Blowfish:

Topological and statistical signatures for quantifying ambiguity in semantic search

Thomas R. Barillot*†
BlackRock
London, UK

Alex De Castro*‡
BlackRock
London, UK



Neighbourhoods of ambiguous queries have, topologically speaking, a blowfish-like appearance. Image courtesy: Dall-E-3.

Abstract

This works reports evidence for the topological signatures of ambiguity in sentence embeddings that could be leveraged for ranking and/or explanation purposes in the context of vector search and Retrieval Augmented Generation (RAG) systems. We proposed a working definition of ambiguity and designed an experiment where we have broken down a proprietary dataset into collections of chunks of varying size - 3, 5, and 10 lines and used the different collections successively as queries and answers sets. It allowed us to test the signatures of ambiguity with removal of confounding factors. Our results show that proxy ambiguous queries (size 10 queries against size 3 documents) display different distributions of homologies 0 and 1 based features than proxy clear queries (size 5 queries against size 10 documents). We then discuss those results in terms increased manifold complexity and/or approximately discontinuous embedding sub-manifolds. Finally we propose a strategy to leverage those findings as a new scoring strategy of semantic similarities.

Prédire N'est Pas Expliquer (Thom [1992]) (To Predict is not to Explain)

— René Thom, 1991

1 Introduction

In the field of natural language processing, semantic matching between sentence embeddings often employs metrics such as Euclidean distance, dot product (inner product), or cosine similarity.

^{*}Both authors contributed equally to this work.

[†]thomas.barillot@gmail.com

[‡]castro.inbox@icloud.com

However, the effectiveness of these measures can be undermined by the fact that the manifold constructed by sentence embeddings is unlikely to be globally smooth, and it may not be locally smooth either. The nature of the semantic manifold, often hypothesized to be low-dimensional based on empirical evidence and intuition, remains elusive yet critical. Literature contains attempts to smooth manifold representation for words and sentences (Hasan and Curry [2017], Kayal [2021], Chu et al. [2023]) but they can be seen as fine-tuning exercises that are by construction limited to a certain domain. The elusive structure of the semantic manifold and its probable discontinuous nature are particularly problematic for ambiguous queries in question-and-answer processes, which may consist in multi-factual or too vague questions. Recent studies have leveraged Topological Data Analysis (TDA, Carlsson and Vejdemo-Johansson [2021]) to interpret the polysemic nature of words (Jakubowski et al. [2020]), indicating local discontinuities within an otherwise smooth word embedding manifold. Building on the concept of polysemy in words, we promote a working and computational definition of ambiguity in sentences based on our experience with vector search and RAG (Gao et al. [2024]) systems in tackling disambiguation problems.

Polysemy refers to a word having multiple meanings, which occurs when a word belongs to more than one semantic domain. Similarly, we examine the *relative ambiguity* in sentences within a corpus. A corpus $\mathcal C$ comprises n sentences, or text chunks, represented as $\mathcal C=\{c_1,...,c_n\}$, with each sentence potentially fitting into one or more of m identified semantic domains $S_{\mathcal C}=\{s_1,...,s_m\}$, where $m\leq n$. From here onwards, the notation $\mathcal C_{/j}$ means the corpus $\mathcal C$ with c_j removed. Also for every chunk c_j we will associate a subset of semantic domains S_{c_j} . A sentence c_j is more ambiguous or polysemic than c_k if $|S_{c_j}|>|S_{c_k}|$.

We now consider a (typically new) sentence or chunk c' with a topic set $S_{c'}$ polysemic relative to a corpus C if it meets any of the following **working definitions**:

• It belongs to more semantic domains than the sentence in the corpus with the fewest domains:

$$|S_{c'}| > \min_c(|S_c|)$$
 with $S_{c'} \subset S_{\mathcal{C}}$

• It introduces at least one semantic domain not present in other sentences:

$$S_{c'} \not\subset S_{\mathcal{C}}$$
 and $S_{c'} \neq \{\emptyset\}$

• It does not belong to any recognized semantic domain:

$$S_{c'} = \{\emptyset\}$$

Notice that c' may or may not pertain to the original corpus \mathcal{C} . In applications, new chunks c' represent unseen incoming queries. Any chunk c_j will be associated with a single semantic domain $S_{c_j} = \{s_k\}$ if the set S has been determined, for example, through clustering. It is important to note that clustering of sentence embeddings has been successfully explored with Local Density Approximation or more recently BERTopic (Grootendorst [2022]) to represent semantic domains (called **topics**). Therefore it follows that relative ambiguity reflects the fact that incoming queries may be associated with multiple pre-existing topics. This can have a degrading impact on RAG systems that depend on clear semantic context for providing answers with high utility for end users. Our goal is to be able to quantify and observe the degree of ambiguity in a sentence c in relation to the corpus \mathcal{C} , so we can mitigate or correct the effects of the additional relative ambiguity when retrieving information or generating new content based on an existing context window.

We do not claim that this is the only possible definition. We welcome feedback from the computational NLP community to understand if there are other practical and computable definitions, and whether they can be proven mathematically or experimentally to be equivalent.

Intuitive definition: An ambiguous sentence is formally defined as a sentence whose embedding satisfies the conditions of connecting to more topics within its corpus than the minimal connections established by other sentences, or not fitting neatly into any predefined topic of the corpus.

Toy Example:

- Consider sentences involving the word *crane*:
 - Sentence 1: "He craned his neck to see the construction site."

- Sentence 2: "The crane lifted the heavy beams effortlessly."
- Sentence 3: "A crane flew over the lake at sunset."
- topics might include {action, machine, bird}.
- Sentence 1 connects to {action}, Sentence 2 to {machine}, and Sentence 3 to {bird}.
- To illustrate polysemy further:
 - "He craned his neck to watch the crane lift near the crane."
- This sentence employs all three topics, making it a polysemic example where the word "crane" connects to {action, machine, bird} simultaneously.

The generalization from word polysemy to sentence ambiguity raises research questions that we propose to explore:

- 1. Does an ambiguous sentence embedding display a distinctive geometric and topological signature within its neighborhood, as observed with polysemic word embeddings?
- 2. And if so, is the signature dependent on a specific training dataset or embedding model architecture?

To address these questions, we conducted a retrieval experiment where we tested queries that either relate to a single or multiple topics covered in the corpus of answers. We designed topological features using TDA toolbox and evaluated their sensitivity to the presence or absence of multiple topics. Key findings include the exploration of the relative nature of sentence ambiguity, demonstrating how the number of concepts contained in the query and corpus affect the occurrence of local manifold discontinuities. We believe our findings can be used to design strategies to improve retrieval in RAG systems.

2 Related works

Detecting and mitigating effects of queries ambiguities in RAG pipelines is the object of recent contributions. In particular CRAG (Yan et al. [2024]) explicitly categorise in {"Correct","Ambiguous","Incorrect"} a retrieved chunk of text for a given query using a T5-large fine tuned model and subsequently triggers different actionables. That work reports a whole retrieval augmentation pipeline and is an elegant solution but does not give insights about explainability of incorrect retrievals in terms of ambiguity and is domain dependent as it relies on a fine-tuned language model. SELF-RAG (Asai et al. [2023]) also evaluates the relevance of each retrieved chunks given a specific query but by directly using the Large Language Model at generation step. This model is fine-tuned to generate additional relevance tokens to the output in order to trigger generation or a new retrieval loop. As for CRAG, this works focuses on producing flags to iterate the retrieval process but does not interrogate the notion of ambiguity or relevance.

It is also worth mentioning that research groups also question widely used scoring methods such as cosine similarity as unique metric for semantic similarity (Zhou et al. [2022], Steck et al. [2024],) mostly due to words frequency induced biases in training (Wannasuphoprasit et al. [2023]). These questions call for an exploration of additional metrics to characterize the semantic search results. In investigating queries ambiguity, we distinctively focus on an understanding of domain-specific polysemy using topological and geometric methods.

3 Methodology

In order to investigate topological signatures of ambiguity in a semantic search context we evaluated several embedding models and we built specific features from the query sentence and answer sentences embedding vectors. We opted for a strategy consisting of choosing document chunks for queries and answers but with different chunk sizes which has the advantage of correlating the amount of topics in queries and answers and therefore limiting potential confounding factors in the observed topological signal. Synthetic queries generation does not guarantee such a degree of control to evidence the meaningful signals which explain why we have not chosen this popular dataset generation.

The generic experimental design shown in figure 1 is the following:

- 1. **Chunk Formation:** Divide dataset \mathcal{D} into chunks of varying sizes based on the number of text lines $(L \in \mathcal{I} = [3, 5, 10])$, creating chunk sets $\{\mathcal{C}_L\}_{L \in \mathcal{I}}$ (table 1).
- 2. **Setup** Use C_x as the corpus C and C_y as queries set Q. Generate embeddings for both Q and C. Index C and generate topics labels via UMAP and HDBSCAN.
- 3. **Retrieval and TDA:** Perform semantic search, retrieve top-k=50 corpus chunks. Filter chunks on the conditions $c_i \subset q_j$ or $c_i \supset q_j$ where c_i is corpus chunk i token sequence and q_j is query j token sequence. Tag queries based on multiplicity of topics they relate to.
- 4. **Evaluation:** Derive homology based features and identify ambiguous from unambiguous queries signatures in those features.

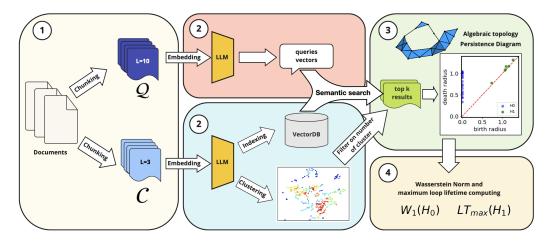


Figure 1: Experimental pipeline: Documents in dataset \mathcal{D} are divided into chunks of varying sizes to create a collection of chunk sets $\{\mathcal{C}_L\}_{L\in\mathcal{I}}$. One set, \mathcal{C}_L , represents the corpus \mathcal{C} and another serves as query set \mathcal{Q} . The topics of the corpus set are defined and used to evaluate the ambiguity of queries from other sets.

3.1 Datasets and Embedding Models

We conducted our experiments using a proprietary dataset from BlackRock (BLK), consists of 26 documents and chunked with variable sizes (table 1). Those documents consist in "how-to" documentation for usage of the BlackRock Aladdin ExploreTM platform. They contain mostly free text, caption of figures (images are not stored) and a few shallow tables (with a maximum of 4 rows). Despite the private and confidential nature of the dataset, the structures and layouts of its the documents are generic enough for the results presented here to be applicable to any other publicly available dataset.

Dataset	chunk size	Nchunks
\mathcal{D}_{BLK}	10	2574
_	5	7192
_	3	22404

Table 1: Overview of Datasets sizes

Embedding Model	embedding size
SBERT-multiqa(Reimers and Gurevych [2019], Huggingface [2022])	768
SBERT-msmarco (Reimers and Gurevych [2019], Huggingface [2021])	-
GTR-T5-XL (Ni et al. [2021])	-

Table 2: Overview of Embedding Models and Datasets

We conducted semantic searches using the cosine similarity function for each model and dataset, and collected the top-k = 50 retrieved answers for each query. We embedded the content of these datasets

using three different embedding models, as detailed in Table 2 (and appendix C), then stored each resulting vector embeddings in separate FAISS (Douze et al. [2024]) vector store indexes. These models were selected due to the diversity of corpora used in their pre-training. None of these models were fine-tuned on our datasets for the experiment.

3.2 Computational proxy for topics

Despite the established references for word-level polysemy, equivalent references for sentences are not feasible due to complexity and dependency on text corpora. However, we can approximate a sentence semantic domain by clustering sentences as **topics** in an embedding space. Below, we outline the steps involved using UMAP for dimensionality reduction and HDBSCAN for clustering. Our approach is identical as the one stated in Grootendorst [2022] and more details justifying this approach on be found there.

Algorithm 1 Identify topics

```
1: Input: Embedding vectors set \{v_E\}, HDBSCAN parameters, UMAP parameters 2: Output: Topic labels 3: procedure GENERATETOPICVECTORS( \{v_E\}, HDBSCAN params, UMAP params) 4: \{v_E'\} \leftarrow \text{UMAP}(\{v_E\}; n_{\text{neighbors}}, \min_{\text{dist}}, n_{\text{comp}}) 5: S_{\mathcal{C}} \leftarrow \text{HDBSCAN}(\{v_E'\}; m_{\text{size}}, m_{\text{samples}}, \text{method}, \epsilon) 6: return labels = \{l_1, l_2, \dots, l_n\} with l_i \in S_{\mathcal{C}}\} 7: end procedure
```

For hyperparameters tuning and additional details, see Appendix D. The quality of the clusters obtained through HDBSCAN is assessed using silhouette scores, which help evaluate the separation distance between the resulting clusters.

Advantages of the Selected Methods: UMAP is preferred over PCA and t-SNE due to its robust handling of local structures and scalability, making it well-suited for large datasets. HDBSCAN is chosen for its effectiveness in identifying clusters with varying densities, which is crucial for the nuanced clustering required in semantic domain analysis. For a comprehensive understanding of UMAP's methodology, refer to the seminal paper McInnes et al. [2020].

3.3 Query neighbourhood

The features employed in our experiment are derived from the relative positions of the query embedding vector \vec{v}_q and the embedding vectors of its k nearest neighbors in the answer corpus, denoted by $\{\Delta \vec{v}_{iq}\}_{i=0}^{k-1}$ where $\Delta \vec{v}_{iq} = \vec{v}_i - \vec{v}_q$.

Each neighbour of the query relative distance is represented as a scaling factor ε_i w.r.t the relative distance with the nearest neighbour with $\varepsilon_0 = 0$.

$$d(i,q) = \frac{\Delta \vec{v}_{iq}}{\Delta \vec{v}_{0q}} \equiv 1 + \varepsilon_i \tag{1}$$

We opted for this representation as it reflects the local neighbours density which the ranking index k does not. We can then tune an arbitrary scaling factor $\varepsilon_1 \leq \varepsilon \leq \varepsilon_k$ and observe the evolution of topological features. The value of k is taken sufficiently high to allow probing both local and global neighbourhood.

Topological features are constructed based on persistent homology of order 0 and 1, denoted as H_0 and H_1 . Following the methodology presented in Jakubowski et al. [2020], we calculate them by first normalizing the nearest neighbor vectors:

$$\frac{\Delta \vec{v}_{iq}}{\|\Delta \vec{v}_{iq}\|}$$

and applying the persistent homology algorithm, as detailed in Appendix A and referenced in Tauzin et al. [2021]. Practically, Wasserstein distance and maximum hole lifetime are chosen to quantify H_0 and H_1 respectively.

3.4 Understanding Wasserstein Distance in Topological Data Analysis

The Wasserstein distance, also known as the earth mover's distance, quantifies the "work" required to transform one distribution into another. In topological data analysis, this measure is particularly useful when applied to persistence diagrams. The diagonal of a persistence diagram ideally represents features (often considered as noise) that appear and disappear at the same scale, thus serving as a baseline for comparison.

For a set of points in a persistence diagram that represent the death times of homology-0 features (e.g., regular lattices), the 1-Wasserstein distance can be conceptualized as the cost of moving this distribution onto the diagonal. This distribution of death times on the y-axis not only reflects the moment each feature disappears but also indicates the granularity of the lattice at which these features resolve. Each y-value, therefore, represents the radius within which a particular regular lattice ceases to exist, providing insights into the scale of data concentration and granularity.

The practical calculation of the 1-Wasserstein for H_0 distance involves:

$$W_1(H_0) = \frac{1}{N-1} \sum_{y \in H_0} |y - \gamma_{\perp}(y)| \tag{2}$$

where N is the number of points in H_0 , and $\gamma_\perp(y)$ is the y-coordinate of the orthogonal projection of each point onto the diagram's diagonal. This calculation measures how far each point deviates from the baseline (diagonal), it implicitly weighs the features by their granularity. Each distance moved in the calculation represents a weighted adjustment of the data's granularity, reflecting varying scales of data clustering and feature resolution. Please see the appendix for a detailed derivation of the Wasserstein distance using the Wasserstein norm for the H_0 distribution, which underscores the importance of granularity in interpreting topological data.

3.5 Maximum loop lifetime

The next homology order H_1 , which captures 1-dimensional loops in the data, is more sensitive to noise and therefore the usage of equation 2 would artificially reduce long-lived loops weights, therefore we opted for representing the longest-lived loop in the persistence diagram:

$$LT_{max}(H_1) = \sup_{(x,y)\in H_1} [|y - \gamma_{\perp}(y)|]$$
 (3)

This increased sensitivity of H_1 features to noise arises from the complexity and the intricate structures they capture. Noise can introduce spurious loops with short lifespans, affecting the persistence of H_1 features. Identifying these loops requires careful scale selection, as noise impacts data at multiple scales, leading to false positives. Additionally, noise alters local density and distribution, creating artificial loops. Computational algorithms for H_1 are sensitive to point distances, and noise perturbs these distances, causing instability. Furthermore, noise can distort the assumed low-dimensional manifold, resulting in incorrect H_1 feature identification. These factors collectively make H_1 more prone to noise than H_0 features.

4 Results

In our experiments, the set C_3 is initially selected as the corpus C, with chunks from C_{10} used as queries set Q. This configuration allows for the inference of query-related topics, as at least one chunk in C_3 contains token sequences that are also present in the query chunks, thereby satisfying the first condition of our ambiguity definition. By design, each corpus chunk encapsulates a single topic, while query chunks can encompass one or multiple topics. We apply topic clustering to discern and isolate queries that retrieve single-topic chunks, directing our focus to those that demonstrate the ambiguity characteristic under scrutiny.

The experimental setup is reversed by selecting C_{10} as the corpus and C_5 as the queries. In this scenario, the queries are generally associated with a single topic, though the potential for the embedding model's semantic bias to link a small chunk query to multiple topics cannot be dismissed. Here too, topic clustering assists in filtering out queries that correlate to multiple topics, ensuring a focused analysis on relevant data points.

To maintain the integrity of our experiments, we stipulate that any retrieved corpus chunk, denoted as $c_i = [w_1^i, \ldots, w_m^i]$, must fully contain or be contained by the query chunk in both experimental configurations $(\mathcal{Q}; \mathcal{C})$: $(\mathcal{C}_{10}; \mathcal{C}_3)$ and $(\mathcal{C}_5; \mathcal{C}_{10})$. This ensures that the semantic coherence between the chunks is adequately maintained, allowing for accurate assessment of query ambiguity and topic distribution within the chunks.

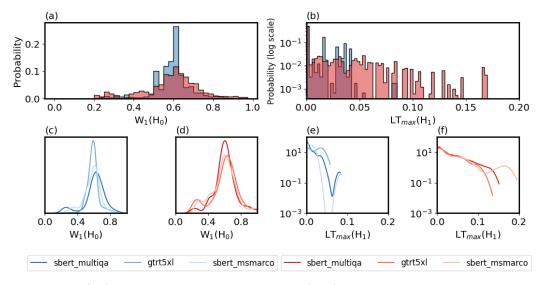


Figure 2: a) $W_1(H_0)$ distributions for pairs of queries-corpus $(\mathcal{Q};\mathcal{C})$ chunk sets $(\mathcal{C}_5;\mathcal{C}_{10})$ and $(C_{10};\mathcal{C}_3)$ in blue and red respectively for neighborhood scale $\varepsilon=0.3$. b) Max H_1 loop lifetime distributions for $(\mathcal{C}_5;\mathcal{C}_{10})$ and $(C_{10};\mathcal{C}_3)$ in blue and red respectively with neighborhood scale $\varepsilon=0.3$. c)-d) $W_1(H_0)$ Distribution breakdown per embedding model for $(C_5;C_{10})$ and $(C_{10};C_3)$ respectively as KDE plots. e)-f) $LT_{max}(H_1)$ Distribution breakdown per embedding model for $(C_5;C_{10})$ and $(C_{10};C_3)$ respectively as KDE plots in log scale.

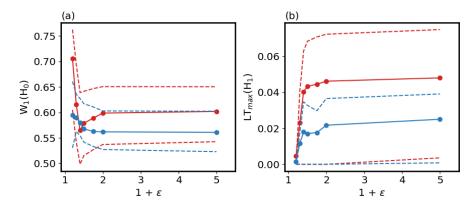


Figure 3: a) Mean $W_1(H_0)$ (plain lines) and 25-75% inter-quantile range (dashed lines) for $(C_5; C_{10})$ and $(C_{10}; C_3)$ in blue and red respectively as a function of the relative neighborhood scaling factor ε . b) equivalent plot for $LT_{max}(H_1)$.

In the experimental pairing of $\mathcal{Q}=\mathcal{C}_5$ and $\mathcal{C}=\mathcal{C}_{10}$, where the predefined conditions of ambiguity are not met, the H_0 signal primarily cluster around a narrow single peak at $W_1(H_0)=0.6$. The absence of significant variation in neighbor arrangements, even with adjustments to the neighborhood scaling factor ε (as shown in figure 3-a), indicates a lack of structured topological features such as pinched manifolds within these queries.

In contrast, the pair $\mathcal{Q}=\mathcal{C}_{10}$ and $\mathcal{C}=\mathcal{C}_3$ demonstrates a significantly broader distribution with the mean shifted to a higher value of $W_1(H_0)\approx 0.7$. This broader spread is a partial indication of a more structured neighborhood. H_1 which capture the lifetime of loops existing in the data points cloud exhibits more persistent elements as well as shown in figure 2-b. As the neighbourhood evolves from local to global (by varying ε), the unambiguous queries also start to exhibit persistent loops.

Those loops can be interpreted as discontinuities that mark separation between different semantic clusters, reinforcing the picture of a pinched manifold that reflects a complex and discontinuous local embedding space surrounding the query chunks. This behavior is observed across different embedding models both for H_0 and H_1 demonstrating that topological signatures of ambiguity are invariant with embedding model training corpus

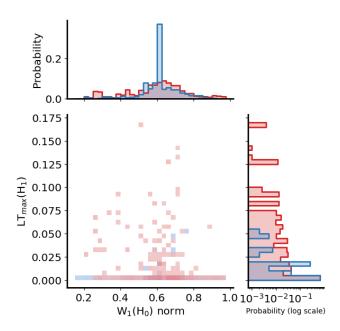


Figure 4: Samples joint distribution of H_0 and H_1 signatures for both unambiguous and ambiguous conditions (blue and red distributions respectively)

Notably, the ambiguous $W_1(H_0)$ distribution progressively narrows and get closer to the unambiguous and as the neighborhood scaling ratio ε increases, the proximity of more topics near the queries leads to an Wasserstein norm that eventually stabilizes to a point beyond which it becomes more challenging to distinguish between structured neighbourhood containing multiple topics and single topic one. Increasing the neighborhood size by tuning ε also leads to more persistent loops in average for both ambiguous and unambiguous queries but with an increased scaling in the mean lifetime of ambiguous queries loops.

Finally the sample joint distribution of $W_1(H_0)$ and $LT_{max}(H_1)$ (figure 4) emphasizes the existence of long lived loops across the range covered by the homology 0 signature. Ambiguous queries contain proportionally more loops with longer lifetimes, helping to disentangle their signal from clear queries one. We can compute the empirical conditional probability density:

$$p(LT_{max}(H_1)|W_1(H_0);\varepsilon) = \frac{p(W_1(H_0), LT_{max}(H_1);\varepsilon)}{p(W_1(H_0);\varepsilon)}$$
(4)

and we observe a non trivial correlation between our homology-0 Wasserstein norm and existence of long lived loops likelihood (figure 5). This observation is unlikely to be a statistical artifact as it is confirmed for $\varepsilon=4.0$ (figure 5-b) where statistics is an order of magnitude higher. In the range $W_1(H_0)\in[0.8,1.0]$, long lived loops are almost nonexistent while the probability of having such long lived loops crosses 50% in the [0.2,0.3] range. We only observe this behaviour for ambiguous queries. For clear ones, despite the existence of a couple of samples that contain loops (see figure 4), the computed conditional probability does not exceed 1%, far from the results observed on ambiguous ones and justifying why it is not reported in figure 5. This reinforces the use of persistent loops to discriminate ambiguous from clear queries especially around the region of $W_1(H_0)=0.6$ where the two connected components distributions overlap strongly.

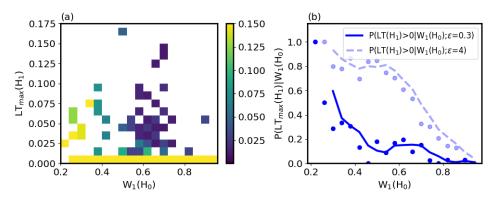


Figure 5: a) Empirical conditional probability density function map $p(LT_{max}(H_1)|W_1(H_0);\varepsilon=0.3)$ for ambiguous queries only. b) Corresponding cumulative probability density function $p(LT_{max}(H_1)>0|W_1(H_0);\varepsilon=0.3)$ and $p(LT_{max}(H_1)>0|W_1(H_0);\varepsilon=4)$ as a function of $W_1(H_0)$.

5 Discussion and perspectives

5.1 Homology encodes ambiguity

Our findings underscore H_0 and H_1 homology signatures captured by $W_1(H_0)$ and $LT_{max}(H_1)$ as powerful tools to differentiate the two sets of query-corpus we explored. The design of our experiment aligns with our formal definition of sentence ambiguity/clarity therefore we conclude that homology capture ambiguity akin it can be used for words polysemy. The apparent similar behaviour of H_0 and H_1 features distributions across embedding models reinforces their alignment with the formal definition of ambiguity. Indeed in our definition, the notion of semantic domain or topic is an abstract concept that inevitably vary from one embedding model to another and that is approximated by clustering. Once the data filtered by neighbouring clusters, the homologies differences appear in all tested models, therefore demonstrating the general nature of homologies signature of ambiguity. In that context, it is important to note that a finer analysis of H_0 and H_1 distribution open the door to a richer set of features for detecting and quantifying ambiguity. An example is counting the number of persistent loops per sample after having filtered out statistically insignificant ones using techniques such as Kernel Density Estimation (KDE) or Distance To a Measure (DTM) as suggested in Chazal et al. [2011], Wasserman [2016]. Nevertheless, given the limited number of nearest neighbours in the vicinity of the query it will likely be challenging to derive statistically meaningful signals from higher order homologies.

5.2 Geometry intuition behind homology signatures

Unambiguous or clear samples appear to distribute narrowly around $W_1(H_0)=0.6$. Given the fact that a single topic cluster is involved in this case, it is possible to envision a simple simulation (detailed in appendix B) in order to link the norm distribution in terms of neighbourhood orientation anisotropy. The observed behaviour from the data and the simulation is an anisotropy of neighbourhood, corresponding to an anisotropy factor $\alpha=15$ (refer to appendix B) for the definition). One topological consequence is that the query is located outside the convex hull created by the (k-1)-dimensional simplex generated by the k neighbours. It appears to be a geometric signature of queries with higher clarity nevertheless not sufficient to fully attribute it. In the context of unknown topics clusters multiplicity in the neighbourhood, our simulation shows that the multiplicity of clusters distributed across orthogonal dimensions and their corresponding effective anisotropies vary significantly for a single $W_1(H_0)$ value. Therefore, if in the regions $W_1(H_0) \in [0.2, 0.4] \cup [0.9, 1.0]$ this single feature is an acceptable signal, it gets more complex between those two intervals and H_1 signature appears then to be needed to discriminate clear from ambiguous queries at various degrees. With similar apparent H_0 feature value, indeed long-lived loops, characterised by $LT_{max}(H_1)$, appear to be more probable in the case of ambiguous queries in our experiment.

The presence such long lived loops could approximate discontinuities in the manifold underlying the neighbourhood in the case of ambiguous queries. It is broadly accepted that Language Models encode embeddings on a low dimensional sub-space. Manifold smoothing strategies for SBERT as well as

recent successes of highly compressed models (Lee et al. [2024], Kusupati et al. [2024]) add empirical evidence to that conjecture. The apparent high degree of anisotropy of the projected neighbourhood of a query onto a sphere S^n seems in agreement with a set of neighbours contained in a much lower dimensional sphere S^m with $m \ll n$, compatible with a semantic manifold vision. Under this assumption, the presence of persistent loops could corresponds to punctures in the manifold. As a consequence, it would constitute semantic discontinuities between the query neighbours, thus reinforcing the idea that the query lies in a manifold singularity. One can note rule out either that the embedding space \mathbb{R}^n of the tested model here contain more than a single low dimensional semantic manifolds and that the query lies at their intersection.

5.3 Perspectives: the BLOWFISH toolbox

We have shown empirical evidence for ambiguity signatures in query-chunk homology derived features. Supported by the experimental results, we believe that those features and others derived from H_0 and H_1 could be used to improve semantic search, particularly as they are embedding model architecture and training data agnostic (to the extent of the models we could test). We propose here a methodology for detecting ambiguity in sentences using homology signatures and kernel density estimation (KDE) under the name of BLOWFISH toolbox. This method allows us to express ambiguity as a probability for effective ambiguity/clarity detection. It builds on evidence from controlled experiments and incorporates several components for enhancing the detection process.

The BLOWFISH toolbox

Components

1. Homology Signatures

Identify homological signatures H_0 and H_1 for a specified scaling factor ε :

$$f_{homology}(\varepsilon)$$

2. Ambiguity Score construction

Construct an ambiguity score based on homology features:

$$P(A|f_{homology}(\varepsilon), \rho_{queries}(\varepsilon), \theta_{embedding}) \equiv \text{KDE}(f_{homology}(\varepsilon), \rho_{queries}(\varepsilon), \theta_{embedding})$$

where P(A|.) is the probability that a query is ambiguous, $\rho_{queries}(\varepsilon)$ is the density of neighbours for a particular scaling factor ε (defined in appendix A.4) and $\theta_{embedding}$ representing the weights of the embedding model.

3. Model Independence

The ambiguity score is empirically independent of the embedding model of the datapoints it is applied to:

$$P(A|f_{homology}(\varepsilon), \rho_{queries}(\varepsilon), \theta_{embedding}) \rightarrow P(A|f_{homology}(\varepsilon), \rho_{queries}(\varepsilon))$$

Notably, the ambiguity score is independent of the embedding model used. This independence enables the use of smaller, cost-effective embedding models for fitting the KDE, which can then be applied to larger models, making the process efficient and scalable. It is also important to emphasize that whereas we ruled out the use of synthetic queries-answers datasets for this work which focuses on identifying signals, we believe that this method could be leveraged to build ambiguity score models with robust statistics.

6 Conclusion and Next Steps

We have designed an experiment to identify and isolate topological signatures of ambiguity in a sentence, itself defined as a generalisation of words polysemy. We identified at least two signals derived from homologies H_0 and H_1 that exhibit different behaviour in ambiguous and unambiguous conditions. In addition to the discriminative utility of those signals they inform us of the geometric and topologic arrangement of query sentences neighbourhood which can be used as another probing tool of Large Language Models representation and structuring of information. As those signature

seem independent of the language model used in our experiment, further investigations will explore how this relationship holds across a broader range of embedding dimensions and models, aiming to refine our understanding of semantic domain analysis in textual data. It is crucial to understand this in order to properly evaluate uncertainty of similarity metrics widely used in semantic search pipelines such as RAG (Gao et al. [2024]).

Acknowledgments

Special thanks to: Javier Makmuri, Carlos Tomei, and Rob Murray-Rust for helpful discussions and varying levels of technical and emotional support. Special thanks to kids and spouses too (writing time is time away from family).

Disclosure. All opinions are our own and don't represent that of BlackRock. Results here are not meant for investment purposes, and authors cannot be held liable to any misuse of the concepts proposed here.

References

- René Thom. Prédire n'est pas expliquer. *Revue Philosophique de la France Et de l'Etranger*, 182(2): 262–265, 1992.
- Souleiman Hasan and Edward Curry. Word re-embedding via manifold dimensionality retention. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 321–326, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1033. URL https://aclanthology.org/D17-1033.
- Subhradeep Kayal. Unsupervised sentence-embeddings by manifold approximation and projection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.
- Yonghe Chu, Heling Cao, Yufeng Diao, and Hongfei Lin. Refined sbert: Representing sentence bert in manifold space, 2023.
- Gunnar Carlsson and Mikael Vejdemo-Johansson. *Topological Data Analysis with Applications*. Cambridge University Press, 2021.
- Alexander Jakubowski, Milica Gašić, and Marcus Zibrowius. Topology of word embeddings: Singularities reflect polysemy, 2020.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.
- Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation, 2024.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.
- Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. Problems with cosine as a measure of embedding similarity for high frequency words, 2022.
- Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In Companion Proceedings of the ACM on Web Conference 2024, WWW '24, page 887–890, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701726. doi: 10.1145/3589335.3651526. URL https://doi.org/10.1145/3589335.3651526.
- Saeth Wannasuphoprasit, Yi Zhou, and Danushka Bollegala. Solving cosine similarity underestimation between high frequency words by 12 norm discounting, 2023.

- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
- Huggingface. multi-qa-mpnet-base-cos-v1. https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-cos-v1, 2022.
- Huggingface. msmarco-distilbert-base-v4. https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4, 2021.
- Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Abrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers, 2021.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library, 2024.
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.
- Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Wojciech Reise, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2021.
- Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures, 12 2011.
- Larry Wasserman. Topological data analysis, 2016.
- Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models, 2024.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning, 2024.
- Michael Kerber, Dmitriy Morozov, and Arnur Nigmetov. Geometry helps to compare persistence diagrams, 2016.
- 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, 2018.
- Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, 1987. ISSN 0377-0427. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Appendices

A Geometric and Topological Foundations in Query Analysis

This appendix delves into the geometric and topological underpinnings essential to our analysis, particularly focusing on the application of homology theories to discern patterns within data clusters, cycles, and cavities. Our investigation employs the giotto-tda library, a versatile tool for Topological Data Analysis (TDA), enabling us to explore complex structures within high-dimensional data spaces. The rationale behind utilizing giotto-tda lies in its robust computational framework for extracting topological features, pivotal for understanding ambiguities in query datasets.

A.1 Understanding Homologies

Homologies offer a window into the intrinsic shape of data, categorized by dimensions:

- 0-Homology focuses on clusters, identifying connected components within the data.
- 1-Homology reveals loops, cycles, and tunnels, offering insights into data's cyclical structure.
- 2-Homology captures spheres or cavities, indicating voids within the data framework.
- 3-Homology, our study's top homology, identifies complex 3D simplices, necessitating at least five points for construction.

A.2 Application of giotto-tda

The giotto-tda library is instrumental in computing these topological features. By employing its Vietoris-Rips complex generator, we analyze the renormalized embeddings of our dataset to produce persistence diagrams. The Vietoris-Rips complex is a simplicial complex that is pivotal for understanding the shape of data. It is constructed by connecting points that are within a certain distance from each other, thus capturing the underlying topological structure of the data at different scales.

These diagrams, in turn, facilitate the examination of homological features such as the birth of homologies, indicative of the scale at which points in a point cloud start to connect, reflecting the overall density and distribution of points.

A.3 Derivation of 1-Wasserstein Norm

The Wasserstein distance is typically used for comparing dissimilarities between distributions. More specifically it has a use case to compare topological spaces from point clouds by the means of quantifying similarities between persistence diagrams (Kerber et al. [2016], Wasserman [2016]) for each homology orders separately. As an example for two persistent diagrams A and B of homology H_0 , the general formula of the Wasserstein distance in the "giotto-tda" (Tauzin et al. [2021]) package is written as:

$$W_p(A, B) = \inf_{\gamma: A \cup \Delta \to B \cup \Delta} \left(\sum_{u \in A \cup \Delta} \| x - \gamma(u) \|_{\infty}^p \right)^{1/p}$$
 (5)

where Δ corresponds to the virtual set of diagonal points in the diagram that ensures that there is a full bijection between A and B, and $\|.\|_{\infty}$ accounts for $\max(|x|,|y|)$ with $(x,y) \in \mathbb{R}^2$ the coordinate of a point in the diagram. In our experiments we use W_1 , homology H_0 so $\max(|x|,|y|) \to |y|$ as all components lie on the vertical axis (x=0,y)) and instead of comparing two diagrams A and B we will calculate the norm for diagram A only which means that we replace B by the empty set $\{\emptyset\}$. Finally we can discard the $\inf \lim C$ condition as the orthogonal projection of any point on (x=0,y) onto the diagonal axis (x,y=x)) is the function that minimizes the distances. Therefore the formula transforms into:

$$W_1(A) = \sum_{y \in A} |y - \gamma_{\perp}(y)| \tag{6}$$

Where $\gamma_{\perp}(y)$ is the y-component of the orthogonal projection $\gamma(y)$ on the diagram diagonal.

A.4 Renormalization

For a given scaling factor ε , the distributions of nearest neighbours are spanning over a different fraction of the top-k retrieval results (with k=50) between unambiguous and ambiguous configuration as shown in figure 6. As much as this density of neighbours is already indicative of whether a query aggregates different topics we want to compare ambiguous from unambiguous queries with the same settings and therefore for each value of ε we limit ourselves to a number of nearest neighbours bounded by $\min(\max(\mathcal{N}_{ambiguous}), \max(\mathcal{N}_{unambiguous}))$ with \mathcal{N} being the set of neighborhood sizes.

In addition those distributions are not uniform because we picked an arbitrary sample of documents to build \mathcal{D}_{BLK} which is why we correct the persistent homology features distributions by weights corresponding to:

$$w = 1/\rho_{queries}(N_{NearestNeighours}, \varepsilon)$$

where $\rho_{queries}(N_{NearestNeigbours}, \varepsilon)$ is the density of queries function of the neighbourhood size and the scaling factor ε

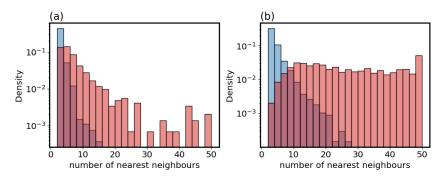


Figure 6: a) Empirical distribution of queries neighbourhood size with scaling factor $\varepsilon=0.3$ for ambiguous and ambiguous configurations in blue and red respectively. b) same distribution for a scaling factor $\varepsilon=4$

A.5 Pseudo-code for Calculating Homology features

Algorithm 2 Calculate Homology Birth Metrics

- 1: **Input:** Renormalized embeddings v_E
- 2: Output: Homology 0 and 1 metrics, $W_1(H_0)$ and $LT_{max}(H_1)$
- 3: **procedure** CALCULATEHOMOLOGYMETRICS(v_E)
- 4: Initialize Vietoris-Rips complex with the renormalized embeddings v_E
- 5: Fit the complex to the embeddings
- 6: Extract the first persistence diagram
- 7: Extract homology (birth,death) ordered pairs from the diagram
- 8: Compute $W_1(H_0)$ using Wasserstein distance
- 9: Compute $LT_{max}(H_1)$ using the lifetime of the longest feature
- 10: end procedure

Incorporating these insights enhances our understanding of the dataset's structure, offering a nuanced view of how topic diversity impacts the geometric and topological features we observe. Through this detailed analysis, we aim to provide a comprehensive framework for assessing query ambiguity, leveraging the sophisticated tools offered by giotto-tda to navigate the complexities of high-dimensional data spaces.

B Geometric intuition of topological features

The 1-Wasserstein norm of connected components (i.e.: $W_1(H_0)$) should carry some information about the anisotropy and the local density of the projected query neighbourhood on the hypersphere

- S^n . Nevertheless it is non-trivial to directly infer geometric information from that quantity. In order to build our intuition we considered a toy model where we define:
 - The dimensionality of the space D
 - The number of neighbours N_{neigh} we want to represent.
 - Their equipartition in $N_{\rm clust}$ clusters.

Together, N_{neigh} and N_{clust} provide a control on the density of the neighbourhood. In addition we define a scaling factor α that biases the sampling of a neighbour vector along a specific dimension i. All neighbours from the same cluster are sampled with the bias along the same dimension. By doing so we can control the anisotropy of the neighbourhood distribution. $W_1(H_0)$ can therefore be calculated for each quadruplet of parameters $(D, N_{\text{neigh}}, N_{\text{clust}}, \alpha)$ by the algorithm:

Algorithm 3 Anisotropic W_1 sampling in the space S^n

```
1: Input: D, N_{\text{neigh}}, N_{\text{clust}}, \alpha
 2: Output: W_1(H_0)
 3: procedure SAMPLEDWASSERSTEIN(D, N_{\text{neigh}}, N_{\text{clust}}, \alpha)
          For i in range(N_{clust}):
 5:
                Define the covariance matrix \Sigma as the Identity matrix I
 6:
                Substitute \Sigma_{ii} = 1 for \Sigma_{ii} = \alpha
                 Draw N_{\text{neigh}}/N_{\text{clust}} sample vectors \vec{v}_k = (x_0, \dots, x_{n-1}) with x_{j=i} sampled from
 7:
     \mathcal{N}(\vec{0}, \Sigma).
                Renormalize each vector \vec{v}_k' = \frac{\vec{v}_k}{||\vec{v}_k||}
 8:
          Compute Vietoris-Rips persistence diagram for the set \{\vec{v}_k'\}_{k=0}^{N_{\text{neigh}}} \to H_0
10:
          Compute W_1(H_0)
11: end procedure
```

and we can observe its behaviour in different neighbourhood configurations in figures 7 and 8.

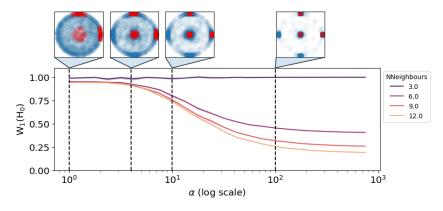


Figure 7: Evolution of the norm $W_1(H_0)$ with scaling coefficient α for D=768, $N_{\text{clust}}=3$ and variable N_{neigh} . On top of the main figure are example of biased distributions along the 3 axes in S^2 with different α values (blue circles). 3 clusters are sampled along those dimensions (red crosses)

Note that biasing the distribution of each cluster along an independent dimension of the space S^n corresponds to a boundary case of a real context where every dimension encodes mainly a single semantic. It also means that the cosine similarity across clusters tends to ~ 0 as α increases. The configuration $(N_{\text{neigh}}=2,N_{\text{clust}}=1)$ is trivial to analyze as $W_1(H_0)$ only represents the scaled distance between the two points, which drops as those two points are sampled with a stronger orientation bias accounting only for the anisotropy of the neighborhood. The same conclusions can be drawn for all configurations with $N_{\text{clust}}=1$.

For all other configurations, we can distinguish three different regions in figure 8:

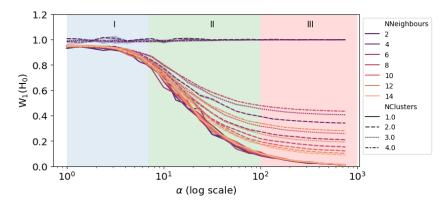


Figure 8: Evolution of the norm $W_1(H_0)$ with scaling coefficient α for D=768 and variable N_{neigh} and N_{clust} .

- Region I: High $W_1(H_0)$ values are achieved irrespective of the number of clusters (N_{clust}) and the number of neighbors (N_{neigh}) . This corresponds to a more uniform sampling (strictly uniform in the case of $\alpha=1$). In a real context, this behavior would be expected when data is distributed uniformly across all dimensions in high-dimensional spaces.
- Region II: This marks a transition regime with a clear drop in the norm. The separation between clusters becomes larger, but intra-cluster distances also shrink (indicating higher cluster density), which significantly impacts $W_1(H_0)$. The neighborhood becomes more structured and anisotropic.
- Region III: This region exhibits a relative stabilization of $W_1(H_0)$ but with high variance between different $(N_{\text{neigh}}, N_{\text{clust}})$ configurations. Increasing the number of clusters raises the norm value, while a higher density within each cluster decreases it. Therefore, anisotropy is mitigated by the increasing number of clusters.

C Embedding Models

The three models we have used, SBERT-multiqa v6, SBERT-msmarco v4, and GTR-T5-XL, vary in architecture and application but all three have same output dimensionality equal to 768. The SBERT-multiqa (Reimers and Gurevych [2019], Huggingface [2022]) is a pretrained multi-qa-mpnet-base-cos-v1 model designed for efficient similarity search in multi-domain question-answering. The SBERT-msmarco (Reimers and Gurevych [2019], Huggingface [2021]) is also sentence-bert model fine-tuned on the MS MARCO dataset (Bajaj et al. [2018]), targeting high relevance in retrieval tasks. Meanwhile, the GTR-T5-XL (Ni et al. [2021]) leverages a large T5 transformer, optimized for general textual retrieval across a wider range of datasets and tasks.

D Hyperparameters of computational proxy for semantic domains

In order to define the semantic domains (or topics) we performed dimensionality reduction followed by clustering using UMAP and HDBSCAN. We chose UMAP because it can retain both local or global topological structure. We performed the reduction down to 2 dimensions and optimised it on the parameter n_neighbors which corresponds to the number of neighbours the model will consider. Intuitively, low values favor local structure projection and high values favor global one. We chose to automate the search of optimal n_neighbors for each document independently by scanning a range $n_{neighbors} \in (2,50]$ and considering the maximum rate of pairwise distance δ_{ij} change:

$$n_{\text{neighbors}}^* = \arg\max_{n_{\text{neighbors}}} \left(\frac{d^2 \left(\sum_{i,j;i \neq j} \delta_{ij} (n_{\text{neighbors}}) - \delta_{ij}(2) \right)}{dn_{\text{neighbors}}^2} \right)$$
 (7)

This approach empirically favours relatively low parameter value (n_neighbors $\simeq 10$) and ensure that the local topological structure of the embeddings is preserved in the reduction process while

Notation	Description
$n_{ m neighbors}$	The number of neighbors considered by the UMAP model.
δ_{ij}	Pairwise distance between points i and j in the reduced dimensional
	space.
$\delta_{ij}(n_{\text{neighbors}})$	Pairwise distance calculated using a specific $n_{\text{neighbors}}$.
$\delta_{ij}(2)$	Pairwise distance when $n_{\text{neighbors}} = 2$, used as a reference point.
arg max	Indicates that we are searching for the value of $n_{\text{neighbors}}$ that maximizes the expression.
$rac{d^2(\cdot)}{dn_{ m neighbors}^2}$	Second derivative with respect to $n_{\text{neighbors}}$, capturing the rate of change.

Table 3: Summary of Notation

scaling easily the experiment to an arbitrary number of documents. The "min_dist" parameter is left at its default value of 0.1. A value too low (\simeq 0.0) would result in collapsing of similar embeddings in single positions in 2D while a high value would lead to a strong spread of data points which would make the optimisation process in 7 impossible.

The HDBSCAN clustering step is also optimised automatically relying on achieving maximum silhouette score (Rousseeuw [1987]) by tuning the min_cluster_size parameter over the range [2, 40]. The parameter min_samples is not optimised independently which means it defaults to the same value as min_cluster_size. All other parameters are left to their default value. minimum cluster size is distributed in the range [6,10] for all documents and outlier chunks (those which don't belong to any cluster) are consistently kept under 5% of the total number of chunks which allow us to say that they don't impact the results presented.