

Review

Exploring the Potential of Topological Data Analysis for Explainable Large Language Models: A Scoping Review

Petar Sekuloski *, Dimitar Kitanovski , Igor Goshev , Kostadin Mishev , Monika Simjanoska Misheva 
and Vesna Dimitrievska Ristovska 

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje, Skopje 1000, North Macedonia; dimitar.kitanovski@finki.ukim.mk (D.K.); kostadin.mishev@finki.ukim.mk (K.M.)
* Correspondence: petar.sekuloski@finki.ukim.mk

Abstract

Large language models (LLMs) have become central to modern artificial intelligence, yet their internal decision-making processes remain difficult to interpret. As interest grows in making these models more transparent and reliable, topological data analysis (TDA) has emerged as a promising mathematical approach for exploring their structure. This scoping review maps the current landscape of research where TDA tools—such as persistent homology and Mapper—are used to examine LLM components like attention patterns, latent representations, and training dynamics. By analyzing topological features across layers and tasks, these methods provide new ways to understand how language models generalize, respond to unfamiliar inputs, and shift under fine-tuning. The review also considers how TDA-based techniques contribute to broader goals in interpretability and robustness, especially in detecting hallucinations, out-of-distribution behavior, and representational collapse. Overall, the findings suggest that TDA offers a rigorous and versatile framework for studying LLMs, helping researchers uncover deeper patterns in how these models learn and reason.

Keywords: topological data analysis; persistent homology; mapper; large language models; explainability; interpretability; robustness; representation learning

MSC: 62R40; 55N31; 68T07; 68T99; 68T50; 68W05

1. Introduction

The prominence of artificial intelligence is largely due to the development of large language models. They have made significant enhancements in translation, summarization, and dialogue systems as they can generate coherent and contextual texts. However, along with such advancements comes a dilemma: the underlying decision-making mechanisms remain a black box. Predicting model failures is nearly impossible, both for users and researchers, and formulating reasons for outputs is equally difficult. This opaqueness stems from decisions intent as well as provides questions for the liability, equity, and social trust scaffolding of the model.

Over the years, many methods have been suggested to make LLMs more transparent. Attention maps, saliency measures, probing tasks, and attribution methods have provided pieces of the puzzle. But these tools usually point out local effects instead of global structures that influence the behavior of the model. As a result, they can give helpful hints without fully addressing bigger questions: How do representations change across



Academic Editor: Jonathan Blackledge

Received: 20 November 2025

Revised: 22 December 2025

Accepted: 4 January 2026

Published: 22 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

layers? How does meaning develop and organize in hidden spaces? And why do models sometimes hallucinate or fail?

A new approach that addresses these issues comes from Topological Data Analysis (TDA). Drawn from algebraic topology, TDA aims to uncover the shape of complex, high-dimensional data. Techniques like persistent homology, Betti numbers, and Mapper graphs reveal stable structural patterns such as clusters, loops, and voids. These patterns stay even when the data is noisy or compressed. When applied to LLMs, these tools offer a way to go beyond token-level explanations and capture the broader shape of embeddings, attention structures, and training dynamics.

Early research in this area has indicated that topological methods can find meaningful patterns in how LLMs process language, from identifying coherent semantic clusters to diagnosing unreliable outputs. Although the field is still developing, these studies suggest that TDA could complement traditional interpretability and explainability techniques by providing a global geometry-focused perspective.

This work is a *scoping review* focused on the *mathematical methodology* of Topological Data Analysis (TDA) as applied to the explainability and interpretability of neural representation spaces in large language models and transformer-based architectures. The objective of this review is not to benchmark language models or evaluate downstream NLP performance, but rather to systematically map, organize, and synthesize the topological constructions, invariants, and stability properties that have been employed to study high-dimensional representations arising in modern language models.

The primary contribution of this scoping review is a *formal mathematical synthesis* of existing TDA-based interpretability approaches. In particular, we provide a principled organization of the literature according to mathematically meaningful criteria, including the homological dimension under consideration, the type of representation manifold analyzed, and the manner in which persistence information is used or aggregated. To our knowledge, no prior review has systematically structured this body of work along these topological and algebraic axes.

While the focus of this review is on large language models, we also include studies based on smaller transformer architectures (e.g., BERT-like models) when their analyses target representation geometries or attention structures that are architecture-independent and transferable to LLM settings. This reflects the fact that the topological constructions reviewed depend on properties of representation spaces rather than on model scale alone.

Accordingly, the emphasis throughout this paper is placed on topological objects such as simplicial complexes, persistence diagrams, Betti numbers, zigzag persistence, and Mapper graphs, and on their mathematical properties, including robustness, metric dependence, scalability, and interpretability. Large language models serve as a motivating application domain in which these mathematical tools are instantiated, rather than as the primary object of evaluation.

Since the work is spread across various venues and methods, it is hard to obtain a complete view. This paper uses a scoping review to outline the current landscape of TDA in LLM explainability and interpretability. Rather than ranking methods, we aim to organize the field, highlight common themes, and pinpoint areas where more research is needed.

Specifically, this review aims to

1. Summarize the TDA techniques that have been used on LLMs, or smaller transformer based models, but transferable to LLMs.
2. Group existing work into themes like attention analysis, latent representations, robustness, and training dynamics, interactive explanations.
3. Highlight challenges and opportunities for future studies.

By bringing these threads together, we hope to demonstrate how topological approaches can make large language models not only more powerful but also more transparent and trustworthy.

The synthesis presented here also reveals several open mathematical challenges, including the limited use of higher-dimensional homology, the role of metric choice in anisotropic embedding spaces, and the absence of formal links between topological instability and optimization dynamics.

2. Methods

This review was designed and conducted in accordance with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. The aim was to provide a comprehensive mapping of the current research landscape on the use of topological data analysis (TDA) for explainability and interpretability in large language models (LLMs). Rather than seeking to answer a narrowly defined question, the scoping review format was chosen to identify the range of approaches, highlight major themes, and point toward gaps and opportunities for future research.

2.1. Protocol and Registration

We followed the PRISMA-ScR checklist to ensure transparency and reproducibility. No formal protocol registration was performed, as existing registries such as PROSPERO focus primarily on biomedical systematic reviews and do not accommodate mathematical or artificial intelligence reviews. Nevertheless, the eligibility criteria, search strategy, screening methods, and synthesis approach were defined in advance and consistently applied.

2.2. Eligibility Criteria

Studies were considered eligible for inclusion if they satisfied the following criteria. First, the work must apply topological data analysis techniques—such as persistent homology, Mapper, or related topological constructions—to the analysis of neural network representations, embeddings, activations, or attention structures. Second, the study must investigate machine learning or deep learning models relevant to modern language modeling, explicitly including transformer-based architectures, large language models, or neural networks whose internal representation structure is transferable to transformer and LLM settings.

In particular, studies focusing on transformer models (e.g., BERT-like architectures or attention-based networks) were included due to their direct relevance to large language models. Additionally, studies analyzing convolutional or model-agnostic neural networks were included when their topological analyses targeted *structural properties* of representation spaces—such as connectivity, cycles, clustering, or geometric complexity—that are largely architecture-independent and therefore transferable to transformer-based and large language models.

Third, the study must present a clear methodological description, empirical analysis, or conceptual framework linking topological features to model behavior, interpretability, robustness, or representation structure. Only works written in English and providing sufficient technical detail to assess the application of topological methods were considered. Studies focusing exclusively on abstract topological theory without application to neural network representations, or applying topological methods solely to non-neural data domains, were excluded.

Importantly, eligibility was determined independently of methodological rigor, computational scalability, or the degree of explicit explainability achieved. No study was excluded on the basis of limited evaluation, partial reproducibility, or implicit rather than

explicit interpretability. These aspects were instead assessed *after inclusion* through a separate quantitative scoring and weighting scheme used to characterize methodological heterogeneity among the included studies (see Appendix A).

2.3. Information Sources

To ensure broad and comprehensive coverage of the literature, we searched six major academic databases: Scopus, Web of Science, IEEE Xplore, SpringerLink, MDPI Journals, and the arXiv preprint server (categories: Mathematics, Computer Science, Artificial Intelligence, and Machine Learning). The search period was defined as January 2018 to August 2025 to capture both the emergence of transformer-based architectures, including large language models, and the increasing application of topological data analysis methods in machine learning.

The database searches were conducted in August 2025. In addition to database searching, a small number of additional records were identified through manual screening of reference lists and related-work sections of relevant articles. Only studies published in English were considered. Both journal articles and conference papers were included, along with relevant preprints from arXiv.

2.4. Search Strategy

The search strategy was designed to combine terms related to topological data analysis, large language models, and explainability or interpretability objectives. A representative Boolean query was formulated as follows:

("Topological Data Analysis" OR "persistent homology" OR "Mapper" OR "Betti numbers") AND ("Large Language Models" OR "transformer" OR "GPT" OR "BERT" OR "LLaMA") AND ("Explainability" OR "Interpretability" OR "Transparency" OR "Robustness")

The query was adapted for each database to account for differences in indexing systems, controlled vocabularies, and available filters, while preserving the same conceptual structure. This approach aimed to balance sensitivity, by capturing a broad range of potentially relevant studies, and specificity, by reducing the inclusion of unrelated mathematical or deep learning works.

2.5. Selection Process

In accordance with the PRISMA-ScR guidelines, the study selection was conducted through a structured multi-stage procedure comprising identification, deduplication, screening, eligibility assessment, and final inclusion. During the identification stage, the literature search retrieved 445 records from database sources and an additional 6 records from other sources, resulting in a total of 451 records. All identified records were imported into the Zotero reference management software, where automated duplicate detection followed by manual verification was performed. This process led to the removal of 241 duplicate entries, leaving 210 unique records for further screening.

In the screening stage, titles and abstracts of the 210 remaining records were examined, and 68 records were excluded for irrelevance to large language models, absence of topological data analysis, or lack of an explicit focus on explainability or interpretability. The remaining 142 records underwent an initial eligibility assessment, during which non-research items (e.g., editorials, surveys without methodological contributions, or duplicate venue versions) were removed. Following this refinement, 99 articles were assessed in full text. Of these, 73 articles were excluded with documented reasons, primarily because they did not employ TDA methods, did not focus on LLM-based architectures, or failed to establish a clear connection between topological analysis and interpretability objectives.

Ultimately, 26 studies satisfied all predefined eligibility criteria and were included in the qualitative synthesis. The complete study selection workflow is summarized in the PRISMA flow diagram (Figure 1).

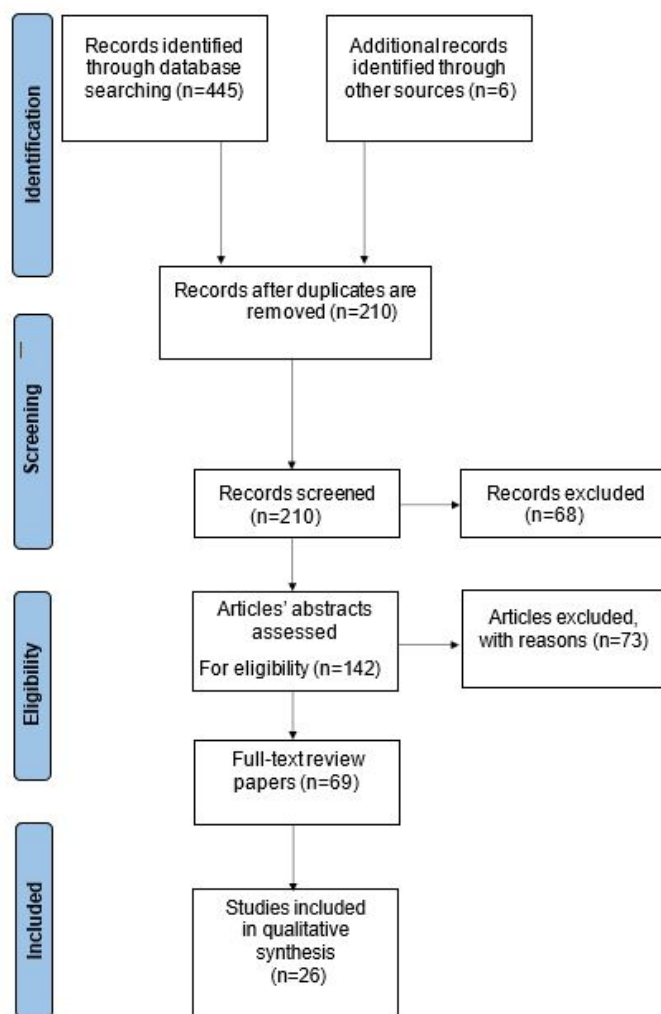


Figure 1. PRISMA flow diagram summarizing the study selection process for the scoping review.

Any disagreements about inclusion during this process were discussed collaboratively until consensus was reached.

2.6. Excluded Studies

Study exclusion was conducted in a structured and transparent manner in accordance with PRISMA-ScR guidelines and was guided by conceptual scope rather than by formal quality assessment or risk-of-bias evaluation. During the title and abstract screening stage, records were excluded if they were clearly outside the scope of this review, including studies unrelated to large language models, those that did not employ topological data analysis, or works focusing exclusively on predictive performance without addressing explainability or interpretability aspects.

In the full-text review stage, a more detailed methodological evaluation was performed. Studies were excluded if they did not apply TDA methods in a substantive manner, were not based on large language models or could not be reasonably transferred to LLM settings, or failed to establish an explicit and interpretable connection between topological analysis and explainability or interpretability objectives. In addition, non-research items such

as editorials, short position papers, or survey articles lacking original methodological contributions relevant to TDA-based explainability were excluded.

A summary of exclusion categories and their frequencies in the full-text stage is provided in Supplementary Table S1.

2.7. Data Extraction

For each study included in the qualitative synthesis, structured data extraction was performed using a predefined charting framework to support systematic comparison and thematic analysis. The extracted attributes were selected to capture both methodological diversity and relevance to explainability and interpretability in large language models. Specifically, for each included study we recorded bibliographic information (authors, publication year and venue), the LLM architecture or model family under investigation (e.g., BERT, GPT, LLaMA), and the specific topological data analysis technique applied (e.g., persistent homology, Betti numbers, Mapper graphs, or topological divergence). In addition, we documented the primary interpretability objective addressed by the study, such as analysis of attention mechanisms, latent space structure, robustness, or training dynamics. Key findings related to explainability and interpretability were summarized, together with any limitations or open challenges explicitly acknowledged by the authors. The extracted information formed the basis for the comparative tables and thematic synthesis presented in the Results section.

Data extraction was performed consistently in all included studies to ensure transparency and comparability.

2.8. Study Quality Considerations and Methodological Transparency

In line with PRISMA-ScR recommendations for scoping reviews, no formal study quality assessment or risk-of-bias evaluation was conducted. Instead, the screening and inclusion process was designed to ensure conceptual relevance, methodological clarity, and transparency of reporting rather than to exclude studies based on quality thresholds.

Article screening was performed independently by four authors at the title, abstract, and full-text levels. Disagreements were resolved through discussion until consensus was reached. Studies were retained if they met the predefined inclusion criteria and demonstrated a clear methodological link between topological data analysis and explainability or interpretability in large language models.

Rather than applying standardized critical appraisal tools, methodological considerations focused on descriptive aspects such as clarity of study design, completeness of reported data, transparency of the analytical approach, and explicit discussion of limitations or potential sources of bias where acknowledged by the original authors. These considerations were used solely to characterize heterogeneity across studies and to inform the qualitative synthesis, not to perform quality-based exclusion or weighting.

3. Theoretical Background

This section introduces the fundamental concepts behind Topological Data Analysis (TDA) and the current landscape of explainability and interpretability in large language models (LLMs). These two domains, while historically distinct, intersect in promising ways for revealing high-dimensional structure in neural representations and attention mechanisms.

3.1. Topological Data Analysis (TDA)

Topological Data Analysis (TDA) refers to a collection of mathematical techniques from algebraic topology designed to study the shape and structure of data. It is particularly effective in capturing global, non-linear patterns in high-dimensional or unstructured datasets [1,2].

3.1.1. Persistent Homology

A fundamental tool in TDA is **persistent homology**, which tracks the evolution of topological features on multiple scales. Let $X \subset \mathbb{R}^d$ be a point cloud derived from the embeddings or hidden states of an LLM. For a scale parameter $\varepsilon > 0$, we define the Vietoris–Rips complex as follows:

$$K_\varepsilon(X) = \{\sigma \subseteq X : \|x_i - x_j\| \leq \varepsilon \forall x_i, x_j \in \sigma\}.$$

As ε increases, we obtain a filtration:

$$K_{\varepsilon_0} \subseteq K_{\varepsilon_1} \subseteq K_{\varepsilon_2} \subseteq \dots,$$

and compute homology groups $H_k(K_\varepsilon)$ to capture k -dimensional features. The k -th Betti number is

$$\beta_k = \text{rank}(H_k(K_\varepsilon)),$$

representing the number of independent k -dimensional cycles at scale ε . The resulting *barcode* or *persistence diagram* records the intervals (b_i, d_i) during which each feature persists:

$$D = \{(b_i, d_i) \mid \text{feature appears at } b_i \text{ and disappears at } d_i\}.$$

By tracking Betti numbers across filtration scales, one obtains Betti curves, which describe the evolution of topological features as a function of the filtration parameter.

When Betti curves are derived from attention-based representations, they can be computed at the level of individual attention heads or combined across heads and layers to obtain a more stable and interpretable summary. Let $\beta_k^{(\ell, h)}(t)$ denote the Betti curve of homological dimension k computed at layer ℓ and attention head h . To reduce variability arising from individual heads, these curves can be averaged across all layers and heads:

$$\bar{\beta}_k(t) = \frac{1}{LH} \sum_{\ell=1}^L \sum_{h=1}^H \beta_k^{(\ell, h)}(t),$$

where L and H denote the number of layers and attention heads, respectively. Such aggregation strategies are reported or implicitly adopted in several of the reviewed studies when summarizing attention-derived topological features, as they help smooth head-specific fluctuations while preserving structural patterns that are consistent across the model. This results in Betti curves that are more robust and better suited for model-level interpretability analyzes. Across the reviewed studies, Betti curves are reported at different levels of granularity. In several works, aggregation across attention heads and layers is implicitly performed to obtain sentence-level or model-level summaries, although the exact aggregation procedure is often not formally specified. Other studies instead analyze Betti curves at the level of individual heads or layers without aggregation, focusing on localized or qualitative structural patterns. In this review, we do not perform aggregation ourselves; instead, we describe common aggregation strategies—such as uniform averaging across heads and layers—as representative procedures that may be employed to obtain stable model-level summaries when required.

Persistent homology requires a notion of distance (or similarity) to construct pairwise relations and build the Vietoris–Rips filtration. Although the Euclidean distance is a common default, transformer embeddings are often anisotropic, and their geometry may be better captured by alternative metrics. In the reviewed literature, this issue is typically handled by adopting cosine distance (which is less sensitive to global scale and anisotropy)

or by applying normalization and whitening procedures prior to distance computation. More generally, one may use a Mahalanobis distance:

$$d_M(x, y) = \sqrt{(x - y)^\top \Sigma^{-1} (x - y)},$$

where Σ is an estimate of the embedding covariance, to account for direction-dependent variability. In this review, we emphasize that the reported topological summaries should be interpreted in conjunction with the chosen metric, since the metric directly influences the resulting simplicial complexes and persistence diagrams.

In this review, the term *topological divergence* is used in a broad sense to describe how topological summaries derived from data are compared. Most commonly, these summaries are persistence diagrams, which can be compared directly using standard distances such as the bottleneck or Wasserstein metrics. In other cases, persistence diagrams are first transformed into vector-based representations, such as persistence landscapes or persistence images, allowing the use of kernel-based or norm-based distances. Throughout the reviewed literature, topological divergence therefore refers to differences measured either directly in the space of persistence diagrams or in induced metric spaces obtained after vectorization, depending on the methodological choice of each study.

3.1.2. Stability of Persistent Homology

Persistent homology enjoys a fundamental stability property that guarantees the robustness of persistence diagrams under small perturbations of the underlying metric space or filtration function. Let $f, g : X \rightarrow \mathbb{R}$ be tame functions inducing two filtrations, with corresponding persistence diagrams denoted by $D(f)$ and $D(g)$. The classical stability theorem for persistent homology states that

$$d_B(D(f), D(g)) \leq \|f - g\|_\infty, \quad (1)$$

where $d_B(\cdot, \cdot)$ denotes the bottleneck distance between the persistence diagrams.

This inequality implies that small perturbations in the input data or representation space lead to proportionally small changes in the resulting persistence diagrams. In the context of transformer-based language models, such perturbations may arise from noise in embedding spaces, variations across layers, or attention-derived point clouds. Consequently, all persistent homology-based descriptors used in the reviewed studies, including persistence lifetimes, Betti curves, and persistence entropy, inherit this robustness property.

3.1.3. Zigzag Persistence

Although persistent homology assumes a monotone filtration, many problems in machine learning involve data sets that grow and shrink, such as when monitoring LLM training dynamics or comparing overlapping attention subgraphs. In such cases, zigzag persistence is used.

Zigzag persistence extends classical persistent homology to non-monotonic filtrations, in which simplicial complexes may be both added and removed along the filtration sequence.

Instead of a nested sequence of complexes, zigzag persistence allows for a sequence of inclusions and deletions:

$$K_0 \longleftrightarrow K_1 \longleftrightarrow K_2 \longleftrightarrow \cdots \longleftrightarrow K_n,$$

where the arrows may point in either direction. Each K_i is a simplicial complex, and the sequence forms a zigzag filtration. Homology groups $H_k(K_i)$ are connected by induced maps corresponding to forward or backward inclusions. The resulting invariants capture

topological features that persist across varying subcomplexes, even when simplices are removed and added.

For such constructions to be well-defined, each forward or backward map must induce a homomorphism between the corresponding homology groups, ensuring functorial consistency of the resulting zigzag module. Under these conditions, zigzag persistence admits an interval decomposition and stability properties at the level of algebraic modules, analogous to those of standard persistent homology.

In the reviewed studies, non-monotonic zigzag filtrations arise from structured and controlled changes in representations, such as layer-wise transitions or feature insertions and removals. These constructions implicitly satisfy the required functorial conditions, ensuring that the resulting zigzag persistence diagrams remain mathematically well-defined and stable.

Zigzag persistence yields barcodes similar to standard persistence, but they reflect features stable across dynamic, non-monotonic changes in the data. This makes it well suited for

- Tracking representational drift during pre-training and fine-tuning;
- Comparing different snapshots of LLM attention structures;
- Capturing topological stability in evolving latent spaces.

3.1.4. Persistence Entropy

Beyond raw persistence diagrams, several studies summarize topological information using scalar descriptors. One commonly used measure is persistence entropy, which quantifies the distribution of persistence lifetimes in a diagram. Given a persistence diagram D with intervals (b_i, d_i) , persistence entropy is defined as

$$H(D) = - \sum_i p_i \log(p_i), \quad p_i = \frac{d_i - b_i}{\sum_j (d_j - b_j)}.$$

This normalization step converts persistence lifetimes into a probability distribution, making persistence entropy independent of the absolute number or scale of topological features in the diagram. As a result, entropy values can be compared across different layers, models, or representations, even when their persistence diagrams have different sizes or levels of complexity. This property is particularly important in large language models, where representational complexity may vary substantially across architectures and layers.

3.1.5. Mapper Algorithm

The Mapper algorithm offers an intuitive, graph-based way to explore the global structure of high-dimensional data. Given a dataset $X \subset \mathbb{R}^d$ and a filter function $f : X \rightarrow \mathbb{R}$, Mapper builds a simplicial graph by examining how data points group together across overlapping regions of the filter space.

The filter function determines how the data are viewed through a lower-dimensional lens and strongly influences the shape of the resulting Mapper graph. Typical choices include projection-based functions, density estimates, or model-derived quantities such as activation norms. The filter function range is divided into overlapping intervals, controlled by a resolution parameter and an overlap percentage. Larger overlaps generally lead to more connected graphs, whereas smaller overlaps emphasize finer, more localized structures.

Within each interval of the cover, the data points are clustered independently and each cluster becomes a node in the Mapper graph. Connections between nodes arise when clusters share data points due to overlapping intervals. The choice of clustering method and distance criterion affects the level of detail and the connectivity of the graph. In the reviewed studies, these choices are typically guided by the goal of producing stable and

interpretable visualizations, with topological patterns assessed through consistency across multiple parameter settings rather than reliance on a single configuration.

The Mapper algorithm is known to be sensitive to the choice of hyperparameters, such as the filter function, the cover resolution, the overlap, and the clustering method. For this reason, stability of Mapper-based explanations is usually approached in a practical rather than theoretical sense. In the reviewed studies, robustness is commonly evaluated by checking whether the main structural patterns of the Mapper graph remain visible when reasonable parameter choices are varied. Explanations are considered reliable when they are supported by consistent structures across multiple Mapper configurations, rather than depending on a single set of hyperparameters.

3.2. Relevance for LLM Interpretability

In the context of large language models (LLMs), topological data analysis tools provide several complementary perspectives:

- **Global structure detection:** Identifying clusters, connected components, and voids in embedding spaces.
- **Attention topology:** Analyzing connectivity patterns in attention graphs using Betti numbers and topological divergence metrics.
- **Robustness signals:** Detecting structural changes under adversarial perturbations or out-of-distribution inputs.
- **Training dynamics analysis:** Monitoring representational drift over time using zigzag persistence.

The mathematical generality and coordinate-free nature of TDA methods make them highly suited for studying the abstract feature spaces of large-scale language models, where traditional Euclidean assumptions often fail.

The mathematical generality and coordinate-free nature of TDA methods make them particularly well suited for studying the abstract feature spaces of large-scale language models, where traditional Euclidean assumptions may not hold. Several studies reviewed in this paper interpret changes in topological descriptors—such as variations in Betti numbers or persistence diagrams—as signals of representation instability or reduced robustness. A natural question is whether such topological changes can be formally related to optimization dynamics during training, for example through correlations with loss gradients $\nabla_{\theta}L$. While this connection is conceptually appealing, existing work largely treats topological summaries as geometric or structural descriptors of representation spaces, rather than quantities directly coupled to gradient-based optimization. Establishing a formal relationship between topological instability (e.g., changes $\Delta\beta_k$) and optimization dynamics would require joint analysis of training trajectories, gradient information, and topological summaries and remains largely unexplored in the current literature. We therefore view this connection as a promising direction for future research, rather than a settled component of current TDA-based interpretability approaches.

Statistical Validation of ID–OOD Topological Differences

To strengthen robustness *interpretations*, differences between ID and OOD topological summaries can be complemented by statistical significance testing. Let $\{X_i\}_{i=1}^{n_{\text{ID}}}$ and $\{Y_j\}_{j=1}^{n_{\text{OOD}}}$ denote representation samples (e.g., embeddings or attention-derived features) from ID and OOD inputs, respectively. Fix a statistic $T(\cdot, \cdot)$ that compares the two groups, such as a persistence-diagram distance (bottleneck or p -Wasserstein), an integrated difference between Betti curves, or a difference in persistence entropy. Compute the following observed test statistic:

$$T_{\text{obs}} = T(\{X_i\}_{i=1}^{n_{\text{ID}}}, \{Y_j\}_{j=1}^{n_{\text{OOD}}}).$$

A permutation test can then be used under a null hypothesis of homology equivalence (i.e., no systematic topological difference between ID and OOD), by pooling all samples, randomly permuting the ID/OOD labels, recomputing T for each permutation, and estimating the p -value as

$$p = \frac{1 + \sum_{b=1}^B \mathbb{I}[T^{(b)} \geq T_{\text{obs}}]}{B + 1},$$

where $T^{(b)}$ is the statistic computed on the b -th label permutation and B is the number of permutations. Alternatively, bootstrap resampling within each group can be used to estimate a confidence interval for T_{obs} (e.g., via the percentile bootstrap), providing an uncertainty-aware comparison. We emphasize that this testing framework is presented as a recommended evaluation procedure rather than a re-analysis of the reviewed studies, since explicit hypothesis testing of topological differences is not commonly reported in the current literature.

3.3. Explainability and Interpretability in Large Language Models (LLMs)

As large language models, achieve impressive performance across NLP tasks, concerns over their opacity and reliability have grown. These models contain hundreds of millions, or even billions, of parameters, making it difficult to understand how specific outputs arise from specific inputs. This has given rise to the field of *explainable AI* (XAI), which seeks to make the behavior of models more understandable and transparent to humans [3,4].

In this context, *explainability* refers to techniques that clarify why a model produced a specific output, while *interpretability* is concerned with how easily a human can understand the internal mechanisms or learned representations of the model.

Current explainability methods for LLMs include the following:

- **Attention analysis:** Visualizing or averaging attention weights between tokens.
- **Saliency-based methods:** Using gradients (e.g., Integrated Gradients) or perturbations to estimate input importance.
- **Probing classifiers:** Training shallow models to test for linguistic properties in hidden layers.
- **Attribution techniques:** Applying SHAP, LIME, or LRP to language inputs.

Despite their utility, these approaches often fall short in key respects. Attention is not always a faithful indicator of model reasoning [5], and saliency methods are sensitive to model perturbations and input noise. Most crucially, these tools are *local*—they describe token-level influences or layer-level properties, but rarely capture the *global structure* of how information flows through the model or how it organizes concepts in latent space.

TDA offers a complementary lens for understanding model behavior. By analyzing the topology of attention graphs, hidden-state manifolds, and feature-space embeddings, one can uncover persistent structures that reflect model decisions, generalization behavior, and even anomalies. Several recent studies have demonstrated that topological signatures can be linked to phenomena such as hallucinations, adversarial robustness, and syntactic acceptability [6–8].

Thus, TDA not only augments existing interpretability methods but may also support the development of new, geometry-aware frameworks for understanding LLMs at scale.

3.4. TDA in Neural Network Explainability/Interpretability: Pre-LLM Applications

Before the emergence of large language models (LLMs), Topological Data Analysis (TDA) had already demonstrated significant promise as a tool for enhancing the explainability and interpretability of deep neural networks (DNNs). In image classification, speech recognition, and conventional NLP, TDA provided rigorous means to explore the internal structure of learned representations—revealing how models organize, transform, and com-

press information. These early studies served as crucial testbeds, confirming the viability of topology-based methods for neural model interpretation.

A foundational peer-reviewed contribution in this direction was introduced by Rieck et al. [2], who proposed *neural persistence* as a complexity measure for feedforward networks. By applying persistent homology to the weight matrices of each layer, they defined a topological invariant that correlates with model generalization. This measure enabled a new class of interpretable diagnostics for neural architecture design and regularization, offering a structural explanation for why some networks overfit while others generalize better.

Wheeler et al. [1] extended this idea with the notion of *activation landscapes*, in which persistent homology is computed on neuron activations across layers. Their results showed that models with similar test accuracy can have markedly different topological profiles in activation space—thus revealing latent organizational properties not captured by traditional metrics. This provided a global and unsupervised tool for model explainability, particularly useful for comparing training regimes, architectural choices, and hyperparameter settings.

Purvine et al. [9] conducted large-scale empirical experiments on convolutional neural networks (CNNs), examining the topological evolution of internal feature maps. Their work showed that persistent topological features corresponded to semantically meaningful input classes and were sensitive to adversarial perturbations. This supported the idea that topological summaries can serve as class-level or even instance-level explanations for internal model behavior, beyond simple attribution scores.

In a complementary direction, Tulchinskii et al. [10] explored the role of TDA in transformer-based speech models. They applied persistent homology to embeddings and attention maps from speech transformers and demonstrated that topological features could differentiate phonetic structures and linguistic content. This not only improved robustness in downstream tasks, but also provided interpretable topological descriptors of linguistic categories, complementing token-level or attention-based explanations.

Together, these peer-reviewed studies provide three critical insights for explainable AI:

1. **Topology reveals global structure:** Persistent homology captures how models organize their internal representations at multiple scales.
2. **TDA supports comparative interpretation:** Topological descriptors enable interpretable comparisons across architectures, training conditions, or datasets.
3. **Cross-domain generalization:** From vision to speech, TDA methods have demonstrated broad applicability for neural interpretability beyond simple visualizations.

These insights laid the groundwork for the extension of TDA-based methods to transformer-based language models. The next section surveys how topological tools are now being used to explain attention dynamics, latent geometry, and output behavior of LLMs.

4. Results

4.1. Study Identification and Selection

The study selection process, including database search, keyword strategy, screening stages, and inclusion criteria, is described in detail in Section 2. In total, 26 studies met the eligibility criteria and were included in the final synthesis. These works represent a diverse range of approaches that apply TDA to enhance the interpretability and explainability of large language models.

Figure 1 presents the PRISMA flow diagram summarizing the identification, screening, eligibility assessment, and inclusion of studies. Following the multi-stage selection process, a total of 26 studies met all inclusion criteria and were included in the qualitative synthesis.

4.2. Characteristics of Included Studies

Consistent with the goals of a scoping review, the analysis presented here focuses on identifying methodological tendencies, structural patterns, and gaps in the literature, rather than comparing task-specific performance or empirical effectiveness between studies.

The 26 included studies exhibit substantial diversity in publication venue, application domain, model architecture, and topological methodology. Table 1 provides a structured overview of the included works, summarizing the analyzed LLM (or transformer architectures), applied TDA techniques, and the primary interpretability objectives addressed. Unlike purely thematic grouping, this structure supports direct cross-study comparison and reveals recurring patterns that characterize current research on topological data analysis (TDA) for explainability and interpretability of language models. The *TDA method type* column distinguishes between full persistent homology computations, approximate variants, PH-derived summaries (such as Betti numbers or curves), and non-PH topological approaches. The *Explainability/Interpretability* column specifies whether topological information is used explicitly for explanation or implicitly for structural, robustness, diagnostic, or visualization purposes; entries labeled methodological denote works that do not directly target explainability but introduce analytical frameworks that may support interpretability in downstream applications. The *Evaluation task* column captures the primary analytical objective addressed by each study at a coarse level. Finally, the *Scalability* column reflects the practical feasibility of the employed approach under commonly used approximations rather than formal asymptotic complexity.

Table 1. Organization of reviewed studies by model family, topological method type, interpretability role, evaluation task, and practical scalability.

Ref.	Model Fam.	TDA Method Type	Explainability/Interpretability	Evaluation Task	Scalability
[11]	LLM	Full PH (layer-wise)	Implicit (Structural)	Latent representation structure analysis	Moderate
[6]	LLM	Approximate PH	Implicit (Robustness)	Stability and robustness analysis	Moderate
[7]	LLM	Full PH (layer-wise)	Implicit (Structural)	Latent representation structure analysis	Low
[8]	Transformer	Approximate PH (attention graphs)	Explicit XAI	Attention structure characterization	Moderate
[5]	Transformer	PH-derived (Betti numbers)	Implicit (Diagnostic)	Authorship and linguistic property analysis	High
[3]	Transformer	PH-derived (Betti numbers)	Implicit (Visualization)	Attention structure characterization	High
[4]	Transformer	PH-derived (Betti curves)	Implicit (Diagnostic)	Authorship and linguistic property analysis	High
[12]	BERT	Full PH	Explicit XAI	Authorship and linguistic property analysis	Low
[13]	BERT	Full PH	Explicit XAI	Latent representation structure analysis	Low
[14]	BERT	Full PH (layer-wise)	Implicit (Structural)	Latent representation structure analysis	Low
[15]	Transformer	Approximate PH	Explicit XAI	Stability and robustness analysis	Moderate
[16]	Transformer	Approximate PH	Explicit XAI	Out-of-distribution sensitivity analysis	Moderate
[17]	Transformer	Full PH	Explicit XAI	Authorship and linguistic property analysis	Low
[18]	Model-agnostic NN	Non-PH (manifold analysis)	Methodological	Methodological assessment	N/A
[19]	BERT	Full PH	Implicit (Structural)	Latent representation structure analysis	Low

Table 1. Cont.

Ref.	Model Fam.	TDA Method Type	Explainability/Interpretability	Evaluation Task	Scalability
[10]	Transformer	Full PH	Implicit (Structural)	Latent representation structure analysis	Low
[20]	Transformer	Approximate PH	Implicit (Robustness)	Out-of-distribution sensitivity analysis	Moderate
[21]	Model-agnostic NN	Full PH	Implicit (Visualization)	Activation space analysis	Low
[22]	Transformer	Approximate PH	Implicit (Diagnostic)	Stability and robustness analysis	Moderate
[23]	Model-agnostic NN	Non-PH (Mapper)	Implicit (Visualization)	Exploratory topological visualization	High
[24]	Model-agnostic NN	Non-PH (RTD)	Implicit (Diagnostic)	Latent representation structure analysis	High
[1]	Model-agnostic NN	Full PH	Methodological	Methodological assessment	Low
[9]	CNN	Full PH	Methodological	Activation space analysis	Low
[2]	Model-agnostic NN	Non-PH (Neural persistence)	Methodological	Methodological assessment	High
[25]	Model-agnostic NN	Full PH	Implicit (Visualization)	Exploratory topological visualization	Low
[26]	LLM	Non-PH (Mapper)	Explicit XAI	Exploratory topological visualization	High

4.3. Methodological Scoring and Heterogeneity Analysis

To quantify methodological variability among the included studies without performing quality-based exclusion, a post hoc methodological scoring and normalization scheme was applied, given in the Appendix A. Each of the 26 included studies was evaluated on seven dimensions: data transparency, pipeline reproducibility, TDA validity, evaluation rigor, interpretability grounding, model scope, and scalability. The total methodological scores resulting span a range of 3 to 14, with an unweighted mean score of 9.35 and a median of 9.5.

Normalizing scores by the maximum observed value ($S_{\max} = 14$) yields synthesis weights used to characterize the heterogeneity of the cross-study. Based on these weights, the weighted mean methodological score is 9.83. The corresponding weighted heterogeneity index is 3.93, which results in a normalized heterogeneity value of 0.020 when compared with S_{\max}^2 . These values indicate moderate methodological heterogeneity in the reviewed corpus.

4.4. Cross-Study Analysis

Across the 26 included studies, persistent-homology-based approaches dominate the literature. Full persistent homology accounts for approximately 42% of the reviewed studies, while an additional subset relies on approximate or PH-derived summaries, resulting in nearly two thirds of the corpus employing persistence-based topology. Among studies using full persistent homology, more than four fifths focus on structural or diagnostic analysis of internal representations, including layer-wise organization, representational collapse, and geometric stability, rather than on explicit or user-facing explanations.

The majority of reviewed studies employ implicit forms of interpretability, such as structural, diagnostic, or visualization-based analysis of internal representations. These approaches describe how language models organize information across layers and how representation structure varies under different conditions, without producing direct explanatory artifacts intended for end users.

Explicit explainability is less prevalent and appears primarily in conjunction with approximate persistent homology or non-PH methods, including Mapper-based abstractions and related graph representations. These studies emphasize abstraction and computational feasibility and are typically applied in contexts requiring human-interpretable or interactive analysis.

Figure 2 summarizes how different topological methods align with interpretability orientations between studies.

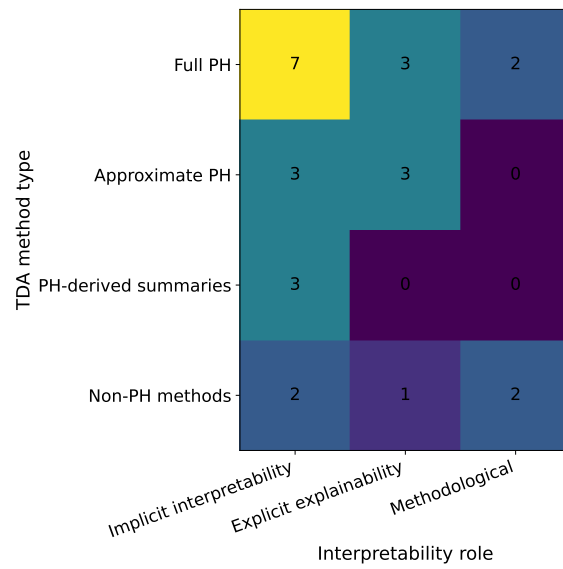


Figure 2. Distribution of reviewed studies by TDA method type and interpretability role. Counts are derived from Table 1 and summarize the cross-study patterns discussed in Section 4.

Scalability further differentiates methodological choices. Nearly three quarters of the reviewed studies exhibit low to moderate scalability, reflecting the computational cost of topological analysis on high-dimensional activations and attention structures. Approaches exhibiting high scalability are largely confined to non-PH or heavily approximated methods.

Although not all reviewed studies are conducted on full-scale large language models, the majority analyze transformer-based architectures such as BERT or generic Transformers. The topological structures examined in these works include attention connectivity patterns, latent representation geometry, and layer-wise activation evolution, all of which arise from architectural properties shared across transformer models.

Figure 3 provides a visual summary of scalability distributions across model families.

Distribution of reviewed studies by model family and scalability

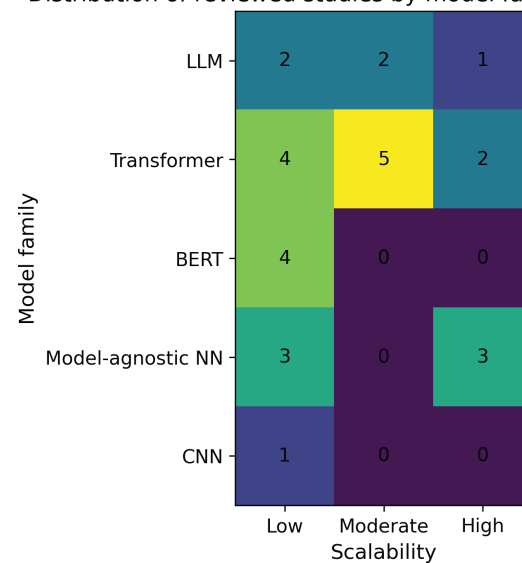


Figure 3. Distribution of reviewed studies by model family and scalability category. Counts are derived from Table 1 and summarize the cross-study patterns reported in Section 4.

Table 1 categorizes the reviewed studies by evaluation task, reflecting the primary analytical purpose for which topological methods are applied. These tasks include representation structure analysis, attention organization, robustness assessment, and activation dynamics, enabling cross-study comparison of analytical intent consistent with the scope of a scoping review.

Figure 4 summarizes how different internal representations are examined across analytical tasks in the reviewed literature.

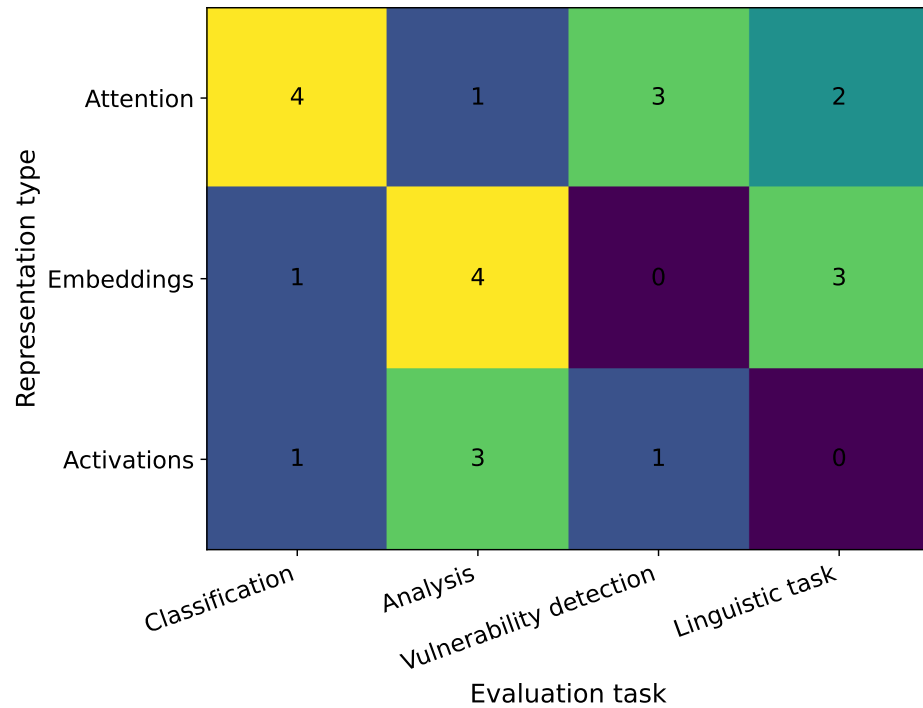


Figure 4. Distribution of reviewed studies by representation type and evaluation task. Counts are derived from Table 1.

Cross-study alignment metrics show that more than 80% of studies employing full persistent homology focus on implicit interpretability, while explicit explainability is predominantly supported by approximate or non-PH methods. In addition, over 70% of studies classified as explicit explainability employ methods with moderate to high scalability.

4.5. Computational Complexity and Scalability Considerations

From a computational perspective, the construction of Vietoris–Rips complexes and the computation of persistent homology are known to exhibit cubic worst-case complexity with respect to the number of points, i.e., $\mathcal{O}(n^3)$ for full persistence computations. This complexity arises from the combinatorial growth of simplices induced by dense pairwise distance matrices and renders the direct application of exact persistent homology impractical for large-scale representation sets typical of modern language models.

To mitigate this limitation, many practical approaches rely on approximation strategies that reduce computational cost while preserving coarse topological structure. Common examples include subsampling or landmark selection, sparsification of distance matrices, truncated filtrations, and witness-based constructions (e.g., lazy witness or weak witness complexes), which restrict simplicial growth by anchoring complexes to a smaller set of representative points. Under suitable assumptions, these approaches can reduce the effective complexity to approximately $\mathcal{O}(n^2)$ or even $\mathcal{O}(n \log n)$, depending on the data structure and filtration strategy employed. Such approximations trade exactness for

computational feasibility and are, therefore, particularly well suited for exploratory or comparative analysis of high-dimensional embeddings.

In addition to the choice of topological construction, the underlying distance metric plays a critical role in high-dimensional embedding spaces. While Euclidean distance is commonly used in persistent homology pipelines, transformer-based representations are often anisotropic, motivating alternative similarity measures such as cosine distance or, in some cases, Mahalanobis-type metrics that account for covariance structure. For example, Vietoris–Rips complexes built using cosine distance may emphasize angular relationships between token embeddings, whereas Euclidean constructions are more sensitive to norm variation. These metric choices further interact with sparsification strategies, as dense or poorly conditioned distance matrices can substantially increase computational burden.

Overall, this formalization of computational complexity is provided at the level of asymptotic upper bounds to contrast exact and approximate topological constructions, rather than to attribute precise algorithmic complexity to individual studies or software implementations.

4.6. Formal Mathematical Taxonomy of TDA-Based Interpretability Methods

To enable a principled comparison, the reviewed studies are organized according to a formal taxonomy based on mathematically meaningful attributes. Let \mathcal{S} denote the set of included studies. Each study $s \in \mathcal{S}$ is characterized by a triplet:

$$\mathcal{T}(s) = (k(s), \mathcal{M}(s), \mathcal{P}(s)),$$

where $k(s)$ denotes the homological dimension emphasized (with most studies focusing on $k = 0$ for connectivity and a smaller subset considering $k = 1$ for cyclic structure), $\mathcal{M}(s)$ specifies the type of representation manifold under analysis (e.g., embedding-based or attention-induced), and $\mathcal{P}(s)$ describes the manner in which persistence lifetimes are used. In particular, $\mathcal{P}(s)$ distinguishes between studies that explicitly emphasize long-lived features as robust structure, those that analyze changes across persistence scales, and those that rely on aggregated or implicit lifetime usage through summary statistics rather than fixed thresholds.

Table 2 provides a mathematical perspective on the reviewed persistent-homology-based studies, highlighting how they differ in terms of homological dimension, underlying data manifold, and the way persistence lifetimes are used or summarized. The table is intended as classification and illustrates that most works rely on aggregated or implicit lifetime usage, with explicit separation of long- and short-lived features appearing only in a limited subset of studies.

Table 2. Mathematical taxonomy of persistent homology-based studies according to homological dimension, data manifold type, and persistence lifetime usage.

Ref.	k	Manifold	Persistence Lifetime Usage
[11]	0,1	LLM embeddings	Long-lived topological features interpreted as stable semantic structure; short-lived features treated as noise
[6]	1	Embeddings	Changes across persistence scales reflect latent instability under adversarial perturbations
[7]	1	Embeddings	Persistent holes interpreted as intrinsic geometric properties of representations
[8]	1	Attention graphs	Aggregated persistence divergence captures structural differences in attention graphs
[5]	0,1	Attention graphs	Aggregated persistence features distinguish attention patterns in synthetic versus human text
[3]	0,1	Attention graphs	Betti curves summarize lifetime evolution without explicit persistence thresholds

Table 2. Cont.

Ref.	k	Manifold	Persistence Lifetime Usage
[4]	1	Attention graphs	Aggregated persistence summaries correlate with syntactic acceptability
[12]	0,1	Attention/embeddings	Aggregated persistence features support explainable linguistic competence
[14]	1	Embeddings	Persistence summaries compare representational stability before and after fine-tuning
[15]	1	Attention graphs	Aggregated persistence summaries reveal systematic vulnerability patterns
[16]	1	Attention/embeddings	Aggregated topological features indicate stylistic consistency
[17]	1	Attention graphs	Aggregated persistence features reflect author-specific attention regularities
[19]	0,1	Embeddings	Aggregated persistence summaries indicate semantic separation in fine-tuned embeddings
[20]	0,1	Embeddings	Shifts in persistence distributions distinguish in- and out-of-distribution samples
[22]	0	Embeddings	Local persistence measures quantify contextual representation stability
[24]	0	Activations	Persistence-based divergence quantifies representation drift
[1]	0,1	Activations	Persistence landscapes provide aggregated summaries of performance-related structure
[21]	0,1	Activations	Topological summaries support qualitative interpretation of activation geometry
[25]	1	Activations	Persistent topological changes explain effects of adversarial training
[2]	0	Weights/activations	Aggregate persistence measures quantify network structural complexity
[10]	0,1	Speech embeddings	Aggregated persistence features correspond to phonetic and prosodic consistency
[9]	0,1	CNN activations	Persistence summaries reveal architectural regularities in deep networks

4.7. Taxonomy-Driven Analysis of TDA-Based Interpretability Methods

The mathematical taxonomies introduced in Tables 1 and 2 allow for a structured analysis of the methodological concentration and diversity in the reviewed literature. In all included studies, the use of low-dimensional homology dominates current practice, with $k = 0$ and $k = 1$ accounting for nearly all applications. Higher-dimensional homological features are almost entirely absent, despite their theoretical relevance for capturing complex structure in high-dimensional representation spaces. Similarly, persistence information is most commonly used in an aggregated or implicit manner, through Betti numbers, Betti curves, or persistence-based summary statistics, rather than through explicit lifetime thresholding or separation of long- and short-lived features.

Examining interactions between taxonomy dimensions reveals systematic relationships between topological methodology and interpretability role. Studies employing full persistent homology are predominantly associated with implicit interpretability, where topological summaries are used for diagnostic, comparative, or structural analysis by researchers. In contrast, explicit explainability is more frequently supported by approximate or non-persistent methods, including Mapper-based abstractions and visualization-oriented approaches, which trade topological completeness for abstraction, interpretability, and computational feasibility. This distinction reflects differences in how topological information is summarized and communicated, rather than differences in model family or application domain.

The taxonomy also highlights underexplored regions of the methodological design space. In particular, the combination of higher-dimensional homology with explicit explainability remains largely unaddressed, as does the development of formally stable

topological summaries beyond persistent homology. In addition, few studies systematically examine the interaction between metric choice, embedding anisotropy, and topological invariants, despite the known sensitivity of simplicial constructions to the underlying distance function. These gaps indicate opportunities for future mathematical development in topology-based interpretability.

Viewed collectively, the taxonomy defines a methodological design space for TDA-based interpretability, in which each study occupies a point determined by homological dimension, representation manifold type, and persistence usage. The reviewed literature occupies only a limited region of this space, suggesting that many theoretically plausible combinations of topological tools and interpretability objectives remain unexplored. As such, the taxonomy serves not only as a descriptive classification but also as an analytical lens for identifying structural imbalances and open directions in the current research landscape.

Descriptions of persistence lifetime usage are intentionally conservative. When studies rely on summary statistics (e.g., Betti numbers, persistence landscapes, entropy, or divergence measures), persistence thresholds are treated as *implicit* rather than explicitly defined. Explicit references to long- or short-lived features are reported only when the original study makes such distinctions.

For clarity, this taxonomy is applied only to the subset of reviewed studies that explicitly employ persistent-homology-based constructions and is not intended to exhaustively cover all topological approaches discussed in this review.

4.8. Thematic Synthesis

For the purposes of this scoping review, the included studies were organized into five thematic groups based on their primary topological focus and interpretability objective. This categorization was used to support cross-study comparison and to highlight recurring methodological and conceptual patterns across the literature.

Specifically, the studies were grouped into the following themes:

- **Topology of attention maps**, focusing on topological characterization of attention mechanisms and attention flow in large language models;
- **Topological data analysis of latent representations and embedding spaces**, examining the geometric and topological structure of learned representations;
- **Topological methods for robustness and out-of-distribution detection**, addressing model stability, failure modes, and generalization behavior;
- **Representation shift and training dynamics**, providing topological perspectives on model evolution and explainability during training;
- **Interactive exploration with explainable Mapper**, emphasizing human-in-the-loop analysis and visual exploration of model behavior.

The results of this thematic analysis are summarized in Table 3, which presents a structured comparison of the included studies according to their thematic group, applied topological methods, and interpretability focus.

Table 3. Summary of included studies grouped by thematic focus.

Study	TDA Method	Model Component	Interpretability Focus
<i>Topology of Attention Maps</i>			
[5]	Betti Numbers	Attention Maps	Synthetic Text Detection
[3]	Betti Numbers	Attention Graphs	Structural Attention Complexity
[4]	Betti Numbers	Attention Maps	Acceptability Judgement
[12]	Persistent Homology	Attention Maps	Syntactic Acceptability
[8]	Topological Divergence	Attention Graphs	Hallucination Detection
[15]	Persistent Homology	Attention Maps	Vulnerability to Attacks

Table 3. Cont.

Study	TDA Method	Model Component	Interpretability Focus
<i>Latent Representations and Embedding Spaces</i>			
[13]	Mapper	Embedding Space	Representation Geometry
[14]	Persistent Homology	Word Embeddings	Layer-wise Abstraction
[7]	Persistent Homology	Latent Space	Semantic Collapse Signals
[6]	Persistent Homology	Embeddings	Adversarial Geometry
[11]	Zigzag Persistence	Activations	Semantic Cluster Evolution
[19]	Persistent Homology	Fine-tuned Representations	Geometry Shift Analysis
[22]	Local Topological Metrics	Latent Space	Dialogue Topology
<i>Robustness and OOD Detection</i>			
[20]	Persistent Homology	Output Representations	OOD via Topology
[24]	Topological Divergence	Representation Space	Representation Stability
[23]	Mapper	Neuron Activations	Global Topology
[18]	TopoSOM	Deep Representations	Manifold Insights
<i>Training Dynamics and Representational Shifts</i>			
[2]	Neural Persistence	Weights/Activations	Layer Complexity
[1]	Persistence Diagrams	Activations	Performance Trends
[9]	Persistent Homology	CNN Activations	Feature Evolution
<i>Interactive Exploration with Explainable Mapper</i>			
[26]	Mapper + Agents	Embedding Space	Interactive Explanation

4.8.1. Topology of Attention Maps in Large Language Models

The attention mechanism is essential to how large language models (LLMs) like BERT and GPT understand context. It determines which words the model focuses on when processing a sentence. Traditionally, researchers have tried to interpret model decisions by visualizing attention scores as heatmaps. However, raw attention values often fail to provide a reliable or accurate explanation of the model's reasoning [5]. To address this issue, recent work has turned to topological data analysis (TDA) as a more structured way to interpret attention behavior.

One notable direction was introduced by Kushnareva et al. [3]. They used Bettis numbers, topological descriptors that count connected components and cycles, to analyze attention maps. By treating attention patterns in each layer as graphs, they uncovered how the model connects information across tokens. Their findings showed that certain topological features are related to linguistic structure and model confidence. This led to a method, which allows for a more comprehensive, layer-wise interpretation of how attention flows through the model.

Building on that, Cherniavskii et al. [4] examined how attention topology reflects whether a sentence is grammatically acceptable. They found that ungrammatical sentences often had "broken" attention structures, such as isolated words or excessive connectivity. These patterns appeared through changes in persistent topological features across layers. This topological perspective offers a way to detect when the model is uncertain internally, even without relying on external labels.

In another direction, Bazarova et al. [8] proposed a topological method to identify when LLMs generate hallucinated or nonsensical text. They compared the persistent homology of attention graphs from hallucinated versus accurate outputs. Their analysis showed that hallucinations often accompany unusually noisy or degenerate attention structures. This provides a way to flag unreliable outputs using only the model's internal topology.

Perez and Reinauer [13] took a broader approach by converting entire attention tensors into topological spaces and analyzing how their features change across model layers. Their research revealed that different attention heads specialize in capturing various linguistic properties, such as syntax, meaning, or reference. Each head shows its own distinct

topological footprint. This creates a more nuanced understanding of how language models process information at different levels of abstraction.

Together, these studies illustrate how TDA contributes new insights to model interpretation:

- It reveals the overall shape and structure of how models process information—not just which word attends to which.
- It operates across layers and heads, making it easier to understand how models develop understanding in stages.
- It can help identify issues like grammatical mistakes or hallucinations, often without needing extra annotations or labels.

By viewing attention as more than just a matrix of scores and as a dynamic, evolving structure, topological methods enable deeper and more faithful explanations of how LLMs reason.

4.8.2. Topological Data Analysis of Latent Representations and Embedding Spaces

Beyond attention patterns, one of the best ways to study what a large language model (LLM) “knows” is to look at the structure of its internal representations, including token embeddings, hidden states, and contextualized vectors. These latent spaces are where the model does most of its reasoning and abstraction. However, because these spaces are high-dimensional and complex, understanding them directly can be challenging. This is where Topological Data Analysis (TDA) provides useful tools. It helps reveal the shape and organization of these spaces, offering both global insights and interpretability.

A key direction was explored in the model TopoBERT by Rathore et al. [19]. The authors investigated how fine-tuning changes the geometry of word embeddings. They employed persistent homology to examine alterations in topological features, including the appearance or disappearance of loops and connected components, after domain-specific adaptations. Their findings indicated that task-relevant fine-tuning caused notable topological reorganization in the latent space. This suggests that persistence diagrams could serve as indicators of learned task structure. This method connects the abstract shape of the model’s space to its behavior, providing a clear summary of what shifts when a model learns.

A related view was presented by Chauhan and Kaul in [14]. The authors utilized persistent homology and TDA visualization to investigate differences across layers of BERT. By embedding activations from each layer and calculating persistence diagrams, they showed how topological complexity changes as the input text moves deeper into the network. Notably, earlier layers maintained more of the local syntactic structure, while deeper layers generated more abstract, low-dimensional embeddings. This layer-wise transformation was captured in an understandable way, clarifying how LLMs derive meaning from form.

Fitz et al. [7] expanded on this by examining persistent topological features in the hidden states of autoregressive models like GPT. They discovered that token representations form significant topological features, including cycles and voids, whose structures are linked to sentence length, coherence, and information density. These features can vanish or become chaotic when the model produces degenerate or off-topic text, offering a geometric signal of semantic breakdown. Such signals provide understandable hints about when and how the model might lose control over generation.

Gardinazzi et al. [11] applied persistent homology to explore how semantic structures develop across LLM layers. They looked at how clusters of similar tokens, such as animals or actions, form or dissolve as they move through the transformer. The persistent topological features reflected this clustering, demonstrating how semantic generalization emerges in the model. This offers a different perspective compared to probing tasks, allowing

researchers to track where and how semantic structure appears and how stable it is under changes or fine-tuning.

Lastly, Zhou et al. [23] used the Mapper algorithm to visualize the latent spaces of neuron activations in LLMs. Mapper creates a simplified graph-based structure that maintains the topological framework of the data. Their findings indicated that various layers and tasks create distinct Mapper topologies, yielding a visual and understandable representation of model behavior. For example, Mapper graphs indicated when a model starts to compress input diversity too early, which could lower interpretability or sensitivity to subtle inputs.

Together, these studies illustrate how TDA can make the complex geometry of high-dimensional LLM representations easier to access and understand:

- **Tracking learning and abstraction:** Persistent features show how structure develops across layers or during fine-tuning.
- **Visualizing conceptual clusters:** TDA methods like Mapper highlight how related concepts are organized in the model.
- **Detecting instability:** Changes in topological complexity might indicate overfitting, degeneration, or task mismatch.

As language models continue to grow, understanding their internal geometry becomes more important. TDA offers a solid mathematical way to study these spaces, helping researchers move beyond black-box behavior to gain a deeper understanding of what models truly learn.

4.8.3. Topological Methods for Robustness and Out-of-Distribution Detection

In the wider context of explainability, it is essential to know when large language models (LLMs) act unexpectedly. This includes situations like domain shifts, adversarial inputs, or incorrect outputs. Understanding this is key to building transparent and trustworthy AI systems. Recently, researchers have explored using topological data analysis (TDA) to identify reliability issues and to gain insights into their causes and mechanisms. By looking at changes in the geometry of internal model representations, TDA offers clear signals of reasoning failures that basic surface metrics might miss.

Pollano et al. [20] applied persistent homology to the hidden states of transformer-based LLMs. They found that out-of-distribution (OOD) inputs show different topological signatures than in-distribution examples. Their method detects changes in the shape and connectivity of activation spaces, which relate to shifts in the model's confidence or semantic consistency. This enables an unsupervised, topologically grounded approach to identifying OOD inputs that goes beyond simple confidence scores. It reveals structural changes within the model's representation space, providing a global explanation of when the model encounters something new

Bazarova et al. [8] proposed a similar method to spot hallucinations in LLM outputs, focusing on the topology of attention patterns. They discovered that hallucinated outputs, often deemed incoherent or fabricated by human reviewers, can be linked to unusually complex or disorganized topologies in attention graphs. These topological changes, analyzed with persistent diagrams and divergence metrics, give clear signals of unreliability. Unlike standard attention visualizations, this method goes beyond individual token weights and captures overall patterns of attention disruption, making it more robust and easier to apply.

Snopov and Golubinskiy [15] examined how adversarial attacks disrupt attention structures and found that successful attacks often cause subtle but noticeable changes in topological complexity. Their findings suggest that vulnerability can be seen as a shift in the geometric structure of information flow. This shift is revealed not through raw logits but through the persistence of topological features across attention layers. This perspective

redefines adversarial sensitivity as a structurally interpretable issue, highlighting areas in the network where reasoning becomes unstable.

Overall, these studies show that TDA can do more than just detect problems; it can offer structural explanations for when and how LLMs fail. By focusing on the overall geometry of representations and attention, topological methods clarify what occurs inside the model during failures. This represents a shift from reactive detection to proactive interpretability, supporting the larger goal of making LLMs both reliable and understandable.

Traditional methods for detecting when models fail—such as using confidence thresholds or measuring sequence-level divergence—often capture only surface-level issues. They tell us *what* went wrong, but not *why*. These approaches can easily miss deeper changes in the internal geometry of model representations. Topological Data Analysis (TDA) offers a different way of looking at the problem. Instead of focusing on outputs alone, it studies the overall shape and structure of the model's reasoning process. By examining how these structures shift, TDA can reveal early signs of instability or confusion that traditional metrics overlook. This perspective turns failure detection into something more insightful: a window into how models think and where that reasoning begins to break down. In doing so, TDA brings us closer to building language models that are not just powerful, but also transparent, reliable, and easier to trust.

TDA-based methods in this area are still evolving, but they present exciting opportunities for future research. Potential extensions include connecting topological signatures with human-rated uncertainty, comparing robustness across different architectures, or adding topology-based diagnostics into model training processes.

While several reviewed studies report qualitative or distance-based robustness signals between ID and OOD samples, formal statistical hypothesis testing is rarely conducted; a generic validation framework is outlined in Section 3.2.

4.8.4. Representation Shift and Training Dynamics: Topological Perspectives on Model Explainability

One of the least understood yet most crucial aspects of large language models (LLMs) is how their internal representations change during training, fine-tuning, or continual learning. These shifts in representation affect performance and greatly impact explainability. A model's ability to generalize, reason, and behave consistently depends on how stable and organized its internal geometry is over time. Topological data analysis (TDA) provides unique tools to capture and measure these changes, offering insights into when and where a model's understanding forms—or fails.

From an interpretability standpoint, the evolution of internal representations across layers and training steps reveals how the model converts raw input into more abstract, task-relevant information. TDA achieves this by treating hidden states and embedding trajectories as high-dimensional geometric objects. We can track and interpret their topological features, such as connected components, loops, and voids, over time.

Gardinazzi et al. [11] looked at persistent topological features in LLMs during inference and fine-tuning, using Zizag persistence. Their study showed that semantic clusters—like animal names, professions, or verbs—form consistent components that appear, merge, or disappear across model depth. From an explainability point of view, this highlights where semantic abstraction starts to develop in the model and how concepts are organized topologically. It also provides a method to localize interpretability. By identifying which layers stabilize meaningful groupings, practitioners can understand better which parts of the network encode specific types of knowledge.

Fitz et al. [7] examined the appearance and disappearance of topological holes (cycles and voids) in LLM hidden states across sequences. They discovered that coherent text inputs created stable topological features, while less coherent or hallucinated outputs often

showed chaotic or collapsed structures. These findings indicate that persistent topological features may serve as structural signatures of semantic consistency. Unlike attention-based saliency, which shows where the model focuses, this method indicates how it maintains structured meaning throughout generation.

Barannikov et al. [24] tackled training dynamics directly by introducing Representation Topology Divergence (RTD). This metric quantifies the topological shift of representations between pre-training and fine-tuning. They found that sudden topological changes often coincided with reduced generalization or domain overfitting—significant concerns for explainability. RTD provides a mathematically grounded indicator of when a model’s internal structure has significantly changed, allowing for the diagnosis of knowledge loss or concept drift without relying on external validation sets.

From a mathematical perspective, Representation Topology Divergence (RTD) is defined through a comparison of persistence-based topological summaries computed from two representations defined on the same input set. Let R_a and R_b denote two representations and let $D(R_a)$ and $D(R_b)$ be the corresponding persistence diagrams obtained via the R-Cross-Barcode construction [24]. RTD is then expressed as

$$\text{RTD}(R_a, R_b) = d(D(R_a), D(R_b)),$$

where d denotes a standard distance on persistence diagrams, such as the bottleneck or Wasserstein distance. Since these distances are non-negative, symmetric, and satisfy the triangle inequality, RTD inherits these properties. At the level of representations, RTD therefore defines a pseudometric, as distinct representations may induce identical persistence diagrams, yielding zero divergence without implying representational equivalence. The formal construction and theoretical justification of RTD are provided in [24], and are not redeveloped here.

Zhou et al. [25] offered more insights by analyzing adversarial training through Mapper-based topological visualizations. Their work demonstrated that adversarial fine-tuning can lead layers to compress diverse inputs too quickly, suggesting a loss of representational flexibility. From an interpretability perspective, this underscores a failure mode where the model might behave predictably but with poor reasoning detail. Topological visualization made these shifts visible, adding explanatory value that traditional accuracy metrics overlook.

The interpretability importance of these studies lies in their ability to

- **Explain when and where semantic abstraction happens:** By identifying stable topological features (e.g., clusters, holes), TDA helps pinpoint conceptual representations in the model’s architecture.
- **Track the degradation or strengthening of structure:** Persistent homology shows if semantic structures are preserved or overwritten during training, fine-tuning, or adversarial adaptation.
- **Support model introspection:** Topological shifts across training epochs provide interpretable metrics for when a model is likely to forget, generalize poorly, or become fragile.

Although several of the reviewed studies suggest that topological features are useful in identifying hallucination-related behavior or increased model uncertainty, this connection is typically discussed qualitatively or comparatively. In particular, explicit numerical correlation coefficients between persistence-based measures (such as persistence lifetimes or persistence entropy) and standard uncertainty indicators like perplexity or output entropy are not reported. The existing evidence therefore points to an association rather than a quantified statistical relationship. At a high level, such a relationship could be

quantified by correlating summary statistics derived from persistence diagrams with token- or sequence-level uncertainty measures across inputs, which remains an open direction for future work.

4.8.5. Interactive Exploration with Explainable Mapper

One of the more exciting and hands-on approaches in the space of topological analysis for LLMs comes from Yan et al. [26], who propose something called *Explainable Mapper*. The idea is simple but powerful: rather than just showing a static topological summary of an embedding space, this framework invites users to **interact** with the Mapper graph. Users can explore clusters, follow edges, and pose hypotheses about why certain regions of the embedding space behave a certain way—are they driven by syntax, semantics, or something else?

What makes this approach especially valuable is the use of *perturbation-based agents* that test these hypotheses automatically. This allows the system to validate whether a proposed explanation holds when input variations are introduced, leading to more reliable and grounded interpretations. If the explanation is robust, it gets flagged as meaningful; otherwise, users are guided to re-evaluate.

This method turns topological summaries into **living tools** for exploration, moving from passive visualization to active, explainable engagement. It is an important step toward more transparent, collaborative, and testable LLM interpretability.

5. Discussion

5.1. Positioning TDA Relative to Existing Interpretability Methods

To situate the findings of this scoping review within the broader landscape of interpretability and explainability for large language models, we begin by contrasting TDA-based approaches with commonly used non-TDA methods. Table 4 summarizes the key distinctions observed in the reviewed studies.

Table 4. Comparison of TDA-based and non-TDA interpretability methods for LLMs.

Aspect	TDA-Based Methods	Non-TDA Methods
Core principle	Capture topological features (e.g., clusters, loops, voids) in high-dimensional representations	Emphasize local importance via gradients, masking, or attribution
Level of analysis	Global structure and geometry across layers or model states	Local saliency at token or neuron level
Model components analyzed	Embedding spaces, hidden states, attention graphs	Gradients, logits, attention weights, neuron activations
Interpretability output	Persistence diagrams, Betti curves, Mapper graphs	Saliency maps, SHAP/LIME scores, attention visualizations
Sensitivity to instability	Detects global structural shifts and collapse	Detects local anomalies but may miss geometric degradation
Typical use cases	Representation analysis, robustness, OOD behavior	Decision-level explanations, feature importance
Scalability	Computationally intensive; mitigated through approximation	Generally lightweight and fast
Explainability depth	Structural, often unsupervised insight	Intuitive, prediction-level explanations

This comparison reflects a pattern consistently observed in the Results: TDA-based and non-TDA approaches address complementary interpretability objectives. While non-TDA methods are well suited for explaining individual predictions, TDA is primarily employed to analyze the organization, stability, and evolution of internal model representations.

5.2. Interpretation of Cross-Study Patterns

The cross-study analysis reported in Section 4 reveals several consistent patterns characterizing current uses of topological data analysis for interpretability and explainability in large language models. Persistent-homology-based methods dominate the literature, with full persistent homology accounting for a substantial portion of reviewed studies and approximate or PH-derived summaries comprising most of the remainder. This concentration indicates that persistence-based topology serves as the principal mathematical foundation for topological analysis of language model representations.

The Results further show a strong alignment between methodological choice and interpretability orientation. Studies employing full persistent homology are overwhelmingly associated with implicit interpretability objectives, including structural characterization of latent spaces, diagnostic analysis of attention connectivity, and assessment of representational stability. Explicit explainability is comparatively rare within this group, suggesting that the expressive richness of full topological information does not readily translate into direct explanatory artifacts.

By contrast, explicit explainability is most frequently observed in studies employing approximate persistent homology or non-PH approaches, such as Mapper-based abstractions and graph-level summaries of attention. These methods reduce the complexity of the underlying topological representation and prioritize interpretive clarity, often at the expense of fine-grained geometric detail. The cross-study distribution therefore indicates that abstraction plays a key mediating role between topological analysis and explainability.

Scalability further differentiates methodological clusters. Studies exhibiting moderate to high scalability rely almost exclusively on approximate or non-PH methods, whereas low scalability is strongly associated with full persistent homology. This pattern indicates that scalability is not orthogonal to interpretability, but actively shapes the form that explanation can take.

The cross-study patterns also reveal that methodological choice is more closely coupled to analytical objective than to model family. Although many studies focus on BERT or other transformer-scale models rather than full-scale LLMs, similar topological techniques are applied across architectures. This suggests that the observed patterns reflect structural properties of transformer representations rather than artifacts of model scale alone.

Taken together, these patterns indicate that the current landscape of TDA-based interpretability is organized around recurring trade-offs rather than a single dominant paradigm. Persistent homology, approximate topology, and non-PH abstractions occupy distinct positions with respect to fidelity, interpretability, and scalability.

5.3. Implicit and Explicit Interpretability as Distinct Result-Driven Modes

One of the most salient patterns emerging from the Results is the clear separation between implicit interpretability and explicit explainability in TDA-based analyzes of large language models. The majority of the reviewed studies—particularly those that employ full persistent homology—are oriented toward implicit interpretability, using topological descriptors to characterize internal structure, diagnose model behavior, or compare representational states, without producing explanations intended for direct human consumption.

This dominance of implicit interpretability is closely related to the nature of topological information itself. Persistent homology captures global, multi-scale properties of high-dimensional spaces that are robust to noise and small perturbations. Although these properties are mathematically expressive and well suited for comparative analysis, they are not naturally aligned with token-level or decision-level explanations typically expected in explainable AI. Consequently, topological summaries are most often interpreted as signals of structural organization, stability, or degradation rather than as standalone explanations.

Explicit explainability, by contrast, is relatively rare in the reviewed corpus and is strongly associated with approximate persistent homology or non-PH methods such as Mapper-based abstractions and graph-level summaries. These approaches introduce additional layers of abstraction that compress or reorganize topological information into forms that are more accessible to human users. The Results show that such abstraction is a necessary condition for explicit explainability in topology-based frameworks, as it mediates between mathematical structure and interpretive clarity.

Although topological descriptors are abstract, reviewed studies typically communicate their effects through visual and structural proxies (e.g., clustering patterns, regime changes, or graph connectivity), rather than through direct semantic explanations, highlighting an inherent semantic gap that remains an open challenge.

Importantly, the distinction between implicit and explicit interpretability does not represent a binary division. Instead, the Results suggest a continuum in which topological methods occupy different positions depending on how their outputs are processed and presented. Full persistent homology tends to reside at the implicit end of this continuum, while approximate and Mapper-based approaches move closer to explicit explainability by prioritizing summarization, visualization, and interaction.

The Results further indicate that interpretability orientation is closely coupled to scalability. Studies supporting explicit explainability are disproportionately represented among methods with moderate to high scalability, whereas low-scalability methods are almost exclusively associated with implicit interpretability. This alignment suggests that computational feasibility and interpretability are jointly constrained, rather than independent design dimensions.

A related limitation concerns the actionability of topology-based explanations. While TDA provides a global and structural view of representation spaces, the reviewed literature generally does not offer explicit inverse mappings from detected topological features (e.g., loops, voids, or regime changes) back to individual input tokens. Instead, any localization is handled indirectly, for example by associating topological anomalies with clusters of representations, specific attention heads, or subsets of inputs that can then be examined using local explainability methods. As such, topology-based approaches are best viewed as complementary to token-level attribution techniques rather than as direct mechanisms for input-level explanation.

5.4. Attention and Representations as Structural Objects

The Results indicate that a substantial portion of TDA-based work focuses on attention mechanisms and latent representations. Rather than treating attention weights or embeddings as isolated numerical quantities, the reviewed studies analyze their induced graph or manifold structure.

This structural perspective enables the identification of patterns—such as fragmentation, over-connectivity, or loss of coherence—that are not consistently visible through raw attention visualizations or probing-based analysis. In this sense, topology is used to study structure first, with the explanation emerging only secondarily through interpretation of that structure.

5.5. Scalability as a Result-Shaping Constraint

Scalability emerges from the Results as one of the most influential factors shaping how TDA-based interpretability methods are designed and applied in practice. Across the reviewed corpus, nearly three quarters of studies are classified as exhibiting low to moderate scalability, reflecting the computational demands of topological analysis on high-dimensional embeddings and dense attention structures. This pattern is observed

consistently across model families and application domains, indicating that scalability constraints are primarily methodological rather than task-specific.

The cross-study analysis shows a strong association between scalability and the type of topological method used. Exact persistent homology is used almost exclusively in low-scalability settings, typically involving restricted subsets of representations, reduced dimensionality, or smaller transformer models. In contrast, studies classified as moderately or highly scalable rely predominantly on approximate persistent homology, PH-derived summaries, or non-PH methods such as Mapper. This distribution indicates that approximation and abstraction are central mechanisms for enabling practical topological analysis at scale.

Importantly, the results indicate that scalability constraints do more than limit computational feasibility; they actively shape interpretability outcomes. Approaches that preserve detailed topological structure tend to support implicit interpretability, focusing on structural diagnostics, robustness analysis, or comparative evaluation. By contrast, methods that achieve higher scalability are more frequently associated with explicit or interactive forms of explainability.

The Results further suggest that scalability influences the level and resolution at which topological analysis is conducted. Lower-scalability approaches often focus on static snapshots or aggregated statistics, whereas more scalable methods enable analysis across larger input sets, multiple layers, or evolving attention graphs. This reinforces the view that scalability functions as a structuring dimension of interpretability rather than a secondary implementation concern.

From a computational perspective, this distinction also reflects a separation between the offline and real-time use cases. Due to the computational cost of exact persistent homology constructions, the reviewed studies primarily apply TDA-based methods in offline or post hoc analysis settings, rather than for continuous real-time monitoring of large language models in production.

5.6. Stability and Robustness Considerations for Mapper-Based Explanations

An additional open challenge concerns the stability of Mapper-based constructions. Several of the reviewed studies employ Mapper for interactive exploration or qualitative analysis of embedding spaces and attention-derived representations, leveraging its ability to produce interpretable graph abstractions. In these works, robustness is typically assessed empirically by examining whether salient Mapper structures (such as clusters, branches, or connectivity patterns) persist across variations in the filter function, cover resolution, overlap parameter, or clustering strategy.

Unlike persistent homology, however, Mapper does not currently admit formal stability theorems guaranteeing invariance under perturbations of the data or hyperparameters. As a result, the reliability of Mapper-based explanations in the reviewed literature relies on consistency across multiple parameter settings rather than on provable robustness. Developing mathematical frameworks that characterize the stability of Mapper graphs under perturbations of the data, filter functions, or covering parameters remains an important direction for future work and would significantly strengthen the theoretical foundations of Mapper-based interpretability.

5.7. Complementarity Rather Than Replacement

Taken together, the Results support an interpretation in which TDA-based and non-TDA interpretability methods serve complementary roles. Non-TDA approaches provide localized, intuitive explanations tied to individual predictions, whereas TDA-based methods capture global organization, stability, and structural change across model components.

Rather than replacing existing explainability techniques, topology contributes an additional structural layer of interpretation that enriches the understanding of large language models while preserving the strengths of established methods.

5.8. Illustrative Analysis Workflow for LLM Representations

To support reproducibility and clarify how the reviewed studies operationalize topological data analysis in practice, we summarize an illustrative analysis workflow linking raw model representations to structured geometric summaries. This workflow is intended as a high-level descriptive template that captures common methodological steps across the literature, rather than as a prescriptive or implementation-specific procedure.

1. **Select the representation of interest.** Identify the internal model representations to be analyzed, such as token embeddings, layer-wise hidden states, intermediate activations, attention-based constructs, or pooled sequence-level representations.
2. **Extract representations from the model.** For a collection of inputs $\{x_i\}_{i=1}^N$, extract the corresponding representations $\{z_{i,\ell}\}$ at one or more layers or components $\ell \in \mathcal{L}$. Depending on the study, these representations can be defined at the token, sequence, or component level.
3. **Preprocess and define the comparison space (optional).** Apply standard preprocessing steps such as normalization, dimensionality reduction, subsampling, or landmark selection to manage scale and computational cost. Specify an appropriate distance or similarity measure to compare representations, noting that this choice influences both the resulting structure and the computational feasibility.
4. **Organize representations into a structured object.** Arrange the representations into a point cloud, similarity graph, or other relational structure suitable for subsequent analysis. The choice of structure depends on whether representations are treated as geometric objects or as elements in a relational system.
5. **Construct multi-scale structures.** Define a family of nested structures across a range of scales, for example via distance-based constructions or thresholded graph representations. These constructions provide a way to summarize how relationships among representations evolve as the analysis scale changes.
6. **Compute summary descriptors.** From the resulting structures, compute summary descriptors that capture structural properties across scales. These may include persistence-based summaries, curve-based statistics, graph abstractions, or other aggregated representations used in the reviewed literature.
7. **Aggregate across components, when applicable.** When analyses are performed across multiple layers, heads, or components, aggregate summaries are generated using averaging or consensus-based approaches to obtain stable, model-level characterizations.
8. **Compare conditions and analyze differences.** Compare summaries across inputs, layers, perturbations, training stages, or data regimes (e.g., in-distribution versus out-of-distribution). Such comparisons are typically used to support diagnostic or comparative analysis rather than to produce direct input-level explanations.

6. Conclusions

This scoping review examines the published work on the use of topological data analysis (TDA) for the interpretability and explainability of large language models and related transformer-based architectures. By reviewing 26 studies published between 2018 and 2025, this review documents how topological techniques such as persistent homology, PH-derived summaries, and non-persistent approaches including Mapper have been applied to analyze embeddings, attention mechanisms, and internal representations.

The central contribution of this scoping review lies in its formal mathematical synthesis of the literature. By organizing existing studies according to the homological dimension, the type of representation manifold and the lifetime usage of persistence, we provide a principled taxonomy that clarifies how different TDA methods relate to each other and highlights the recurring methodological patterns across applications. This structure enables comparison across studies that otherwise would remain fragmented across venues and application domains.

The reviewed literature shows that TDA is most commonly used to support interpretability at a structural level. Topological descriptors are used primarily to characterize global organization, stability, and variation in internal model representations, rather than to explain individual predictions. In this respect, TDA-based analyses are typically used alongside, rather than in place of, established explainability methods such as attention visualization, attribution, or probing.

In all studies, a clear distinction is observed between implicit interpretability and explicit explainability. Full persistent homology is predominantly associated with implicit interpretability, where topological summaries are used for diagnostic or comparative analysis by researchers. Explicit explainability appears less frequently and is reported primarily in studies that apply approximation, abstraction, or visualization-based techniques. This pattern reflects differences in how topological information is summarized and communicated, rather than differences in the model family or the application domain.

Scalability is consistently identified as a practical constraint in the reviewed work. Exact topological constructions are generally applied to restricted subsets of representations or smaller models, while approximate and non-persistent methods are used in settings requiring greater computational feasibility. These constraints are reflected in both both methodological choices and the types of interpretability supported by the studies.

7. Future Directions

As we have noted in this review, the use of topological data analysis (TDA) in the context of large language model (LLM) interpretability remains in its infancy. Although the studies we reviewed have demonstrated promise, they have also underscored the fact that the domain continues to be in its nascent stage.

One of the most pressing issues is scale. Most of the work to date has focused on not-so-large models, while today's LLMs have billions of parameters. If TDA is to have a lasting impact, persistent homology and Mapper, among other topological techniques, will need to be modified such that they can process models like GPT-4 or LLaMA-3. That will most likely require TDA to devise clever approximations, more efficient algorithms, and distributed computing.

Future research should continue advancing toward interactive TDA pipelines tailored for LLM interpretability—particularly those that combine Mapper-style topological summaries with automated explanation agents and mechanisms for user feedback. These approaches go beyond static visualization; they offer a dynamic way to probe, question, and better understand model behavior.

The framework introduced in *Explainable Mapper* [26] exemplifies this emerging direction. It allows users to engage with embedding spaces in real time, formulate hypotheses, and rely on verification agents to test those intuitions in a structured manner. This type of workflow turns topological tools into active components of model interpretability, rather than passive summaries.

A promising direction for future work is to explore how topology-based explanatory summaries can be formalized in a compositional manner using category theory. In this perspective, internal representations may be treated as objects in a category \mathbf{Emb} , while

representation updates induced by network layers or architectural blocks act as morphisms. Topological summaries could then be obtained via a functor

$$\mathcal{T} : \mathbf{Emb} \rightarrow \mathbf{Pers},$$

mapping embeddings to persistence modules or diagrams.

Functoriality may ensure that the topological characterization of composed transformations is consistent with the composition of their constituent parts, suggesting a principled notion of compositional interpretability. When combined with stability results from persistent homology, such a framework could support structural and diagnostic explanations that remain comparable across layers, model variants, or training stages, while being robust to controlled perturbations. Further investigation of categorical constructions, such as natural transformations or sheaf-based models, may help formalize these ideas and clarify their implications for interpretability in large language models.

When combined with stability results from persistent homology, this perspective suggests that controlled changes in representations induce bounded changes in their topological summaries. As a result, topology-based explanations may offer structural and diagnostic insight that is comparable across layers, model variants, or training stages, while remaining robust to implementation-level perturbations.

Improving scalability without sacrificing explanatory fidelity remains an important direction for future research. Promising avenues include the development of streaming or windowed topological summaries, adaptive subsampling strategies guided by representation uncertainty, and tighter integration between topological methods and representation compression or dimensionality reduction techniques. Progress along these directions will be essential for extending TDA-based explainability beyond exploratory analysis toward more operational and scalable use in large language models.

Extending such systems to handle larger-scale language models and real-world, task-specific settings—such as in legal, clinical, or multilingual domains—could significantly enhance transparency, robustness, and ultimately, trust in LLM-driven applications. Making these tools accessible and interpretable not only for researchers but also for practitioners and domain experts will be a key step forward.

Also, an important open direction is the development of principled inverse mappings that connect global topological descriptors to localized input-level explanations, enabling tighter integration between topology-based analysis and token-level explainability methods.

Overall, this review provides a descriptive synthesis of existing approaches at the intersection of TDA and language model interpretability. It clarifies how topological methods are currently used, the interpretability roles they tend to support, and the practical limitations reported in the literature, without assessing comparative effectiveness or proposing normative evaluation criteria.

From a mathematical perspective, several open problems emerge from this synthesis. These include the limited exploration of higher-dimensional homology in language model representations, the development of metric-aware or functorial formulations of attention-induced complexes, and the absence of formal links between topological instability and optimization dynamics during training. Addressing these challenges would strengthen the theoretical foundations of topology-based interpretability and clarify the limits of current methods.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/math14020378/s1>. Table S1: Categories of full-text exclusion and corresponding frequencies.

Author Contributions: Conceptualization, P.S., M.S.M. and V.D.R.; methodology, P.S., D.K., I.G. and K.M.; software, P.S., D.K. and I.G.; validation, P.S., D.K., I.G., K.M. and M.S.M.; formal analysis, P.S., I.G. and V.D.R.; investigation, P.S., I.G., K.M., M.S.M. and V.D.R.; resources, P.S., D.K. and I.G.; data curation, P.S. and I.G.; writing—original draft preparation, P.S., D.K. and I.G.; writing—review and editing, P.S., I.G., K.M., M.S.M. and V.D.R.; visualization, P.S., D.K., I.G. and K.M.; supervision, V.D.R.; project administration, M.S.M. and K.M.; funding acquisition, M.S.M. All authors have read and agreed to the published version of the manuscript substantially to the work reported.

Funding: Funded by the European Union under Horizon Europe (project ChatMED grant agreement ID: 101159214).

Data Availability Statement: No new data were created or analyzed in this study.

Acknowledgments: Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Methodological Bias, Normalization, and Heterogeneity Analysis

Appendix A.1. Purpose

This appendix describes the quantitative framework used to assess methodological bias and heterogeneity among the studies included after the application of the eligibility criteria. The objective of this analysis is not to rank or exclude studies, but to provide a transparent, reproducible characterization of methodological dispersion within the reviewed corpus, consistent with PRISMA 2020 recommendations for scoping reviews.

Appendix A.2. Scoring Scheme and Normalization

Each included study was evaluated across seven methodological dimensions: data transparency (D), pipeline reproducibility (R), TDA validity (T), evaluation rigor (E), interpretability grounding (I), model scope (M), and scalability (Sc). Each dimension was scored on a discrete ordinal scale $\{0, 1, 2\}$, where higher values indicate stronger methodological support.

For each study i , a total methodological score was computed as

$$S_i = D_i + R_i + T_i + E_i + I_i + M_i + Sc_i, \quad S_i \in [0, 16].$$

To enable cross-study comparison without introducing exclusion thresholds, scores were normalized using the maximum observed score in the corpus,

$$S_{\max} = \max_i S_i.$$

The normalized score was then defined as

$$\tilde{S}_i = \frac{S_i}{S_{\max}}, \quad \tilde{S}_i \in [0, 1].$$

These normalized values were used exclusively for weighting and heterogeneity estimation and did not affect study inclusion or qualitative synthesis.

Appendix A.3. Per-Study Methodological Scoring

Table A1. Per-study methodological scoring used for heterogeneity analysis.

Reference	D	R	T	E	I	M	Sc	S
[11]	2	1	2	1	1	2	1	10
[6]	1	1	2	2	1	2	1	10
[7]	0	1	2	1	1	2	0	7
[8]	2	2	2	2	2	1	1	12
[5]	2	1	2	2	1	1	2	11
[3]	2	1	2	1	1	1	2	10
[4]	2	1	2	2	1	1	2	11
[12]	2	1	2	2	2	1	0	10
[13]	2	1	2	2	2	1	0	10
[14]	2	1	2	1	1	1	0	8
[15]	2	2	2	2	2	1	1	12
[16]	2	2	2	2	2	1	1	12
[17]	0	1	2	2	2	1	0	8
[18]	0	1	1	1	0	0	0	3
[19]	2	2	2	2	1	1	0	10
[10]	2	1	2	1	1	1	0	8
[20]	2	2	2	2	1	1	1	11
[21]	2	2	2	1	1	0	0	8
[22]	1	1	2	1	1	1	1	8
[23]	2	2	1	1	1	0	2	9
[24]	2	2	1	2	1	0	2	10
[1]	1	1	2	1	0	0	0	5
[9]	2	2	2	2	0	0	0	8
[2]	2	2	1	2	0	0	2	9
[25]	2	2	2	1	1	0	0	8
[26]	2	2	2	2	2	2	2	14

Appendix A.4. Weighting and Heterogeneity Computation

Normalized scores \tilde{S}_i were used as synthesis weights,

$$w_i = \tilde{S}_i = \frac{S_i}{S_{\max}}$$

The weighted mean methodological score is defined as

$$\bar{S}_w = \frac{\sum_i w_i S_i}{\sum_i w_i}$$

Methodological heterogeneity was quantified using a weighted variance,

$$H = \frac{\sum_i w_i (S_i - \bar{S}_w)^2}{\sum_i w_i}$$

For comparability across corpora, a normalized heterogeneity index may be reported as

$$H_{\text{norm}} = \frac{H}{S_{\max}^2}$$

Appendix A.5. Interpretation

This framework enables a quantitative characterization of methodological dispersion among the included studies. Observed heterogeneity primarily reflects differences in data availability, interpretability grounding, model scope, and scalability, while the underlying application of topological methods remains comparatively consistent across the corpus.

References

1. Wheeler, M.; Astudillo, R.; Bubenik, P. Activation Landscapes as a Topological Summary of Neural Network Performance. In Proceedings of the IEEE Big Data 2021, Orlando, FL, USA, 5–18 December 2021; pp. 1–10.
2. Rieck, B.A.; Togninalli, M.; Bock, C.; Moor, M.; Horn, M.; Gumbsch, T.; Borgwardt, K.M. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. In Proceedings of the ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
3. Kushnareva, L.; Piontkovski, D.; Piontkovskaya, I. Betti Numbers of Attention Graphs Is All You Really Need. *arXiv* **2022**, arXiv:2207.01903. [[CrossRef](#)]
4. Cherniavskii, D.; Tulchinskii, E.; Mikhailov, V.; Proskurina, I.; Kushnareva, L.; Artemova, E.; Barannikov, S.; Piontkovskaya, I.; Piontkovski, D.; Burnaev, E. Acceptability Judgements via Examining the Topology of Attention Maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 88–107. [[CrossRef](#)]
5. Kushnareva, L.; Cherniavskii, D.; Mikhailov, V.; Artemova, E.; Barannikov, S.; Bernstein, A.; Piontkovskaya, I.; Piontkovski, D.; Burnaev, E. Artificial Text Detection via Examining the Topology of Attention Maps. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 635–649.
6. Fay, A.; García-Redondo, I.; Wang, Q.; Dubossarsky, H.; Monod, A. Holes in Latent Space: Topological Signatures under Adversarial Influence. *arXiv* **2025**, arXiv:2505.20435. [[CrossRef](#)]
7. Fitz, S.; Romero, P.; Schneider, J.J. Hidden Holes: Topological Aspects of Language Models. *arXiv* **2024**, arXiv:2406.05798. [[CrossRef](#)]
8. Bazarova, A.; Yugay, A.; Shulga, A.; Ermilova, A.; Volodichev, A.; Plev, K.; Belikova, J.; Parchiev, R.; Simakov, D.; Savchenko, M.; et al. Hallucination Detection in LLMs via Topological Divergence on Attention Graphs. *arXiv* **2025**, arXiv:2504.10063. [[CrossRef](#)]
9. Purvine, E.; Brown, D.; Jefferson, B.; Joslyn, C.; Praggastis, B.; Rathore, A.; Shapiro, M.; Wang, B.; Zhou, Y. Experimental Observations of the Topology of Convolutional Neural Network Activations. In Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023), Washington, DC, USA, 7–14 February 2023; AAAI Press: San Francisco, CA, USA, 2023. [[CrossRef](#)]
10. Tulchinskii, E.; Kuznetsov, K.; Kushnareva, L.; Cherniavskii, D.; Barannikov, S.; Piontkovskaya, I.; Nikolenko, S.; Burnaev, E. Topological Data Analysis for Speech Processing. In Proceedings of the Interspeech 2023, Dublin, Ireland, 20–24 August 2023.
11. Gardinazzi, Y.; Panerai, G.; Viswanathan, K.; Ansuini, A.; Cazzaniga, A.; Biagetti, M. Persistent Topological Features in Large Language Models. *arXiv* **2024**, arXiv:2410.11042. [[CrossRef](#)]
12. Proskurina, I.; Piontkovskaya, I.; Artemova, E. Can BERT eat RuCoLA? Topological data analysis to explain. In Proceedings of the 9th Workshop on Slavic Natural Language Processing, Dubrovnik, Croatia, 6 May 2023; pp. 123–137.
13. Perez, I.; Reinauer, R. The Topological BERT: Transforming Attention into Topology for Natural Language Processing. *arXiv* **2022**, arXiv:2206.15195. [[CrossRef](#)]
14. Chauhan, J.; Kaul, M. BERTops: Studying BERT Representations under a Topological Lens. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8. [[CrossRef](#)]
15. Snopov, P.; Golubinskiy, A.N. Vulnerability Detection via Topological Analysis of Attention Maps. *arXiv* **2024**, arXiv:2410.03470. [[CrossRef](#)]
16. Uchendu, A.; Le, T.; Lee, D. TopFormer: Topology-Aware Authorship Attribution of Deepfake Texts with Diverse Writing Styles. In Proceedings of the ECAI 2024, Santiago de Compostela, Spain, 19–24 October 2024; pp. 1446–1454.
17. Sakurai, W.; Takayama, M.; Asada, K.; Kurosawa, K. Authorship Attribution by Attention Pattern of BERT with Topological Data Analysis and UMAP. In Proceedings of the IEEE International Conference on Artificial Intelligence in Information and Communication (ICAII), Osaka, Japan, 18–21 February 2025; pp. 1–6.
18. Magai, G. Deep Neural Network Architectures from the Perspective of Manifold Learning. In Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence (PRAI), Qingdao, China, 27–29 October 2023; pp. 1–8.
19. Rathore, A.; Zhou, Y.; Srikumar, V.; Wang, B. TopoBERT: Exploring the Topology of Fine-Tuned Word Representations. *Inf. Vis.* **2023**, *22*, 186–208. [[CrossRef](#)]
20. Pollano, A.; Chaudhuri, A.; Simmons, A. Detecting Out-of-Distribution Text Using Topological Features of Transformer-Based Language Models. *arXiv* **2023**, arXiv:2311.13102.
21. Rathore, A.; Chalapathi, N.; Palande, S.; Wang, B. TopoAct: Visually Exploring the Shape of Activations in Deep Learning. *Comput. Graph. Forum* **2021**, *40*, 111–125. [[CrossRef](#)]
22. Ruppik, B.M.; Heck, M.; van Niekerk, C.; Vukovic, R.; Lin, H.; Feng, S.; Zibrowius, M.; Gasic, M. Local Topology Measures of Contextual Language Model Latent Spaces with Applications to Dialogue Term Extraction. In Proceedings of the SIGDIAL 2024, Kyoto, Japan, 18–20 September 2024; pp. 344–356.

23. Zhou, Y.; Jenne, H.; Brown, D.R.; Shapiro, M.; Jefferson, B.; Joslyn, C.; Henselman-Petrusek, G.; Praggastis, B.; Purvine, E.; Wang, B. Comparing Mapper Graphs of Artificial Neuron Activations. In Proceedings of the Topological Data Analysis and Visualization Workshop (TopoInVis), Graz, Austria, 22 October 2023; pp. 41–50.
24. Barannikov, S.; Trofimov, I.; Balabin, N.; Burnaev, E. Representation Topology Divergence: A Method for Comparing Neural Network Representations. *arXiv* **2021**, arXiv:2103.08749.
25. Zhou, Y.; Zhou, Y.; Ding, J.; Wang, B. Visualizing and Analyzing the Topology of Neuron Activations in Deep Adversarial Training. In Proceedings of the TAG-ML 2023, Honolulu, HI, USA, 23–29 July 2023; pp. 134–145.
26. Yan, X.; Sevastjanova, R.; van der Ben, S.; El-Assady, M.; Wang, B. Explainable Mapper: Charting LLM Embedding Spaces Using Perturbation-Based Explanation and Verification Agents. *arXiv* **2025**, arXiv:2507.18607. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.