-words removed, and (iii) randomly sample 10 queries per cluster to provide additional context for annotators. These summaries are exported as Markdown and as a spreadsheet for manual annotation by the authors.

Following manual annotation, we merge clusters that we judge to represent the same query intent (most notably, multiple US state-specific "what county is [place] in" clusters fragmented by named entities rather than intent), yielding a final taxonomy of 88 categories.

We group the categories into nine broad themes – Statistical, Temporal, "POIs and Commercial", Administrative, Physical, Cultural, Historical, Biographical, and Events – for organisational convenience rather than as a theoretically grounded hierarchy. We experimented with alternative groupings but found that meaningful structure resides in the leaf categories themselves; any higher-level grouping risks implying a cleaner taxonomy than the data supports.

The resulting taxonomy is exported as a parent–child JSON graph for visualisation.

4 Results

4.1 Gold dataset

Of the 5,000 sampled queries, 628 received at least one "geospatial" vote from the LLM, of which the vast majority (547) were unanimously labelled "geospatial" across all five LLM runs. We manually verified 561 true weak positives and identified 67 false weak positives.

We also randomly sampled and manually verified 572 weak negatives, identifying 7 false weak negatives. Thus, the gold dataset consists of 568 positives and 632 negatives (total size of $1,200$).

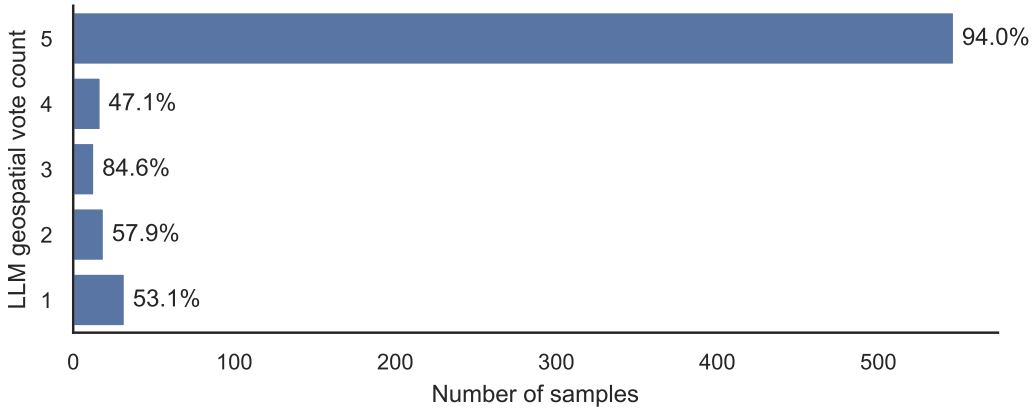

Figure 2 Human verification of weak geospatial labels across LLM vote counts. Bars indicate the number of queries receiving 1-5 spatial votes across repeated LLM runs; percentages denote the proportion confirmed as geospatial after manual verification.

Figure 2 shows human agreement with the weak LLM labels as a function of the number of "geospatial" votes (between 1 and 5). The distribution is heavily skewed toward unanimous predictions, suggesting that most weakly-positive queries are unambiguous for the Llama-3.1 model under the prompt and the definition adopted. The intermediate bins (2-4 votes) contain few queries, so their confirmation rates should not be read as a monotonic trend. Unsurprisingly, the more obvious geospatial queries, such as "what county is springfield, or" (*613049*) or "which region is egypt located" (*1018537*) typically received all 5 positive votes; more borderline cases, such as "biggest house in the world price" (*53174*) or "where do people think dolphins live" (*971381*) produced inconsistent votes across runs, with the model split between geospatial and non-geospatial judgements.

To assess inter-annotator agreement, the second and third authors independently re-annotated 200 samples drawn from the gold set. The lead, second and third authors all agreed on 179 of 200 samples (89.5%); of those, 107 were negative and 72 positive. Pairwise Cohen's κ was 0.88 (lead vs second author), 0.84 (lead vs third), and 0.84 (second vs third), indicating "almost perfect" agreement [26]. As expected, disagreements concentrated on borderline cases. For example, only the lead author read "va" as Virginia in the query "va tax contact" (*535301*); "cricket wireless address" (*112731*) was labelled as geospatial by the lead and second authors but not the third; "biggest house in the world price" (*53174*) was labelled as geospatial by the lead and third authors, but not the second.

4.2 SetFit classifier

When trained on just 200 samples (105 negative and 95 positive) and tested on 800 samples (421 negative and 379 positive) of the gold dataset, the SetFit binary classifier achieves an accuracy of 0.934 (95% bootstrap CI $0.915 - 0.950$) and an F_1 score of 0.930 (95% bootstrap CI $0.909 - 0.947$), producing 27 false negatives and 26 false positives. The majority of misclassified samples are borderline geospatial and can be argued both ways; for example, "biggest snakes in world" (FP, *53491*), "when is the good time to see northern lights" (FP,

952893), "address for expresspcb" (FN, *11492*), or "how wide is the base of the great wall" (FN, *1175805*). This result is particularly notable given the small size of the training set and near-equal class distribution in both training and test sets, making F_1 a meaningful performance indicator rather than a result of class imbalance.

We further verify the classifier's robustness by sense-checking its behaviour on made-up, semantically complex queries and short phrases, a subset of which is demonstrated in Table 1.

Table 1 Some examples of correctly classified, semantically complex made-up queries and phrases (*not* part of MS MARCO) to sense-check the trained classifier's behaviour.

Correctly classified as geospatial	Correctly classified as non-geospatial
nearest hospital	near impossible
countries bordering ukraine	borderline invisible
distance from paris to berlin	keep your distance
restaurants along the m1	stop it m8
where should i live	where can i escape mentally
flood risk in this area	not my area of expertise
is it a long flight london to kaunas	a long way to go in my career

After retraining the classifier on the full gold dataset consisting of 632 negative and 568 positive cases, we run it on the entire MS MARCO dataset consisting of 1,010,916 queries, and identify 181,827 spatial queries, or 18.0% of the entire dataset.

An analysis of the most frequent first words in each class (see Table 2) reveals that *what* dominates both geospatial and non-geospatial queries (29.6% and 36.1% respectively), *how* appears in the top-3 for both (11.6% and 17.9%), while *where* ranks second in geospatial queries (15.8%) but only 14th in non-geospatial queries (0.8%). The overlap between the two vocabularies is striking: 14 of the top-20 first words appear in both lists.

Table 2 Top-20 most frequent first words in geospatial and non-geospatial MS MARCO queries.

Geospatial						Non-geospatial					
#	Word	%	#	Word	%	#	Word	%	#	Word	%
1	what	29.6	11	cost	1.0	1	what	36.1	11	define	1.0
2	where	15.8	12	population	0.9	2	how	17.9	12	definition	1.0
3	how	11.6	13	what's	0.9	3	who	3.7	13	cost	0.9
4	average	3.4	14	does	0.6	4	is	3.0	14	where	0.8
5	when	3.3	15	can	0.5	5	when	2.6	15	do	0.7
6	weather	2.8	16	most	0.5	6	can	2.1	16	the	0.6
7	is	2.5	17	largest	0.5	7	why	1.8	17	are	0.6
8	which	1.9	18	distance	0.5	8	which	1.8	18	meaning	0.5
9	who	1.8	19	temperature	0.5	9	does	1.3	19	what's	0.4
10	why	1.1	20	the	0.4	10	average	1.2	20	causes	0.4

4.3 Grid search for clustering

The top-10 UMAP and HDBSCAN configurations by DBCV, which produced at least 10 clusters, are shown in Table 3.

Table 3 Top-10 UMAP and HDBSCAN hyper-parameter configurations by DBCV score, restricted to those producing at least 10 clusters. *Dims* is UMAP output dimensionality, *Neigh* is UMAP `n_neighbors`, *Min cluster size* and *Min samples* are HDBSCAN's `min_cluster_size` and `min_samples` parameters; *Noise* is the proportion of queries assigned to the noise ('−1') cluster.

Config ID	UMAP		HDBSCAN		DBCV	Noise	Num clusters	Median cluster size
	Dims	Neigh	Min cluster size	Min samples				
93	15	25	200	40	0.358	0.359	102	449.5
79	15	10	100	50	0.341	0.359	182	246
81	15	10	200	40	0.337	0.388	122	476
22	5	25	200	100	0.334	0.389	91	467
21	5	25	200	40	0.333	0.362	105	469
95	15	25	200	200	0.324	0.417	77	553
9	5	10	200	40	0.319	0.392	122	443.5
45	10	10	200	40	0.317	0.374	114	475
23	5	25	200	200	0.315	0.426	79	485
76	15	10	50	25	0.312	0.418	415	119

Table 4 reports the consistency of the top-10 configurations across six repeated runs with different random seeds. Six configurations are stable, with standard deviation for both DBCV and noise fraction below 0.050.

Table 4 Consistency of top-10 configurations across six repeated runs with various random seeds $\{0, 1, 2, 3, 4, 42\}$, reporting minimum, mean, maximum, and standard deviation of the DBCV scores (1=perfect), noise fraction, and number of produced clusters.

Config ID	DBCV				Noise				Num clusters			
	Min	Mean	Max	Std	Min	Mean	Max	Std	Min	Mean	Max	Std
93	0.245	0.364	0.712	0.176	0.000	0.332	0.432	0.165	3	88	109	41.9
45	0.097	0.363	0.811	0.242	0.000	0.245	0.400	0.190	4	77	115	56.3
23	0.279	0.326	0.416	0.049	0.341	0.409	0.464	0.040	64	75	82	6.4
21	0.257	0.310	0.367	0.040	0.358	0.389	0.415	0.027	100	108	114	6.2
81	0.178	0.306	0.358	0.069	0.001	0.244	0.388	0.189	4	77	122	56.6
95	0.227	0.289	0.348	0.044	0.410	0.432	0.470	0.026	72	76	82	3.4
79	0.247	0.282	0.341	0.033	0.344	0.395	0.442	0.038	173	195	208	14.3
76	0.238	0.279	0.312	0.031	0.412	0.421	0.443	0.011	406	421	459	19.4
9	0.092	0.271	0.412	0.109	0.000	0.253	0.406	0.196	4	79	122	58.3
22	0.231	0.269	0.334	0.040	0.389	0.434	0.461	0.033	87	96	101	5.4

Based on the consistency analysis, configurations #21 and #23 are superior; we pick configuration #21 as it achieves a lower mean share of noise (0.389 vs 0.409 in configuration #23) and a higher number of clusters (mean 108 vs 75) despite having a slightly lower mean DBCV (0.310 vs 0.326). For the seed, we pick 42, whose DBCV falls near the configuration's median across runs (we deliberately avoid selecting the highest-DBCV seed to prevent cherry-picking).

4.4 Final clustering

The radial hierarchy chart in Figure 3 illustrates 88 categories (collapsed from 105 initial clusters), grouped into nine broad themes.

Examples of clusters' representative queries, top-20 most frequently occurring terms, and random samples of 10 queries are illustrated in Table 5. Table 6 in the Appendix shows resulting labelled clusters with sizes and representative queries. We make all cluster-level data available on GitHub[15].

[15] `https://github.com/ilyankou/ms-marco-geospatial`/blob/main/interim/cluster_review.md

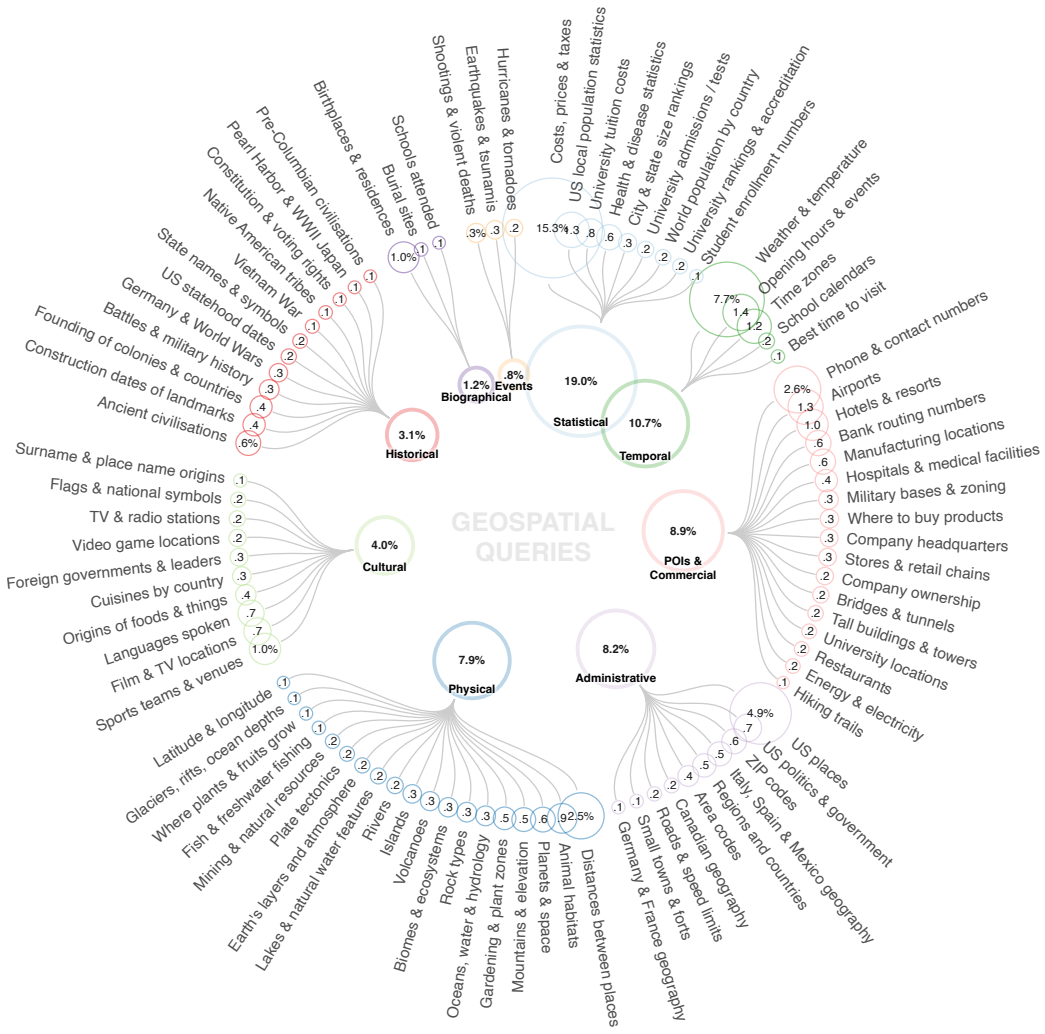

Figure 3 Geospatial query clusters grouped into nine themes. Unclustered (noise) queries representing 36.2% of identified geospatial queries are not shown.

5 Discussion

5.1 Characteristics of the taxonomy

This study identifies 181,827 geospatial queries in MS MARCO, representing 18.0% of the dataset, which is a substantially higher proportion than the 6.17% of queries labelled as *Location* in the original MS MARCO annotations [5]. This discrepancy reflects the breadth of our geospatial definition, which captures implicitly location-anchored queries such as weather conditions, prices, and biomes – categories that may not be caught by keyword-based approaches.

18 of 105 initial clusters (*#55*, and *#88–#104* in Table 6 in the Appendix) largely represent variants of the query "what county is [place] in", differentiated primarily by US state abbreviation. This fragmentation is partly an HDBSCAN artefact as embeddings likely treat state abbreviations, such as "tx" or "ny", as distinguishing features. We group these

Table 5 To derive initial cluster labels, we looked at each cluster's representative query, top-20 most frequent terms, and ten randomly sampled queries.

	Cluster #48	**Cluster #55**
No. samples	344	1015
Representative query	what is the tallest building in the world	what is the county of florida
Top-20 terms	tower, building, pisa, tower pisa, tallest, leaning, leaning tower, tall, tallest building, empire state, floors, building world, empire, state building, world, towers, khalifa, burj, burj khalifa, skyscraper	fl county, fl, county, florida, florida county, beach, palm, beach fl, population, florida located, jacksonville, fl population, county florida, beach florida, located, county palm, springs, orange, county jacksonville, naples
Ten sample queries	largest moving structure in the world · what is the pyramid shaped building on the skyline of san francisco · leaning tower of pisa facts · how many floors is tower one · how many floors does the tallest building in sydney have · leaning tower locale · how many feet is the world's tallest building · how tall is victory tower · how many floors in the empire state building. · where is cornerstone condos located	homes for sale st. cloud fl. · where is dunedin florida · where is naples manor fl · what county is atlantic beach fl in · what county is harbor springs in · where is legoland florida located · what county is ft. lauderdale · homes for sale jacksonville nc · where is citrus florida located · what county is holt fl
Cluster label	Tall buildings & towers	County membership and places in Florida → US places
Theme	POIs & Commercial	Administrative

fragmented clusters into a single *US places* category for Figure 3. HDBSCAN treats 36.2% of identified geospatial queries as noise (not shown in Figure 3), a substantial but unsurprising proportion given the scale and messiness of the real-world dataset. Natural-language web search queries are inherently heterogeneous, and some degree of noise is expected and acceptable in density-based clustering applied at this scale. The *Planets & space* cluster (1,045 largely astronomical queries, which our definition excludes) likely reflects the binary classifier not having been exposed to such queries during training; representing only 0.6% of all geospatial queries, its effect on our analysis is negligible.

We use nine broad themes – Statistical, Temporal, "POIs and Commercial", Administrative, Physical, Cultural, Historical, Biographical, and Events – for organisational convenience rather than as a theoretically grounded hierarchy. We experimented with alternative groupings but found that meaningful structure resides in the leaf categories themselves; any higher-level grouping risks implying a cleaner taxonomy than the data supports. Because we report theme shares over the full set of geospatial queries, including noise, the percentages are conservative; the relative ranking of themes holds only insofar as noise is not concentrated in particular intents, which we do not test here.

The taxonomy reveals that geospatial web search is dominated by transactional and practical lookups rather than by either administrative or physical geography. The Statistical theme alone accounts for 19.0% of geospatial queries and the Temporal theme a further 10.7%. The *Costs, prices & taxes* (15.3%) cluster, with queries such as "what is the average household income of lake oswego" (*1152138*) or "what is currency in brazil" (*736719*), is comparable in size to the Physical (7.9%) and Administrative (8.2%) themes combined. Users

overwhelmingly turn to web search for place-anchored facts about money, time, and weather; questions about landforms, hydrology, or jurisdictional boundaries – traditional GIS territory – form a comparatively small share of expressed information needs.

5.2 Implications for system design

A large share of the taxonomy consists of clusters whose answers are open, stable, and objective. For example, *Distances between places* (e.g., "how far is it from hartford to new london", *231318*) and *ZIP codes* (e.g., "what is burlington,ky zip code", *726469*) are answerable deterministically using geoparsing, spatial databases and a traditional GIS toolkit, including distance calculations and containment operations. Query categories such as *Regions and countries* (e.g., "what is the continent is algeria located", *1151645*), *Bridges & tunnels* (e.g., "where does golden gate bridge connect to", *973130*) or *Volcanoes* (e.g., "what type of volcano is mt.santorini", *915164*) are largely answerable from structured knowledge contained in geographical knowledge graphs. Such types of queries do not require generative or probabilistic systems to be answered correctly. This suggests that a hybrid retrieval architecture of routing objectively answerable geospatial queries to structured geodata sources while reserving generative models for more complex or subjective queries could improve both accuracy and reliability compared to treating all geospatial queries uniformly.

Categories such as *Best time to visit* (e.g., "most affordable times to visit hawaii", *456340*), *University rankings & accreditation* (e.g., "is codarts a good school?", *406730*), and *Origins of foods & things* (e.g., "where does spaghetti bolognese come from", *973905*) are more subjective as their answers depend on personal preferences and the source of information used to answer them. These are the clusters where Anglophone-Western geographic and cultural biases in both the MS MARCO dataset and any downstream system are most likely to surface. Systems trained to answer such queries risk encoding and amplifying the biases of their training data under the guise of objective geospatial information. This risk is sharpest in conversational interfaces, which can deliver an evaluative or unverified answer with the same fluency as a deterministic lookup [20].

Temporal or POI-specific categories such as *Opening hours & events* (e.g., "what day does the garrett pool open", *616654*), *Phone & contact numbers* (e.g., "northampton county tax department pa phone number", *1092161*), or *Restaurants* (e.g., "what restaurant serves both mexican and steak", *1109148*) demand answers that are not only geographically specific but temporally volatile: opening hours, phone numbers, and menus change, and businesses close and relocate much more frequently than coastlines redraw themselves. Unlike physical geography, which changes slowly, institutional information goes out-of-date quickly, making it particularly challenging for static retrieval systems to serve reliably without access to up-to-date data. The stakes are uneven across clusters: an outdated restaurant menu or opening time may cause minor inconvenience, whereas obsolete information about hospital locations or emergency services could have serious consequences. The *Weather & temperature* category (e.g., "what is the weather in nashville tomorrow", *853623*), and the Events theme, accounting for 0.8% of identified geospatial queries and covering *Hurricanes & tornadoes* (e.g., "where there storms last night", *1001188*), *Earthquakes & tsunamis* (e.g., "was there a earthquake in alaska", *541104*), and *Shootings & violent deaths* (e.g., "who is the woman that beat up woman at stop light in houston texas", *1042188*), highlight the relationship between spatial and temporal information needs, which is another argument for hybrid architectures that can handle APIs, news and live data feeds in addition to static geographic knowledge.

The most significant implication concerns emerging natural-language interfaces, including conversational chatbot search. If nearly one in five web search queries is geospatial, these systems require robust geospatial reasoning for a substantial fraction of real user traffic. Existing geographic benchmarks for LLMs have largely focused on factual and location reasoning [6, 28, 37, 30], with comparatively less attention to the transactional, temporal, and subjective queries that dominate our taxonomy. Costs and prices, opening hours, and travel recommendations are examples of query categories where retrieval-augmented and real-time architectures matter most, yet they are largely absent from current evaluation benchmarks. Our taxonomy offers a more representative target for both system design and benchmarking, grounded in what users actually ask rather than what is easiest to test.

5.3 Limitations and future work

MS MARCO was collected prior to 2018 and reflects search behaviour on Bing at that time; the distribution of geospatial query types has likely shifted as search interfaces, user habits, and geospatial services have evolved. The dataset over-represents Anglophone North American users, and this is reflected directly in the taxonomy, where *US places* (4.9%) accounts for vastly more place-specific queries than *Italy, Spain & Mexico geography*, *Canadian geography*, and *Germany and France geography* clusters combined (0.8%). Future work should seek to validate and extend the taxonomy using more geographically diverse query corpora.

The original MS MARCO dataset specifically excludes queries "with navigational and other intents" [5], meaning that a distinct and practically significant category of geospatial information need is not represented in our taxonomy. Our 18.0% is therefore best read as a lower bound, since routing and directions (likely to be a significant geospatial category, even in web search) are excluded by construction.

The corpus reflects web search habits and as such consists of short and typically single-intent queries, which constrains the taxonomy. Longer, multi-intent interactions typical of conversational AI systems, voice assistants, or chatbots fall outside what this analysis captures. In such systems, geospatial intent may be implicit, distributed across dialogue turns, or conditioned on prior context in ways that single-query classification cannot detect. Whether the underlying information needs remain stable across interfaces, or shift as users adapt to conversational systems, is an open empirical question that we aim to address in future work.

Several directions emerge from this work. First, the taxonomy clusters could be linked to existing geospatial datasets and spatial operations (for example, mapping distance queries to routing APIs, or county membership queries to administrative boundary datasets), providing a concrete bridge between user intent and existing GIS infrastructure. Second, the binary geospatial classifier could be extended into a multi-class router that directs queries to category-specific retrieval pipelines. The unclustered "noise" queries, representing over a third of all identified geospatial queries, also deserve further study; repeating the full clustering pipeline on this subset alone may reveal latent structure that the current parameter settings were too coarse to resolve. Finally, the distinction between objective and subjective geospatial queries merits a more formal investigation, as it has direct implications for system design, bias mitigation, and the appropriate use of generative versus deterministic systems.

6 Conclusion

Geospatial web search queries are acts of spatial cognition. When someone types "in which county is north hollywood california" (*394665*) or "cost of living in turkmenistan" (*105311*), they are navigating a conceptually structured space of places, attributes, and relationships.

Our finding that 18.0% of MS MARCO queries are geospatial under a semantically broad definition is evidence that geospatial thinking pervades everyday information behaviour at a scale that toponym-based studies in particular have underestimated.

The derived taxonomy shows that this everyday spatial cognition diverges from what GIS has traditionally modelled. Questions about landforms, hydrology, and physical geography are outnumbered by transactional and practical lookups about costs, weather, and opening hours. This has direct consequences for systems designed to serve such queries.

The taxonomy also surfaces a more fundamental split: queries about distances, ZIP codes, or county membership presuppose that place has determinate answers; queries about the best place to live or the right season to visit presuppose that place is evaluative and culturally inflected. These two modes – geospatial information as fact versus geospatial information as judgement – demand different infrastructure and raise different risks. Distinguishing them is, we argue, a precondition for handling geographic knowledge responsibly.

References

1 Mebarka Allaoui, Mohammed Lamine Kherfi, and Abdelhakim Cheriet. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In Abderrahim El Moataz, Driss Mammass, Alamin Mansouri, and Fathallah Nouboud, editors, *Image and Signal Processing*, pages 317–325, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-51935-3_34.

2 Dimo Angelov. Top2Vec: Distributed Representations of Topics, August 2020. arXiv:2008.09470.

3 David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, USA, January 2007. Society for Industrial and Applied Mathematics. URL: https://dl.acm.org/doi/10.5555/1283383.1283494.

4 Ira Assent. Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery*, 2(4):340–350, 2012. doi:10.1002/widm.1062.

5 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. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, October 2018. doi:10.48550/arXiv.1611.09268.

6 Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Are Large Language Models Geospatially Knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23, pages 1–4, New York, NY, USA, December 2023. Association for Computing Machinery. doi:10.1145/3589132.3625625.

7 Ricardo J. G. B. Campello, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 10(2):e1343, 2020. doi:10.1002/widm.1343.

8 Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-Based Clustering Based on Hierarchical Density Estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer. doi:10.1007/978-3-642-37456-2_14.

9 Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960. doi:10.1177/001316446002000104.

10 Tim Cresswell. *Place: An Introduction*. John Wiley & Sons, August 2014.

11 Alistair J. Edwardes and Ross S. Purves. A Theoretical Grounding for Semantic Descriptions of Place. In J. Mark Ware and George E. Taylor, editors, *Web and Wireless Geographical Information Systems*, pages 106–120, Berlin, Heidelberg, 2007. Springer. doi:10.1007/978-3-540-76925-5_8.

12 Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, pages 49–56, New York, NY, USA, April 2008. Association for Computing Machinery. doi:10.1145/1367798.1367806.

13 Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models, 2024. doi:10.48550/arXiv.2407.21783.

14 Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022. doi:10.48550/arXiv.2203.05794.

15 Stefan Hahmann and Dirk Burghardt. How much information is geospatially referenced? Networks and cognition. *International Journal of Geographical Information Science*, 27(6):1171–1189, June 2013. doi:10.1080/13658816.2012.743664.

16 Ehsan Hamzei, Haonan Li, Maria Vasardani, Timothy Baldwin, Stephan Winter, and Martin Tomko. Place Questions and Human-Generated Answers: A Data Analysis Approach. In *Proceedings of the 22nd AGILE Conference on Geographic Information Science*, pages 3–19, 2019. doi:10.1007/978-3-030-14745-7_1.

17 Ehsan Hamzei, Stephan Winter, and Martin Tomko. Initial Analysis of Simple Where-Questions and Human-Generated Answers. In Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart, editors, *LIPIcs, Volume 142, COSIT 2019*, volume 142, pages 12:1–12:8. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.COSIT.2019.12.

18 Ehsan Hamzei, Stephan Winter, and Martin Tomko. Place facets: a systematic literature review. *Spatial Cognition & Computation*, 20(1):33–81, January 2020. doi:10.1080/13875868.2019.1688332.

19 Andreas Henrich and Volker Luedecke. Characteristics of geographic information needs. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, GIR '07, pages 1–6, New York, NY, USA, November 2007. Association for Computing Machinery. doi:10.1145/1316948.1316950.

20 Ilya Ilyankou, Stefano Cavazzi, and James Haworth. The Scenic Route to Deception: Dark Patterns and Explainability Pitfalls in Conversational Navigation, March 2026. doi:10.48550/arXiv.2603.14586.

21 Ilya Ilyankou, Aldo Lipani, Stefano Cavazzi, Xiaowei Gao, and James Haworth. Do Sentence Transformers Learn Quasi-Geospatial Concepts from General Text? *GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, 2024, Glasgow, Scotland*, 2024.

22 Ilya Ilyankou, Meihui Wang, Stefano Cavazzi, and James Haworth. Quantifying Geospatial in the Common Crawl Corpus. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '24, pages 585–588, New York, NY, USA, November 2024. Association for Computing Machinery. doi:10.1145/3678717.3691286.

23 Rosie Jones, Wei V. Zhang, Benjamin Rey, Pradhuman Jhala, and Eugene Stipp. Geographic intention and modification in web search. *International Journal of Geographical Information Science*, 22(3):229–246, March 2008. doi:10.1080/13658810701626186.

24 Sergios-Anestis Kefalidis, Dharmen Punjani, Eleni Tsalapati, Konstantinos Plas, Maria-Aggeliki Pollali, Pierre Maret, and Manolis Koubarakis. The question answering system GeoQA2 and a new benchmark for its evaluation. *International Journal of Applied Earth Observation and Geoinformation*, 134:104203, November 2024. doi:10.1016/j.jag.2024.104203.

25 Werner Kuhn, Ehsan Hamzei, Martin Tomko, Stephan Winter, and Haonan Li. The semantics of place-related questions. *Journal of Spatial Information Science*, 1(23):157–168, 2021. doi:10.5311/JOSIS.2021.23.161.

26 J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

27 Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. Geographic Question Answering: Challenges, Uniqueness, Classification, and Future Directions. *AGILE: GIScience Series*, 2:1–21, June 2021. doi:10.5194/agile-giss-2-8-2021.

28 Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. GeoLLM: Extracting Geospatial Knowledge from Large Language Models, February 2024. arXiv:2310.06213 [cs]. doi:10.48550/arXiv.2310.06213.

29 Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. doi:10.48550/arXiv.1802.03426.

30 Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geographic Information Systems (GIS) Exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '23, pages 85–94, New York, NY, USA, November 2023. Association for Computing Machinery. doi:10.1145/3615886.3627745.

31 Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-Based Clustering Validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, pages 839–847, 2014. doi:10.1137/1.9781611973440.96.

32 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive Text Embedding Benchmark, March 2023. doi:10.48550/arXiv.2210.07316.

33 D. Punjani, K. Singh, A. Both, M. Koubarakis, I. Angelidis, K. Bereta, T. Beris, D. Bilidas, T. Ioannidis, N. Karalis, C. Lange, D. Pantazi, C. Papaloukas, and G. Stamoulis. Template-Based Question Answering over Linked Geospatial Data. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, pages 1–10, Seattle WA USA, November 2018. ACM. doi:10.1145/3281354.3281362.

34 Ross S. Purves, Stephan Winter, and Werner Kuhn. Places in Information Science. *Journal of the Association for Information Science and Technology*, 70(11):1173–1182, 2019. doi:10.1002/asi.24194.

35 Mansi Radke, Prarthana Das, Kristin Stock, and Christopher B. Jones. Detecting the Geospatialness of Prepositions from Natural Language Text. In Sabine Timpf, Christoph Schlieder, Markus Kattenbeck, Bernd Ludwig, and Kathleen Stewart, editors, *14th International Conference on Spatial Information Theory (COSIT 2019)*, volume 142 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:8, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.COSIT.2019.11.

36 Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics. doi:10.18653/v1/D19-1410.

37 Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. GPT4GEO: How a Language Model Sees the World's Geography, May 2023. arXiv:2306.00020 [cs]. doi:10.48550/arXiv.2306.00020.

38 Mark Sanderson and Janet Kohler. Analyzing geographic queries. In *Proceedings of the Workshop on Geographic Information Retrieval. 27th Annual International ACM SIGIR Conference*, 2004.

39 Benny Shanon. Answers to where-questions. *Discourse Processes*, 6(4):319–352, 1983. doi:10.1080/01638538309544571.

40 Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.135.

41 Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient Few-Shot Learning Without Prompts, September 2022. doi:10.48550/arXiv.2209.11055.

42 Haiqi Xu, Ehsan Hamzei, Enkhbold Nyamsuren, Han Kruiger, Stephan Winter, Martin Tomko, and Simon Scheider. Extracting interrogative intents and concepts from geo-analytic questions. *AGILE: GIScience Series*, 1:1–21, July 2020. doi:10.5194/agile-giss-1-23-2020.

A LLM prompt for weak labelling

Listing 1 Prompt for Llama-3.1 to label 5,000 sampled queries as geospatial/non-geospatial.

```
Classify the search query as geospatial (true) or not (false).

A query is geospatial if it requires qualitative or quantitative geographic
knowledge of Earth-bound features to be answered. This is usually the case if the
    query involves:
- A geographic entity (named place on Earth: city, country, river, POI, address)
- A geographic concept (place type: city, lake, mountain, park, building)
- A spatial relation (near, within, north of, between, borders, crosses, distance)

Non-geospatial: anatomical, microscopic, astronomical, fictional, or abstract
'where' questions; queries needing no geographic knowledge.

Output only: true or false.

Examples:
  * How far is Brighton from London -> true
  * what does square from greece mean -> false
  * Capital of France -> true
  * What does 'Dutch courage' mean -> false
  * Population of Piraeus -> true
  * Where is the epimysium found -> false
  * Restaurants near Hyde Park -> true
  * Where did the name Missouri come from -> false
  * Where does Paris Hilton live -> true
  * What is a river -> false
  * What city is Ebright Azimuth in -> true
  * is the water underneath a hurricane calm -> false
  * how many state are in the us -> true
  * Where is the SSID -> false
  * where are new franchises needed -> true
  * which side is the us flag put -> false
  * what languages are spoken in israel -> true
  * Where is Hogwarts -> false
  * most western point in portugal -> true
  * What is Greek yoghurt -> false

Query: <SEARCH QUERY>
Answer:
```

B Clusters with representative queries

Table 6 Resulting clusters with final labels and representative queries.

#	Size	Label	Representative query *(query ID)*
-1	65823	Unclustered (HDBSCAN noise)	what is the url for the state bank of the lakes *(852466)*
0	696	Area codes	area code numbers united states *(26408)*
1	1054	ZIP codes	what's the zip code *(933999)*
2	13966	Weather & temperature	what's the temperature outdoors *(933263)*
3	27761	Costs, prices & taxes	what is the current minimum wage for ny *(1151495)*
4	526	Earthquakes & tsunamis	what is been the biggest earthquake *(722978)*
5	1344	US politics & government	how many representatives for each state *(294437)*
6	596	Military bases & zoning	what military base is the largest in the us *(878154)*
7	298	State names & symbols	when was the state nickname chosen *(962316)*
8	1296	Film & TV locations	where was where the heart is filmed *(1004279)*
9	217	Schools attended	what college did trump attend *(597060)*
10	1744	Birthplaces & residences	where was mlk born *(1002963)*
11	505	City & state size rankings	what is the largest city in the US by area *(827081)*
12	818	Hospitals & medical facilities	university of iowa hospitals and clinics patient *(532623)*
13	316	Canadian geography	toronto is in what province *(522784)*
14	441	Hurricanes & tornadoes	where does a hurricane occur *(972347)*
15	1406	University tuition costs	usu tuition cost per semester *(534985)*
16	1096	Bank routing numbers	us bank routing numbers *(533522)*
17	4641	Phone & contact numbers	state street service desk number *(503076)*
18	246	Best time to visit	when is the best time to visit the bahamas *(952523)*
19	287	Roads & speed limits	types of roads and their speed limits *(529594)*
20	2228	Time zones	what time zone is in *(905691)*
21	2571	Opening hours & events	what time does shoe carnival close *(904553)*
22	201	Hiking trails	how long is the trail to hike the narrows *(265295)*
23	300	Energy & electricity	what energy source is used most by the usa *(657321)*
24	206	Latitude & longitude	latitude is measured from where *(438015)*
25	301	Restaurants	what restaurant are open *(891272)*
26	307	University locations	which state is harvard university located *(1019701)*
27	449	Company ownership	who owns the ritz carlton chain *(1045915)*
28	1801	Hotels & resorts	where is the resort that has the rooms over the water *(997370)*
29	476	Company headquarters	where is headquarters located *(984270)*
30	1797	Sports teams & venues	where is the college football playoff *(995090)*
31	375	TV & radio stations	what networks stream live tv *(881802)*
32	367	School calendars	what day does school start in florida *(616616)*
33	2368	Airports	what is closest airport *(731434)*
34	4494	Distances between places	how far is it between two cities *(231276)*
35	1082	Manufacturing locations	where does ford manufacture *(973047)*
36	469	Stores & retail chains	how many shops does walmart have *(295729)*
37	526	Where to buy products	Where Can I Buy Cruex *(8025)*
38	249	Burial sites	name of where dead bodies are buried *(461873)*
39	562	Shootings & violent deaths	how many deaths from the shooting in fl *(1097258)*
40	321	University rankings & accreditation	top ranked universities in us *(522650)*
41	247	Student enrollment numbers	how many students does osu have *(297340)*
42	391	University admissions / tests	average gpa acceptance at unr *(36767)*
43	250	Small towns & forts	what part of maine is portland in *(884763)*
44	237	Surname & place name origins	where does the name ireland come from *(974631)*
45	533	Cuisines by country	what are the names of two traditional foods in argentina *(571955)*
46	661	Origins of foods & things	where do chickens come from originally *(1141679)*
47	390	Bridges & tunnels	longest bridge in *(1174000)*
48	344	Tall buildings & towers	what is the tallest building in the world *(849724)*
49	526	Islands	what island is off the map *(865244)*

Table 6 (continued).

#	Size	Label	Representative query *(query ID)*
50	368	Lakes & natural water features	what is the largest freshwater lake in the united states *(827162)*
51	255	Native American tribes	where native american came from *(1000891)*
52	214	Pre-Columbian civilisations	what ancient civilizations were in mexico *(553539)*
53	396	Rivers	which river is the longest *(1139538)*
54	948	Mountains & elevation	which mountain is the tallest *(1013640)*
55	1015	US places	what is the county of florida *(813209)*
56	480	Foreign governments & leaders	what type of government does mexico have currently *(912604)*
57	328	Flags & national symbols	what do the colors of the flag say about the nation *(625172)*
58	440	Video game locations	where do i find quarried stone skyrim *(970929)*
59	1046	Ancient civilisations	what is the ancient name of the area *(804856)*
60	1080	Health & disease statistics	how many sick people in the united states *(295798)*
61	345	World population by country	what the world population *(903847)*
62	615	Oceans, water & hydrology	most of the water on earth is found where *(458406)*
63	210	Glaciers, rifts, ocean depths	what is the deepest part of the ocean *(814391)*
64	2403	US local population statistics	what is the population of wv *(1034203)*
65	811	Construction dates of landmarks	what year was the statue of liberty constructed *(928604)*
66	291	Mining & natural resources	where is silver ore found *(992908)*
67	245	Fish & freshwater fishing	what fish are in lake nottely *(1117541)*
68	271	Vietnam War	when did the us go to vietnam war *(942467)*
69	1182	Languages spoken	what languages are spoken *(872702)*
70	909	Regions and countries	asia where is it located *(27347)*
71	314	Plate tectonics	when oceanic crust and continental crust meet at the plate boundary *(954075)*
72	1045	Planets & space	what is the nearest planet to earth *(836079)*
73	330	Earth's layers and atmosphere	what layer of the earth is on the surface *(872996)*
74	561	Volcanoes	what type of eruption formed the volcano in this photograph *(912074)*
75	598	Rock types	granite is what type of rock *(196556)*
76	239	Germany & France geography	germanys geographic location *(194599)*
77	979	Italy, Spain & Mexico geography	what is italy's location *(761674)*
78	221	Pearl Harbor & WWII Japan	when did the japanese attack pearl harbor *(941810)*
79	236	Constitution & voting rights	how many states voted to ratify the constitution *(296971)*
80	236	Where plants & fruits grow	where do pomegranates grow *(971421)*
81	842	Gardening & plant zones	what planting zone am i *(887313)*
82	579	Biomes & ecosystems	place where desert are found and their plants *(475668)*
83	1594	Animal habitats	where are most of the animals located *(965819)*
84	794	Founding of colonies & countries	dates that colonies were founded quizlet *(115607)*
85	334	US statehood dates	what year did states become states *(926788)*
86	612	Battles & military history	what was the last major battle of the civil war *(921122)*
87	514	Germany & World Wars	what did the construction of the berlin wall do to germany *(619964)*
88	433	US places	in what county is washington wi *(394062)*
89	262	US places	what county in minneapolis mn *(602024)*
90	258	US places	lansing mi is in what county *(435687)*
91	1241	US places	austin texas is in what county *(29756)*
92	474	US places	what county richmond va *(615048)*
93	205	US places	what county is atlanta *(602628)*
94	258	US places	what county is kansas city mo in *(607886)*
95	456	US places	what county is new york ny *(610266)*
96	295	US places	what county is cambridge ma *(603814)*
97	467	US places	in what county is columbus, oh *(393920)*
98	382	US places	what county is chicago in illinois *(604215)*
99	555	US places	what county in lebanon pa in *(602014)*
100	213	US places	what county is jersey city nj *(607788)*
101	750	US places	what county is farmville nc in *(605858)*
102	342	US places	what county nashville tn in *(615017)*
103	869	US places	valencia is in what county in ca *(1089047)*
104	381	US places	what county portland in *(615046)*