Geometric and Topological Structure of LLM Responses under Prompt Variations*

Denys Symonov[1,*,†], Oleksandr Palagin[1,†] and Yehor Symonov[1,†]

[1] V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences (NAS) of Ukraine, Academician Glushkova Avenue 40, 03187, Kyiv, Ukraine

Abstract

The paper considers the problem of formalizing the stability of a large language model to variations in the formulation of a query as a property of its response space. A geometric-topological approach is proposed, within which the set of responses is presented as a metric space of embedding vectors, and its structure is analyzed using persistent homology. On this basis, an integral indicator of topological stability and a system of partial indices are introduced, reflecting component integrity, topological complexity, and sensitivity to query variations. A typology of the model's behavior modes is formulated, which connects topological invariants with the nature of response generation. Experimental verification demonstrates that the deterministic generation mode corresponds to a compact and topologically integral structure, while stochasticity leads to its disorganization. It is established that semantic variations affect mainly the geometry of the response space, while structural changes cause its fragmentation and a decrease in topological stability. The results obtained demonstrate the possibility of quantitative analysis of LLM behavior modes and create a basis for further research into their reliability.

Keywords

large language models (LLM), prompt sensitivity, embedding space, representation learning, topological data analysis, persistent homology, model robustness, semantic and structural variations, CEUR-WS

1. Introduction

Modern large language models (LLMs), including transformer-based architectures such as GPT, PaLM, and LLaMA, achieve strong performance in text generation tasks but remain highly sensitive to prompt formulation. Minor semantic or structural perturbations can induce substantial changes in content, reasoning, and compositional structure, complicating the assessment of robustness and reproducibility. Existing approaches to prompt sensitivity rely on embedding-based similarity measures, statistical descriptors, and clustering techniques [1-3]. While these methods provide local estimates of output variability, they fail to capture the global organization of the response space, including its connectivity and higher-order structure, and remain sensitive to metric choice and analysis scale. As a result, they do not identify invariant properties under admissible perturbations. Topological data analysis (TDA) offers a principled framework by modeling the response set as a geometric object with multi-scale structure. Persistent homology enables the extraction of topological invariants, such as connected components and cycles, that are stable under noise and local deformations [4-5].

This paper introduces a geometric–topological framework for analyzing LLM response stability under semantic and structural prompt perturbations. The approach embeds the response set into a metric space and applies persistent homology to characterize its organization. An integrated stability index, defined via the persistence of homological features, together with partial descriptors of connectivity and higher-order structure, enables quantitative assessment of regime-dependent behavior and sensitivity to perturbation types..

* The Ninth International Workshop on Computer Modeling and Intelligent Systems (CMIS-2026), May 5, 2026, Zaporizhzhia, Ukraine

1,* Corresponding author.

† These authors contributed equally.

denys.symonov@gmail.com (D. Symonov); palagin_a@ukr.net (O. Palagin); e.symonov@gmail.com (Y. Symonov)

0000-0002-6648-4736 (D. Symonov); 0000-0003-3223-1391 (O. Palagin); 0009-0008-2581-2001 (Y. Symonov)

2. Problem statement

Modern large language models demonstrate high quality of text generation, however, their behavior remains sensitive to variations in query formulation. Minor changes in the structure or semantics of the prompt can lead to significantly different responses, which complicates the assessment of the reliability and reproducibility of results. Existing approaches are mostly based on local similarity metrics or heuristic estimates that do not account for the global organization of the response space. The lack of generalized invariant characteristics limits the possibilities of systematic analysis of the stability of models. In this regard, the task of building a formalized approach that allows describing the structure of the response set as a geometrically and topologically organized object, as well as determining quantitative indicators of its integrity and sensitivity to query variations, arises.

3. Goal and objectives of the study

The goal of the study is to build a formalized model for quantitative analysis of the stability of the response space of a large language model to semantic and structural variations of the query based on geometric and topological invariants of embedding representations. It is intended to establish a connection between the topological organization of the response set and the model generation modes.

To achieve this goal, the following tasks are solved: to formalize the response space as a metric structure; to introduce a consistent system of geometric characteristics; to apply persistent homology to highlight stable topological features; to construct an integral indicator of topological stability and a system of partial indices; to formulate a typology of the model behavior modes; to carry out experimental verification and to assess the sensitivity of the proposed indicators to different types of query variations.

4. Formalization of the problem and basic definitions

Consider the space of text queries Q and the space of answers Y. An element $q \in Q$ is interpreted as a sequence of tokens from some dictionary Σ, i.e. [1]

$$q = (w_1, \ldots, w_n), w_i \in \Sigma. \quad (1)$$

The language model is considered as an operator for generating an answer to a query given by the mapping $f: Q \to Y$ where $f(q)$ is the text answer generated by the model for the query q [2]. Thus, the operator f associates an element of the answer space Y with each query $q \in Q$.

Each answer $y \in Y$ is considered as an object of the semantic space for which a geometric representation is allowed. For this purpose, a mapping

$$\varphi: Y \to R^d, \quad (2)$$

is introduced that maps the text of the answer into a vector of embeddings of fixed dimension.

The further analysis concerns not a single answer $f(q)$, but a set of answers that arise when varying the formulation of the same query. Let $q_0 \in Q$ be the basic query. Then the set of variations $Q(q_0) \subset Q$ is considered, which contains queries that correspond to the same task as q_0, but differ in formulation or structure [3].

Accordingly, the set

$$Y(q_0) = \{f(q): q \in Q(q_0)\}, \quad (3)$$

forms a family of model responses, the topological structure of which is the object of further analysis.

In the following discussion, two types of query transformations are distinguished [4]. Semantic transformations change the linguistic form of the query without modifying its content; they include reformulation, lexical substitution, synonymous reconstruction, and other syntactic transformations. Structural transformations change the compositional organization of the query, in particular, the rearrangement of instructions, the decomposition of complex conditions, and the addition or removal of service formulations or formatting elements [5].

Let $T_{sem} \subset \{T : Q \to Q\}$ be the set of semantic transformations of the query, i.e., such mappings T that change the linguistic form of the query, but not its content, and $T_{str} \subset \{T : Q \to Q\}$ let be the set of structural transformations that change the compositional organization of the query without changing its functional task. Then the set of variations of the basic query can be represented as

$$Q(q_0) = \{\, T(q_0) \mid T \in T_{sem} \cup T_{str} \,\}. \tag{4}$$

Therefore, each element $q \in Q(q_0)$ is obtained by applying an admissible transformation T to the base query q_0.

Let the set of query variations $Q(q_0)$ be given an equivalence relation $q_i \sim q_j$, which means that queries q_i and q_j implement the same target problem and differ only in admissible semantic or structural modifications. In such a case, it is natural to require that the model responses be invariant with respect to such transformations [6].

Formally, the model is considered stable on the set $Q(q_0)$, if for equivalent queries $q_i \sim q_j$ the condition is satisfied:

$$d_Y\big(f(q_i), f(q_j)\big) \le \varepsilon, \tag{5}$$

where d_Y is the distance in the response space Y; $\varepsilon > 0$ is the admissible deviation threshold.

Since the definition of the natural metric in the space of text responses Y is a non-trivial problem, in the following we use the distance induced in the representation space:

$$d_Y(y_i, y_j) := \| \varphi(y_i) - \varphi(y_j) \|, \tag{6}$$

where $\varphi : Y \to R^d$ is the mapping of response embeddings [7].

Condition (6) specifies a local interpretation of stability, i.e. small changes in the formulation of the query should not cause a significant shift in the response in the semantic space. However, even with pairwise proximity of the responses, the set $Y(q_0)$ may have a complex global structure. In particular, variations in the query may lead to the splitting of the response set into several components of connectivity, the appearance of cavities or fragmentation of the structure. Therefore, the paper considers the concept of stability, which refers to the preservation of the topological organization of the response set. The key object of analysis is the topological structure of the set $\varphi(Y(q_0)) \subset R^d$ that is formed under the action of semantic and structural variations of the query. Therefore, the research task is to establish whether a large language model preserves the topological structure of the set $Y(q_0)$ under small perturbations of the query, or whether such perturbations cause its restructuring. To analyze this property, the persistent homology apparatus is used further, which allows detecting and quantitatively describing the topological stability of the response space.

5. Geometric representation of the LLM response space

To proceed to the geometric analysis of the response space, the embedding mapping $\varphi : Y \to R^d$ introduced in (2) is used. Each response $y_i \in Y$ refers to a representation vector [8]:

$$z_i = \varphi(y_i) \in R^d. \tag{7}$$

For a set of responses $Y(q_0)$, the set of their vector representations is considered:

$$Z = \{ z_i = \varphi(y_i) \mid y_i \in Y(q_0) \} \subset R^d. \tag{8}$$

The set Z (8) is interpreted as a finite set of points in the space R^d, representing the model responses generated by semantic and structural variations of the basic query q_0. The geometric configuration of this set reflects the nature of the dispersion of responses, possible clustering, and structural breaks in the response space [9].

For a quantitative analysis of the proximity of responses on the set Z, we introduce the metric $d: Z \times Z \to R_+$. For example, for in practical tasks for embedding-representations, Euclidean distance

$$d_E(z_i, z_j) = \| z_i - z_j \|_2, \tag{9}$$

and cosine distance

$$d_{\cos}(z_i, z_j) = 1 - \frac{\langle z_i, z_j \rangle}{\| z_i \| \ \| z_j \|}. \tag{10}$$

In the following, it is assumed that the selected metric is consistent with the geometry of the embedding space and reflects the semantic proximity of the responses [10]. Under this assumption, the pair (Z, d) is considered as a metric space. On this space local neighborhoods, neighborhood graphs, and topological complexes are further constructed, which are used for topological analysis of the response set by the persistent homology method.

For the set of vector representations Z (8), it is advisable to introduce basic geometric characteristics that describe the configuration of points in the representation space. These quantities allow us to quantitatively assess the dispersion of responses, their local grouping, and the differences between responses generated by different types of query variations.

The spatial dispersion of the set Z can be determined using the average pairwise distance:

$$\acute{d}(Z) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} d(z_i, z_j), \tag{11}$$

where d is the metric introduced in (9) - (10).

To describe the deviation of points from their average position, the center of the cloud is introduced:

$$\acute{z} = \frac{1}{N} \sum_{i=1}^{N} z_i, \tag{12}$$

and the corresponding dispersion

$$V(Z) = \frac{1}{N} \sum_{i=1}^{N} \| z_i - \acute{z} \|^2. \tag{13}$$

The quantity (13) characterizes the degree of concentration of points in the representation space.

Local grouping of points is characterized by the intracluster compactness index. Let $Z_k \subset Z$ be the subset of points that forms a local cluster, and c_k is its center. Then the compactness of the cluster is defined as:

$$C(Z_k)=\frac{1}{|Z_k|}\sum_{z\in Z_k}\|z-c_k\|^2, \tag{14}$$

where $c_k=\frac{1}{|Z_k|}\sum_{z\in Z_k} z$ [11].

To compare the responses generated by different types of query variations, the distance between the corresponding subsets is considered. Let $Z_{\text{sem}}\subset Z$, $Z_{\text{str}}\subset Z$ be the subsets of points corresponding to semantic and structural transformations. Then the intergroup distance can be defined as the distance between their centers

$$D_{\text{inter}}(Z_{\text{sem}},Z_{\text{str}})=\|\acute{z}_{\text{sem}}-\acute{z}_{\text{str}}\|, \tag{15}$$

where $\acute{z}_{\text{sem}}=\frac{1}{|Z_{\text{sem}}|}\sum_{z\in Z_{\text{sem}}} z$, $\acute{z}_{\text{str}}=\frac{1}{|Z_{\text{str}}|}\sum_{z\in Z_{\text{str}}} z$.

The introduced characteristics (11)-(15) specify the basic geometric description of the configuration of the points of the set Z. They are used as auxiliary indicators of the structure of the response space before proceeding to further topological analysis.

6. Topological analysis of the response space

On the metric space (Z,d), introduced in section 2, a filtering of Vietoris-Rips complexes is constructed, parameterized by the scale parameter $\varepsilon>0$ [12]. For each value ε, a complex $VR_\varepsilon(Z)$ with a set of vertices Z is defined. For an arbitrary subset $\{z_{i_0},z_{i_1},\ldots,z_{i_k}\}\subset Z$, a k-simplex belongs $VR_\varepsilon(Z)$ if and only if the condition is satisfied:

$$d(z_{i_r},z_{i_s})\le\varepsilon, 0\le r<s\le k. \tag{16}$$

Therefore, the complex $VR_\varepsilon(Z)$ encodes the proximity relation between the points of the set Z on the scale ε. Since for $\varepsilon_1\le\varepsilon_2$ holds the inclusion

$$VR_{\varepsilon_1}(Z)\subseteq VR_{\varepsilon_2}(Z), \tag{17}$$

then the family $\{VR_\varepsilon(Z)\}_{\varepsilon>0}$ forms filtration [13]. Accordingly, the multiscale construction (16)-(17) is used to analyze the topological structure of the set Z.

For each complex $VR_\varepsilon(Z)$, topological invariants are introduced:

$$\beta_k(\varepsilon)=\text{rank } H_k(VR_\varepsilon(Z)), k\ge 0, \tag{18}$$

where H_k is the variable denoting the k-th homology group of the corresponding complex [14].

In the following analysis, the main attention is paid to the invariants β_0 and β_1. The number $\beta_0(\varepsilon)$ determines the number of connectivity components and characterizes the degree of fragmentation of the response space. The number $\beta_1(\varepsilon)$ corresponds to the number of one-dimensional cycles and reflects the presence of cavities or closed structures in the configuration of points. If necessary, the invariant $\beta_2(\varepsilon)$, which characterizes higher-order cavities, can also be considered [15, 16].

However, the Betti numbers (β_0,β_1), calculated for a single value of ε, give only a static description of the structure. For multiscale analysis, persistent homology is used, which tracks the emergence and disappearance of topological features along the filtration, i.e.:

$$VR_{\varepsilon_1}(Z)\subseteq VR_{\varepsilon_2}(Z)\subseteq\cdots\subseteq VR_{\varepsilon_m}(Z), \varepsilon_1<\varepsilon_2<\cdots<\varepsilon_m. \tag{19}$$

In this scheme, each topological feature of dimension k is characterized by the moment of birth b and the moment of disappearance δ, which determine the interval of its existence $[b, \delta)$.

The set of such intervals forms a barcode-representation. Each homologous class corresponds to a horizontal segment of length

$$l = \delta - b, \tag{20}$$

which is called the persistence of the corresponding topological feature.

An equivalent way of representation is a persistence diagram - a set of points

$$D_k = \{(b_i, \delta_i)\}_{i=1}^{m_k}, \tag{21}$$

where each point corresponds to one k-dimensional topological feature. The distance of a point from the diagonal $\delta = b$ characterizes the duration of existence of the corresponding homologous class [17].

Thus, barcode and persistence diagrams allow separating persistent topological structures from short-lived artifacts. Features with high persistence are interpreted as invariant features of the geometry of the response space, while short intervals are usually associated with local fluctuations or noise effects of the embedding-representation. Therefore, persistent homology provides a multiscale description of the topological structure of a set Z and allows us to detect its persistent geometric features induced by query variations.

The topological characteristics obtained from complex $VR_\varepsilon(Z)$ filtering can be interpreted in terms of the behavior of a large language model when the query formulation changes. If the β_0-persistence contains one persistent connectivity component, this means that most of the responses generated by changes in the basic query q_0 belong to a single topologically connected space, which indicates the invariance of the model to local changes in the formulation. Instead, the appearance of several persistent connectivity components indicates a splitting of the response space and may mean that similar query formulations activate different generation modes. Topological features with low persistence reflect local fluctuations in the space of embedding representations and usually have no structural significance. The appearance of one-dimensional cycles, fixed through β_1-persistence, indicates a more complex organization of the response space and can be interpreted as a manifestation of several similar, but not identical, generation modes. Thus, topological analysis allows us to describe the structure of the response space and serves as the basis for the further introduction of integral indicators of persistence [18].

7. Stability indicators and interpretation of topological regimes

To generalize the results of persistent analysis, it is advisable to introduce an integral indicator of the topological stability of the response space. Let for $k = \{0,1\}$ the set of persistence intervals have the form:

$$I_k = \{[b_r^{(k)}, \delta_r^{(k)})\}_{r=1}^{m_k}, l_r^{(k)} = \delta_r^{(k)} - b_r^{(k)}, \tag{22}$$

where $l_r^{(k)}$ is the duration of the existence of the corresponding topological feature.

The sets of lengths

$$L_0 = \{l_r^{(0)}\}_{r=1}^{m_0}, L_1 = \{l_r^{(1)}\}_{r=1}^{m_1}, \tag{23}$$

characterize the stability of the connectivity components and one-dimensional cycles in the filtering of complexes $VR_\varepsilon(Z)$. The quantities (23) allow distinguishing structurally significant topological features from unstable local effects caused by fluctuations in embedding representations.

The integral indicator of topological stability is defined as a functional

$$S_{top} = F(\beta_0, \beta_1, L_0, L_1), \tag{24}$$

which aggregates the main characteristics of persistent homology and reflects the degree of integrity of the response space induced by variations of the basic query q_0 [19]. Unlike individual Betti (β_0, β_1) numbers or individual persistence intervals, the quantity S_{top} gives a generalized quantitative assessment of the topological stability of the model.

The indicator is constructed so that it acquires large values in the presence of one dominant long-lived component of connectivity and decreases with increasing topological disorganization of the response space, in particular due to its fragmentation, the appearance of numerous short-lived topological features, or increased cyclic complexity.

To quantitatively describe fragmentation, we order the β_0-intervals in decreasing order of their lengths

$$l_{(1)}^{(0)} \geq l_{(2)}^{(0)} \geq \cdots \geq l_{(m_0)}^{(0)}. \tag{25}$$

Then the relative mass of non-dominant components can be given by:

$$P_{frag} = \frac{\sum_{r=2}^{m_0} l_{(r)}^{(0)}}{\sum_{r=1}^{m_0} l_{(r)}^{(0)} + \eta}, \eta > 0. \tag{26}$$

If the response space is topologically complete, one component of connectivity dominates and $P_{frag} \to 0$. In the case of the appearance of several stable components, this value increases, reflecting the splitting of the response space.

Topological features with low persistence reflect local instabilities. For a given threshold $\tau > 0$, we introduce the fraction of such features:

$$P_{noise} = \frac{1}{m_0 + m_1}\left(\sum_{r=1}^{m_0} 1_{\{l_r^{(0)} < \tau\}} + \sum_{s=1}^{m_1} 1_{\{l_s^{(1)} < \tau\}}\right). \tag{27}$$

The increase in the value (27) indicates that the topological structure of the set Z contains a significant number of unstable elements sensitive to minor changes in the query formulation.

The complexity of the topological organization of the response space is primarily associated with the appearance of one-dimensional cycles, which can be estimated through the normalized total β_1-persistence [20]:

$$P_{comp} = \frac{\sum_{s=1}^{m_1} l_s^{(1)}}{\sum_{r=1}^{m_0} l_r^{(0)} + \eta}. \tag{28}$$

The increase in the value (28) means that the configuration of points in the representation space acquires an increasingly complex, heterogeneous structure.

Taking into account the characteristics (26)-(28), the integral indicator can be given as:

$$S_{top} = \exp(-\lambda_1 P_{frag} - \lambda_2 P_{noise} - \lambda_3 P_{comp}), \lambda_1, \lambda_2, \lambda_3 > 0, S_{top} \in (0, 1]. \tag{29}$$

If $S_{top} \approx 1$, then the response space is topologically coherent; a decrease S_{top} indicates an increase in fragmentation, local instability, or structural multimodality. Thus, S_{top} is interpreted as an integral assessment of the model's stability to query variations.

It is advisable to detail the integral indicator S_{top} through a system of partial indices that reflect various aspects of the topological organization of the response space. Within the framework of this work, three indices are considered: component stability, topological complexity and sensitivity to query variations.

Let for $k \in \{0,1\}$ be given a set of persistent intervals (22) with corresponding interval lengths $l_r^{(k)}$. To separate structurally significant features, a threshold $\tau > 0$ is introduced; features with $l_r^{(k)} \geq \tau$ are considered as stable. The component stability index characterizes the degree of topological integrity of the response space. For this, we introduce the number of stable connectivity components:

$$n_0^{(\tau)} = \sum_{r=1}^{m_0} 1_{\left\{l_r^{(0)} \geq \tau\right\}}. \tag{30}$$

Then the index is defined as

$$I_{comp} = \frac{1}{n_0^{(\tau)}}, \tag{31}$$

where values close to unity correspond to the presence of one dominant component, and a decrease in the indicator reflects the fragmentation of the response space.

The index of topological complexity refers to the role of one-dimensional cycles and is determined through the normalized total β_1-persistence:

$$I_{compl} = \frac{\sum_{r=1}^{m_1} l_r^{(1)}}{\sum_{s=1}^{m_0} l_s^{(0)} + \eta}, \eta > 0. \tag{32}$$

The growth of the index (32) indicates the complexity of the topological structure and the presence of several competing generation modes. The sensitivity index reflects the difference between the topological structures generated by semantic and structural variations of the query [21].

Let D_k^{sem} and D_k^{str}, $k \in \{0,1\}$, then:

$$I_{sens} = \alpha \; d_B\left(D_0^{sem}, D_0^{str}\right) + (1-\alpha) \; d_B\left(D_1^{sem}, D_1^{str}\right), \alpha \in [0,1]. \tag{33}$$

Small values I_{sens} correspond to topologically similar configurations, while the growth of the index indicates an increased sensitivity of the model to the type of variations of the query.

Thus, the index I_{comp} characterizes the integrity of the response space, I_{compl} characterizes its topological complexity and I_{sens} characterizes the differential response of the model to different types of perturbations. Together, these indices give a structured description of the stability of the LLM and can be used both separately and as components of the integral indicator S_{top}.

The introduced indicators S_{top}, I_{comp}, I_{compl}, I_{sens} allow us to move from describing individual topological characteristics to classifying the behavior modes of the model in the response space. The classification is based on three agreed aspects: the integrity of the response space, the level of its topological complexity, and the sensitivity to the type of query variations.

Let the threshold values be given

$$0<\theta_c^{(2)}<\theta_c^{(1)}\le 1, 0<\theta_l^{(1)}<\theta_l^{(2)}, 0<\theta_s^{(1)}<\theta_s^{(2)}, \tag{34}$$

where $\theta_c^{(1)}, \theta_c^{(2)}$ are threshold levels of the component stability index I_{comp}, separating the regions of high, intermediate and low integrity of the response space; $\theta_l^{(1)}, \theta_l^{(2)}$ are threshold values of the topological complexity index I_{compl}, which determine the transition from simple to complex topological organization; $\theta_s^{(1)}, \theta_s^{(2)}$ are thresholds of the sensitivity index I_{sens}, which distinguish weak, moderate and high dependence of the topology on the type of query variations.

Then three typical regimes are distinguished.

1. Stable regime, characterized by the conditions:

$$I_{comp}\ge\theta_c^{(1)}, I_{compl}\le\theta_l^{(1)}, I_{sens}\le\theta_s^{(1)}, \tag{35}$$

and corresponds to high values of S_{top}.

2. Moderately sensitive regime, characterized by the conditions:

$$\theta_c^{(2)}\le I_{comp}<\theta_c^{(1)}, \theta_l^{(1)}<I_{compl}\le\theta_l^{(2)}, \theta_s^{(1)}<I_{sens}\le\theta_s^{(2)}, \tag{36}$$

and which is characterized by the appearance of additional structures, limited fragmentation and moderate difference between topological representations.

3. Topologically unstable regime, characterized by the conditions:

$$I_{comp}<\theta_c^{(2)}, I_{compl}>\theta_l^{(2)}, I_{sens}>\theta_s^{(2)}, \tag{37}$$

and corresponds to low values of S_{top}.

Thus, the proposed typology interprets topological characteristics as operational features of LLM behavior: a stable mode corresponds to invariance, moderately sensitive to limited dependence, and topologically unstable to structural splitting of the response space.

8. Description of the experiments

The experiment is aimed at assessing the stability of the geometric and topological structure of the response space of the language model under different types of query variations. Three categories of queries are considered: informational, logical, and instructional. The study included three stages: (1) comparison of deterministic and stochastic generation modes; (2) analysis of semantic query variations obtained by paraphrasing and lexical-syntactic changes; (3) analysis of structural variations associated with changing the order of instructions and the formulation of conditions. For each query variant, model responses were generated and converted into embedding vectors. Based on them, pairwise cosine distances were calculated, Vietoris–Rips complexes were constructed, and topological characteristics were determined, in particular the number of stable components H_0 and the topological stability indicator S_{top}. The obtained indicators were used to compare the impact of different types of query variations on the structure of the model's response space.

9. Results of the experiments

This section presents the results of experimental verification of the proposed approach. To analyze the model response space, two generation modes are compared. The first is a deterministic mode without random sampling (temperature ≈ 0), in which repeated model responses for the same query should be as similar as possible. The second is a stochastic mode using temperature sampling, which allows for variability in formulations [22]. Such a comparison allows us to assess

how the geometric and topological structure of the response space changes depending on the level of stochasticity of generation.

To assess the basic geometry of the model response space, aggregated matrices of pairwise cosine distances between the embedding representations of responses generated in the deterministic and stochastic modes were constructed. The results are shown in 1.

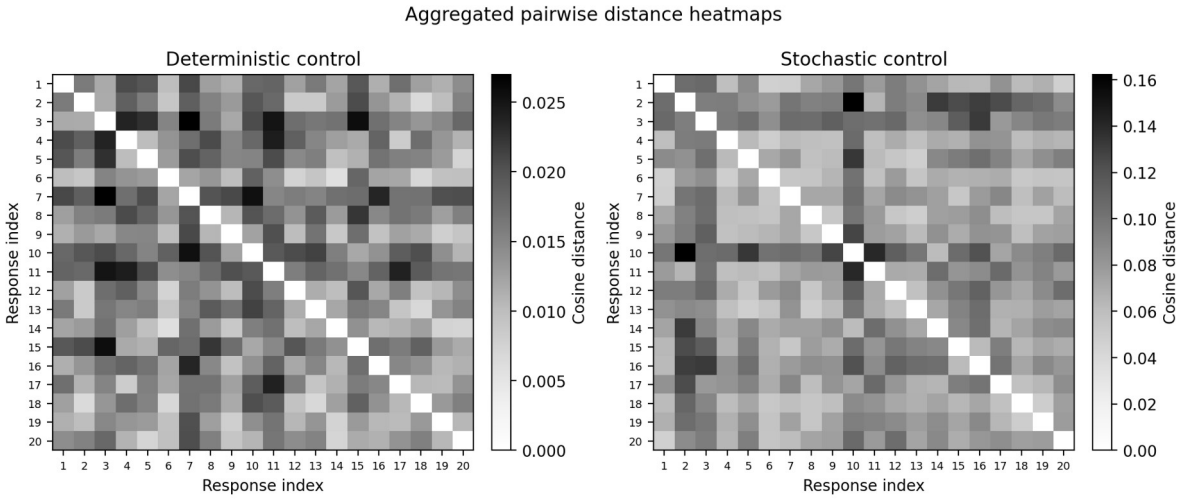

Figure 1: Pairwise embedding distances of LLM responses under different generation regimes

1 displays the aggregated matrices of pairwise cosine distances between the embedding representations of the model responses for the two generation modes. In the deterministic mode, most of the distances are in the interval of approximately $0.005 - 0.03$, which indicates the compactness of the response space and high consistency of repeated generations of the same query. In the stochastic mode, the spread is much larger: the distance values reach approximately $0.04 - 0.16$, which reflects the increase in the variability of text formulations. At the same time, there are no clear block structures on the heat maps that could indicate the formation of separate response clusters. Therefore, even under stochastic generation, the response space remains globally connected, which is consistent with the results of further topological analysis.

In order to quantitatively compare the response space in different generation modes, generalized metrics were calculated that describe their geometric dispersion and topological stability. The results are shown in 2.

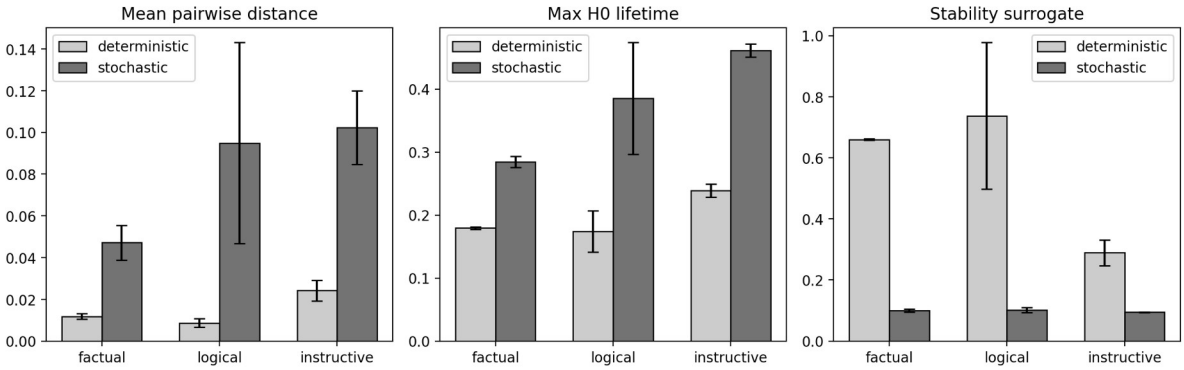

Figure 2: Response Space Geometry and Topology Across Generation Regimes

2 presents a comparison of three quantitative indicators characterizing the geometry and topological properties of the model response space: the average pairwise cosine distance between

embedding vectors, the maximum lifetime of the connectivity component H_0 in persistent homology, and the stability indicator. In the deterministic mode, the average pairwise distance remains small and lies within approximately $0.009 - 0.024$ which corresponds to the compact placement of responses in the space of embedding representations. In the stochastic mode, the values of this metric increase noticeably and reach approximately $0.047 - 0.102$, which reflects the greater variability of the formulations. A similar trend is observed for the component H_0 lifetime: in the stochastic mode, it is about $0.28 - 0.46$, while in the deterministic mode it is in the interval $0.17 - 0.24$. At the same time, the stability indicator is significantly higher for the deterministic mode (approximately $0.65 - 0.74$) and decreases to $0.09 - 0.10$ in the stochastic mode, which indicates a greater geometric dispersion of the response space in the presence of stochastic sampling.

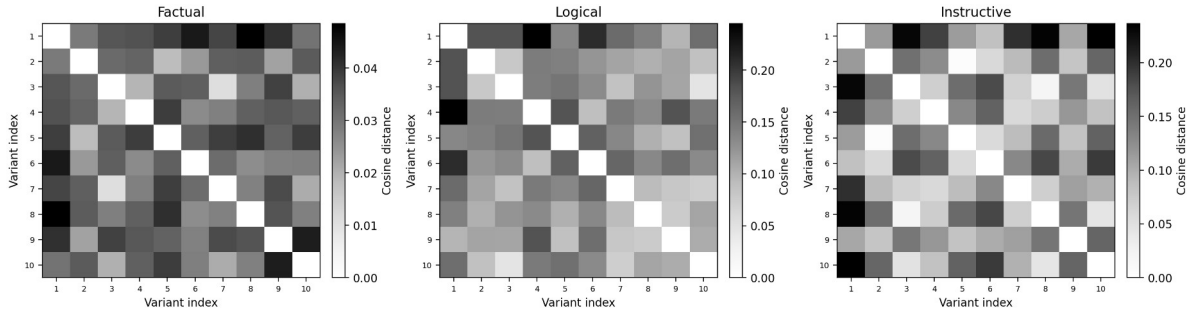

Figure 3: Pairwise embedding distances of LLM responses under semantic prompt variations

3 shows the aggregated matrices of pairwise cosine distances between the embedding vectors of responses for the three categories of queries. For actual queries, most of the distances remain low (mostly $d \approx 0.01 - 0.05$), which indicates a compact geometry of the response space and high stability of the result under semantic variations of the formulation. For logical and instructive queries, a larger spread of values (up to $d \approx 0.20 - 0.24$), which indicates an increased sensitivity of the model to changes in the query formulation. At the same time, the matrices do not demonstrate clearly separated block structures, which means the absence of stable subclusters of responses and the preservation of the overall connectivity of the response space even in the presence of semantic perturbations of the query.

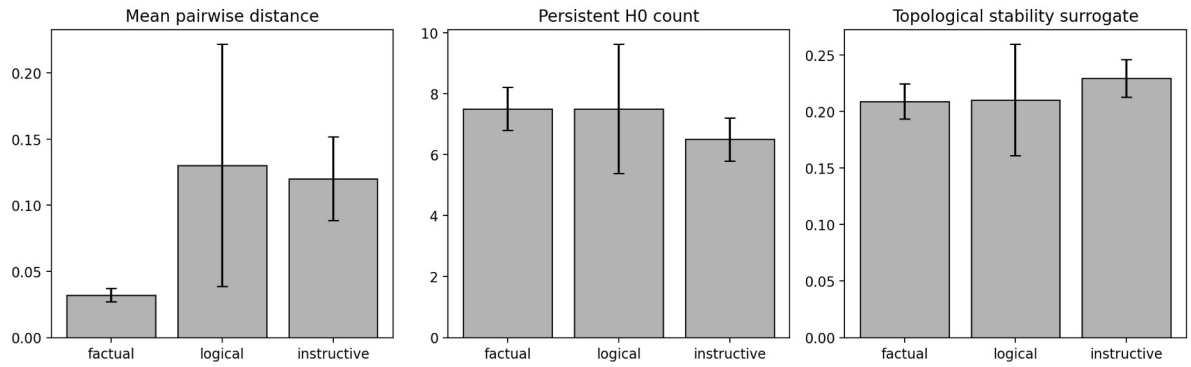

Figure 4: Response Space Geometry and Topology Across Semantic Prompt Variations

4 shows a comparison of three indicators characterizing the geometric and topological properties of the model's response space under semantic variations of the query. The smallest value of the average pairwise cosine distance is observed for factual queries and is about 0.03, while for logical and instructive queries it increases to approximately $0.12 - 0.13$. This indicates that the

responses to factual queries remain more compact in the space of embedding representations, while logical and instructive formulations exhibit higher sensitivity to semantic paraphrasing. The number of stable connectivity components H_0 in all three categories remains close, within approximately $6.5-7.5$, which does not give grounds to speak of a sharp fragmentation of the response space. At the same time, the values of the integral indicator of topological stability change slightly and lie within $0.21-0.23$. This gives grounds to believe that semantic variations of the query noticeably affect the geometric compactness of the responses, but do not cause a significant violation of their global topological integrity.

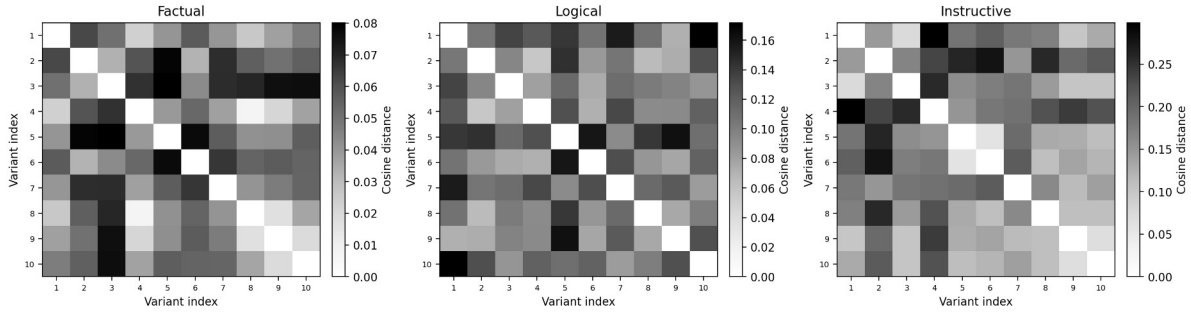

Figure 5: Pairwise embedding distances of LLM responses under structural prompt variations

As shown in 5, structural variations in the prompt lead to a significant change in the geometry of the response space. For information retrieval prompts, the values of pairwise cosine distances mostly remain relatively small and mostly do not exceed approximately $0.07-0.08$, which corresponds to a more compact configuration of the embedding space. For logical prompts, the distances increase and reach approximately $0.15-0.17$, while for instruction-following prompts they can exceed 0.25, which indicates a much greater dispersion of responses. This difference in values means that structural modifications of the prompt can significantly change the geometric configuration of the model's response space, with the effect being most pronounced for more complex cognitive types of prompts.

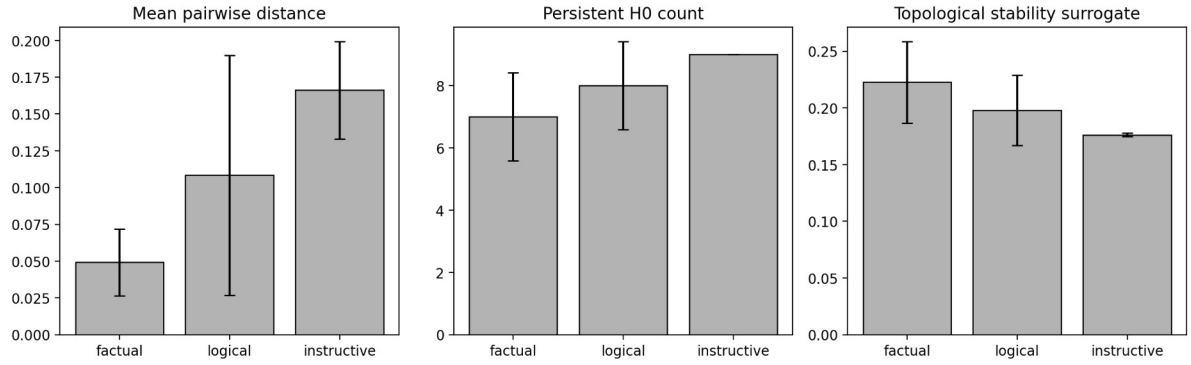

Figure 6: Geometric and topological metrics of the response space under structural prompt variations

As can be seen from 6, structural variations in the query lead to a gradual increase in the geometric dispersion of the response space. The average pairwise cosine distance increases from approximately 0.05 for information-seeking prompts to approximately 0.11 for reasoning prompts and approximately $0.16-0.17$ for instruction-following prompts. Similarly, the number of stable components H_0 increases: from approximately 7 to approximately 9, which indicates a gradual fragmentation of the embedding representation space. At the same time, the integral indicator of

topological stability S_{top} shows the opposite trend and decreases from approximately 0.22 to 0.17. The combination of these indicators shows that structural changes in the query formulation noticeably affect the topological configuration of the model's response space, and the effect is most pronounced for instructional tasks.

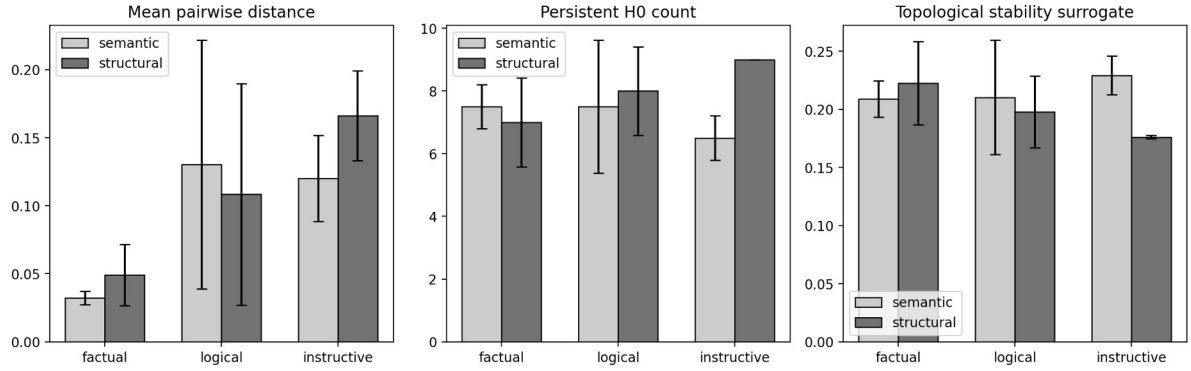

Figure 7: Response-space sensitivity to semantic and structural prompt perturbations

As can be seen from 7, structural variations in the query formulation on average cause a greater geometric dispersion of the response space than semantic paraphrasing. In particular, for instruction-following prompts, the average pairwise cosine distance increases from approximately 0.12 in the case of semantic changes to approximately $0.16-0.17$ for structural modifications. A similar trend is demonstrated by the number of stable components H_0, which increases from approximately 6.5 to 9, which indicates an increase in the fragmentation of the embedding space. At the same time, the values of the integral indicator of topological stability S_{top} for structural variations decrease (from approximately 0.23 to $0.17-0.18$), which is consistent with the assumption of a greater sensitivity of the model to changes in the instruction structure than to semantic paraphrasing.

Table 1
Response-Space Geometric and Topological Statistics Across Categories and Regimes

Prompt category	Regime	$d_{\cos}$	$sd(d_{\cos})$	H_0	$sd(H_0)$	S_{top}	$sd(S_{top})$	ΔS_{top}
factual	control	0.01	0.001	1.0	0.000	0.66	0.002	0.0
factual	semantic	0.03	0.004	7.5	0.707	0.21	0.015	-0.45
factual	structural	0.05	0.022	7.0	1.414	0.22	0.035	-0.43
logical	control	0.01	0.002	1.0	1.414	0.74	0.339	0.0
logical	semantic	0.13	0.091	7.5	2.121	0.21	0.049	-0.52
logical	structural	0.11	0.081	8.0	1.414	0.20	0.031	-0.53
instructive	control	0.02	0.006	5.0	1.414	0.29	0.059	0.0
instructive	semantic	0.12	0.031	6.5	0.707	0.23	0.016	-0.05
instructive	structural	0.17	0.033	9.0	0.000	0.18	0.001	-0.11

Note. $d_{\cos}$ – mean cosine distance between response embeddings; H_0– number of persistent connected components; S_{top} – topological stability indicator; ΔS_{top} – change relative to control.

1 contains the generalized values of the geometric and topological characteristics of the model's response space for three query categories and three perturbation regimes. In the control regime, the response space is characterized by low geometric scatter ($d_{\cos} = 0.008 - 0.024$) and minimal fragmentation ($H_0 = 1$ for factual and logical queries). In the case of semantic variations, the mean cosine distances increase to $0.032 - 0.130$, and the number of stable components increases to $H_0 \approx 6.5 - 7.5$, which indicates the appearance of additional structures in the response space. Structural variations lead to a further increase in geometric scatter (to $d_{\cos} = 0.166$) and the number of stable components ($H_0 \to 9$). At the same time, the value of the topological stability indicator S_{top} decreases relative to the control regime ($\Delta S_{top} < 0$), with the largest deviations observed for logical queries. The results obtained are consistent with the assumption of a greater sensitivity of the response space topology to structural changes in the query formulation compared to semantic variations.

10. Conclusions

The results establish practical guidelines for LLM development and deployment. Model design should account not only for local performance metrics but also for the global structural stability of the response space, which can be quantified using topological indicators. Semantic perturbations primarily affect geometric compactness, whereas structural perturbations induce fragmentation and regime shifts; thus, maintaining prompt structure is critical for stable generation. The proposed indicators provide a basis for reliability assessment and detection of unstable operating regimes.

The approach represents the response set in an embedding space, enabling its treatment as a metric structure where geometric characteristics capture dispersion and local clustering. Persistent homology provides a multi-scale description of topology and identifies stable invariants under perturbations.

An integrated topological stability index S_{top}, defined via the persistence of homological features, aggregates information on component structure, local instabilities, and cyclic organization. Complementary partial indices decompose this measure into interpretable components of model behavior.

Experimental results demonstrate the discriminative capacity of the proposed indicators. Deterministic generation yields compact and topologically coherent structures, whereas stochasticity increases dispersion and reduces S_{top}. Semantic perturbations primarily affect geometry without substantial topological change, while structural perturbations induce fragmentation, increase the number of persistent components, and degrade topological stability.

The resulting regime typology is consistent with empirical observations: stable regimes correspond to topologically coherent and invariant response structures, whereas unstable regimes are characterized by fragmentation and sensitivity to structural perturbations. The proposed framework thus provides a quantitative basis for analyzing LLM stability and identifying critical behavioral regimes..

Acknowledgements

The work was supported by the state budget research project "To develop evolutionary modeling methods for complex systems aimed at analyzing the dynamics of self-organized structures under conditions of uncertainty" (state registration number 0126U000358) of the V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences (NAS) of Ukraine.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to translate research notes and results from Ukrainian to English. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication

References

[1] X. Liu, Q. Song, Q. Zhou, H. Du, S. Xu, W. Jiang, W. Zhang, X. Jia, Focusing on Language: Revealing and Exploiting Language Attention Heads in Multilingual Large Language Models, Proc. AAAI Conf. Artif. Intell. 40.38 (2026) 32195–32203. doi:10.1609/aaai.v40i38.40492.

[2] V. Maniappan, R. Praghaadeesh, G. Bharathi Mohan, R. Prasanna Kumar, Small Language Models: An Advancing Efficient Open-Source Alternatives to Large Language Models, in: Lecture Notes in Networks and Systems, Springer Nature Singapore, Singapore, 2025, pp. 321–334. doi:10.1007/978-981-96-7505-0_25.

[3] L. Ngweta, K. Kate, J. Tsay, Y. Rizk, Towards LLMs Robustness to Changes in Prompt Format Styles, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 529–537. doi:10.18653/v1/2025.naacl-srw.51.

[4] J. Li, S. Papay, R. Klinger, Are Humans as Brittle as Large Language Models?, in: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 2130–2155. doi:10.18653/v1/2025.ijcnlp-long.116.

[5] S. Harvan, J. Kopčan, M. Šuppa, A. Findor, The Brittle Compass: Navigating LLM Prompt Sensitivity in Slovak Migration Media Discourse, in: The First Workshop on Advancing NLP for Low-Resource Languages, Incoma Ltd. Shoumen, BULGARIA, 2025, pp. 88–101. doi:10.26615/978-954-452-100-4-010.

[6] Y. Moros Daval, How Can Large Language Models Be More Reliable?, Proc. AAAI/ACM Conf. AI, Ethics, Soc. 8.3 (2025) 2911–2912. doi:10.1609/aies.v8i3.36790.

[7] H. Canot, P. Durand, E. Frenod, Anisotropic Shear Metrics for Persistent Homology and Their Application to Convective Systems, Int. J. Topol. 3.1 (2026) 6. doi:10.3390/ijt3010006.

[8] W. Zadrozny, Topological Data Analysis in Natural Language Processing -- A Tutorial, Int. FLAIRS Conf. Proc. 36 (2023). doi:10.32473/flairs.36.133337.

[9] D. Symonov, O. Palagin, B. Zaika, Dynamical Clustering via Neural Vector Fields with Attractor-Based Structure, in: Workshop “Intelligent information technologies" UkrProg-IIT`2025 co-located with 15th International Scientific and Practical Programming Conference UkrPROG’2025, 2025, pp. 189–201. URL: https://ceur-ws.org/Vol-4049/paper16.pdf.

[10] N. Rair, A. Goupil, V. Vrabie, E. Chochoy, When Annotators Disagree, Topology Explains: Mapper, a Topological Tool for Exploring Text Embedding Geometry and Ambiguity, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 8468–8491. doi:10.18653/v1/2025.emnlp-main.426.

[11] D. Symonov, O. Palagin, Y. Symonov, B. Zaika, Automated Design of Complex Systems Using Generative Models, in: Artificial Intelligence Technologies and Data Science 2025 (IT\&I-WS: AITDS 2025), Vol. 4158, CEUR-WS.org, 2025, pp. 38–50. URL: https://ceur-ws.org/Vol-4158/Paper04.pdf.

[12] S. Basu, N. Cox, Harmonic Persistent Homology, SIAM J. Appl. Algebra Geom. 8.1 (2024) 189–224. doi:10.1137/22m1518761.

[13] M. Herick, M. Joachim, J. Vahrenhold, Adaptive approximation of persistent homology, J. Appl. Comput. Topol. (2024). doi:10.1007/s41468-024-00192-7.

[14] Z. Su, X. Liu, L. B. Hamdan, V. Maroulas, J. Wu, G. Carlsson, G.-W. Wei, Topological data analysis and topological deep learning beyond persistent homology: a review, Artif. Intell. Rev. (2025). doi:10.1007/s10462-025-11462-w.

[15] Z. Zhang, Y. Sun, Y. Liu, L. Jiang, Z. Li, Persistent Homology Combined with Machine Learning for Social Network Activity Analysis, Entropy 27.1 (2024) 19. doi:10.3390/e27010019.

[16] M. D. S. Hopp, V. Labatut, A. Amalvy, R. Dufour, H. Stone, H. Jach, K. Murayama, Persistent Homology of Topic Networks for the Prediction of Reader Curiosity, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 28121–28132. doi:10.18653/v1/2025.acl-long.1364.

[17] Obayashi, Stable volumes for persistent homology, J. Appl. Comput. Topol. (2023). doi:10.1007/s41468-023-00119-8.

[18] B. M. Ruppik, M. Heck, C. van Niekerk, R. Vukovic, H.-c. Lin, S. Feng, M. Zibrowius, M. Gasic, Local Topology Measures of Contextual Language Model Latent Spaces with Applications to Dialogue Term Extraction, in: Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Stroudsburg, PA, USA, 2024, pp. 344–356. doi:10.18653/v1/2024.sigdial-1.31.

[19] U. Bauer, A. M. Medina-Mardones, M. Schmahl, Persistent homology for functionals, Commun. Contemp. Math. (2023). doi:10.1142/s0219199723500554.

[20] P. Sekuloski, D. Kitanovski, I. Goshev, K. Mishev, M. S. Misheva, V. D. Ristovska, Exploring the Potential of Topological Data Analysis for Explainable Large Language Models: A Scoping Review, Mathematics 14.2 (2026) 378. doi:10.3390/math14020378.

[21] R. Lavery, A. Jurek-Loughrey, L. Bai, Combining Topological Signature with Text Embeddings: Multi-Modal Approach to Fake News Detection, in: 2024 35th Irish Signals and Systems Conference (ISSC), IEEE, 2024. doi:10.1109/issc61953.2024.10603336.

[22] M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3, Proc. Natl. Acad. Sci. 120.6 (2023). doi:10.1073/pnas.2218523120.