Hallucination Detection in LLMs with Topological Divergence on Attention Graphs

Alexandra Bazarova[1], Andrei Volodichev[1], Aleksandr Yugay[1], Andrey Shulga[1], Alina Ermilova[1], Konstantin Polev[2], Julia Belikova[2], Rauf Parchiev [2], Dmitry Simakov[2], Maxim Savchenko[2], Andrey Savchenko[2,3], Serguei Barannikov[*1,4], Alexey Zaytsev[*1],

[1]Applied AI Institute, [2]SB AI Lab,[3] HSE University, [4]CNRS, Universite Paris Cite

Correspondence: bazarovaai.239@gmail.com

Abstract

Hallucinations remain a critical challenge for large language models (LLMs), particularly in Retrieval-Augmented Generation (RAG) settings where models may generate outputs unsupported by the provided context. To address this, we introduce TOHA, a TOpology-based HAllucination detector, 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 in RAG settings reveals consistent patterns: higher divergence values in specific attention heads correlate with unfaithful outputs, independent of the dataset. Extensive experiments — including evaluations on question-answering and summarization tasks — show that our approach achieves state-of-the-art or competitive results across several benchmarks while requiring minimal annotated data and computational resources. Our findings indicate that the topological structure of attention matrices provides an efficient and robust metric for assessing the correctness of LLM's responses. Our source code is publicly available: https://github.com/sb-ai-lab/TOHA.

1 Introduction

Large language models (LLMs) have advanced significantly in recent years, with applications across various fields (Chkirbene et al., 2024). To ensure factual reliability, modern LLMs are often combined with the retrieval-augmented generation (RAG) technique, which integrates relevant external knowledge from diverse databases directly into the generation process (Lewis et al., 2020).

Despite these improvements, LLMs remain prone to producing so-called *hallucinations*, i.e., content that is factually or contextually incorrect (Huang et al., 2023; Li et al., 2024). Detecting hallucinations is crucial for the safe deployment of LLMs in sensitive fields, as erroneous outputs may seriously erode user trust. An effective detector would therefore expand the scope of LLM applications while mitigating risks (Gao et al., 2024).

Though many methods address this problem (Sahoo et al., 2024; Shorinwa et al., 2025), they often face significant practical constraints, such as the scarcity of annotated datasets (Zhang et al., 2023) required for supervised methods (Sky et al., 2024; Orgad et al., 2025), the high computational cost of generating multiple additional samples (Chen et al., 2024; Hou et al., 2025), or the inability of LLMs' output probabilities to fully represent the model's true uncertainty (Fadeeva et al., 2024; Shelmanov et al., 2025).

These challenges can be addressed by leveraging LLMs' internal states, which are informative for the hallucination-detection problem (Azaria and Mitchell, 2023; Sriramanan et al., 2024; Gekhman et al., 2025). We introduce TOHA (a TOpology-based HAllucination detector), a training-free method designed for the RAG scenario that leverages the structure of LLM attention maps to identify hallucinations. Our method requires minimal annotated data while avoiding the computational overhead of multiple generations, which makes TOHA both data- and compute-efficient.

Our core insight is that the prompt-response interconnections within an LLM's attention mechanism reveal the extent to which a response is faithful to the original prompt. TOHA formalizes this idea by analyzing attention graphs — complete weighted graphs derived from LLM attention maps, the representation prior used for topological data analysis (TDA) in NLP (Kushnareva et al., 2021; Tulchinskii et al., 2023). Unlike existing attention-based methods, which treat all attention heads as equally important (Sriramanan et al., 2024; Binkowski et al., 2025) or ignore the geometric structure of attention maps (Sun et al., 2025; Vazhentsev et al., 2025), our approach leverages a

[*]Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15449–15470

small set of attention heads selected by a topological criterion designed to identify hallucinations.

Specifically, TOHA is based on the $\mathrm{MTop\text{-}Div}_G(R, P)$, our adaptation of Manifold Topology Divergence (Barannikov et al., 2021) for the graph setting, which quantifies the dissimilarity between the prompt (P) and response (R) token sets in the attention graph. We demonstrate that this score not only reflects the geometry but also measures the informational novelty of the response relative to the prompt, making it well-suited for hallucination detection in RAG scenarios.

By analyzing divergence values across attention heads, we identified a subset of heads that consistently yield higher scores for hallucinated samples (Figure 2), irrespective of the dataset. TOHA's final hallucination score is the average $\mathrm{MTop\text{-}Div}_G(R, P)$ value from these "hallucination-aware" heads. Some of these heads are associated with copying behavior, which is aligned with prior findings (Sun et al., 2025).

Here are our main **contributions**:

- We introduce TOHA, a training-free framework for hallucination detection in the RAG scenario that leverages the topology of LLM attention graphs. Our method operates up to an order of magnitude faster than methods of comparable quality and requires minimal annotated data.
- Central to TOHA is $\mathrm{MTop\text{-}Div}_G(R, P)$, an adaptation of Manifold Topology Divergence for graphs, which quantifies the topological dissimilarity between prompt and response sets of tokens. We demonstrate that this score characterizes the novelty of the response relative to the prompt, making it highly effective for detecting hallucinations in RAG systems.
- By analyzing the proposed score, we discover the existence of "hallucination-aware" attention heads, which consistently yield greater divergence values for hallucinated samples across different datasets. This finding ensures TOHA's efficiency and strong cross-domain transferability: averaging $\mathrm{MTop\text{-}Div}_G(R, P)$ over just 10 heads is sufficient for robust hallucination detection.
- Our experiments show that TOHA consistently matches or exceeds state-of-the-art performance across multiple benchmarks when applied to open-source LLMs of varying sizes, including 7B- and 13B-parameter models.

2 Background

2.1 Attention Map as a Weighted Graph

Modern LLMs are mainly based on the self-attention mechanism (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:

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

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

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

and each entry $w_{ij} = W_{ij}(Q, K)$ captures how strongly token i attends to token j, with greater values indicating closer relationship. We consider decoder-only LLMs, for which the attention maps are lower triangular.

An attention map can be reframed as a complete undirected weighted graph G with edge weights $1 - w_{ij}$ to represent pseudo-distances between tokens. We call G an *attention graph*. It naturally partitions into prompt (P) and response (R) tokens (see Figure 1b). We analyze the topological relationships between these node subsets, which, as we assume, should be indicative of hallucinations in the RAG scenario.

2.2 Manifold Topology Divergence

The Manifold Topology Divergence, or MTop-Div, was proposed in (Barannikov et al., 2021). $\mathrm{MTop\text{-}Div}(M, N)$ quantifies the difference between two data manifolds $\mathcal{M}$ and $\mathcal{N}$, approximated by point clouds M and N. It is computed as the sum of interval lengths in the $\mathrm{Cross\text{-}Barcode}(M, N)$, a set of intervals representing topological features distinguishing N from $M \cup N$. A larger divergence indicates a greater topological difference between the manifolds. Appendix B contains more details on TDA.

3 Method

This section introduces the $\mathrm{MTop\text{-}Div}_G(R, P)$ score, which is designed to quantify the topological divergence between the prompt and response

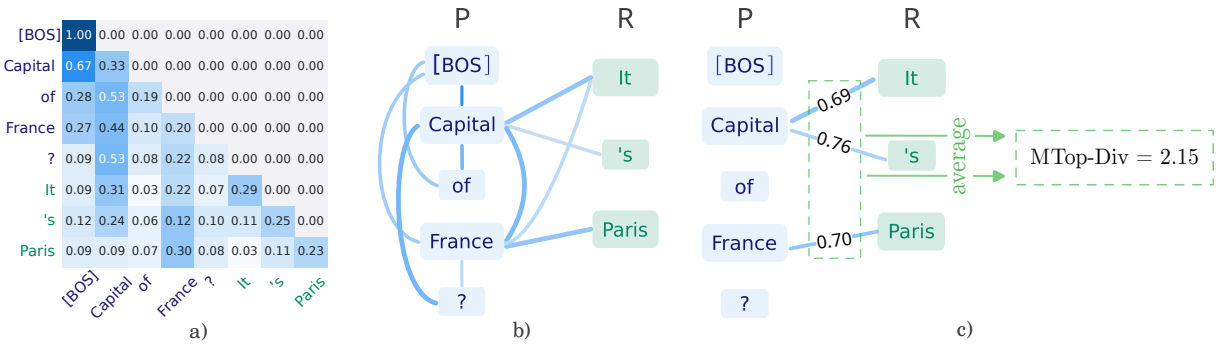

Figure 1: a) An attention map. Blue and green denote the prompt and response tokens, respectively. b) The corresponding attention graph G. Prompt tokens P are located on the left, response tokens R — on the right. To keep the figure neat, we only plot edges with an attention score of at least 0.15. c) The minimum spanning forest attaching R to P and the corresponding MTop-Div value.

subgraphs in attention maps. We demonstrate that not only can it be interpreted as a geometric characteristic of the attention graph, but also as an information-theoretic measure of the novelty of the response in relation to the prompt. With $\text{MTop-Div}_G(R, P)$, we identify "hallucination-aware" heads that consistently separate hallucinated samples from grounded, irrespective of the dataset. This finding allows us to formulate the TOHA algorithm (Algorithm 1), which computes a final hallucination score by averaging the divergence values from these specific heads.

3.1 MTop-Div for Attention Graphs: Definition

Let R and P be the response and prompt vertex sets in an attention graph G. After zeroing edge weights between the P nodes, we compute the 0-th order homology barcode $\mathcal{B}_0$ of the Vietoris-Rips simplicial complex of the modified graph. Essentially, this barcode tracks the evolution of connected components as we threshold the graph edges (see Appendix B for details).

Our proposed topological divergence is a sum of the lengths of $\mathcal{B}_0$ intervals:

$$\text{MTop-Div}_G(R, P) = \sum_{[b_i, d_i] \in \mathcal{B}_0} |d_i - b_i|.$$

This score can be interpreted from two perspectives: geometric (as the length of a minimal spanning forest in the attention graph) and information-theoretic (as a measure of the response novelty in the space induced by query-key matrices). We explore these two interpretations in the following subsections.

3.2 $\text{MTop-Div}_G(R, P)$ and the Geometry of the Attention Graph

Formally, we can prove the following property (see proof in Appendix C):

Proposition 3.1. *Consider an attention graph G with vertex set V_G and its complementary vertex subsets P, R, where $P \cup R = V_G$ and $P \cap R = \varnothing$. $\text{MTop-Div}(R, P)$ value equals the length of the minimal spanning forest (MSF) attaching R to P.*

This proposition's geometric meaning is illustrated in Figure 1. While the connections between prompt nodes (Figure 1b) are semantically and syntactically meaningful, we hypothesize they primarily introduce noise for hallucination detection (see Section 4.4 for the corresponding experiment). After setting these distances to zero, we construct an MSF on the modified graph (Figure 1c). The sum of the lengths of the remaining edges in this MSF is precisely the value of $\text{MTop-Div}_G(R, P)$.

3.3 $\text{MTop-Div}_G(R, P)$ as a Topology-Based Novelty Score

Consider an attention graph G. It can be interpreted as a non-metric space, with "distances" induced by the attention weights. By Proposition 3.1, $\text{MTop-Div}_G(R, P)$ equals the length of the minimum spanning forest attaching the response tokens R to the prompt tokens P. Therefore,

$$\text{MTop-Div}_G(R, P) \geq L_{\text{MST}}(R \cup P) - L_{\text{MST}}(P),$$

since adding an MST on P to such a forest yields a spanning tree on $R \cup P$.

This shows that $\text{MTop-Div}_G(R, P)$ is at least the increase in MST length obtained when the response tokens are added to the prompt tokens. Re-

call that MST length is commonly used as a proxy for geometric dispersion (Müller et al., 2012), more precisely, the entropy estimate given the MST of length L on points $\mathcal{X}$ is the following:

$$H_{MST}(\mathcal{X}) = d \log L - (d-1) \log n + \log \beta_d,$$

where $|\mathcal{X}| = n$, d is the intrinsic dimensionality of the data, and β_d is some data-independent constant. Therefore larger values of $\text{MTop-Div}_G(R, P)$ can be interpreted as indicating greater structural novelty of the response relative to the prompt. Thus, our $\text{MTop-Div}_G(R, P)$ should be an effective statistic for identifying hallucinated responses.

3.4 Hallucination-Aware Heads

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:

$$\Delta_{ij} = \frac{1}{|S_{\text{hallu}}|} \sum_{s \in S_{\text{hallu}}} d_{ij}(s) - \frac{1}{|S_{\text{gr}}|} \sum_{s \in S_{\text{gr}}} d_{ij}(s), \tag{3}$$

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

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

Figure 2 displays sample differences Δ_{ij} across three datasets, with each marker representing some attention head. We discovered that the same four (for Mistral-7B) and three (for Llama-2-7B) heads, highlighted in pink, demonstrate similar behavior across the datasets: they consistently appear in the upper-right corner, indicating strong separation between hallucinated and grounded samples, irrespective of the dataset. This finding suggests that there are specific attention heads somehow "aware" of the presence of hallucination, which is captured by the proposed $\text{MTop-Div}_G(R, P)$.

3.5 TOHA

The existence of "hallucination-aware" attention heads underlies our method, which is detailed in Algorithm 1. TOHA uses two annotated probe sets (S_h for hallucinations, S_g for grounded samples) to rank attention heads by their separation capability Δ_{ij} and select the top N_{opt} ones with the combined probe size $|S_h \cup S_g|$ kept small (see Figure 3). During testing, the final hallucination score for test samples T equals the average topological divergence from the selected N_{opt} heads. For computational efficiency, we limit N_{opt} to a maximum of $N_{\max} = 10$ in all experiments.

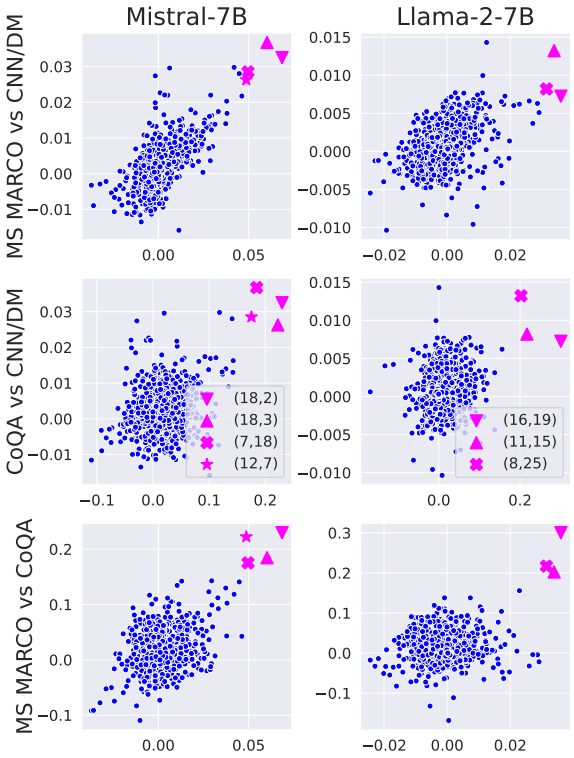

Figure 2: Δ_{ij} values for ij-th heads. Vertical axis corresponds to the difference on dataset (B), horizontal — to the one on dataset (A). The heads that separate samples best are highlighted in pink; their (layer, head) positions are reflected in the legend.

4 Results

4.1 Experiment Setting

Datasets and models. We consider five tasks that evaluate question answering and summarization abilities of LLMs: two benchmarks of RAGTruth (Niu et al., 2023) — long-form QA dataset MS MARCO and summarization dataset CNN/DM; conversational QA dataset CoQA (Reddy et al., 2019), reading comprehension dataset SQuAD (Rajpurkar et al., 2016), and extreme summarization dataset XSum (Narayan et al., 2018). For more details, see Appendix D.

We employ five popularly adopted open-source LLMs with accessible inner states: LLaMA-2-7B-chat, LLaMA-2-13B-chat, LLaMA-3.1-8B-Instruct, Mistral-7B-Instruct-v0.1, and Qwen2.5-

Algorithm 1 TOHA Algorithm

Require: $d_{ij}(s)$, S_h, S_g, $V = S_h \cup S_g$, T, $N_{\max}$
Ensure: Hallucination scores $\{p_s\}$ for $s \in T$

function HEADSSELECTION
  for each head h_{ij} **do**
    Calculate Δ_{ij} according to eq. (3)
  end for
  $H \leftarrow$ sort all heads by Δ_{ij} (descending)
  $N_{\text{opt}}, \text{AUROC}_{\max} \leftarrow 1, 0$
  Initialize $p_s \leftarrow 0$ for all $s \in V$
  for $N = 1$ **to** $N_{\max}$ **do**
    for each $s \in V$ **do**
      $p_s \leftarrow \frac{N-1}{N} p_s + \frac{1}{N} d_{h_N}(s)$
    end for
    auroc $\leftarrow \text{AUROC}(\{y_s\}_{s\in V}, \{p_s\}_{s\in V})$
    if auroc $> \text{AUROC}_{\max}$ **then**
      $\text{AUROC}_{\max} \leftarrow$ auroc; $N_{\text{opt}} \leftarrow N$
    end if
  end for
  return N_{opt}
end function

function PREDICTION(N_{opt})
  for each $s \in T$ **do**
    $p_s \leftarrow \frac{1}{N_{\text{opt}}} \sum_{i=1}^{N_{\text{opt}}} d_{h_i}(s)$
  end for
  return $\{p_s\}$
end function

7B-Instruct. As the RAGTruth dataset does not contain responses for LLaMA-3.1-8B and Qwen-2.5-7B, we conducted experiments on SQuAD, CoQA, and XSum for these models.

Baselines. We compare TOHA with a comprehensive set of eight baselines: uncertainty-based perplexity (Ren et al., 2023) and max entropy (Fadeeva et al., 2024); inner states-based ReDeEP (Sun et al., 2025), HaloScope (Du et al., 2024), and LLM-Check (Sriramanan et al., 2024); consistency-based semantic entropy (Farquhar et al., 2024), EigenScore (Chen et al., 2024), and SelfCheckGPT (Manakul et al., 2024).

Implementation details. The reported results are averaged over 5 runs with different data splits, using test sets comprising 25% of the data and a fixed validation set size of 100, following the methodology of HaloScope (Du et al., 2024). To ensure a fair comparison, we consider two implementations of the latter: a standard setting with 20 generations (see Tables 6-7), and an efficiency-comparable one with minimal number of generations — 1 for SelfCheckGPT, 5 for semantic entropy and EigenScore (Tables 1-2). Appendix E provides additional implementation details.

4.2 Results

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 TOHA significantly outperforms uncertainty-based baselines and matches the quality of consistency-based approaches, achieving notable improvements of 11.7% on the challenging long-form QA MS MARCO dataset for Mistral-7B and 21.6% on the conversational QA dataset CoQA for LLama-2-7B.

We rigorously validate the statistical significance of our results using a critical difference diagram with Wilcoxon-Holm post-hoc analysis (Ismail Fawaz et al., 2019). The results provided in Tables 12-13 (Appendix A) confirm that while many baseline methods remain statistically indistinguishable from one another, TOHA achieves the top overall rank (1.67) and its performance improvements are statistically significant compared to every other evaluated method ($p \leq 0.0016$).

To evaluate TOHA's robustness to data distribution change, we conducted transfer experiments on Mistral-7B, see Figure 3(c) for the results. TOHA exhibits strong transferability: across the XSum and CNN/DM datasets, performance falls within the method's standard deviation (see the statistical significance analysis in Table 14), while remaining competitive on the other datasets (Table 1).

Evaluation on a multi-hop dataset. To validate TOHA in a more realistic setting, we consider an additional experiment using the HotpotQA (Yang et al., 2018) dataset. It consists of questions that require knowledge from multiple supporting documents — much like real-world queries, which rarely rely on a single source of information. The results are provided in Table 3. TOHA demonstrates superior performance to baselines, confirming its effectiveness "in the wild".

Efficiency. Figure 4(a) shows that TOHA is approximately seven times faster than SelfCheckGPT with a *single* additional generation. Given that SelfCheckGPT typically requires 10-20 generations, this makes TOHA over 70 times faster in practice. Furthermore, TOHA's runtime is close to the

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]	✗	0.63 ± 0.04	0.51 ± 0.04	0.86 ± 0.02	0.71 ± 0.04	0.66 ± 0.04
Semantic entropy [2]	✗	0.54 ± 0.03	0.51 ± 0.04	0.83 ± 0.02	0.70 ± 0.03	0.56 ± 0.03
EigenScore [3]	✗	0.54 ± 0.04	0.50 ± 0.06	0.74 ± 0.02	0.71 ± 0.04	0.58 ± 0.04
HaloScope [4]	✓	0.57 ± 0.08	0.51 ± 0.10	0.62 ± 0.08	0.92 ± 0.07	0.62 ± 0.02
LLM-Check [5]	✓	0.49 ± 0.03	0.49 ± 0.03	0.60 ± 0.01	0.50 ± 0.03	0.58 ± 0.04
Perplexity [6]	✓	0.45 ± 0.01	0.54 ± 0.02	0.54 ± 0.03	0.81 ± 0.05	0.54 ± 0.06
Max entropy [7]	✓	0.68 ± 0.04	**0.60 ± 0.07**	0.73 ± 0.00	0.75 ± 0.05	**0.71 ± 0.02**
ReDEEP [8]	✓	0.54 ± 0.02	0.47 ± 0.06	0.59 ± 0.03	0.45 ± 0.05	0.63 ± 0.04
TOHA (ours)	✓	**0.76 ± 0.04**	**0.60 ± 0.09**	**0.89 ± 0.01**	**0.96 ± 0.01**	0.66 ± 0.05
LLama-2-7B						
SelfCheckGPT [1]	✗	0.59 ± 0.03	**0.60 ± 0.03**	0.66 ± 0.03	0.57 ± 0.03	0.64 ± 0.05
Semantic entropy [2]	✗	0.53 ± 0.03	0.51 ± 0.03	0.76 ± 0.01	0.73 ± 0.03	0.61 ± 0.04
EigenScore [3]	✗	0.55 ± 0.03	0.53 ± 0.04	0.61 ± 0.03	0.75 ± 0.02	0.63 ± 0.02
HaloScope [4]	✓	0.51 ± 0.05	0.48 ± 0.05	0.61 ± 0.04	0.67 ± 0.04	0.57 ± 0.07
LLM-Check [5]	✓	0.44 ± 0.02	0.49 ± 0.06	0.60 ± 0.03	0.49 ± 0.01	0.61 ± 0.01
Perplexity [6]	✓	0.54 ± 0.04	0.44 ± 0.03	0.74 ± 0.02	0.46 ± 0.03	0.56 ± 0.09
Max entropy [7]	✓	**0.65 ± 0.04**	0.59 ± 0.06	0.65 ± 0.03	0.73 ± 0.04	0.56 ± 0.03
ReDEEP [8]	✓	0.54 ± 0.04	0.52 ± 0.04	0.72 ± 0.04	0.42 ± 0.08	0.54 ± 0.06
TOHA (ours)	✓	**0.65 ± 0.02**	0.56 ± 0.02	**0.90 ± 0.01**	**0.87 ± 0.04**	**0.68 ± 0.05**
LLaMA-2-13B						
SelfCheckGPT [1]	✗	0.58 ± 0.04	**0.58 ± 0.05**	0.77 ± 0.02	0.64 ± 0.03	0.60 ± 0.04
Semantic entropy [2]	✗	0.57 ± 0.04	0.54 ± 0.03	0.76 ± 0.04	0.65 ± 0.03	0.60 ± 0.03
EigenScore [3]	✗	0.56 ± 0.04	0.47 ± 0.04	0.57 ± 0.03	0.57 ± 0.02	0.52 ± 0.06
HaloScope [4]	✓	0.54 ± 0.09	0.51 ± 0.04	0.57 ± 0.03	0.55 ± 0.02	0.55 ± 0.07
LLM-Check [5]	✓	0.49 ± 0.06	0.56 ± 0.05	0.57 ± 0.02	0.57 ± 0.07	0.57 ± 0.07
Perplexity [6]	✓	0.54 ± 0.04	0.46 ± 0.07	0.62 ± 0.03	0.45 ± 0.02	0.49 ± 0.05
Max entropy [7]	✓	0.62 ± 0.03	0.53 ± 0.06	0.66 ± 0.03	0.78 ± 0.02	0.59 ± 0.04
ReDEEP [8]	✓	0.62 ± 0.06	0.48 ± 0.05	0.73 ± 0.02	0.48 ± 0.07	0.58 ± 0.08
TOHA (ours)	✓	**0.67 ± 0.04**	0.56 ± 0.05	**0.92 ± 0.02**	**0.88 ± 0.05**	**0.66 ± 0.03**

lightweight entropy baseline; with a more efficient low-level implementation, it has the potential to approach the cost of a single forward pass while achieving higher accuracy than other inexpensive techniques.

4.3 Analysis of Hallucination-Aware Attention Heads

Attention patterns and copying behavior. For hallucination-aware heads, we analyzed MSF's patterns that distinguish hallucinated from grounded samples. A key finding is that for these heads, hallucinated samples frequently exhibit strong attention to the first token, whereas grounded samples tend to attend to the first token less. An example is provided in Figure 5. Since this pattern is a known behavior of *induction*, or *token copying* heads (Feucht et al., 2025) — which default to the first token when unable to find previous occurrences of the current token pattern in the context (Elhage et al., 2021) — we decided to explore the relationship between these special heads and hallucination-aware ones. We ranked all attention heads in LLama-2-7B and Mistral-7B based on their token copying (induction) scores, following the method of (Feucht et al., 2025). For the subset of heads frequently identified by TOHA (appearing in $\geq 20\%$ of runs across all datasets), their copying ranks were recorded. The results (Figure 6) reveal that hallucination-aware heads are often also among the model's top-25 copiers, a finding that aligns with prior work (Sun et al., 2025).

4.4 Ablation Studies

Why zero out the distances between prompt tokens? We hypothesize that the connections

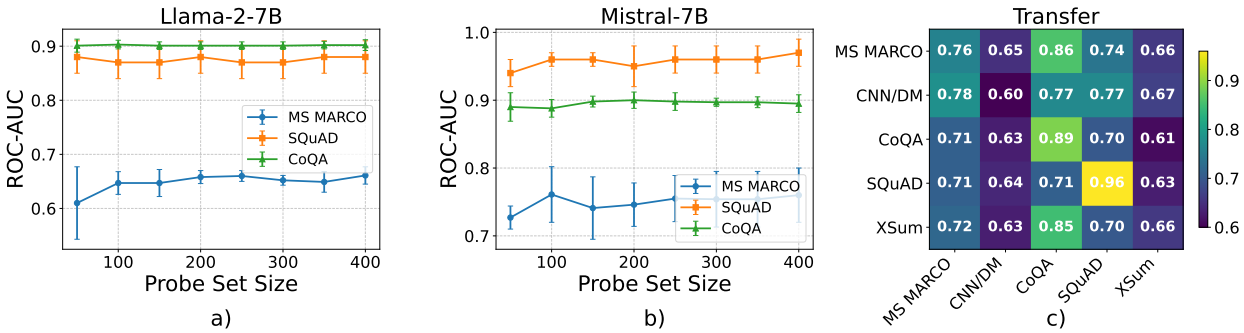

Figure 3: (a)-(b): Detection quality dependence on the size of a probe set, models: Mistral-7B (left), LLama-2-7B (right). (c) Generalizability between the datasets, model: Mistral-7B. The vertical axis corresponds to the origin of the probe set, and the horizontal axis to the test dataset.

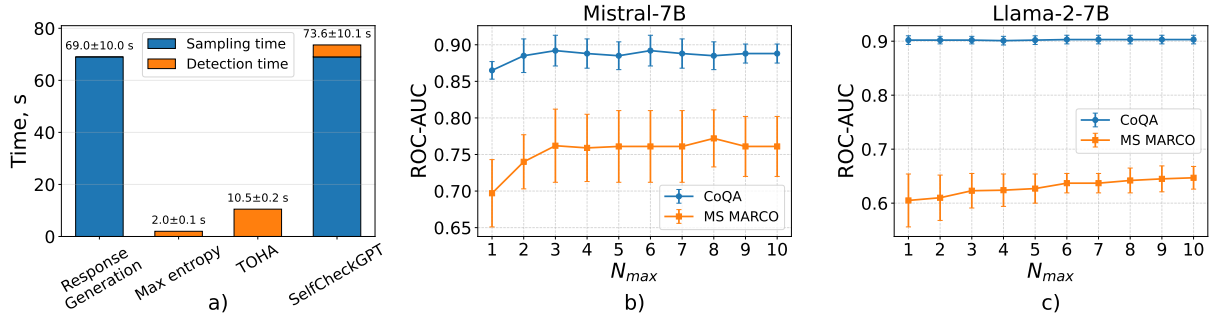

Figure 4: a): Inference time comparison (seconds) for various methods evaluated on 16 MS MARCO samples using Mistral-7B. SelfCheckGPT measurement includes one additional generated answer per sample. b)-c): ROC-AUC performance of TOHA across different numbers of selected attention heads (N_{max}) on Mistral-7B and Llama-2-7B.

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.	CoQA	SQuAD	XSum
LLaMA-3.1-8B				
SelfCheckGPT [1]	✗	**0.91 ± 0.01**	0.65 ± 0.05	**0.68 ± 0.05**
Semantic entropy [2]	✗	0.78 ± 0.02	0.55 ± 0.07	0.47 ± 0.04
EigenScore [3]	✗	0.80 ± 0.02	0.45 ± 0.05	0.50 ± 0.03
HaloScope [4]	✓	0.67 ± 0.08	0.84 ± 0.07	0.55 ± 0.06
LLM-Check [5]	✓	0.54 ± 0.06	0.49 ± 0.05	0.52 ± 0.04
Perplexity [6]	✓	0.51 ± 0.05	0.68 ± 0.02	0.65 ± 0.03
Max entropy [7]	✓	0.82 ± 0.03	0.60 ± 0.02	0.65 ± 0.03
ReDEEP [8]	✓	0.58 ± 0.08	0.39 ± 0.04	0.62 ± 0.06
TOHA (ours)	✓	0.84 ± 0.01	**0.87 ± 0.03**	0.65 ± 0.05
Qwen2.5-7B				
SelfCheckGPT [1]	✗	0.69 ± 0.01	0.74 ± 0.02	**0.69 ± 0.02**
Semantic entropy [2]	✗	0.76 ± 0.04	0.57 ± 0.04	0.64 ± 0.03
EigenScore [3]	✗	0.78 ± 0.03	0.55 ± 0.04	0.55 ± 0.04
HaloScope [4]	✓	0.66 ± 0.13	0.75 ± 0.04	0.57 ± 0.07
LLM-Check [5]	✓	0.53 ± 0.07	0.54 ± 0.06	0.54 ± 0.02
Perplexity [6]	✓	0.39 ± 0.03	0.65 ± 0.03	0.53 ± 0.05
Max entropy [7]	✓	**0.85 ± 0.02**	0.47 ± 0.05	0.60 ± 0.06
ReDEEP [8]	✓	0.37 ± 0.10	0.56 ± 0.03	0.67 ± 0.03
TOHA (ours)	✓	0.79 ± 0.05	**0.77 ± 0.02**	**0.69 ± 0.03**

within the prompt contribute little to hallucination detection; therefore, our method is designed to disregard them. To validate this architectural choice, we considered an ablation study, where $\mathrm{MTop\text{-}Div}_G(R, P)$ was replaced by the MST length of the complete graph in Algorithm 1. The

Table 3: ROC-AUC (↑) values on the HotpotQA dataset. Best results are highlighted in **bold**, and the second best are underlined.

Method	Single gen.	Mistral-7B	Llama-2-13B
SelfCheckGPT	✗	0.70 ± 0.06	0.63 ± 0.04
Semantic entropy	✗	0.70 ± 0.05	0.70 ± 0.04
EigenScore	✗	0.68 ± 0.04	0.67 ± 0.04
HaloScope	✓	0.60 ± 0.06	0.50 ± 0.01
LLM-Check	✓	0.48 ± 0.03	0.56 ± 0.04
Perplexity	✓	0.55 ± 0.06	0.49 ± 0.04
Max entropy	✓	0.62 ± 0.04	0.69 ± 0.05
ReDEEP	✓	0.49 ± 0.04	0.62 ± 0.05
TOHA (ours)	✓	**0.71 ± 0.08**	**0.80 ± 0.03**

results showing the effectiveness of the proposed approach are in Table 4.

What about other attention map-based features? To demonstrate that our topology-based approach offers an advantage over conventional attention map features, we compared them to MTop-Div in the supervised setting. Table 9 (Appendix A) shows that a classifier trained on $\mathrm{MTop\text{-}Div}_G(R, P)$ achieves the best performance, confirming that our score captures unique patterns in an attention map that standard approaches miss.

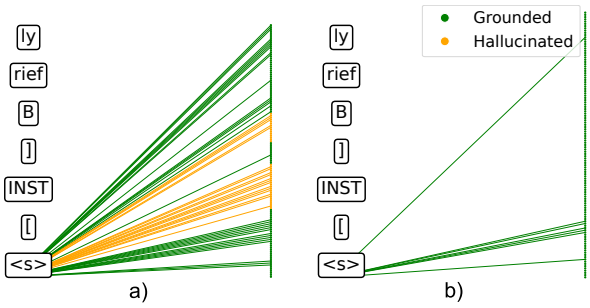

Figure 5: Attention to the first token (<s> in this example) for (a) a hallucinated generation and (b) a grounded one. Green highlights edges and nodes corresponding to grounded tokens, while yellow indicates hallucinated tokens. Model: Mistral-7B.

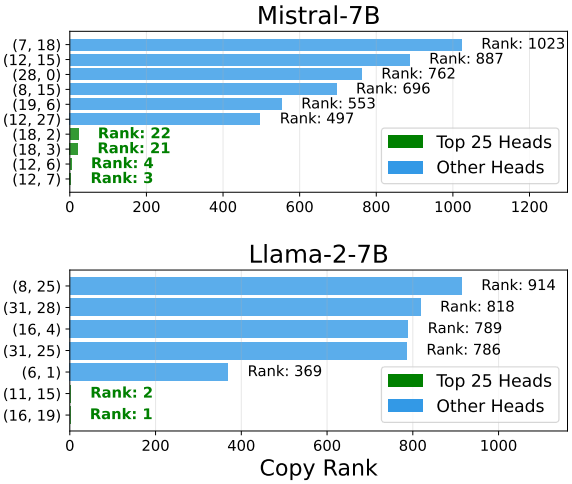

Figure 6: Copying ranks of hallucination-aware attention heads (lower ranks indicate stronger copying behavior). The row labels (X, Y) correspond to X-th layer and Y-th head.

How large should the probe sets be? To evaluate TOHA's sensitivity to probe set size, we conducted a sensitivity study. The results shown in Figure 3 confirm TOHA's robustness: even with only 50 samples, performance does not drop significantly and remains mostly stable as the probe set size increases.

Can TOHA operate without any annotated data? As demonstrated in Section 4.3, many of the attention heads naturally selected by TOHA exhibit strong inductive (token copying) behavior. Crucially, identifying these heads does not require any labeled hallucination data (Feucht et al., 2025). To evaluate TOHA in an extreme zero-shot setting, we completely bypassed the probe set selection and simply averaged the divergence scores from the model's top-4 copying heads ($\text{TOHA}_{\text{Copy Heads}}$).

Table 4: ROC-AUC (↑) of Algorithm 1 with MST length of the complete graph vs $\text{MTop-Div}_G(R, P)$. Best results are highlighted in **bold**.

Dataset	MST length	$\text{MTop-Div}_G(R, P)$
Mistral-7B		
CoQA	0.75 ± 0.03	$\mathbf{0.90 \pm 0.01}$
MS MARCO	0.38 ± 0.03	$\mathbf{0.65 \pm 0.02}$
LLama-2-7B		
CoQA	0.60 ± 0.03	$\mathbf{0.89 \pm 0.01}$
MS MARCO	0.37 ± 0.02	$\mathbf{0.76 \pm 0.04}$

The results in Table 5 confirm that this fully unsupervised variant is highly competitive, demonstrating that TOHA can be effectively deployed even when zero annotated data is available.

How many attention heads do we need? To evaluate the sensitivity of our method to the hyperparameter N_{max}, we performed an ablation study for values from 1 to 10. As shown in Figure 4, TOHA achieves strong detection performance even when $N_{\max} = 1$, which underscores the effectiveness of our topological approach.

5 Related Works

Existing hallucination detection methods face strict trade-offs (Zhang et al., 2023; Huang et al., 2023; Wang et al., 2024). Consistency-based approaches (Manakul et al., 2024; Chen et al., 2024; Qiu and Miikkulainen, 2024; Nikitin et al., 2024; Hou et al., 2025) are robust but computationally expensive due to multiple generation passes. Surface-level metrics like perplexity (Fadeeva et al., 2024; Malinin and Gales, 2021) are efficient yet limited, as they ignore the model's rich internal representations (Gekhman et al., 2025). Conversely, hidden-state classifiers (Azaria and Mitchell, 2023; Sky et al., 2024; Zhou et al., 2025) capture these representations but require extensive annotated data and suffer from poor task transferability (Sky et al., 2024). Even semi-supervised alternatives like HaloScope (Du et al., 2024) still demand large volumes of unannotated outputs.

Attention map-based methods represent a promising yet underdeveloped direction. Current techniques either (i) rely on large annotated datasets (Chuang et al., 2024; Binkowski et al., 2025), (ii) exploit only simplistic attention graph properties like self-loop weights (Sriramanan et al., 2024), or (iii) disregard attention map geometry entirely, using mechanistic scores instead (Sun et al.,

Method	Single generation	LLaMA-2-7B		Mistral-7B	
		CoQA	MS MARCO	CoQA	MS MARCO
SelfCheckGPT [1]	✗	0.66 ± 0.03	0.59 ± 0.03	$\mathbf{0.86 \pm 0.02}$	0.63 ± 0.04
Semantic entropy [2]	✗	$\underline{0.76 \pm 0.01}$	0.53 ± 0.03	$\underline{0.83 \pm 0.02}$	0.54 ± 0.03
EigenScore [3]	✗	0.61 ± 0.03	0.55 ± 0.03	0.74 ± 0.02	0.54 ± 0.04
HaloScope [4]	✓	0.61 ± 0.04	0.51 ± 0.05	0.62 ± 0.08	0.57 ± 0.08
LLM-Check [5]	✓	0.60 ± 0.03	0.44 ± 0.02	0.60 ± 0.01	0.49 ± 0.03
Perplexity [6]	✓	0.74 ± 0.02	0.54 ± 0.04	0.54 ± 0.03	0.45 ± 0.01
Max entropy [7]	✓	0.65 ± 0.03	$\mathbf{0.65 \pm 0.04}$	0.73 ± 0.04	$\mathbf{0.68 \pm 0.04}$
ReDeEP [8]	✓	0.72 ± 0.04	0.54 ± 0.04	0.59 ± 0.03	0.54 ± 0.02
$\text{TOHA}_{\text{Copy Heads}}$ (ours)	✓	$\mathbf{0.89 \pm 0.01}$	$\underline{0.62 \pm 0.04}$	0.77 ± 0.01	$\underline{0.66 \pm 0.04}$

Table 5: ROC AUC (↑) evaluation of TOHA using only the top-4 copying heads (zero annotated data) compared to baselines. The best results in each column are highlighted in **bold**, and the second best are underlined.

2025). Thus, training-free methods that fully leverage the rich structural information encoded in attention graphs remain underexplored.

6 Conclusion

This paper introduces TOHA, a novel hallucination detection method based on the topological structure of attention maps. Central to TOHA is the $\text{MTop-Div}_G(R, P)$, our adaptation of Manifold Topology Divergence (Barannikov et al., 2021) for the graph setting, which, as we demonstrate, serves as a measure of the response novelty in relation to the prompt. This property makes the proposed divergence well-suited for hallucination detection in RAG scenarios.

By analyzing divergence values, we identified a subset of "hallucination-aware" attention heads that reliably distinguish hallucinated from grounded samples across datasets. TOHA computes the final hallucination scores by averaging the topological divergences from these heads. Further investigation reveals that some of these heads are associated with copying behavior, which aligns with prior work (Sun et al., 2025).

Extensive experiments show that TOHA is a robust alternative to existing approaches, matching or surpassing state-of-the-art baselines. Moreover, our method is both data- and compute-efficient: just 50 annotated samples suffice for reliable detection, and inference runs several times faster than comparable methods of similar quality. Crucially, we validate TOHA's transferability, demonstrating its robustness to shifts in data distribution — a key 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 well-suited for practical applications.

Acknowledgements

The authors wish to thank our colleague Ilya Kuleshov for his time and the many insightful discussions regarding this research. The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4F0002 and the agreement №139-10-2025-033.

Limitations

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

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.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. 2025. Hallucination detection in llms using spectral features of attention maps. *arXiv preprint arXiv:2502.17598*.

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. 2021. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963.

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*.

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.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

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.

Sheridan Feucht, Eric Todd, Byron C Wallace, and David Bau. 2025. The dual-route model of induction. In *Second Conference on Language Modeling*.

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.

Zorik Gekhman, Eyal Ben-David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in LLMs. In *Second Conference on Language Modeling*.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. 2025. A probabilistic framework for llm hallucination detection via belief tree propagation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3076–3099.

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*.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963.

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.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Andreas C. Müller, Sebastian Nowozin, and Christoph H. Lampert. 2012. *Information Theoretic Clustering Using Minimum Spanning Trees*, page 205–215. Springer Berlin Heidelberg.

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. Association for Computational Linguistics.

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*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *ICLR*.

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.

Artem Shelmanov, Maxim Panov, Roman Vashurin, Artem Vazhentsev, Ekaterina Fadeeva, and Timothy Baldwin. 2025. Uncertainty quantification for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4, Vienna, Austria. Association for Computational Linguistics.

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.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).

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*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *ICLR*.

Christopher Tralie, Nathaniel Saul, and Rann Bar-On. 2018. Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, 3(29):925.

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.

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*.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and 1 others. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *arXiv preprint arXiv:2505.20045*.

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.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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 Additional Experimental Results

A.1 Multiple Generations-based Methods

To provide a complete comparison, we considered an implementation of consistency-based methods with 20 additional generations. The results are demonstrated in Tables 6-7. We can see that TOHA still remains superior to the baselines, achieving top performance in most experiments.

A.2 Comparison with ReDeEP

To provide a comprehensive comparison with ReDeEP (Sun et al., 2025), which was evaluated on the entire RAGTruth dataset in the original paper, we conducted a similar evaluation of TOHA under the same conditions. As shown in Table 8, TOHA not only outperforms ReDeEP but also demonstrates greater robustness to data distribution shifts, exhibiting less significant performance degradation on separate benchmarks of RAGTruth (Tables 1-2).

A.3 Alternative Attention-based Features for Hallucination Detection

During preliminary experiments for an attention map-based hallucination detector, we evaluated a range of topological and traditional features. As standard topological features, we employed 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 (Kushnareva et al., 2021). As for traditional attention-based features, we

Table 6: ROC AUC (↑) of multiple generations-based methods (with 20 additional samples) and TOHA. 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]	✗	0.67 ± 0.03	0.59 ± 0.04	**0.93 ± 0.01**	0.83 ± 0.02	**0.71 ± 0.04**
Semantic entropy [2]	✗	0.53 ± 0.03	0.52 ± 0.03	0.86 ± 0.01	0.74 ± 0.01	0.63 ± 0.02
EigenScore [3]	✗	0.49 ± 0.03	0.53 ± 0.05	0.78 ± 0.02	0.77 ± 0.04	0.59 ± 0.05
TOHA (ours)	✓	**0.76 ± 0.04**	**0.60 ± 0.09**	0.89 ± 0.01	**0.96 ± 0.01**	0.66 ± 0.05
LLama-2-7B						
SelfCheckGPT [1]	✗	0.60 ± 0.04	**0.60 ± 0.03**	0.77 ± 0.02	0.78 ± 0.02	0.67 ± 0.04
Semantic entropy [2]	✗	0.56 ± 0.03	0.49 ± 0.03	0.79 ± 0.01	0.77 ± 0.02	0.63 ± 0.04
EigenScore [3]	✗	0.57 ± 0.04	0.52 ± 0.06	0.61 ± 0.03	0.78 ± 0.02	0.65 ± 0.03
TOHA (ours)	✓	**0.65 ± 0.02**	0.56 ± 0.02	**0.90 ± 0.01**	**0.87 ± 0.04**	**0.68 ± 0.05**
LLaMA-2-13B						
SelfCheckGPT [1]	✗	0.61 ± 0.05	**0.60 ± 0.06**	0.88 ± 0.02	0.75 ± 0.04	0.61 ± 0.04
Semantic entropy [2]	✗	0.60 ± 0.03	0.52 ± 0.03	0.77 ± 0.04	0.70 ± 0.02	0.62 ± 0.03
EigenScore [3]	✗	0.58 ± 0.04	0.48 ± 0.05	0.59 ± 0.03	0.60 ± 0.02	0.54 ± 0.05
TOHA (ours)	✓	**0.67 ± 0.04**	0.56 ± 0.05	**0.92 ± 0.02**	**0.88 ± 0.05**	**0.66 ± 0.03**

Table 7: ROC AUC (↑) of multiple generations-based methods (with 20 additional samples) and TOHA. The best results for each model are highlighted in **bold**, and the second best are underlined.

Method	Single gen.	CoQA	SQuAD	XSum
LLaMA-3.1-8B				
SelfCheckGPT [1]	✗	**0.95 ± 0.01**	0.78 ± 0.03	**0.75 ± 0.03**
Semantic entropy [2]	✗	0.82 ± 0.03	0.54 ± 0.06	0.46 ± 0.05
EigenScore [3]	✗	0.84 ± 0.01	0.56 ± 0.03	0.48 ± 0.03
TOHA (ours)	✓	0.84 ± 0.01	**0.87** ± 0.03	0.65 ± 0.05
Qwen2.5-7B				
SelfCheckGPT [1]	✗	0.75 ± 0.02	**0.77 ± 0.02**	**0.72 ± 0.03**
Semantic entropy [2]	✗	0.83 ± 0.05	0.58 ± 0.04	0.69 ± 0.05
EigenScore [3]	✗	0.83 ± 0.05	0.56 ± 0.03	0.48 ± 0.03
TOHA (ours)	✓	0.79 ± 0.05	**0.77 ± 0.02**	0.69 ± 0.03

Table 8: Performance comparison between ReDEEP and TOHA on the entire RAGTruth dataset.

Method	Llama-2-7B	Llama-2-13B
ReDeEP [8]	0.68 ± 0.02	0.77 ± 0.01
TOHA (ours)	**0.70 ± 0.04**	**0.80 ± 0.02**

Table 9: ROC-AUC values of supervised classifiers on top of various sets 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		
Standard 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**

used sparsity ratio, attention entropy, and spectral norm (Kobayashi et al., 2020; Vig and Belinkov, 2019; Ji et al., 2021). We also considered Wasserstein distances between the persistent diagrams (Chazal and Michel, 2021) of the prompt and response subgraphs as an alternative measure of their similarity. Finally, we analyzed average attention to the first token and found that hallucination-aware heads often attend in the presence of hallucinations (see Section 4).

To identify the most informative features for hallucination detection, we trained $L1$-regularized supervised classifiers on concatenated features from all layers and heads and compared them to classifiers trained only on $\mathrm{MTop\text{-}Div}_G(R, P)$ values under the same conditions. The results are presented in Table 9.

While the classifier based on $\mathrm{MTop\text{-}Div}_G(R, P)$ 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 that aggregates divergence values from only a subset of the "hallucination-aware" attention heads.

A.4 Evaluation of Alternative Head Selection Metrics

Additionally, we investigated alternative attention map-based scores, such as entropy, spectral norm, and the Wasserstein distance between the persistent diagrams of prompts and responses, for selecting specialized attention heads. Following the pipeline of the Algorithm 1, we computed the average distances between hallucinated and grounded samples using alternative scores. The results, presented in Figure 10, reveal that the heads that separate samples best for MS MARCO do not generalize to the CoQA dataset. This suggests that our proposed $\mathrm{MTop\text{-}Div}_G(R, P)$ metric is better suited for the task compared to existing solutions.

A.5 Feature-Level Comparison with LookBack Lens

To further validate the representational power of our proposed $\mathrm{MTop\text{-}Div}_G(R, P)$ features, we compared them against the attention-ratio features introduced in Lookback Lens (Chuang et al., 2024). As Lookback Lens is fundamentally a supervised method, we established a fair, comparable setting by isolating its attention-ratio features and evaluating them using the same attention-head selection procedure as TOHA. The results of this comparison are provided in Table 10. TOHA consistently outperforms the attention-ratio features across all evaluated models and datasets, confirming the superior effectiveness of topological features over standard attention metrics for hallucination detection.

Table 10: Performance comparison between TOHA and attention-ratio features (derived from LookBack Lens) evaluated under identical settings.

Method	CoQA	XSum
Llama-2-7B		
Attention ratio	0.76 ± 0.05	0.59 ± 0.05
TOHA (ours)	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.68 \pm 0.05}$
Qwen2.5-7B		
Attention ratio	0.71 ± 0.07	0.58 ± 0.04
TOHA (ours)	$\mathbf{0.79 \pm 0.05}$	$\mathbf{0.69 \pm 0.03}$

A.6 Comparison with Linear Probes under Matched Data Constraints

To ensure a fair comparison with supervised baseline approaches, we evaluated the performance of linear probes trained on datasets with exactly 100 samples—matching the probe set size used by TOHA. We conducted these experiments across all five datasets (MS MARCO, CNN/DM, CoQA, SQuAD, and XSum) using the LLaMA-2-13B model.

The corresponding results are detailed in Table 11. While linear probes achieve the second-best performance in two out of the five experiments, they are significantly outperformed by TOHA across the board. This highlights the superior sample efficiency and robustness of our topology-based method when operating under constrained data availability.

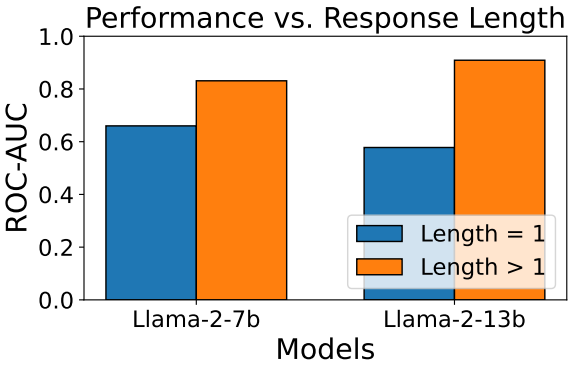

Figure 7: Comparison of ROC-AUC scores (↑) for single-word versus multi-word model responses. Dataset: SQuAD.

A.7 Performance Analysis on Out-of-Design Cases

Additionally, we analyzed TOHA's performance on extremely short responses consisting of a single word. The results are presented in Figure 7. While its performance drops, as expected when the response graph degenerates to a single vertex, it remains non-random. This indicates that TOHA retains some predictive power even for these challenging, out-of-design cases.

As for the TOHA's performance on longer responses, the results for the MS MARCO, CNN/DM, and XSum datasets demonstrate that as the response length grows, the hallucination signals become less distinguishable for all the considered methods, as hallucinated tokens comprise a very small part of the response. However, TOHA remains superior to baselines, confirming that our topology-based approach is more effective than existing methods for complex long-form responses.

Table 11: ROC AUC (↑) evaluation of a linear probe trained on exactly 100 samples alongside TOHA and other baseline methods. The best results for the model are highlighted in **bold**, and the second best are underlined.

Method	Single generation	MS MARCO	CNN/DM + Recent News	CoQA	SQuAD	XSum
LLaMA-2-13B						
SelfCheckGPT [1]	✗	0.58 ± 0.04	$\mathbf{0.58 \pm 0.05}$	0.77 ± 0.02	0.64 ± 0.03	0.60 ± 0.04
Semantic entropy [2]	✗	0.57 ± 0.04	0.54 ± 0.03	0.76 ± 0.04	0.65 ± 0.03	0.60 ± 0.03
EigenScore [3]	✗	0.56 ± 0.04	0.47 ± 0.04	0.57 ± 0.03	0.57 ± 0.02	0.52 ± 0.06
HaloScope [4]	✓	0.54 ± 0.09	0.51 ± 0.04	0.57 ± 0.03	0.55 ± 0.02	0.55 ± 0.07
LLM-Check [5]	✓	0.49 ± 0.06	0.56 ± 0.05	0.57 ± 0.05	0.57 ± 0.07	0.57 ± 0.07
Perplexity [6]	✓	0.54 ± 0.04	0.46 ± 0.07	0.62 ± 0.03	0.45 ± 0.02	0.49 ± 0.05
Max entropy [7]	✓	0.62 ± 0.03	0.53 ± 0.06	0.66 ± 0.03	0.78 ± 0.02	0.59 ± 0.04
ReDEEP [8]	✓	0.62 ± 0.06	0.48 ± 0.05	0.73 ± 0.02	0.48 ± 0.07	0.58 ± 0.08
Linear probe	✓	0.62 ± 0.03	0.55 ± 0.05	0.65 ± 0.03	0.78 ± 0.01	0.58 ± 0.05
TOHA (ours)	✓	$\mathbf{0.67 \pm 0.04}$	0.56 ± 0.05	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.88 \pm 0.05}$	$\mathbf{0.66 \pm 0.03}$

A.8 Statistical Significance of the Results

To rigorously evaluate the robustness and consistency of our results, we conducted a statistical significance analysis of the performance differences across all experimental settings. Specifically, we applied a non-parametric Wilcoxon signed-rank test followed by Holm's step-down procedure to correct for multiple comparisons. The outcomes of this post-hoc analysis are presented in Tables 12 and 13.

As detailed in Table 12, the pairwise p-values confirm that TOHA yields a statistically significant difference against every single baseline evaluated. Furthermore, the aggregate rankings in Table 13 show that TOHA achieves the best overall average rank (1.67) and stands alone at the top of the hierarchy, clearly distinguishable from all other methods.

A.9 T-test Analysis for Transferability Experiments

As previously noted, TOHA's performance on the CNN/DM and XSum datasets in transfer settings falls within the method's standard deviation. To statistically confirm this observation, we calculated p-values using a t-test for the means of two independent samples. The results, provided in Table 14, fully support our claim: all p-values are well above the 0.05 threshold, confirming that the performance variations in transfer settings are not statistically significant.

B Topological Data Analysis: Background

Simplicial complexes. A simplicial complex S is a collection of simplices such that every face of

Table 12: Pairwise p-values from the Wilcoxon-Holm post-hoc analysis. Methods are considered statistically different after thresholding the p-value by the normalized Holm correction. Notice that TOHA yields a "True" significance against every single baseline evaluated.

Method 1	Method 2	p-value	Stat. diff.
Haloscope	TOHA	9.5×10^{-7}	True
LLM-Check	SelfCK	9.5×10^{-7}	True
LLM-Check	TOHA	9.5×10^{-7}	True
ReDEEP	TOHA	9.5×10^{-7}	True
Sem. Entropy	TOHA	9.5×10^{-7}	True
EigenScore	TOHA	1.9×10^{-6}	True
Perplexity	TOHA	1.9×10^{-6}	True
Entropy	LLM-Check	5.2×10^{-5}	True
ReDEEP	SelfCK	6.7×10^{-5}	True
LLM-Check	Sem. Entropy	1.0×10^{-4}	True
Perplexity	SelfCK	2.9×10^{-4}	True
Entropy	TOHA	1.0×10^{-3}	True
EigenScore	Entropy	1.6×10^{-3}	True
SelfCK	TOHA	1.6×10^{-3}	True
Entropy	ReDeEP	2.5×10^{-3}	False
Haloscope	LLM-Check	2.9×10^{-3}	False
EigenScore	SelfCK	3.8×10^{-3}	False
Entropy	Perplexity	4.9×10^{-3}	False
Haloscope	Perplexity	1.8×10^{-2}	False
EigenScore	LLM-Check	1.9×10^{-2}	False
ReDeEP	Sem. Entropy	2.2×10^{-2}	False
SelfCK	Sem. Entropy	2.2×10^{-2}	False
Haloscope	SelfCK	2.6×10^{-2}	False
Perplexity	Sem. Entropy	4.6×10^{-2}	False
Entropy	Haloscope	5.0×10^{-2}	False
Entropy	Sem. Entropy	1.7×10^{-1}	False
EigenScore	Sem. Entropy	2.6×10^{-1}	False
Haloscope	Sem. Entropy	3.2×10^{-1}	False
Haloscope	ReDeEP	3.4×10^{-1}	False
EigenScore	Perplexity	3.7×10^{-1}	False
EigenScore	ReDeEP	3.9×10^{-1}	False
LLM-Check	ReDeEP	4.1×10^{-1}	False
Perplexity	ReDeEP	4.5×10^{-1}	False
LLM-Check	Perplexity	5.9×10^{-1}	False
Entropy	SelfCK	8.9×10^{-1}	False
EigenScore	Haloscope	1.0×10^{0}	False

Table 13: Average rank of each method across evaluated settings and the number of baselines from which it is statistically indistinguishable (derived from the Wilcoxon-Holm post-hoc analysis). Lower values (↓) indicate better performance. TOHA achieves the top rank and is statistically distinct from all other evaluated approaches.

Method	Rank ↓	# Indistinguishable ↓
TOHA (Ours)	**1.67**	**0**
SelfCK	3.00	3
Entropy	3.62	3
Semantic Entropy	4.95	3
HaloScope	5.71	3
EigenScore	5.74	3
ReDEEP	6.38	2
Perplexity	6.81	2
LLM-Check	7.12	0

Table 14: p-values of the independent two-sample t-test for the transferability experiments.

Probe Set	CNN/DM	XSum
MS MARCO	0.272	1.000
CNN/DM	1.000	0.746
CoQA	0.499	0.134
SQuAD	0.373	0.346
XSum	0.515	1.000

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 discrete analogs of topological spaces.

Vietoris-Rips simplicial complex. 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. 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.

Filtrations. 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

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

Persistent homology. 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 $\{H_k(S_i) \,|\, 0 \leq i \leq n\}$ 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 8). 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).

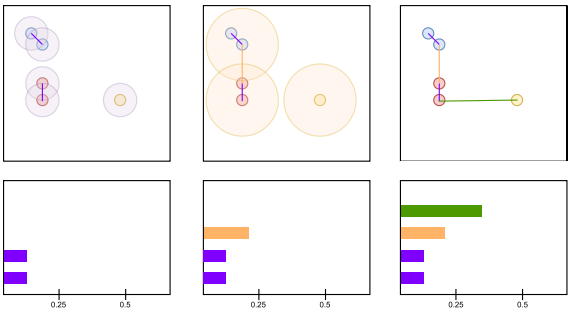

Figure 8: H_0 barcode construction. As the threshold increases, the separate connected components merge, resulting in the death of topological features. The horizontal axis is a sequence of thresholds ε, and each horizontal bar corresponds to a single feature.

C Properties of $\mathrm{MTop\text{-}Div}_G(R, P)$

Proof of Proposition 3.1. 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 the P-vertex subset equal zero. Excluding the zero-weight edges, this edge set coincides with the minimal spanning forest that attaches the vertex set R to the P vertices. □

Other properties of MTop-Div for attention graphs. Here we present some other properties of our proposed $\mathrm{MTop\text{-}Div}_G(R, P)$.

Proposition C.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 $P \cap R = \varnothing$.*

- *The divergence itself is bounded by*

$$0 \leq \mathrm{MTop\text{-}Div}_G(R, P) \leq |R|. \quad (4)$$

- ***(Stability.)*** *If the weights of G change by no more than ε, then the corresponding $\mathrm{MTop\text{-}Div}(R, P)$ changes by no more than $\delta = \varepsilon|R|$.*

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

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

- ***(Connection with hallucinations.)*** *The normalized divergence value $\frac{1}{|R|}\mathrm{MTop\text{-}Div}(R, P) = 0$ iff the MSF attaches every response token to a prompt token by a subtree with attention weights $= 1$.*

Proof of Proposition C.1.

1. This property is immediately obtained from the properties of an attention map: all its weights lie between 0 and 1. □

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

$$\mathrm{MTop\text{-}Div}(R, P) = \sum_{e \in \mathrm{MSF}(R,P)} w(e). \quad (5)$$

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

$$\delta = \varepsilon \cdot \#\mathrm{edges}(\mathrm{MSF}(R, P)) = \varepsilon \cdot |R|.$$

Consider the second case. Denote by $\mathrm{MSF}_{\mathrm{prev}}$ the original MSF, by $\mathrm{MSF}_{\mathrm{new}}$ — the MSF after the change; let w be the edge weight function before the change, $\hat{w}$ — after the change. The following inequalities hold:

$$\hat{w}(\mathrm{MSF}_{\mathrm{new}}) \leq \hat{w}(\mathrm{MSF}_{\mathrm{prev}}); \quad (6)$$

$$w(\mathrm{MSF}_{\mathrm{prev}}) - \delta \leq \hat{w}(\mathrm{MSF}_{\mathrm{prev}}) \leq \leq w(\mathrm{MSF}_{\mathrm{prev}}) + \delta; \quad (7)$$

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

$$w(\mathrm{MSF}_{\mathrm{new}}) \geq w(\mathrm{MSF}_{\mathrm{prev}}). \quad (9)$$

From (6)-(7) follows that $\hat{w}(\mathrm{MSF}_{\mathrm{new}}) < w(\mathrm{MSF}_{\mathrm{prev}}) + \delta$; from (8)-(9) follows that $\hat{w}(\mathrm{MSF}_{\mathrm{new}}) \geq w(\mathrm{MSF}_{\mathrm{prev}}) - \delta$. □

3. 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 surjectivity 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. □4. Follows obviously from the MSF formula for $\mathrm{MTop\text{-}Div}(R, P)$ and attention map properties. □

Intuition behind MTop-Div properties and hallucination detection. The stability property guarantees that similar attention patterns yield similar hallucination scores, making the metric's behavior consistent and predictable. The exact sequence property formalizes the geometric intuition behind our metric, which measures the strength of the response's connection to the prompt through multi-scale topological features of the attention graph. In the last property, we present an "ideal" case: if a model "knows what to look at" — each token in the response attends to some token in the prompt with an attention weight equal to 1 — MTop-Div would be equal to 0, indicating zero uncertainty.

D 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 dataset.

D.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 15). For SQuAD and XSum, responses are generated using a zero-shot approach. In contrast, for CoQA, we generate 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, culminating in 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 GPT-4o with an LLM's output, preceded by an instruction (prompt). 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 17.

For human annotation, we asked three team members with at least upper-intermediate English proficiency to independently annotate approximately 100 samples from each dataset. The human annotation consistency metrics are provided in Table 19. We selected samples for which all annotators reached consensus and treated these annotations as 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 18, 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, v0.4.6)

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 15: 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 16: 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 17: 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	Average
CoQA	Accuracy (↑)	0.809 ± 0.017	0.861 ± 0.015	0.742 ± 0.003	0.795 ± 0.009	0.831 ± 0.025	0.808
	Precision (↑)	0.849 ± 0.021	0.911 ± 0.007	0.771 ± 0.003	0.828 ± 0.011	0.860 ± 0.012	0.844
	Recall (↑)	0.871 ± 0.004	0.877 ± 0.019	0.877 ± 0.013	0.877 ± 0.005	0.893 ± 0.027	0.879
SQuAD	Accuracy (↑)	0.831 ± 0.003	0.857 ± 0.018	0.857 ± 0.008	0.872 ± 0.003	0.854 ± 0.007	0.854
	Precision (↑)	0.813 ± 0.002	0.831 ± 0.028	0.845 ± 0.021	0.850 ± 0.011	0.847 ± 0.007	0.837
	Recall (↑)	0.796 ± 0.008	0.839 ± 0.010	0.823 ± 0.023	0.858 ± 0.018	0.813 ± 0.017	0.826

Table 18: Classification metrics of GPT-4o annotation for CoQA and SQuAD with human labels considered as true. The table shows metric scores for different prompt variants and the average score across all variants.

Pair	Accuracy	Pearson-r
1 vs 2	0.783	0.6
1 vs 3	0.812	0.631
2 vs 3	0.826	0.653
Mean	0.807	0.628
Std	0.022	0.027

Table 19: Cross-annotator consistency metrics.

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 are shown in Table 20. The number of samples in the datasets varies across models, as we sought to maintain a balance between hallucinated and grounded responses, ensure sample cleanliness, 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. Additionally, we provide statistics on the response length distribution for each dataset we collected (Figure 9).

Model	CoQA		SQuAD		XSum	
	Hal.	Grounded	Hal.	Grounded	Hal.	Grounded
Mistral-7B	776	776	311	389	301	448
LLaMA-2-7B	375	375	440	258	239	507
LLaMA-2-13B	279	384	314	436	208	522
LLaMA-3.1-8B	356	350	350	400	243	407
Qwen2.5-7B	171	218	423	423	194	556

Table 20: Datasets statistics. Number of hallucinated and grounded samples of each model.

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 SelfCheckGPT, we used its NLI-based version.
- For LLM-Check, we considered its white-box attention score modification, as it works in a setting similar to ours.
- The topological divergences were calculated using `ripser` library (Tralie et al., 2018), MIT license.
- For the RAGTruth dataset, the model settings were aligned with the original paper.

All experiments were carried out on an 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 the MSR-LA license;

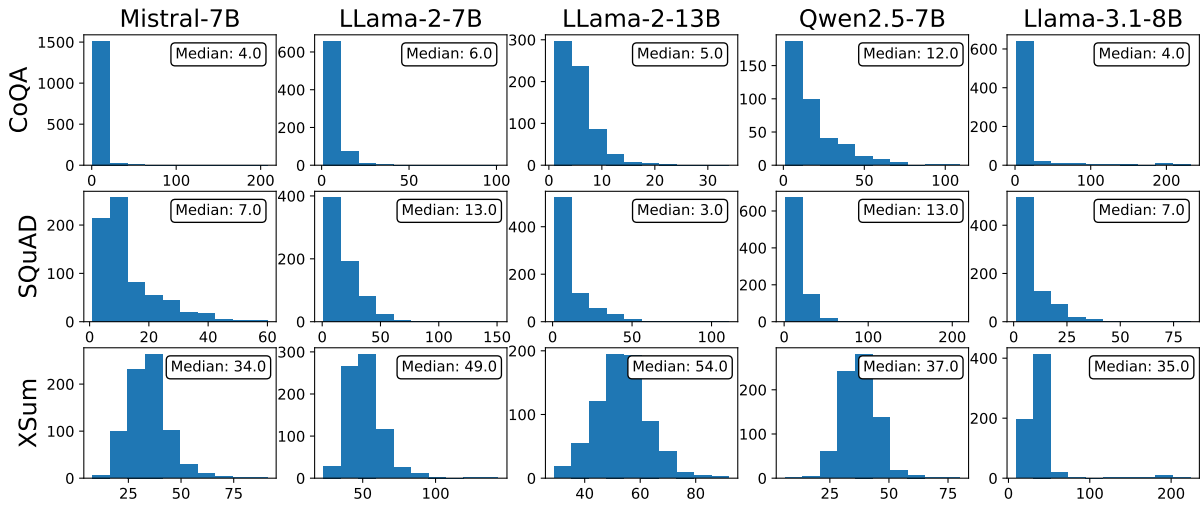

Figure 9: The distributions of response lengths (in words).

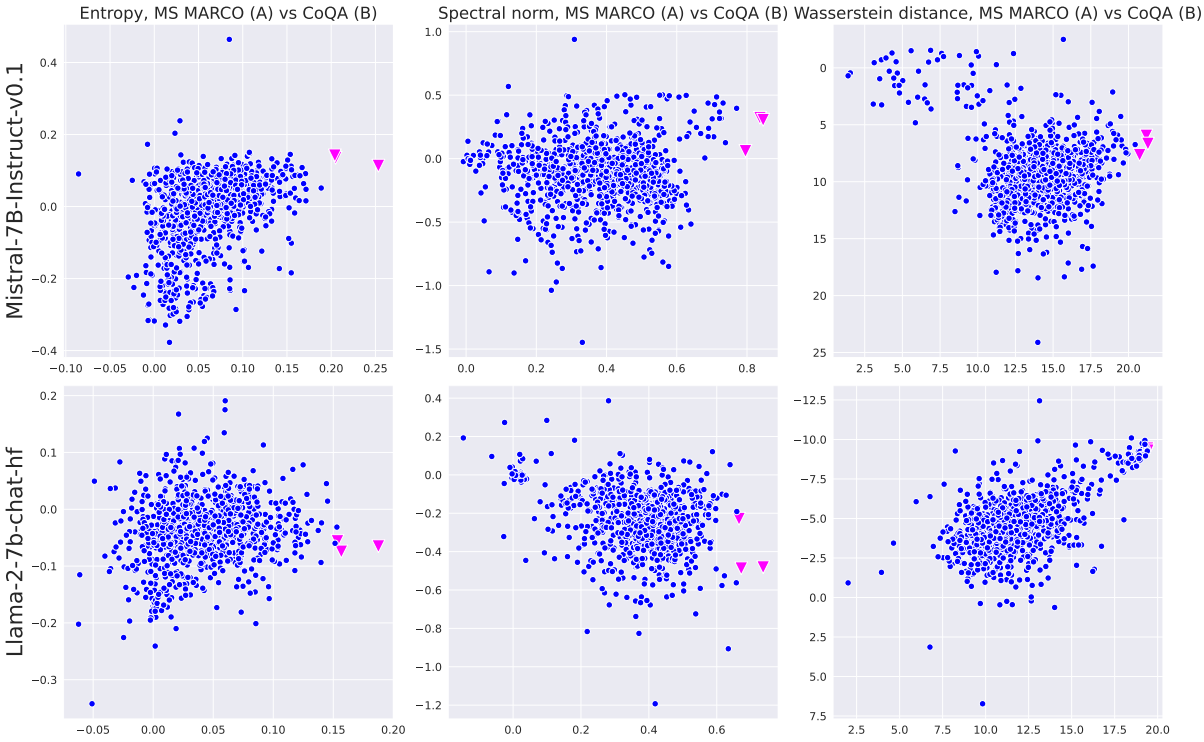

Figure 10: Δ_{ij} values for ij heads, MS MARCO and CoQA. Vertical axis corresponds to the difference in the dataset (B), horizontal axis to the difference in the dataset (A). The heads that separate samples best are highlighted in pink.

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. The SQuAD dataset is licensed under CC BY-SA 4.0. The RAGTruth dataset comes under the MIT license. The XSum dataset comes under the 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 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 framed 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.