Hallucination Detection in LLMs with Topological Divergence on Attention Graphs

Anonymous ACL submission

Abstract

Hallucination, i.e., generating factually incorrect content, remains a critical challenge for large language models (LLMs). We introduce TOHA[1], a TOpology-based HAllucination detector in the RAG setting, which leverages a topological divergence metric to quantify the structural properties of graphs induced by attention matrices. Examining the topological divergence between prompt and response subgraphs reveals consistent patterns: higher divergence values in specific attention heads correlate with hallucinated outputs, independent of the dataset. Extensive experiments — including evaluation on question answering and summarization tasks — show that our approach achieves state-of-the-art or competitive results on several benchmarks while requiring minimal annotated data and computational resources. Our findings suggest that analyzing the topological structure of attention matrices can serve as an efficient and robust indicator of factual reliability in LLMs.

1 Introduction

Large language models (LLMs) have progressed significantly in recent years, finding applications in various fields (Chkirbene et al., 2024). However, these models are prone to generate so-called *hallucinations*, i.e., content that is factually or contextually incorrect (Huang et al., 2023). Detecting hallucinations is crucial for safely deploying LLMs in sensitive fields since erroneous outputs may seriously harm user trust. An effective detector would therefore expand the scope of LLM applications while mitigating risks (Gao et al., 2024).

Multiple methods address this problem (Huang et al., 2023; Sahoo et al., 2024), though many face significant practical constraints. A common limitation is the reliance on large annotated datasets (Sky et al., 2024; Azaria and Mitchell, 2023; Chuang et al., 2024), which are rarely available publicly (Zhang et al., 2023) and require extensive annotation effort for each new model released. Another popular approach depends on generating multiple samples for scoring (Manakul et al., 2024; Chen et al., 2024; Farquhar et al., 2024), which increases computational costs substantially.

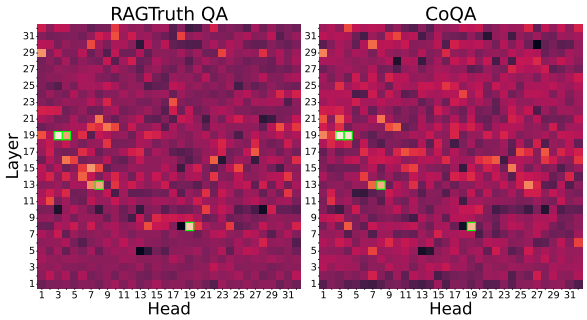

Figure 1: Difference between average topological divergence values for hallucinated and grounded samples per attention head/layer, evaluated on MS MARCO and CoQA datasets. A lighter color corresponds to a greater difference. Green frames highlight the heads that separate samples best. The same attention heads assign greater divergence values to the hallucinated samples in both datasets. Model: Mistral-7B-Instruct-v0.1.

We address these challenges by introducing TOHA (TOpology-based HAllucination detector), a training-free method for the retrieval-augmented generation (RAG) setting (Gao et al., 2023). Following the prior work (Du et al., 2024), TOHA requires minimal annotated data (just 50 annotated samples suffice for reliable detection, see Figure 4) while avoiding the computational overhead of multiple generations, making it both data- and compute-efficient.

The core insight behind TOHA is that hallucinated answers tend to have weaker connections to the context than grounded ones in the RAG setting. TOHA formalizes this by analyzing attention

[1]The code of the proposed method and the considered baselines is available at https://anonymous.4open.science/r/tda4hallu-BED5

graphs — complete graphs induced by LLM attention maps, a representation previously used in other NLP tasks (Kushnareva et al., 2021; Tulchinskii et al., 2023). Unlike prior works that rely on simplistic classifiers over basic graph properties (Proskurina et al., 2023a; Cherniavskii et al., 2022), TOHA advances the approach by computing the topological dissimilarity between the attention subgraphs of the model's response and the given context. This dissimilarity, measured via our adaptation of Manifold Topology Divergence (Barannikov et al., 2021) for the graph setting, quantifies the "strength" of context-response ties: higher values indicate weaker links and thus likely hallucinations. We prove key stability properties for this metric, ensuring its reliability as a hallucination score.

Through analysis of divergence patterns across different heads, we identified a subset of attention heads that consistently assign higher divergence scores to hallucinated samples (see Figure 1), revealing their implicit "awareness" of hallucinations. TOHA utilizes the average divergence values from these specific heads as hallucination scores. Crucially, these heads exhibit consistent behavior across different datasets, enabling strong transferability of our method.

Our main contributions can be summarized the following:

- We propose TOHA, a training-free method based on the topological divergences of attention graphs. While efficient — TOHA operates up to an order of magnitude faster than methods of comparable quality and requires minimal annotated data — our method demonstrates strong in-domain performance and maintains domain transferability across different tasks.

- The existence of hallucination-aware attention heads is discovered: calculating topological divergences from just six specific heads is enough for reliable hallucination detection, irrespective of the dataset.

- Our experiments show TOHA consistently matches or exceeds state-of-the-art performance on all benchmarks when applied to modern open-source LLMs of varying scales (7B to 13B parameters).

2 Background

2.1 Attention matrix as a weighted graph

Modern LLMs are mainly based on the self-attention mechanism, introduced in (Vaswani et al., 2017). Let $X \in \mathbb{R}^{n\times d}$ be a matrix consisting of d-dimensional representations of n tokens, W_Q, W_K, $W_V \in \mathbb{R}^{d\times d}$ be trainable projection matrices. Given a set of queries $Q = XW_Q \in \mathbb{R}^{n\times d}$, a set of keys $K = XW_K \in \mathbb{R}^{n\times d}$, and corresponding values $V = XW_V \in \mathbb{R}^{n\times d}$, the attention mechanism calculates a weighted sum of the values as follows:

$$\text{Attention}(Q, K, V) = W(Q, K)V, \quad (1)$$

where $W(Q, K)$ is an attention matrix

$$\text{W} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \quad (2)$$

and each entry w_{ij} in it captures how strongly token i attends to token j, $i \geq j$ for a decoder, with larger w_{ij} indicating closer relationship.

An attention matrix W can be reframed as a complete weighted graph G where tokens are vertices and weights w_{ij} represent the strength of connections between them. From the perspective of topological data analysis, however, it is more convenient to consider these weights as pseudo-distances rather than correlation measures. Hence, we reassign the edge weights of such a graph to equal $1 - w_{ij}$, creating what we call *attention graphs*.

In the generation process, the vertices of such graphs naturally partition into two distinct subsets: prompt tokens P and response tokens R generated by the model (see Figure 2b). This split allows us to analyze the topological relationships between input and output content.

2.2 Manifold Topology Divergence

One way to compare two data manifolds, $\mathcal{M}$ and $\mathcal{N}$, approximated by point clouds M and N is the $\text{MTop-Div}(M, N)$ topological measure (Barannikov et al., 2021). This divergence is based on the $\text{Cross-Barcode}(M, N)$, which is a set of intervals $\{(b_i, d_i)\}_{i=1}^n$ corresponding to "births" and "deaths" of independent topological features that distinguish the point cloud N from the union $M \cup N$. The further $\text{Cross-Barcode}(M, N)$ is from an empty set, the more the data manifold $\mathcal{N}$ differs from $\mathcal{M}$ in its topological structure. To measure the distance

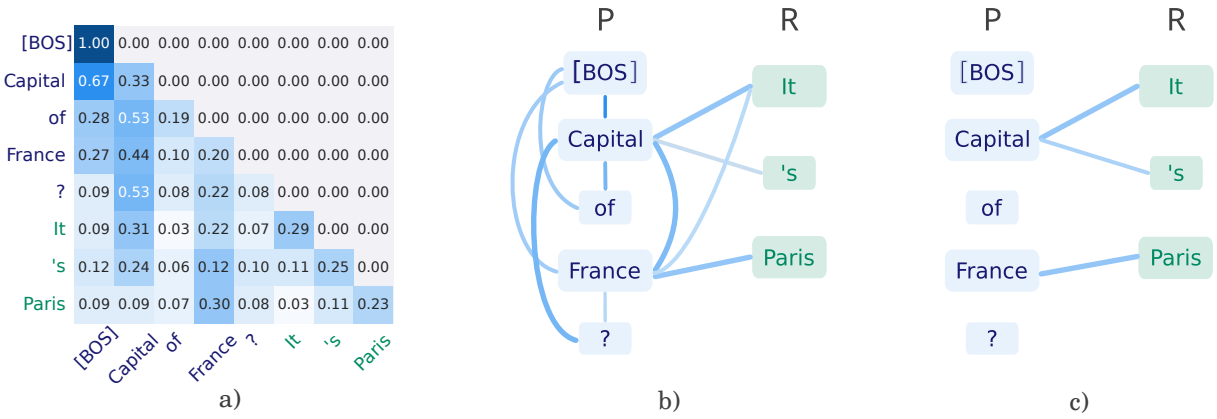

Figure 2: a) An attention map. Blue and green denotes the prompt and response tokens, respectively. b) The corresponding attention graph G. Prompt tokens P are located on the left, response tokens R — on the right. To keep figure neat, we only plot the edges with an attention score of no less than 0.15. c) The minimum spanning forest attaching R to P.

from an empty set, the sum of lengths of intervals in $\text{Cross-Barcode}(M, N)$ was taken in loc. cit.:

$$\text{MTop-Div}(M, N) = \sum_{i=1}^{n} |d_i - b_i|.$$

3 Method

Given an attention matrix for the (prompt + response) text, we construct the attention graph, imitating a data manifold of the text, and study its relation with the weighted subgraph, imitating the data submanifold of the prompt. We measure the topological divergence between these graphs, assuming responses that are consistent with the context would result in lower divergence values.

3.1 MTop-Div for attention graphs

The $\text{MTop-Divergence}(M, N)$ measure was originally developed for data manifolds, relying on metric space properties to some extent. We adapt this concept to quantify divergence between attention graphs and their subgraphs, where traditional metric axioms do not apply, while preserving the measure's comparative utility.

Let R and P be the sets of response and the prompt vertices in the attention graph G. We set to zero the edge weights between the P vertices, denote $w_{(R\cup P)/P}$ the resulting matrix of edge weights, and define $\text{Cross-Barcode}_i(R, P)$ as the i-th homology barcode of the Vietoris-Rips simplicial complex $VR_\alpha(G, w_{(R\cup P)/P})$. We define $\text{MTop-Div}(R, P)$ as the total sum of interval lengths in $\text{Cross-Barcode}_0(R, P)$, where we consider the 0-dimensional homology group H_0. In this context, we can prove that this score is equivalent to the total edge length of the minimum spanning forest (MSF) connecting R to P.

Properties of MTop-Div for attention graphs. Here, we only list the properties relevant to hallucination detection; for more properties and proofs, see Appendix A.

Proposition 3.1. *The following holds for any attention graph G with vertex set V_G and its complementary vertex subsets P, R, where $P \cup R = V_G$ and $P \cap R = \varnothing$.*
1. ***(Formula.)*** $\text{MTop-Div}(R, P)$ *value equals the length of the MSF attaching R to P.*
2. ***(Stability.)*** *If the weights of G change by no more than ε, then the corresponding $\text{MTop-Div}(R, P)$ changes by no more than $\delta = \varepsilon|R|$.*
3. ***(Connection with hallucinations.)*** *The normalized divergence value $\frac{1}{|R|}\text{MTop-Div}(R, P) = 0$ iff the MSF attaches every response token to a prompt token by a subtree with attention weights $= 1$.*

The stability property guarantees that similar attention patterns yield similar hallucination scores, making the metric's behavior consistent and predictable. The latter captures the intuitive relationship between divergence and response quality: well-grounded responses (closely tied to the context) produce small divergence values, while hallucinations (occurring when evidence is missing) lead to smaller attention weights and consequently larger divergence values. Together, these properties enable the metric to reliably measure how strongly a response connects to its context while remaining

robust to minor attention variations.

3.2 Hallucination-aware heads

We hypothesize, inspired by prior investigations in LLM interpretability (Voita et al., 2019; Gould et al., 2024), that particular attention heads exhibit distinct patterns related to hallucinations. To identify such heads, we analyzed head-specific topological divergences as follows.

Denote by h_{ij} the j-th attention head from the layer i. For the specific data sample s and head h_{ij}, let G^s_{ij} be the corresponding attention graph, P^s_{ij}, R^s_{ij} — its prompt and response vertex subsets.

We examined typical values of the average distance between hallucinated and grounded training examples for different heads and layers:

$$\Delta_{ij} = \frac{1}{|S_{\text{hallu}}|} \sum_{s \in S_{\text{hallu}}} d_{ij}(s) - \frac{1}{|S_{\text{gr}}|} \sum_{s \in S_{\text{gr}}} d_{ij}(s),$$

where S_{hallu} stands for all hallucinated samples from the training set, S_{gr} stands for all grounded training samples, and

$$d_{ij}(s) = \frac{1}{|R^s_{ij}|} \text{MTop-Div}(R^s_{ij}, P^s_{ij}).$$

The sample differences obtained for three datasets are displayed in Figure 3. Each dot represents an individual attention head, with its x-coordinate indicating its Δ_{ij} value on dataset (A) and its y-coordinate the corresponding value on dataset (B). For each dataset pair, we highlight the top three most separating heads for dataset (A) in pink. Notably, these heads consistently appear in the upper-right corner of the plot, indicating that they also exhibit strong separation on dataset (B). This observation suggests that these heads are inherently attuned to hallucination patterns, regardless of the dataset.

3.3 TOHA

The existence of universal hallucination patterns in the attention heads underlies our efficient method TOHA, detailed in Algorithm 1. It uses two small, annotated probe sets, S_h (containing hallucinated samples) and S_g (containing grounded samples), to rank model heads by their separation capability based on their Δ_{ij} values, where ij denotes a head index, and select the most relevant ones. In our experiments, the combined size of the probe sets is kept small (see Figure 4 for the number of the required samples analysis). During testing, hallucination scores are computed as the average topological divergence from the top N_{opt} heads, where N_{opt} is a hyperparameter tuned on the set $V = S_h \cup S_g$. For computational efficiency, we limit N_{opt} to a maximum of 6 in all experiments.

Algorithm 1 TOHA algorithm

Require:
$d_{ij}(s)$ — topological divergences for samples;
S_h, S_g — probe sets;
$V = S_h \cup S_g$ — validation set;
T — test set;
$N_{\max}$ — max number of selected heads.

procedure TOHA HEADS SELECTION
  $\Delta_{ij} \leftarrow \frac{1}{|S_h|} \sum_{s \in S_h} d_{ij}(s) - \frac{1}{|S_g|} \sum_{s \in S_g} d_{ij}(s)$
  $H \leftarrow \text{sort\_descending}(h_{ij}, \text{key} = \Delta_{ij})$
  $N, N_{\text{opt}} \leftarrow 1, 1$
  $H_{\text{subset}} \leftarrow \varnothing$ ▷ Optimal heads set.
  $\text{AUROC}_{\max} \leftarrow 0$
  $p_s = 0, s \in V$ ▷ Hallucination scores.
  while $N \leq N_{\max}$ **do**
    $H_{\text{subset}} \leftarrow H_{\text{subset}} \cup \{h_N\}$
    for $s \in V$ **do**
      $p_s \leftarrow \frac{N-1}{N} p_s + \frac{1}{N} d_{h_N}(s)$
    end for
    $\text{auroc} \leftarrow \text{AUROC}(\{y_s\}_{s \in V}, \{p_s\}_{s \in V})$
    if $\text{auroc} > \text{AUROC}_{\max}$ **then**
      $\text{AUROC}_{\max} \leftarrow \text{auroc}$
      $N_{\text{opt}} \leftarrow N$
    end if
    $N \leftarrow N + 1$
  end while
end procedure

procedure TOHA PREDICTION
  for $s \in T$ **do** ▷ Prediction on the test set.
    $p_s \leftarrow \frac{1}{N_{\text{opt}}} \sum_{i=1}^{N_{\text{opt}}} d_{h_i}(s)$
  end for
end procedure

4 Experiments

Datasets. We evaluated our approach on four datasets: RAGTruth (Niu et al., 2023) (we considered its' two separate benchmarks: QA based on MS MARCO (Nguyen et al., 2016) and summarization based on CNN/DM (Nallapati et al., 2016)) combined with news articles from an unnamed news platform), CoQA (Reddy et al.,

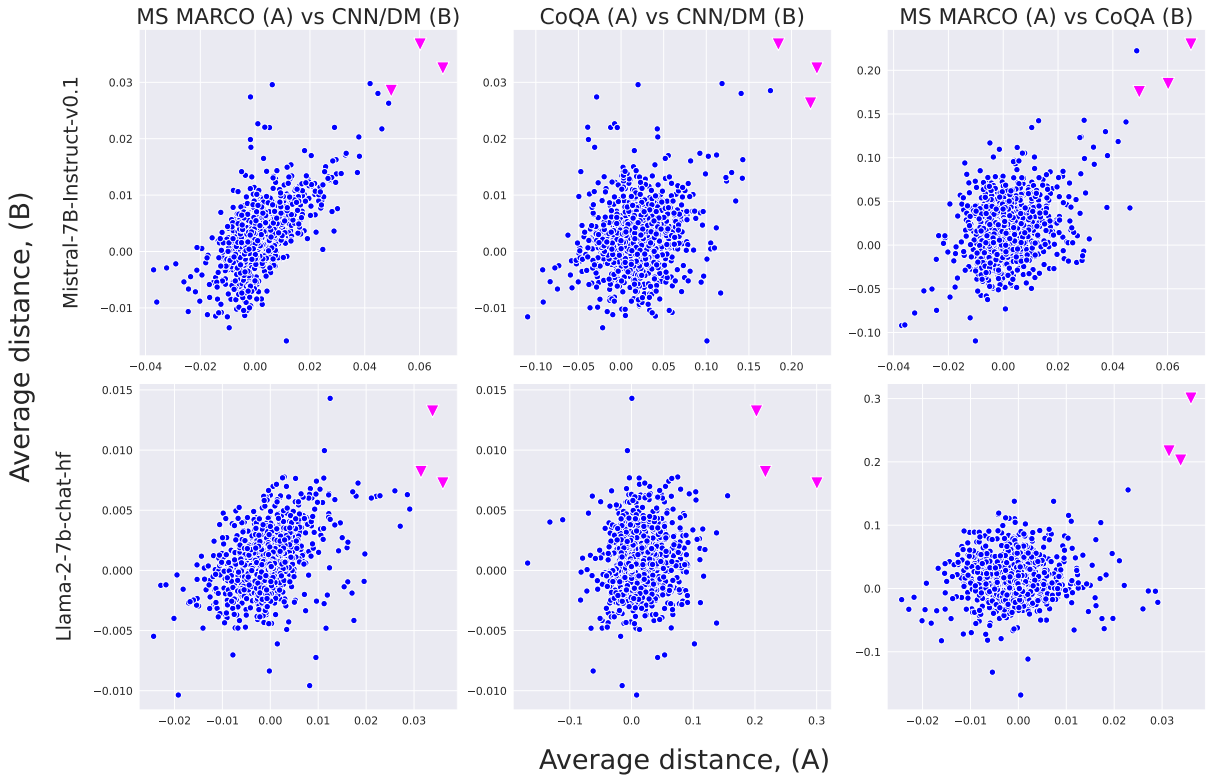

Figure 3: Δ_{ij} values for ij-th heads. Vertical axis corresponds to the difference on dataset (B), horizontal — to the one on dataset (A). The heads that separate samples best are highlighted in pink. Model names for a row are on the left side, datasets: MS MARCO, CNN/DM + Recent News, CoQA.

2019), SQuAD (Rajpurkar et al., 2016), and XSum (Narayan et al., 2018). The RAGTruth dataset consists of manually annotated responses of several LLMs in the RAG setting. The annotations are word-level; we, in turn, predict response-level labels, considering a response hallucinated if it contains at least one hallucination span. For the three latter datasets, we sampled LLM responses and annotated them automatically with GPT-4o (Hurst et al., 2024). Consistency with Human-GPT-4o label (Table 7) validated this approach, matching prior work (Bavaresco et al., 2024); see Appendix C for more details.

Models. We used five popular open-source LLMs: LLaMA-2-7B-chat, LLaMA-2-13B-chat, LLaMA-3.1-8B-Instruct, Mistral-7B-Instruct-v0.1, and Qwen2.5-7B-Instruct. Note that the RAGTruth dataset does not contain responses for LLaMA-3.1-8B and Qwen-2.5-7B; therefore, we only conducted experiments on SQuAD, CoQA, and XSum for these models.

Baselines. We compare TOHA with seven baselines: perplexity (Ren et al., 2023), max entropy (Fadeeva et al., 2024), Haloscope (Du et al., 2024), LLM-Check (Sriramanan et al., 2024), semantic entropy (Farquhar et al., 2024), EigenScore (Chen et al., 2024), and SelfCheckGPT (Manakul et al., 2024). Appendix E provides information on implementation details.

Main results. The results of our experiments are provided in Tables 1–2. We evaluate TOHA against state-of-the-art hallucination detection methods and demonstrate its competitive performance, consistently securing first or second place across most benchmark datasets. TOHA significantly outperforms uncertainty-based baselines and matches the accuracy of consistency-based approaches, achieving a notable 13.4% improvement on the challenging MS MARCO dataset that includes long and detailed model responses. While SelfCheckGPT emerges as TOHA's closest competitor, it relies on additional generations, incurring substantial computational overhead. Consistency-based methods exhibit a complexity of $\sim Kn^2$, where K is the number of additional generations and n is the tokens count. In contrast, TOHA operates with $\sim n^2 + N_{\text{opt}} n \log n$ complexity. Here, n^2 reflects the standard inference cost for transformer-based models, and $N_{\text{opt}} n \log n$ arises from comput-

Table 1: ROC AUC (↑) of hallucination detection techniques for three LLMs. The best results for each model are highlighted in **bold**, and the second best are underlined.

Method	Single generation	MS MARCO	CNN/DM + Recent News	CoQA	SQuAD	XSum
Mistral-7B						
SelfCheckGPT [1]	✗	0.62 ± 0.03	**0.63 ± 0.04**	**0.93 ± 0.02**	0.82 ± 0.03	**0.71 ± 0.03**
Semantic entropy [2]	✗	0.54 ± 0.06	0.57 ± 0.04	0.84 ± 0.02	0.71 ± 0.05	0.65 ± 0.05
EigenScore [3]	✗	0.58 ± 0.02	0.55 ± 0.04	0.73 ± 0.02	0.54 ± 0.04	0.57 ± 0.03
Haloscope [4]	✓	0.56 ± 0.05	0.54 ± 0.05	0.84 ± 0.02	**0.97 ± 0.03**	0.59 ± 0.05
LLM-Check [5]	✓	0.5 ± 0.06	0.57 ± 0.03	0.62 ± 0.04	0.56 ± 0.06	0.57 ± 0.03
Perplexity [6]	✓	0.67 ± 0.03	0.62 ± 0.04	0.77 ± 0.01	0.45 ± 0.06	0.66 ± 0.03
Max entropy [7]	✓	0.66 ± 0.05	0.6 ± 0.05	0.73 ± 0.02	0.73 ± 0.04	**0.71 ± 0.03**
TOHA (ours)	✓	**0.76 ± 0.04**	**0.63 ± 0.04**	0.89 ± 0.02	0.84 ± 0.01	0.61 ± 0.03
LLama-2-7B						
SelfCheckGPT [1]	✗	0.6 ± 0.02	**0.62 ± 0.05**	0.78 ± 0.04	0.59 ± 0.05	**0.70 ± 0.06**
Semantic entropy [2]	✗	0.54 ± 0.05	0.54 ± 0.04	0.76 ± 0.03	0.63 ± 0.06	0.63 ± 0.05
EigenScore [3]	✗	0.54 ± 0.02	0.52 ± 0.04	0.68 ± 0.03	0.53 ± 0.05	0.63 ± 0.06
Haloscope [4]	✓	0.53 ± 0.06	0.48 ± 0.02	0.74 ± 0.04	0.58 ± 0.04	0.58 ± 0.08
LLM-Check [5]	✓	0.46 ± 0.02	0.49 ± 0.03	0.6 ± 0.03	0.58 ± 0.05	0.58 ± 0.08
Perplexity [6]	✓	0.66 ± 0.01	0.57 ± 0.04	0.73 ± 0.03	0.57 ± 0.09	0.58 ± 0.05
Max entropy [7]	✓	0.66 ± 0.03	0.57 ± 0.05	0.71 ± 0.04	0.64 ± 0.02	0.56 ± 0.05
TOHA (ours)	✓	**0.67 ± 0.02**	0.57 ± 0.05	**0.88 ± 0.04**	**0.9 ± 0.04**	0.63 ± 0.02
LLaMA-2-13B						
SelfCheckGPT [1]	✗	0.57 ± 0.04	**0.60 ± 0.04**	0.86 ± 0.04	0.78 ± 0.03	0.61 ± 0.05
Semantic entropy [2]	✗	0.61 ± 0.03	0.51 ± 0.05	0.75 ± 0.05	0.72 ± 0.02	0.64 ± 0.04
EigenScore [3]	✗	0.56 ± 0.04	0.49 ± 0.04	0.45 ± 0.04	0.51 ± 0.04	0.55 ± 0.02
Haloscope [4]	✓	0.55 ± 0.05	0.50 ± 0.01	0.65 ± 0.04	0.54 ± 0.03	0.59 ± 0.01
LLM-Check [5]	✓	0.44 ± 0.04	0.57 ± 0.04	0.54 ± 0.03	0.53 ± 0.05	0.63 ± 0.04
Perplexity [6]	✓	0.65 ± 0.02	0.59 ± 0.02	0.58 ± 0.04	0.5 ± 0.05	0.58 ± 0.05
Max entropy [7]	✓	0.58 ± 0.03	0.55 ± 0.06	0.70 ± 0.06	0.78 ± 0.04	0.56 ± 0.04
TOHA (ours)	✓	**0.69 ± 0.01**	0.54 ± 0.04	**0.93 ± 0.03**	**0.92 ± 0.02**	**0.66 ± 0.05**

Table 2: ROC AUC (↑) of hallucination detection techniques. The best results for each model are highlighted in **bold**, and the second best are underlined.

Method	Single gen.	SQuAD	CoQA	XSum
LLaMA-3.1-8B				
SelfCheckGPT [1]	✗	0.79 ± 0.05	0.76 ± 0.07	**0.81 ± 0.02**
Semantic entropy [2]	✗	0.58 ± 0.04	**0.83 ± 0.05**	0.49 ± 0.05
EigenScore [3]	✗	0.52 ± 0.07	0.82 ± 0.06	0.49 ± 0.05
Haloscope [4]	✓	**0.85 ± 0.02**	0.56 ± 0.07	0.55 ± 0.04
LLM-Check [5]	✓	0.48 ± 0.04	0.5 ± 0.08	0.56 ± 0.03
Perplexity [6]	✓	0.82 ± 0.02	0.69 ± 0.04	0.61 ± 0.04
Max entropy [7]	✓	0.5± 0.04	0.53 ± 0.05	0.47 ± 0.01
TOHA (ours)	✓	**0.85 ± 0.02**	0.73 ± 0.01	0.63 ± 0.04
Qwen2.5-7B				
SelfCheckGPT [1]	✗	0.62 ± 0.06	**0.84 ± 0.04**	**0.74 ± 0.04**
Semantic entropy [2]	✗	0.68 ± 0.03	0.77 ± 0.06	0.65 ± 0.04
EigenScore [3]	✗	0.67 ± 0.04	0.66 ± 0.09	0.51 ± 0.04
Haloscope [4]	✓	0.59 ± 0.05	0.71 ± 0.05	0.58 ± 0.05
LLM-Check [5]	✓	0.5 ± 0.05	0.54 ± 0.1	0.55 ± 0.05
Perplexity [6]	✓	0.63 ± 0.06	0.67 ± 0.08	0.65 ± 0.05
Max entropy [7]	✓	0.73 ± 0.05	0.74 ± 0.05	0.52 ± 0.08
TOHA (ours)	✓	**0.8 ± 0.02**	0.69 ± 0.05	0.69 ± 0.02

ing topological divergences for only N_{opt} attention heads — a small subset of the model's total heads. Comparison of the best-performing baselines (Figure 6) confirms that TOHA reduces inference time by an order of magnitude compared to methods of similar quality.

Compared to Haloscope (Du et al., 2024), which operates with limited annotated data and large-scale unlabeled data, TOHA not only delivers superior performance but also eliminates the need for unannotated generations for hyperparameter tuning and a separate classifier training. This makes TOHA less data-dependent and more practical for real-world applications.

Another interesting comparison is with LLM-Check (Sriramanan et al., 2024), which also uses attention maps to compute hallucination scores—specifically, by averaging the log determinant of attention maps from a single pre-selected layer. However, our TOHA achieves superior performance, demonstrating that not all attention heads contribute equally to hallucination detection. By employing a topology-based head selection strategy, we significantly enhance detection quality.

Generalizability to different data distributions. From a deployment perspective, hallucination detection methods must remain robust to shifts in input data distribution, given the inherent diversity of real-world user queries. To evaluate TOHA's

robustness in this regard, we conducted transfer experiments on Mistral-7B (see Figure 4a). The results highlight TOHA's strong transferability: for the XSum and CNN/DM datasets, performance changes in transfer settings fall within the standard deviation. For the remaining datasets, TOHA maintains competitive performance compared to baseline methods (Table 1), demonstrating its adaptability to diverse data distributions.

How large should the probe sets be? As previously mentioned, TOHA requires only a small set of samples to identify "hallucination-aware" attention heads. To assess its sensitivity to probe set size, we conducted an ablation study (Figure 4). The results demonstrate TOHA's robustness to limited annotated data: even with just 50 samples, performance does not drop significantly and mostly remains stable as the probe set size increases.

What do hallucination patterns look like? As detailed in Section 3, the topological divergences we employ characterize the MSF connecting the vertices R of the response to the vertices P of the prompt. For hallucination-aware heads, we analyzed MSF patterns distinguishing hallucinated and grounded samples. A key finding is that hallucinated samples frequently exhibit strong attention to the <s> token, whereas grounded samples tend to attend to <s> less (Figure 5).

To verify the significance of <s> in hallucination detection, we conducted an ablation experiment: after removing <s> from the texts, we recomputed the TOHA hallucination scores for the selected hallucination-aware heads. The results (Table 3) show a significant performance drop, confirming that attention to <s> is a critical indicator of hallucination. This finding aligns with prior work demonstrating the influential role of <s> in LLM mechanisms (Barbero et al., 2025). However, using the average attention to <s> alone as a hallucination score proves insufficient (see Table 9 in Appendix D). In contrast, our proposed score, which incorporates the intricate structure of attention maps, demonstrates far greater discriminative power for this task.

Table 3: TOHA performance with and without <s> token. TOP-1 results are highlighted with **bold font**.

Dataset	with <s>	w/o <s>
Mistral-7B		
MS MARCO	**0.67** $\pm$ **0.02**	0.56 $\pm$ 0.02
CoQA	**0.88** $\pm$ **0.02**	0.32 $\pm$ 0.03
LLaMA-2-7B		
MS MARCO	**0.76** $\pm$ **0.04**	0.66 $\pm$ 0.03
CoQA	**0.89** $\pm$ **0.02**	0.56 $\pm$ 0.04

5 Related works

Hallucination detection methods. The problem of hallucinations in LLMs has attracted significant attention recently (Zhang et al., 2023; Huang et al., 2023; Wang et al., 2024). Consistency-based methods (Manakul et al., 2024; Chen et al., 2024; Kuhn et al., 2023; Qiu and Miikkulainen, 2024; Nikitin et al., 2024) that use the diversity of multiple LLM responses as a hallucination score offer robust detection but impose significant computational overhead. Surface-level techniques like perplexity and logit entropy (Fadeeva et al., 2024; Malinin and Gales, 2021) analyze model confidence directly from output distributions — efficient but limited in detection capability as they neglect the model's rich internal representations. Hidden states-based classifiers (Azaria and Mitchell, 2023; Sky et al., 2024; Zhou et al., 2025) require extensive annotated datasets, which are scarce in the public domain (Zhang et al., 2023). This issue was partially addressed by the Haloscope (Du et al., 2024), which leverages unlabeled data "in the wild" with minimum annotated data needed for hyperparameter selection. Attention map-based methods represent a promising yet underdeveloped direction. Current techniques either rely on large labeled data, e.g., Lookback Lens (Chuang et al., 2024), or exploit only simple attention graph properties, such as self-loop weights in LLM-Check (Sriramanan et al., 2024). This leaves a critical research gap: training-free methods that fully leverage the rich structural information encoded in attention relationships remain underexplored.

Topological Data Analysis (TDA) in NLP. Topological Data Analysis (TDA) is a mathematical framework for extracting multi-scale structural patterns from data using principles from topology and computational geometry (Chazal and Michel, 2017; Hensel et al., 2021). Recent years have seen growing interest in applying TDA to natural language processing (NLP) tasks to study textual structural properties (Uchendu and Le, 2024). For example, (Tulchinskii et al., 2024) leveraged persistent homology to estimate the intrinsic dimensionality of CLS embeddings for detecting

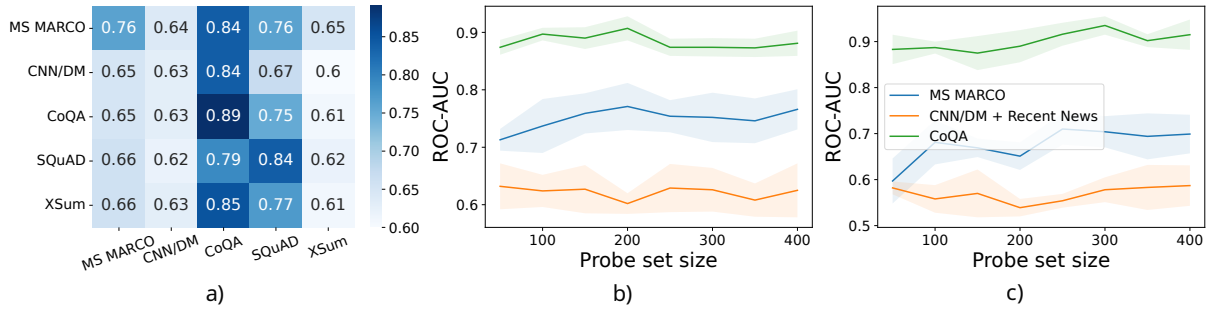

Figure 4: (a) Generalizability between the datasets, model: Mistral-7B-Instruct. The vertical axis corresponds to the origin of the probe set, the horizontal axis to the test dataset. (b)-(c): Detection quality dependence on the size of a probe set, models: Mistral-7B (left), LLama-2-7B (right).

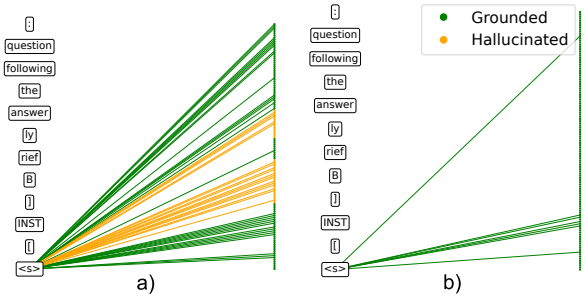

Figure 5: Attention to <s>: a) hallucinated sample and b) grounded one. Green color denotes edges and nodes that correspond to grounded tokens of a generation, yellow color — hallucinated ones. Model: Mistral-7B.

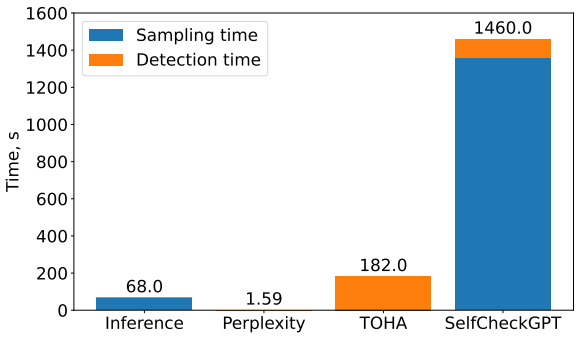

Figure 6: Comparison of methods' inference time in seconds. The measurements were obtained for 16 random samples from MS MARCO, model: Mistral-7B. For the SelfCheckGPT, 20 additional answers were generated for each sample.

machine-generated text. Other work has demonstrated the utility of topological features derived from transformer attention matrices — treated as weighted graphs — for diverse NLP applications. These include uncertainty quantification (Kostenok et al., 2023) and grammatical acceptability classification (Proskurina et al., 2023b), where topological features extracted from the attention graphs were used as input to train auxiliary classifiers.

6 Conclusion

This paper introduces TOHA, a novel hallucination detection method based on the topological divergence of attention maps. At its core, TOHA leverages our key observation that specific attention heads exhibit consistent patterns during hallucinations — regardless of the dataset. TOHA computes hallucination scores by averaging the topological divergences from these heads, and we formally prove several stability properties to ensure these scores are reliable. Additionally, we explored the behaviour of "hallucination-aware" heads, discovering that the attention to <s> token plays an important role in their discriminative ability. This importance of <s> aligns well with prior work (Barbero et al., 2025).

Extensive experiments show that TOHA is a robust alternative to existing approaches, matching or surpassing state-of-the-art baselines like SelfCheckGPT (Manakul et al., 2024). Notably, TOHA is significantly more efficient, operating up to an order of magnitude faster than methods of comparable quality. We further validate TOHA's transferability, demonstrating its robustness to shifts in data distribution — a critical advantage for real-world deployment, where LLM inputs are far more diverse and complex than benchmark examples.

In summary, TOHA delivers state-of-the-art detection performance while combining efficiency and solid generalizability, making it particularly suited for practical applications.

Limitations

While TOHA demonstrates strong performance and efficiency, several limitations warrant discussion.

RAG scenario. While TOHA operates effectively in the RAG scenario under the assumption that the provided context contains the correct answer, we recognize that this condition may not always hold in real-world applications. This limitation points to an important direction for future research, where the method could be extended to handle cases of incomplete or unreliable context knowledge.

Model-specific dependencies. TOHA's effectiveness relies on identifying "hallucination-aware" attention heads, which may vary across LLM architectures. While our experiments cover popular open-source models (e.g., LLaMA, Mistral), further validation is needed for proprietary or larger models (e.g., GPT-4, Claude).

Multimodal extensions. The current framework operates solely on text. Adapting TOHA to multimodal settings (e.g., vision-language models) would require redefining attention graphs across heterogeneous data modalities.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

S. A. Barannikov. 1994. The framed Morse complex and its invariants.

Serguei Barannikov, Ilya Trofimov, Grigorii Sotnikov, Ekaterina Trimbach, Alexander Korotin, Alexander Filippov, and Evgeny Burnaev. 2021. Manifold topology divergence: a framework for comparing data manifolds. *Advances in neural information processing systems*, 34:7294–7305.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Razvan Pascanu, and 1 others. 2025. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *CoRR*, abs/2406.18403.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

Frédéric Chazal and Bertrand Michel. 2017. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability judgements via examining the topology of attention maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 88–107, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. 2024. Large language models (LLM) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234. IEEE.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436.

Xuefeng Du, Chaowei Xiao, and Sharon Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37:102948–102972.

Herbert Edelsbrunner and John Harer. 2010. *Computational Topology: An Introduction*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for

large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhengjie Gao, Xuanzi Liu, Yuanshuai Lan, and Zheng Yang. 2024. A brief survey on safety of large language models. *Journal of computing and information technology*, 32(1):47–64.

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2024. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*.

Felix Hensel, Michael Moor, and Bastian Rieck. 2021. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Tianchu Ji, Shraddhan Jain, Michael Ferdman, Peter Milder, H. Andrew Schwartz, and Niranjan Balasubramanian. 2021. On the distribution, sparsity, and inference-time quantization of attention values in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4147–4157.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Elizaveta Kostenok, Daniil Cherniavskii, and Alexey Zaytsev. 2023. Uncertainty estimation of transformers' predictions via topological analysis of the attention matrices. *arXiv preprint arXiv:2308.11295*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2024. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2023. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396*.

Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. 2023a. Can BERT eat RuCoLA? Topological data analysis to explain. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*.

Irina Proskurina, Ekaterina Artemova, and Irina Piontkovskaya. 2023b. Can bert eat rucola? topological data analysis to explain. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 123–137.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2023. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

C.H.-Wang Sky, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4401–4420.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Eduard Tulchinskii, Kristian Kuznetsov, Daniil Cherniavskii, Serguei Barannikov, Sergey Nikolenko, and Evgeny Burnaev. 2023. Topological data analysis for speech processing. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 311–315.

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Sergey Nikolenko, Evgeny Burnaev, Serguei Barannikov, and Irina Piontkovskaya. 2024. Intrinsic dimension estimation for robust detection of ai-generated texts. *Advances in Neural Information Processing Systems*, 36.

Adaku Uchendu and Thai Le. 2024. Unveiling topological structures in text: A comprehensive survey of topological data analysis applications in NLP. *arXiv preprint arXiv:2411.10298*.

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529.

Simon Zhang, Mengbai Xiao, and Hao Wang. 2020. Gpu-accelerated computation of vietoris-rips persistence barcodes. In *36th International Symposium on Computational Geometry (SoCG 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. 2025. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*.

A Topological data analysis: background

A simplicial complex S is a collection of simplices such that every face of a simplex $\sigma \in S$ is also in S. Simplices are the higher-dimensional generalizations of triangles; a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and so forth. Formally, given a finite set X, an n-simplex σ is an $(n+1)$ subset of X. Simplicial complexes are fundamental objects in algebraic and combinatorial topology, serving as a discrete analog to topological spaces.

The Vietoris-Rips complex $VR_\varepsilon(X)$ of a weighted graph $G = (V_G, E_G)$ with distance threshold $\varepsilon > 0$ is defined as follows:

$$\mathrm{VR}_\varepsilon(G) = \left\{ \sigma \subseteq V_G \;\middle|\; \forall\, v_i, v_j \in \sigma,\ w(e_{ij}) \leq \varepsilon \right\},$$

where w is the edge weight function associated with G.

Homology groups H_k are invariants used in algebraic topology to study the topological properties of a space. Let $C_k(S)$ denote vector space over $\mathbb{Z}/2\mathbb{Z}$, with the basis consisting of k-dimensional simplices of S. Elements of C_k are called chains. Formally, homology groups are derived from a chain complex $(C_\bullet, \partial_\bullet)$, which is a sequence of C_k connected by boundary maps ∂_k:

$$C_\bullet : \cdots \to C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} \cdots,$$
$$\partial_k \circ \partial_{k+1} = 0.$$

The k-th homology group H_k is defined as the quotient of the group of k-cycles (chains whose boundary is zero) by the group of k-boundaries (chains that are the boundary of a $(k+1)$-chain). Mathematically, this is expressed as:

$$H_k(S) = Z_k(S)/B_k(S),$$

where $Z_k = \ker \partial_k = \{c \in C_k \,|\, \partial_k(c) = 0\}$ and $B_k = \operatorname{im} \partial_{k+1} = \{\partial_{k+1}(c) \,|\, c \in C_{k+1}\}$ is the group of k-boundaries. The elements of $H_k(S)$ represent various k-dimensional topological features in S. Elements of a basis in $H_k(S)$ correspond to a set of basic topological features.

A filtration of simplicial complexes $\mathcal{F}$ is a family of nested simplicial complexes:

$$\mathcal{F} : \varnothing \subseteq S_1 \subseteq S_2 \subseteq \cdots \subseteq S_n = S,$$

where each S_k is a simplicial complex itself. In practice, the filtrations of simplicial complexes are usually obtained for sequences of increasing thresholds $0 < \varepsilon_1 < \cdots < \varepsilon_n$. For example, simplicial complexes $VR_{\varepsilon_i}(X)$ form a filtration

$$\mathcal{F}_{VR}(X) : \varnothing \subseteq VR_{\varepsilon_1}(X) \subseteq VR_{\varepsilon_2}(X) \subseteq \ldots$$
$$\subseteq VR_{\varepsilon_n}(X) = VR(X).$$

As the threshold ε increases, new topological features (e.g., connected components, holes) can appear and disappear. The persistent homology tool tracks the dynamics of these topological features. Formally, the k-th persistent homology of S is the pair of sets of vector spaces $\{H_k(S_i) \,|\, 0 \le i \le n\}$ and maps f_{ij}, where $f_{ij} : H_k(S_i) \to H_k(S_j)$ is a map induced by the embedding $S_i \subseteq S_j$. Each persistent homology class in this sequence is "born" at some S_i and "dies" at some S_j or never dies (Barannikov, 1994). This birth-death process of a basic set of independent topological features can be visualized as the set of intervals $[\varepsilon_{\text{birth}}, \varepsilon_{\text{death}}]$ called barcode (see Figure 7). The features with 0 lifespans are typically excluded. The horizontal axis is a sequence of thresholds ε, and each horizontal bar corresponds to a single feature. We begin with $|X| = m$ connected components (all of them are "born"), and as ε increases, their pairs are merged (each merge corresponds to a "death" of a feature). The $0-$th barcode construction procedure is equivalent to Kruskal's algorithm for minimum spanning tree (MST), the bars in the barcode correspond to the edges in the MST of X (Tulchinskii et al., 2023).

B MTop-Div on graphs properties

Basic properties of MTop-Div for attention graphs. Now we consider specific properties for our adaptation of $\mathrm{MTop\text{-}Div}(R, P)$.

Proposition B.1. *The following holds for any attention graph G and its complementary vertex subsets $P, R \subset V_G$.*

- $\mathrm{MTop\text{-}Div}(R, P)$ *value equals the length of the MSF attaching R to P.*
- *Let the natural norm on the cross-barcodes be defined as follows:*

$$\|\text{Cross-Barcode}_0\|_B = \max_{[b_j, d_j] \in \text{Cross-Barcode}_0} (d_j - b_j). \quad (3)$$

The norm of $\mathrm{Cross\text{-}Barcode}_0(R, P)$ *lays in the interval* $[0,\, 1]$*:*

$$0 \le \|\text{Cross-Barcode}_0(R, P)\|_B \le 1. \quad (4)$$

- *The divergence itself is bounded by*

$$0 \le \mathrm{MTop\text{-}Div}(R, P) \le |R|. \quad (5)$$

The second and third statements are immediately obtained from the properties of an attention matrix: all its weights lie between 0 and 1.

The following property formalizes the intuition behind our metric — it measures the strength of the response's connection to the prompt through multiscale topological features of the attention graph.

Proposition B.2. ***(Exact sequence.)*** *For any α, the following sequence of natural maps of homology groups is exact*

$$(\mathbb{Z}/2\mathbb{Z})^{|P|} \xrightarrow{r_2} H_0(VR_\alpha(G)) \xrightarrow{r_1}$$
$$\xrightarrow{r_1} H_0(VR_\alpha(G, w_{(R \cup P)/P})) \xrightarrow{r_0} 0.$$

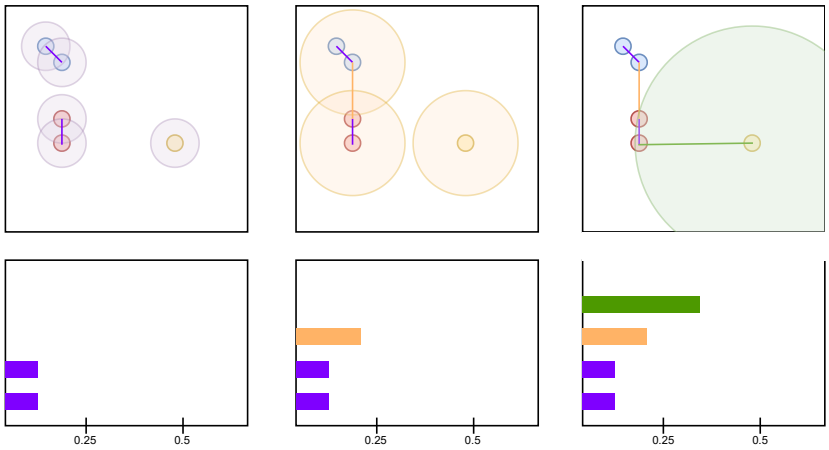

Figure 7: H_0 barcode construction. As the threshold increases, the separate connected components merge, resulting in the death of topological features. The horizontal axis is a sequence of thresholds ε, and each horizontal bar corresponds to a single feature.

Proof of Proposition B.2.
We have to check the definition of the exact sequence: $\mathrm{Ker}(r_i) = \mathrm{Im}(r_{i+1})$. For a pair r_0, r_1, it is equivalent to the surjectvity of r_1. The H_0 homology group of a graph corresponds to the connected components of the graph. The set of edges $E^{\leq\alpha}_{(G,w)} = \{e \in E_G | w_e \leq \alpha\}$ is always a subset in the analogous set of the weighted graph $(G, w_{(R\cup P)/P})$ with all weight edges between P vertices set to zero. Therefore, the map r_1 between their connected components is surjective. Similarly, the kernel of the map r_1 is spanned by the differences of two connected components, which are merged after adding some of the edges between P vertices, and any such difference lies in the image of the map r_2. Also, any two vertices from P belong to the same connected component in the graph $(G, w_{(R\cup P)/P} \leq \alpha)$, hence the image of r_2 is in the kernel of r_1. Therefore, the considered sequence is exact indeed. □

Proof of Proposition 3.1.
1. The $0-$th Cross-Barcode coincides with the set of edges in the minimal spanning tree of the weighted graph G with all the weights within P vertex subset equal zero. Excluding the zero weight edges, this edge set coincides with the minimal spanning forest attaching the vertex set R to P vertices. □

2. Denote by $\mathrm{MSF}(R, P)$ the minimum spanning forest attaching R to P. Note that we have properties B.1, so

$$\text{MTop-Div}(R, P) = \sum_{e\in \mathrm{MSF}(R,P)} w(e). \quad (6)$$

Therefore, we have to show that the weight of $\mathrm{MSF}(R, P)$ does not change significantly when all weights are changed by no more than ε.

There are two possibilities: 1) after a change, all MSF edges remain the same, or 2) some edges are replaced with other edges. In the first case, it is obvious that the total sum of edge weights changes by no more than $\delta = \varepsilon \cdot \#\text{edges}(\mathrm{MSF}(R, P)) = \varepsilon \cdot |R|$. Consider the second case. Denote by $\mathrm{MSF}_{\text{prev}}$ the original MSF, by $\mathrm{MSF}_{\text{new}}$ — the MSF after the change; let w be the edge weight function before the change, $\hat{w}$ — after the change. The following inequalities hold:

$$\hat{w}(\mathrm{MSF}_{\text{new}}) < \hat{w}(\mathrm{MSF}_{\text{prev}}); \quad (7)$$

$$w(\mathrm{MSF}_{\text{prev}}) - \delta \leq \hat{w}(\mathrm{MSF}_{\text{prev}}) \leq \leq w(\mathrm{MSF}_{\text{prev}}) + \delta; \quad (8)$$

$$w(\mathrm{MSF}_{\text{new}}) - \delta \leq \hat{w}(\mathrm{MSF}_{\text{new}}) \leq \leq w(\mathrm{MSF}_{\text{new}}) + \delta; \quad (9)$$

$$w(\mathrm{MSF}_{\text{new}}) \geq w(\mathrm{MSF}_{\text{prev}}). \quad (10)$$

From (7)-(8) follows that $\hat{w}(\mathrm{MSF}_{\text{new}}) < w(\mathrm{MSF}_{\text{prev}}) + \delta$; from (9)-(10) follows that $\hat{w}(\mathrm{MSF}_{\text{new}}) \geq w(\mathrm{MSF}_{\text{prev}}) - \delta$. □

3. Follows obviously from the MSF formula for $\text{MTop-Div}(R, P)$ and attention map properties.

C Datasets

SQuAD (Rajpurkar et al., 2016) and CoQA (Reddy et al., 2019) are widely used English question-answering benchmarks that have facilitated the development of hallucination detection datasets (Kuhn et al., 2023; Manakul et al., 2024).

Similarly, XSum (Narayan et al., 2018), a dataset of news articles with one-sentence summaries, is commonly employed in hallucination detection research for abstractive summarization (Shi et al., 2024; Cao et al., 2022). To assess LLM performance, we used GPT-4o to annotate responses to questions sourced from SQuAD, CoQA, and summarization tasks from XSum.

C.1 Data Generation & Annotation

Generation. We generate responses from a language model (LLM) for the considered datasets, employing different prompting strategies for each dataset while keeping these strategies consistent across models (see prompt examples in Table 4). For SQuAD and XSum, responses are generated using a zero-shot approach. In contrast, for CoQA, we create queries in a few-shot manner without providing specific instructions, following (Lin et al., 2024): each sample consists of a passage and a series of question-answer pairs, concluding with a final question that the model is expected to answer.

Annotation: automated vs human. We treat hallucination detection as a binary classification problem; our target indicates whether a hallucination is present anywhere in the model's response. Two approaches to annotating model generations were considered: 1) automated annotation using an LLM (in our case, GPT-4o), and 2) manual annotation by human experts.

During the automated annotation process, we provide an LLM's output preceded by an instruction (prompt) to GPT-4o. In this prompt, GPT-4o is asked to determine whether the output contains hallucinations, and we expect a single-word response of either "Yes" or "No." An example of such an instruction for the question answering task is shown in Table 6.

For human annotation, we asked three team members with at least upper-intermediate English proficiency to independently annotate approximately 100 samples from each dataset. We selected samples where all annotators reached a consensus and considered these annotations the ground truth hallucination labels.

To further evaluate GPT-4o, we conducted automatic annotation using several variations of prompts, each reformulating the task for GPT-4o, including zero-shot and few-shot versions. We then compared these annotations to the actual hallucination labels. The results, presented in Table 7, demonstrate a consistent alignment between GPT-4o's annotations and those made by humans, regardless of the specific prompt. This consistency confirms the robustness of our approach to the exact form of instruction.

Based on these findings, we prefer automated annotation as a cost-effective and efficient alternative to human experts.

Annotation: general pipeline. CoQA and SQuAD contain questions paired with ground-truth answers. To minimize false positives in labeling, we employed a two-step verification process:

1. Rouge-L scoring: we computed Rouge-L scores (using the `evaluate` library) between the model's response and the ground-truth answers.

2. Substring matching: we checked whether any ground-truth answer was a substring of the response.

Responses with a Rouge-L score of 1 (exact match) were labeled as grounded. Those meeting both of the following criteria were flagged as potential hallucinations:

- Rouge-L score ≤ 0.3 (following (Kuhn et al., 2023));
- no ground-truth answer appears as a substring.

These candidate hallucinations were then reviewed by GPT-4o, and only confirmed cases were finally labeled as hallucinations.

For XSum, where reference summaries are more complex than the ground truth answers in SQuAD/CoQA, we bypassed Rouge-L filtering and relied solely on GPT-4o for annotation.

Detailed statistics for each dataset can be seen in Table 8. The number of samples in the datasets varies across models, as we tried to maintain a balance of hallucinated and grounded responses, ensure sample cleanness, and minimize mislabeling. The procedure outlined above selects a different number of objects in a sample depending on the quality of the model's responses.

D Other experiment results

D.1 Alternative attention map-based features for hallucination detection

In our preliminary experiments for developing an attention maps-based hallucination detector, we trained classifiers using topological features previously applied to other NLP tasks (Kushnareva

SQuAD	CoQA
Given the context, answer the question in a brief but complete sentence. Note that your answer should be strictly based on the given context. In case the context does not contain the necessary information to answer the question, please reply with "Unable to answer based on given context". *Context:* Once upon a time, in a quiet village, there lived a kind old baker named Henry. He was known for his delicious bread and warm smile. One day, a traveler arrived, tired and hungry, and Henry welcomed him with a fresh loaf. *Question:* Who was known for baking delicious bread? *Answer:*	Once upon a time, in a quiet village, there lived a kind old baker named Henry. He was known for his delicious bread and warm smile. One day, a traveler arrived, tired and hungry, Henry welcomed him with a fresh loaf. *Q:* What was Henry known for? *A:* Baking delicious bread. *Q:* What else? *A:* Warm smile. *Q:* How did the traveler feel when he arrived? *A:* Tired and hungry. *Q:* What did Henry give the traveler?

Table 4: Examples of prompts used during generation for CoQA and SQuAD (we add additional delimiter spaces and formatting not present in actual prompts for better readability). SQuAD contains instructions followed by context and questions. In CoQA, the prompt has only a contextual passage followed by a question-and-answer series, with the last question being the actual one.

XSum
Please annotate potentially hallucinated model-generated summaries in the following settings. I will provide a reference text and a model-generated summary of this text. You will judge whether the given model-generated summary contains hallucinations. Answer "Yes" if the summary contains hallucinations, "No" if it does not, and "N/A" if you cannot decide. Do NOT give any extra explanations.

Table 5: The prompt used during generation for the XSum dataset (we add additional delimiter spaces and formatting not present in actual prompts for better readability).

You are an AI assistant specialized in detecting hallucinations in question-answering tasks.
Your job is to analyze the given context, question, and generated answer to identify whether the answer contains any hallucinations. Examples:

Example 1.
Context:
The city of Paris is the capital of France. It is known for its iconic landmarks like the Eiffel Tower and Notre Dame Cathedral.
The city is situated in the northern part of the country, near the Seine River.
Question: Is Paris the capital of Germany?
Generated answer: Yes, Paris is the capital of Germany.
Hallucination: Yes.

Example 2.
Context:
The city of Paris is the capital of France.
It is known for its iconic landmarks like the Eiffel Tower and Notre Dame Cathedral.
The city is situated in the northern part of the country, near the Seine River.
Question: Is Paris the capital of Germany?
Generated answer: No, Paris is not the capital of Germany. According to the context, Paris is the capital of France.
Hallucination: No.

You should determine if the answer contains hallucinations according to the hallucination types above.
If you cannot decide if the generated answer is a hallucination, write "N/A." as the answer.
The answer you give MUST be ONLY "Yes.", "No." or "N/A."; do NOT give ANY explanation.

Table 6: Example of annotation prompt passed to GPT-4o (we add additional delimiter spaces and formatting not present in actual prompts for better readability).

Prompt number		1	2	3	4	5
CoQA	Accuracy (↑)	0.809 ± 0.017	0.861 ± 0.015	0.742 ± 0.003	0.795 ± 0.009	0.831 ± 0.025
	Precision (↑)	0.849 ± 0.021	0.911 ± 0.007	0.771 ± 0.003	0.828 ± 0.011	0.860 ± 0.012
	Recall (↑)	0.871 ± 0.004	0.877 ± 0.019	0.877 ± 0.013	0.877 ± 0.005	0.893 ± 0.027
SQuAD	Accuracy (↑)	0.831 ± 0.003	0.857 ± 0.018	0.857 ± 0.008	0.872 ± 0.003	0.854 ± 0.007
	Precision (↑)	0.813 ± 0.002	0.831 ± 0.028	0.845 ± 0.021	0.850 ± 0.011	0.847 ± 0.007
	Recall (↑)	0.796 ± 0.008	0.839 ± 0.010	0.823 ± 0.023	0.858 ± 0.018	0.813 ± 0.017

Average	Accuracy (↑)	Precision (↑)	Recall (↑)
CoQA	0.808	0.844	0.879
SQuAD	0.854	0.837	0.826

Table 7: Classification metrics of GPT-4o annotation for CoQA and SQuAD with human labels considered actual annotation. The top table shows metric scores for different variants of prompts used. The bottom table shows the metric scores averaged across all prompt variants.

Model	CoQA		SQuAD		XSum	
	Hal.	Grounded	Hal.	Grounded	Hal.	Grounded
Mistral-7B	776	776	311	389	301	448
LLaMA-2-7B	375	375	357	235	239	507
LLaMA-2-13B	279	384	314	436	208	522
LLaMA-3.1-8B	189	200	350	400	243	407
Qwen2.5-7B	124	183	215	249	194	556

Table 8: Datasets statistics. Number of hallucinated and grounded samples of each model.

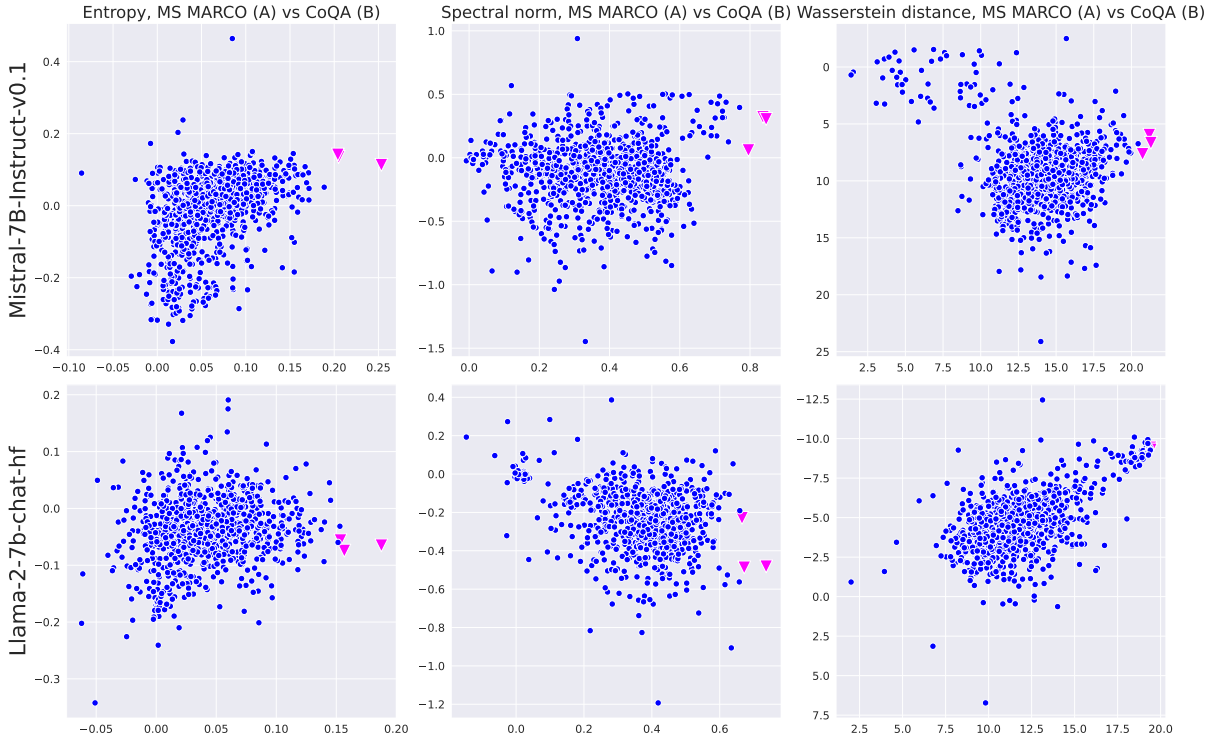

Figure 8: Δ_{ij} values for ij-th heads, MS MARCO vs CoQA. Vertical axis corresponds to the difference on dataset (B), horizontal — to the one on dataset (A). The heads that segregate samples best are highlighted in pink. Model names for a row are on the left side.

et al., 2021; Cherniavskii et al., 2022), as well as traditional attention map characteristics. As standard topological features, we considered barcode-based features, such as the sum of bar lengths in persistence diagrams, and naive topological features, including the average vertex degree in attention graphs. For traditional attention-based features, we used sparsity ratio, attention entropy, and spectral norm (Kobayashi et al., 2020; Vig and Belinkov, 2019; Ji et al., 2021). We also considered Wassersein distances between the persistent diagrams (Edelsbrunner and Harer, 2010) of a context and a response subgraphs as an alternative way to assess their similarity. Finally, we analyzed the average attention to <s> token as we discovered that hallucination-aware heads often attend to it when

Table 9: ROC-AUC values of supervised classifiers on top of various set of features. TOP-1 results are highlighted with **bold font**, while TOP-2 are underlined.

Features	MS MARCO	CoQA
Mistral-7B		
Standard topological	0.67	0.69
Sparsity ratio	0.66	0.7
Entropy	0.75	0.77
Wasserstein	0.77	0.73
Spectral norm	0.73	0.72
Attention to <s>	0.65	0.61
MTop-Div	**0.86**	**0.98**
LLaMA-2-7B		
Naive topological	0.69	0.7
Sparsity ratio	0.49	0.61
Entropy	0.38	0.68
Wasserstein	0.73	0.6
Spectral norm	0.49	0.64
Attention to <s>	0.62	0.64
MTop-Div	**0.75**	**0.96**

the hallucination is present (see 4).

To determine the most informative features for hallucination detection, we trained supervised classifiers on concatenated features from all layers and heads and compared them to classifiers trained only on MTop-Div values under the same conditions. The results are presented in Table 9.

While the classifier based on MTop-Div values significantly outperforms alternative approaches, computing these values across all layers and attention heads is highly computationally expensive. To address this, we developed TOHA — a more efficient alternative by aggregating MTop-Div values from only a subset of the "hallucination-aware" attention heads in the model.

D.2 Other metrics for the head selection procedure

We also investigated alternative attention-map-based scores — including entropy, spectral norm, and the Wasserstein distance between the persistent diagrams of prompts and responses — for selecting specialized attention heads.Following the same pipeline as in TOHA, we computed the average distances between hallucinated and grounded samples using alternative proximity metrics. The results, presented in Figure 8, reveal that the most segregating heads for MS MARCO do not generalize to the CoQA dataset. This suggests that our proposed MTop-Div metric is better suited for the task compared to existing solutions.

E Implementation details

In this section, we describe the key implementation choices.

- For EigenScore, we used the last token representation to embed sentences, as suggested in (Chen et al., 2024). We took outputs from the 16th layer, since middle layers were shown to contain the most factual information (Sky et al., 2024; Azaria and Mitchell, 2023).
- For methods that rely on multiple generations, we generated 20 samples per input, following recommendations from (Manakul et al., 2024; Chen et al., 2024).
- For SelfCheck-GPT, we used its NLI-based variant.
- For LLM-Check, we considered its white-box attention score modification, as it works in a setting similar to ours.
- For Haloscope, we reserved 150 samples for hyperparameter tuning and treated the rest as unlabeled data. For the CoQA dataset, we added several thousand extra generations during training, as in the original paper. Other datasets were used as-is, since their sizes are comparable with those in Haloscope's experiments on the TruthfulQA dataset.
- For TOHA's head selection, we similarly used 150 annotated samples. The topological divergences were calculated using `ripser++` library (Zhang et al., 2020), MIT license.

In our experiments, we used 60/15/25 train/val/test split. All the obtained results were averaged over five runs. All experiments were carried out using NVidia L40.

F Use of scientific artifacts & AI assistants

CoQA contains passages from seven domains under the following licenses: Literature and Wikipedia passages are shared under CC BY-SA 4.0 license; Children's stories are collected from MCTest which comes with MSR-LA license; Middle/High school exam passages are collected from RACE which comes with its own license;

News passages are collected from the DeepMind CNN dataset which comes with Apache license. SQuAD dataset comes under CC BY-SA 4.0 license. RAGTruth dataset comes under MIT license. XSum dataset comes under MIT license.

We used all the artifacts as it was intended by the corresponding licenses. No personal information or offensive content is contained in the considered datasets.

The original text of this paper was spell- and grammar-checked and slightly smoothed out using Grammarly.

G Potential risks

1. Ethical risks from deployment: overconfidence in TOHA's scores could lead to unchecked LLM outputs in high-stakes scenarios (e.g., healthcare). TOHA should be frame as a "warning system" rather than a definitive filter, and advocate for human review.

2. Attention manipulation attacks: adversarial prompts could artificially alter attention patterns, evading detection.