TOPOLOGICAL HALLUCINATION DETECTORS ON DECODER ATTENTION REDUCE TO FIRST-ORDER STATISTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Persistent homology (PH) of attention graphs underlies a family of hallucination detectors, including TOHA (Bazarova et al., 2026). We show that in decoder-only models these topological features reduce to *first-order* attention statistics — quantities read off single rows or columns of the attention matrix, such as a row maximum or one token's attention to another — either exactly or up to one computable number. For any symmetric dissimilarity and vertex s we define the *coning defect* δ_s; classical cone and stability arguments then place the whole Vietoris–Rips barcode within bottleneck distance δ_s of a "star diagram" read off the distances to s, so every stable vectorization is a function of one column of the attention matrix, up to δ_s. The same argument applies to TOHA: its prompt–response topological divergence equals one minus the response tokens' mean largest prompt attention whenever a prompt-level coning defect vanishes, and lies within that defect otherwise. Across 16 settings (7 open models of 1.1B–7.62B; TruthfulQA, HaluEval, and on-policy TriviaQA) the two-sided bound holds without violation on 20,580,160 per-head graphs and collapses to an exact equality on 57–88% of them, including in two models with no first-token sink, where the statistic is the largest prompt attention rather than sink attention. Adding TOHA's score to that first-order counterpart changes AUC by at most 0.007, and supervised probes on the two are equivalent in every setting. On 838,356 layer graphs there are no bound violations; in every model with a first-token sink, 36–98% of graphs are exactly coned with empty H_1, and length-normalized 0D persistence keeps a median per-layer rank correlation of at least 0.98 with sink attention. For detection, 0D PH, sink-deflated PH and the per-head coning defect add at most 0.009 AUC over cheap non-topological probes, and TOHA at most 0.013. We release the extractor, the evaluator, and per-layer features for a subset of settings.

1 INTRODUCTION

A growing class of hallucination detectors views a transformer's attention matrix as a weighted graph, and then summarize it using persistent homology (PH), especially the barcodes of its connected components (H_0) and loops (H_1) along a Vietoris–Rips filtration (Kushnareva et al., 2021; Cherniavskii et al., 2022; Perez & Reinauer, 2022; Kostenok et al., 2023; Bazarova et al., 2026; Samaga et al., 2026). Their intuition is that perhaps global features in the flow of information could be detected by topology, which cannot be captured by statistics on tokens alone. In a parallel development, it has been shown that decoder-only language models devote a high proportion of their attention to the first token, known as the attention sink (Xiao et al., 2024; Sun et al., 2024; Cancedda, 2024; Gu et al., 2025; Barbero et al., 2025). Binkowski et al. (2026) recently showed that sinks carry much of the signal attention-based detectors use: exactly for the spectral LapEigvals (Binkowski et al., 2025), whose Laplacian entries are sink scores minus self-attention, and qualitatively for TOHA, whose minimum spanning tree sink nodes are "likely" to enter. Which part of a barcode the sink fixes, and by how much the rest can deviate, was open; Theorem 1 and Proposition 1 answer this exactly.

We show that, for topological detectors built from the Vietoris–Rips persistence or minimum spanning tree of a single attention graph — the standard construction, and TOHA's — the two literatures largely study the same object; zigzag persistence across layers and learned graph models are out of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

scope (§6). When one vertex s is closer to every other vertex than those vertices are to each other, every Vietoris–Rips sublevel set is a cone with apex s: higher homology vanishes and the minimum spanning tree (which computes H_0) is the star around s. A sink puts real attention graphs there, or close to it, so in most layers their "topology" is a function of each token's attention to the sink, and elsewhere it deviates by at most one computable number, the coning defect. How close is an empirical matter. It is closest at short-form QA lengths; on long retrieved contexts the defect grows and the exact case becomes rare, though the ordering is almost unchanged (Appendix K). The same holds for TOHA, whose divergence is a first-order statistic plus a computable correction.

Contributions.

1. **A whole-barcode sink reduction** (§3). The ingredients are classical: a flag complex with a dominating vertex is a cone (Barmak & Minian, 2012; Boissonnat & Pritam, 2020), and persistence is stable (Cohen-Steiner et al., 2007; Chazal et al., 2009; 2014). We combine them into one-sided bounds in a single computable number. For any symmetric dissimilarity and vertex s, bars of dimension $k \geq 1$ have length at most δ_s, and 0D total persistence is sandwiched around the star weight (Theorem 1). The whole barcode then lies within bottleneck distance δ_s of the star diagram ($\delta_s/2$ of a shifted one), so all stable vectorizations are equivalent to a single column of distances (Corollary 1). For causal attention, δ_0 has a closed form (Corollary 2).
2. **The published topological detector is a first-order detector** (§3, §5.4). Proposition 1: The topological divergence of TOHA is one minus the average max-prompt attention of the response tokens on prompt-coned heads or is within a computable defect of it. The qualitative observation by Binkowski et al. (2026) is made exact. There are no violations of the identity on 20,580,160 real head graphs, and adding TOHA to its first-order counterpart changes AUC by 0.000 to +0.007.
3. **Verification of the whole-graph reduction** (§5.1, Appendix D). There are no bound violations in 838,356 graphs, and graphs are exactly coned in 36–98% of cases in sink-dominated models. In a sink dose-response study, we explain how 1D features become weak because of coning, while H_1 is empty on all coned graphs.
4. **A detection study using strong baselines** (§5.3). We compare 0D/1D PH, sink features, sink-deflated PH, the coning defect, and TOHA against length, lexical, log-probability, LLM-Check (Sriramanan et al., 2024), Lookback-Lens (Chuang et al., 2024), and hidden-state probes. We report results from TruthfulQA, HaluEval, and an on-policy TriviaQA benchmark using grouped cross-validation, early- and late-fusion incremental tests, and equivalence tests.

2 Background and Related Work

Topology of attention. Kushnareva et al. (2021) and Cherniavskii et al. (2022) introduced barcodes of symmetrized BERT attention graphs as features for artificial-text detection and acceptability; Perez & Reinauer (2022) and Kostenok et al. (2023) extended them to classification and uncertainty estimation. TOHA (Bazarova et al., 2026) calculates an MST-based distance between prompt and response subgraphs for hallucination detection, HalluZig (Samaga et al., 2026) relies on zigzag persistence across layers, and CHARM (Frasca et al., 2026) learns a message-passing model over attention graphs. Gardinazzi et al. (2025) study zigzag persistence of hidden representations. All 0D summaries in this family are MST statistics (Fact 1; Gower & Ross, 1969). Our results concern what those statistics measure in causal models.

Attention sinks. Xiao et al. (2024) observed that streaming LMs keep the first tokens' key-values because a large share of attention lands there; Sun et al. (2024) tied sinks to massive activations, Cancedda (2024) to low-rank "dark signals", and Gu et al. (2025) characterized when they emerge during pre-training. Barbero et al. (2025) argued they prevent over-mixing; related "no-op" attention to special tokens was observed in BERT (Clark et al., 2019) and vision transformers (Darcet et al., 2024). Binkowski et al. (2026) use sinks with large-norm values directly as a hallucination signal (SinkProbe) and show that earlier detectors rely on them, exactly for LapEigvals (whose Laplacian eigenvalues, for causal attention, are its diagonal entries) and qualitatively for LLM-Check, Lookback-Lens and TOHA. For topological detectors we make this exact, and for TOHA the reduced statistic is max-prompt attention, which is sink attention only in sink-dominated heads.

Hallucination detection. Beyond topology, detectors use sampling consistency (Manakul et al., 2023; Farquhar et al., 2024), internal probes (Azaria & Mitchell, 2023; Burns et al., 2023; Marks & Tegmark, 2024; Li et al., 2023b; Kossen et al., 2024; Orgad et al., 2025; Du et al., 2024), covariance spectra (Chen et al., 2024), attention-kernel log-determinants (Sriramanan et al., 2024), and context-vs-generated attention ratios (Chuang et al., 2024). Recent audits find that several benchmarks admit near-perfect detection from surface form alone (Hussain & Kantarcioglu, 2026; Janiak et al., 2025), echoing annotation artifacts in NLI (Gururangan et al., 2018; McCoy et al., 2019). We include lexical controls and an on-policy benchmark.

3 Theory: Attention sinks cone the Vietoris–Rips filtration

Let $D \in \mathbb{R}^{N\times N}_{\geq 0}$ be symmetric with zero diagonal. The Vietoris–Rips filtration is $K_\epsilon = \{\sigma \subseteq [N] : \max_{u,v\in\sigma} D_{uv} \leq \epsilon\}$; $\mathrm{dgm}_k(D)$ is the barcode of $H_k(K_\epsilon;\mathbb{F})$ over a field, and $P_k(D)$ is its total persistence (sum of finite bar lengths). For attention, $A^{(l)}$ is the head-averaged, row-stochastic, lower-triangular attention matrix of layer l and, following the literature, $D = 1 - \max(A, A^\top)$ with zero diagonal.

Fact 1 (Gower & Ross, 1969)**.** *All 0D bars are born at 0, and their finite death times are, as a multiset, the edge weights of a minimum spanning tree (MST) of D. In particular $P_0(D) = w(\mathrm{MST}(D))$ is the total persistence (sum of finite bar lengths) of $\mathrm{dgm}_0(D)$.*

Definition 1 (Coning defect, star weight)**.** For $s \in [N]$, the *coning defect* is $\delta_s(D) = \max_{u\neq v,\, u,v\neq s} \big[\max(D_{su}, D_{sv}) - D_{uv}\big]_+$ (zero if $N < 3$), and the *star weight* is $S_s(D) = \sum_{u\neq s} D_{su}$. We call D *coned at s* if $\delta_s(D) = 0$, i.e. s is at least as close to both endpoints of every edge as they are to each other.

Theorem 1 (Sink-coning reduction)**.** *For every symmetric D and every vertex s, with $\delta = \delta_s(D)$:*

(a) every bar of $\mathrm{dgm}_k(D)$, $k \geq 1$, has length at most δ; hence $P_k(D) \leq \delta \cdot |\mathrm{dgm}_k(D)|$;
(b) $S_s(D) - (N-2)\,\delta \leq P_0(D) \leq S_s(D)$, and $\max_u D_{su} - \delta \leq \max \mathrm{dgm}_0^{fin}(D) \leq \max_u D_{su}$;
(c) if D is coned at s, then $H_k(K_\epsilon) = 0$ for all $k \geq 1$ and all ϵ, the star at s is an MST, and $P_0(D) = S_s(D)$.

Proof sketch (full proof in Appendix A). Let $D'_{uv} = \max(D_{uv}, D_{su}, D_{sv})$ for $u, v \neq s$ and $D'_{su} = D_{su}$. Then $D \leq D' \leq D + \delta$ entrywise and D' is coned at s. Since D' is coned, any simplex σ with $s \notin \sigma$ and $|\sigma| \geq 2$ has the same filtration value as $\sigma \cup \{s\}$, and a vertex not yet joined to s has no edges; so each K'_ϵ is a cone with apex s plus isolated points, and $H_k(K'_\epsilon) = 0$ for $k \geq 1$. The inclusions $K_b \subseteq K'_{b+\delta} \subseteq K_{b+\delta}$ factor the persistence map $H_k(K_b) \to H_k(K_{b+\delta})$ through zero, so no bar outlives δ. For (b), the star is a spanning tree of D (upper bound) and an MST of D'. Every spanning tree has an edge at s, where $D' = D$, so passing from D' to D lowers a tree's weight by at most $(N-2)\delta$; the bottleneck value is 1-Lipschitz in the sup-norm. The ingredients are standard: Lemma 1 is the dominating-vertex (cone) argument behind strong and edge collapses (Barmak & Minian, 2012; Boissonnat & Pritam, 2020), and the last step is a one-sided form of the interleaving argument behind persistence stability (Chazal et al., 2009). □

The bound in (a) is attained: if $D_{su} = 1$ for all u and four further vertices form a 4-cycle with side $\frac{1}{2}$ and diagonals 1, then $\delta_s = \frac{1}{2}$ and H_1 has the single bar $[\frac{1}{2}, 1)$. On real attention the bounds are not vacuous (Table 4): the longest H_1 bar reaches $0.9999\,\delta_0$, and (b) is uninformative on at most 5.5% of non-coned graphs.

Theorem 1 controls summaries. The same comparison controls the whole barcode. Call $\Sigma_s(D)$ the *star diagram*: in degree 0 the bars $[0, D_{su})$, $u \neq s$, plus one infinite bar, and nothing in degree $k \geq 1$.

Corollary 1 (The whole barcode is the star, up to δ_s)**.** *In every degree $k \geq 1$, $d_B\big(\mathrm{dgm}_k(D), \emptyset\big) \leq \delta_s/2$. In degree 0, the j-th smallest finite death of D lies in $[D_{s(j)} - \delta_s,\, D_{s(j)}]$, where $D_{s(j)}$ is the j-th smallest entry of $(D_{su})_{u\neq s}$, and β_0 of K_ϵ lies between $1 + \#\{u : D_{su} > \epsilon + \delta_s\}$ and $1 + \#\{u : D_{su} > \epsilon\}$. Hence $d_B\big(\mathrm{dgm}(D), \Sigma_s(D)\big) \leq \delta_s$, and $d_B\big(\mathrm{dgm}(D), \Sigma^\delta_s(D)\big) \leq \delta_s/2$ for the shifted star diagram Σ^δ_s with finite 0D deaths $(D_{s(j)} - \delta_s/2)_+$; both diagrams are equal to $\mathrm{dgm}(D)$ when $\delta_s = 0$. Every vectorization that is L-Lipschitz in d_B (e.g. persistence landscapes;*

Bubenik, 2015) is therefore within $L\delta_s/2$ of a function of the column $(D_{su})_u$ and the number δ_s, and within $L\delta_s$ of a function of the column alone. Vectorizations that are stable in 1-Wasserstein distance (e.g. persistence images; Adams et al., 2017) are within $L\big((N-2)\delta_s + \frac{1}{2}\delta_s \sum_{k\geq 1} |\mathrm{dgm}_k(D)|\big)$ of a function of the column. On coned graphs, every vectorization (such as persistence entropy or Betti curves) is precisely such a function.

Corollary 2 (Causal attention)**.** *Let A be lower-triangular and row-stochastic, $D = 1 - \max(A, A^\top)$, and $s = 0$ the first token. Then $D_{0u} = 1 - A_{u0}$, so $S_0 = (N-1)(1-\bar{m})$ with $\bar{m} = \frac{1}{N-1}\sum_{u\geq 1} A_{u0}$ the mean sink attention, and*

$$\delta_0 = \max_{1\leq v<u} \big[A_{uv} - \min(A_{u0}, A_{v0})\big]_+.$$

Thus $P_0 \in [(N-1)(1-\bar{m}) - (N-2)\delta_0,\ (N-1)(1-\bar{m})]$ and all $H_{k\geq 1}$ bars have length at most δ_0.

Remark 1 (Choice of distance)**.** The graph is exactly coned iff, for all $1 \leq v < u$, $A_{uv} \leq A_{u0}$ and $A_{uv} \leq A_{v0}$: no token attends to an earlier non-sink token more than either of the two attends to the sink. Coning is an order condition, so it is invariant under any entrywise strictly increasing transform f of D with $f(0) = 0$ (e.g. $-\log$ of the symmetrized attention, or ranks without ties, in place of $1 - A$). The set of exactly coned graphs, and the vanishing of their higher homology, does not depend on how attention is turned into a distance. Only the magnitude of δ_s on non-coned graphs does. (A merely non-decreasing f, such as clipping, can create coning.)

Corollary 2 makes four testable predictions: **(C1)** in coned layers 0D total persistence equals $(N-1)(1-\bar{m})$, so after dividing by $N-1$ it is an affine function of sink attention; **(C2)** H_1 is empty whenever $\delta_0 = 0$; **(C3)** 1D persistence can only be non-zero where the sink is violated, with bars no longer than δ_0; **(C4)** removing the sink and renormalizing (deflation) exposes whatever non-sink topology exists.

Random weights and directed complexes. For i.i.d. weights, on the other hand, 1D total persistence is known to scale linearly with N (Hiraoka & Shirai, 2017; Hino & Kanazawa, 2019), and we give an elementary explicit bound (Proposition 2, Appendix A): $\mathbb{E}[P_1] \geq \frac{3N}{8} - \sqrt{3N} + O(1)$. We prove that the directed flag complex of the digraph is not a way out for causal attention: the digraph is a transitive tournament, so its weighted directed flag complex is isomorphic to $\mathrm{VR}(D)$ as a filtered complex (Proposition 3, Appendix A.4).

The published topological detector reduces to a first-order statistic. TOHA (Bazarova et al., 2026) scores an attention head by the topological divergence between the response tokens R and the prompt tokens P. Its per-head score is $d(R,P) = \mathrm{MTopDiv}(R,P)/|R|$, where MTopDiv is the total 0D persistence of the head's graph $D = 1 - \max(A, A^\top)$ after all prompt–prompt distances are set to 0 (equivalently, the minimum spanning forest attaching R to P). TOHA averages d over at most ten heads chosen on labeled data. For $u \in R$ let $\pi_u = \max_{v\in P} A_{uv}$ be u's largest attention to any prompt token.

Proposition 1 (TOHA is max-prompt attention, up to a defect)**.** *Let A be lower-triangular and row-stochastic, and let $P = \{0, \ldots, p-1\}$ and $R = \{p, \ldots, N-1\}$. With $\delta_P = \max_{w<u\in R} \big[A_{uw} - \min(\pi_u, \pi_w)\big]_+$,*

$$\frac{1}{|R|}\sum_{u\in R}(1-\pi_u) \;-\; \frac{|R|-1}{|R|}\,\delta_P \;\leq\; d(R,P) \;\leq\; \frac{1}{|R|}\sum_{u\in R}(1-\pi_u),$$

with equality when $\delta_P = 0$. Moreover $1 - \pi_u \leq 1 - A_{u0}$, with equality iff the first token is among u's most-attended prompt tokens. On a head where this holds for every $u \in R$ and $\delta_P = 0$, $d(R,P) = 1 - \frac{1}{|R|}\sum_{u\in R} A_{u0}$, the response tokens' mean sink attention.

Proof sketch. Zeroing the prompt block merges P at scale 0, so the remaining finite 0D bars are the MST weights of the quotient graph on $\{p^*\} \cup R$, where $D_{p^*u} = \min_{v\in P} D_{uv} = 1 - \pi_u$ (Fact 1). Apply Theorem 1(b) at the apex p^*; as in Corollary 2, its coning defect is δ_P and its star weight is $\sum_u (1-\pi_u)$. Divide by $|R|$. □

Binkowski et al. (2026) note that sink nodes are "likely to appear in the MST" of MTopDiv. Proposition 1 makes this exact, and it indicates where the topology can enter: only through δ_P, which is non-zero only when a response token attends to an earlier response token more than one of the two attends to its best prompt token. We test the proposition on real heads, and TOHA against its first-order counterparts, in §5.4.

4 Experimental setup

Models and benchmarks. We use Qwen2.5-1.5B/3B/7B-Instruct, Phi-3-mini-4k-instruct, TinyLlama-1.1B-Chat, SmolLM-1.7B-Instruct, and Mistral-7B-Instruct-v0.2: 7 open models spanning 1.1B–7.62B parameters, with additional Mistral-7B analysis in Appendix H. Mistral is run twice, with its hand-written `[INST]` string and with its own chat template; as a consistency check the two differ by one response token and agree to 0.023 AUC on every bank except answer-only PH (0.071; Appendix B). **TruthfulQA** (Lin et al., 2022): all 817 questions, each paired with its best answer and first incorrect answer (1,634 rows) under each model's chat format (Appendix B). **HaluEval QA** (Li et al., 2023a): the first 1,000 (Qwen2.5-3B) or 500 (Qwen2.5-1.5B) knowledge-grounded questions with right and hallucinated answers. **TriviaQA on-policy**: 2,000 random validation questions (Joshi et al., 2017); each model answers greedily (short-phrase instruction, ≤ 24 tokens); answer labeled hallucinated iff it does not include a normalized gold alias; features are extracted from the model's own answer (accuracies in Table 7). All models run in bfloat16 with eager attention.

Features. Per layer (head-averaged A): 0D (max lifetime, total persistence); 1D (max lifetime, total persistence, and the birth and the death of the longest bar, each divided by N); Sink (S_0 and $\max_u D_{0u}$, i.e. exactly the two quantities Corollary 2 says 0D reduces to); Deflated PH (0D+1D after deleting token 0 and renormalizing rows); Answer-only PH (induced subgraph on answer tokens); the coning defect (δ_0 and $\min_s \delta_s$); row entropy and row max. Per head: 0D (MST total and max via a dense Prim's algorithm), sink mass, the coning defect δ_0 of the head's graph (Defect$_h$), entropy, Lookback ratio (Chuang et al., 2024), and the LLM-Check attention score (Sriramanan et al., 2024). TOHA (per head, separate forward pass, same rows): the exact score d of Bazarova et al. (2026) via a dense Prim's algorithm on the prompt-contracted graph, the $\bar{\pi}$-score, the response sink score, and δ_P (Proposition 1). Other: answer-token log-probability and predictive-entropy statistics; mean-pooled answer hidden states at the middle layer; TF-IDF word and character n-grams of the answer text (Lexical). Only the 1D summaries are length-normalized: 0D and Sink enter the classifier raw, so the two banks the reduction pairs carry the same length information, with Len scored separately as a control.

Protocol. All numbers use one evaluator: standardized features and L_2 logistic regression, with $C = 1$ for banks of at most 64 dimensions and otherwise $C \in \{0.01, 0.1, 1\}$ chosen by inner grouped 3-fold CV; widening that grid to $\{10^{-4}, \ldots, 10^{2}\}$ helps the topological banks more (up to 0.061) than the non-topological ones (0.015) and reverses none of the orderings we claim (Appendix B). The outer loop is 10-fold `StratifiedGroupKFold` grouped by question, repeated over 3 seeds (42–44), with out-of-fold probabilities averaged. We test incremental value in two ways. Early fusion ("X|Y") concatenates the banks. Late fusion ("X$\oplus$Y") stacks the two banks' out-of-fold logits in a second-level logistic regression under the same CV, so that a small bank cannot be drowned out by a wide one. Differences in pooled AUC use a 10,000-resample question-level cluster bootstrap. We call two banks equivalent if the 90% interval lies inside ± 0.015 AUC (two one-sided tests, TOST, at $\alpha = 0.05$), and report TOST power. Paired benchmarks also report within-pair accuracy. The theory checks use all (example, layer) graphs. Ripser (Bauer, 2021) represents its filtration values as 32-bit floats, so we check the bar-length and max-death bounds at tolerance 10^{-5}, and the P_0 sandwich, a sum of up to N such values, at 10^{-3}.

5 Results

5.1 The reduction holds on real attention

Table 1 checks Theorem 1 on every layer-level graph of every example: 838,356 graphs in total. There are 0 violations of any bound. In the 13 settings whose models place 0.42 to 0.75 of attention

Table 1: **Theorem 1 on real attention** ($s = 0$). *viol.*: bound violations; *coned*: share with $\delta_0 = 0$; $H_1{=}\emptyset \,|\, \text{coned}$: C2 predicts 100%. ρ: median over layers of the Spearman correlation across examples of P_0 with S_0 (raw) and of $P_0/(N{-}1)$ with $1 - \bar{m}$ ($/N$, length removed).

Benchmark	Model	$\bar{N}$	graphs	viol.	sink attn.	coned	$H_1{=}\emptyset \,\|\, \text{coned}$	ρ_{raw}	$\rho_{/N}$
TruthfulQA	Qwen2.5-3B	44.2	58,824	0	0.50	62.2%	100.0%	1.000	1.000
TruthfulQA	Qwen2.5-1.5B	44.2	45,752	0	0.53	64.5%	100.0%	1.000	1.000
TruthfulQA	Phi-3-mini	29.0	52,288	0	0.69	88.4%	100.0%	1.000	1.000
TruthfulQA	TinyLlama-1.1B	42.3	35,948	0	0.53	36.2%	100.0%	1.000	0.982
TruthfulQA	SmolLM-1.7B	32.4	39,216	0	0.02	0.0%	–	0.967	0.412
TruthfulQA	Mistral-7B	33.0	52,288	0	0.75	98.2%	100.0%	1.000	1.000
TruthfulQA	Mistral-7B (chat)	34.0	52,288	0	0.75	98.2%	100.0%	1.000	1.000
TruthfulQA	Qwen2.5-7B	44.2	45,752	0	0.01	0.0%	–	0.976	0.513
HaluEval	Qwen2.5-3B	126.9	72,000	0	0.42	52.7%	100.0%	1.000	1.000
HaluEval	Qwen2.5-1.5B	128.1	28,000	0	0.45	62.5%	100.0%	1.000	1.000
TriviaQA (on-policy)	Qwen2.5-3B	68.6	72,000	0	0.48	63.0%	100.0%	1.000	1.000
TriviaQA (on-policy)	Qwen2.5-1.5B	68.8	56,000	0	0.52	50.6%	100.0%	1.000	1.000
TriviaQA (on-policy)	Phi-3-mini	46.1	64,000	0	0.69	87.1%	100.0%	1.000	1.000
TriviaQA (on-policy)	TinyLlama-1.1B	72.0	44,000	0	0.50	36.4%	100.0%	1.000	0.991
TriviaQA (on-policy)	Mistral-7B	58.7	64,000	0	0.64	65.8%	100.0%	1.000	1.000
TriviaQA (on-policy)	Qwen2.5-7B	68.3	56,000	0	0.01	0.0%	–	0.986	0.489

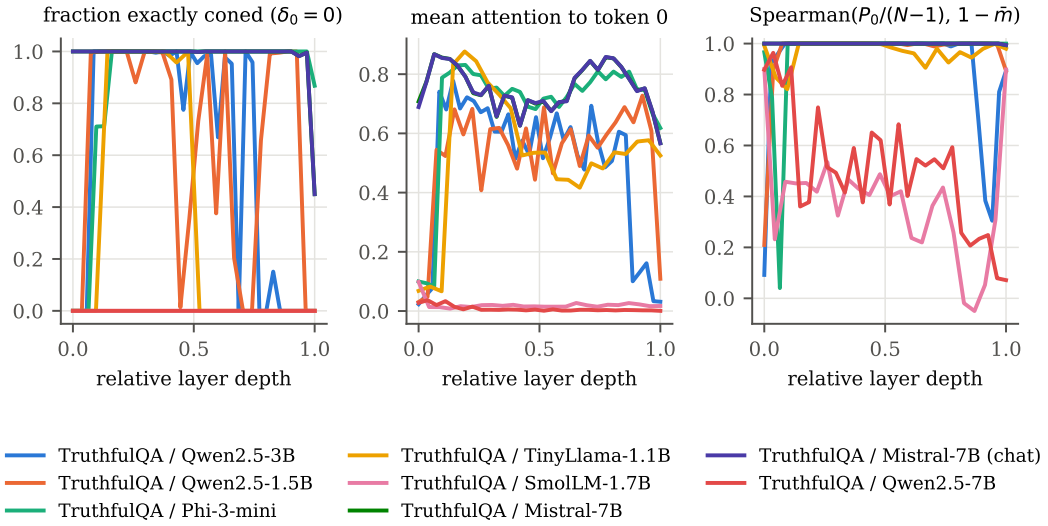


Figure 1: **Layerwise structure (TruthfulQA).** Coning (left) tracks sink attention (middle); Qwen2.5-7B and SmolLM-1.7B are never coned at $s = 0$. Right: length-normalized rank correlation of P_0 with sink attention, near 1 in sink models except some open layers.

on token 0, 36–98% of graphs are exactly coned ($\delta_0 = 0$). On each of these graphs, H_1 is empty as required by prediction C2. For comparison, δ_0 and 1D total persistence have Spearman correlation 0.07 to 0.89 across graphs (C3); the low end is degenerate rather than weak, being the 2 settings above 90% coned, where 99.99% of graphs have no 1D bars. Elsewhere it is at least 0.32. Both P_0 and the star weight $S_0 = (N{-}1)(1 - \bar{m})$ depend on length, so their raw per-layer rank correlation (median $\geq$0.999) depends in part on length. Dividing both by $N - 1$ gives the median per-layer rank correlation between $P_0/(N{-}1)$ and $1{-}\bar{m}$ at least 0.98 for each sink setting (C1). Length-normalized 0D total persistence, up to rank, is thus one minus the mean attention paid to the first token in most layers. The single worst individual layer reaches only -0.14 to 0.99 per setting. Figure 1 locates coning: it happens in the middle of the network, where attention to the sink is high. The first few layers are never sink-dominated, and in some models (TinyLlama, the Qwen models) neither is a block of late layers; these layers are rarely coned. There the length-normalized correlation can fall, as Theorem 1(b) allows once $(N - 1)\delta_0$ is comparable to the spread of S_0 across examples.

Models without a first-token sink as a contrast. We have 2 models with essentially no attention on token 0, Qwen2.5-7B and SmolLM-1.7B, with 0.01 to 0.02 attention there. None of their graphs are coned at $s = 0$, so Theorem 1 at the first token gives them no useful guarantee. The two are not alike: allowing any apex, 0% of SmolLM-1.7B's graphs are coned, so no vertex cones any of

Table 2: **Pooled out-of-fold ROC-AUC** (seed SD $\leq$ 0.008). TQA/HE/TrQA: TruthfulQA/HaluEval/on-policy TriviaQA. Lex.: answer TF-IDF; LogP: log-probability; LLM-C.: LLM-Check; Sink: $(S_0, \max_u D_{0u})$ per layer; subscript h: per head; Lookb.: Lookback; Hidden: mid-layer probe (1,536 to 4,096 dims); Non-topo.: LogP+LLM-C.+Lookb.+Hidden+row statistics. Best in bold; full table in Appendix M.

Setting	Lex.	LogP	LLM-C.	Sink	0D	1D	Sink$_h$	0D$_h$	Lookb.	Hidden	Non-topo.
TQA Qw3B	.820	.630	.693	.737	.736	.660	.851	.853	.897	.913	**.926**
TQA Qw1.5B	.820	.638	.666	.723	.726	.586	.830	.838	.851	.916	**.917**
TQA Phi-3	.820	.578	.653	.651	.646	.526	.871	.868	.892	.933	**.936**
TQA TinyLl.	.820	.637	.614	.643	.654	.592	.808	.806	.850	.883	**.885**
TQA SmolLM	.820	.588	.604	.611	.583	.543	.720	.800	.863	.894	**.899**
TQA Mis7B	.820	.641	.652	.685	.684	.501	.903	.893	.933	.956	**.959**
TQA Mis7B (chat)	.820	.651	.653	.688	.688	.501	.907	.900	.934	.958	**.960**
TQA Qw7B	.820	.674	.667	.749	.679	.610	.821	.850	.922	.933	**.950**
HE Qw3B	.981	.993	.745	.829	.811	.710	.958	.971	.998	1.00	**1.00**
HE Qw1.5B	.979	.995	.791	.820	.789	.747	.922	.949	.999	1.00	**1.00**
TrQA Qw3B	.596	.839	.688	.717	.732	.601	.820	.807	.861	.841	**.870**
TrQA Qw1.5B	.609	.823	.676	.707	.716	.574	.767	.768	.824	.855	**.873**
TrQA Phi-3	.635	.877	.676	.707	.710	.546	.826	.807	.840	.867	**.883**
TrQA TinyLl.	.740	.724	.688	.720	.720	.659	.814	.775	.823	.827	**.842**
TrQA Mis7B	.677	.728	.686	.696	.699	.611	.812	.775	.829	.833	**.841**
TrQA Qw7B	.594	.847	.725	.714	.668	.582	.801	.798	.879	.854	**.881**

them, against 10% of Qwen2.5-7B's, whose attention is strongly concentrated, just not on token 0 (§5.4). One model has no cone structure to find; for the other, $s = 0$ is the wrong apex. Their H_1 is non-empty on 55% to 87% of graphs, and the raw $\rho(P_0, S_0)$ is high (0.967 to 0.986), but nearly all of that comes from shared lengths: the length-normalized correlation is 0.41 to 0.51. Both differ from the sink-dominated models in weights, data and template, so this is a contrast, not a controlled intervention; the within-model intervention is deflation (C4, below).

5.2 What topology adds beyond the sink

If 0D persistence reduced exactly to the star, a classifier on the two sink features would match one on the two 0D features in every layer. Table 22 ("Sink vs. 0D") shows the two banks differ by -0.015 to +0.031 AUC across the 13 sink settings evaluated, and are statistically equivalent within ± 0.015 in 8 of them (the TOST power of this comparison is modest, so non-equivalence is not evidence of a difference). Adding 0D to the sink bank gains 0.000 to +0.028.

The residual lives exactly where the theory says it can. We split layers without using labels: coned layers ($\geq 90\%$ of graphs coned) and open layers ($< 50\%$). Across the 12 sink settings, adding the coned layers' 0D features to Sink changes AUC by -0.001 to +0.004 (equivalent within ± 0.015 in 12 of 12), whereas the open layers account for 0.000 to +0.027 (Table 20). Almost everything 0D adds beyond sink attention comes from the layers where the sink does not dominate.

Per head, the reduction is tighter. Per-head sink mass and per-head 0D are far stronger than their head-averaged versions (Table 2), but adding per-head MST total ($P_{0,h}$) to per-head sink mass changes AUC by only -0.001 to +0.011, which is equivalent within ± 0.015 in 13 of the 13 sink settings with per-head features. The gain is small but not zero: its 95% interval lies above 0 in 6 of them (at most 0.011; Table 22).

The coning defect is a signal, but not a new one. Theorem 1 identifies δ_0 as the unique locus where topology may diverge from the sink, so we evaluate it as a signal (Table 21). Head-averaged, it is weak (AUC 0.50 to 0.67). Per head, it is strong (0.69 to 0.96), with a small improvement when stacked on top of per-head sink mass (+0.004 to +0.029 AUC; 95% interval above 0 in 16 of 16 settings). Stacking on top of the non-topological probes yields smaller gains (0.000 to +0.009; above 0 in 5, equivalent within ± 0.015 in 16 of 16).

Deleting the sink exposes some, redundant, structure. Deflation deletes token 0, so in Qwen2.5-7B and SmolLM-1.7B it removes a non-sink token rather than the apex; those settings are excluded from the C4 numbers below, and Appendix L re-runs the test deflating the apex itself. Deflated PH (C4) gives a modest gain of +0.029 to +0.067 over SINK on TruthfulQA (95% interval excludes 0 in 6 of 6 sink settings) but less on the on-policy benchmark (+0.002 to +0.043; 2 of 5). The same holds for answer-only PH (Appendix M). Attention graphs do contain some non-sink geometry, particularly on teacher-forced text. But this geometry is captured by non-topological probes (§5.3).

One-dimensional persistence. In most real graphs there are no 1D bars at all (Table 1), unlike i.i.d. weights, whose 1D persistence grows linearly. Adding 1D to 0D changes AUC by -0.008 to +0.033, equivalent within ± 0.015 in 9 of 13 sink settings; under a strong sink a loop is visible to H_1 only if it breaks the cone (Figure 4).

5.3 HALLUCINATION DETECTION AGAINST STRONG BASELINES

Table 2 places attention-graph topology among standard detectors. Three patterns hold across benchmarks.

Cheap non-topological probes dominate topology. On TruthfulQA and the on-policy benchmark, the mean-pooled hidden-state probe (0.83 to 0.96) and per-head Lookback ratios (0.82 to 0.93) outperform all topological banks, and combined with log-probability, LLM-Check, and row statistics score 0.84 to 0.96. They are also far wider than the topological banks (Table 2), so this is a claim about what is cheaply available, not about equal capacity; the incremental tests below hold the wide bank fixed. With this combination, adding 0D PH changes AUC from 0.000 to +0.001 (equivalent within ± 0.015 in 16 of the 16 settings with per-head features), and adding sink-deflated PH changes AUC from -0.001 to +0.002 (equivalent in 16 of 16). To check whether early fusion hides a small bank inside a $\sim$2,000-dimensional one, we repeat with late fusion. Stacking the 0D score onto the non-topological score changes AUC by 0.000 to +0.003 (equivalent in 16 of 16 settings; the 95% interval lies above 0 in 0 of them), and stacking the deflated-PH score changes it by 0.000 to +0.005 (equivalent in 16 of 16). Nested stacking, which refits the first level inside each outer fold and so removes the usual stacking leak, gives 0.000 to +0.002 for 0D$\oplus$NT and 0.000 to +0.002 for deflated PH$\oplus$NT (both equivalent in all 11 settings with raw features), 0.000 to +0.013 for TOHA$\oplus$NT, and moves no comparison by more than 0.002 (Table 8).

Which cheap probe wins depends on whether the text is teacher-forced. Log-probability is weak (0.58 to 0.67) on TruthfulQA, which reflects the benchmark's focus on imitative falsehoods. 0D PH (0.58 to 0.74) outscores 7 of 8 models, significantly for 5. On-policy, log-probability is strong (0.72 to 0.88) and significantly beats 0D (0.67 to 0.73) for 4 of 6 models. Single-teacher-forced benchmarks cannot be used to draw conclusions about topology vs. confidence.

Resolution per-head trumps topology. Head-averaged features are weaker than per-head features of any type, and per-head entropy lies within -0.016 to +0.044 AUC of per-head 0D (Appendix M).

5.4 THE PUBLISHED TOPOLOGICAL DETECTOR IS FIRST-ORDER

Scope. TOHA was designed for retrieval-augmented generation; only HaluEval supplies a context passage here, so the AUC comparisons below are off-distribution for TOHA and not a reproduction of its reported accuracy. Proposition 1 and head-level identity checks hold regardless.

We test Proposition 1 on every head of every example: 20,580,160 head graphs from 16 settings (Table 3). We run TOHA's head-selection algorithm (Bazarova et al., 2026, Alg. 1) in cross-validation on the whole training fold, for TOHA's score d and the two first-order scores: the $\bar{\pi}$-score $\frac{1}{|R|}\sum_u(1-\pi_u)$ and the response sink score $1-\bar{a}_0 = \frac{1}{|R|}\sum_u(1-A_{u0})$.

The identity holds on real heads. The sandwich has 0 violations, and our d matches the authors' released MTOP-DIV code to 1.2×10^{-8} on 21,696 head graphs (2 settings; Appendix F). Head graphs are exactly prompt-coned in 57–88% of cases (largest d–$\bar{\pi}$ gap $6 \cdot 10^{-8}$). Median per-head rank correlation $\rho(d, \bar{\pi})$ across examples is 0.954 to 0.998. Against the response sink score it is 0.78 to 0.99 in sink models, but -0.16 to -0.09 in Qwen2.5-7B and SmolLM-1.7B (Table 14), which lack a first-token sink: there d tracks max-prompt attention just as tightly while mildly anti-correlating

Table 3: **TOHA against its first-order counterparts** (Proposition 1). $\delta_P{=}0$: share of exactly prompt-coned head graphs, where d equals the $\bar{\pi}$-score; ρ: median per-head rank correlation. AUCs: TOHA's head selection on each of the three scores, supervised probes on all heads of each, and the non-topological bank. Marks as in Table 22; $\oplus$: late fusion; diagnostics in Table 14.

Setting	$\delta_P{=}0$	$\rho(d,\bar{\pi})$	TOHA	TOHA$[\bar{\pi}]$	TOHA$[\bar{a}_0]$	TOHA$_{\text{sup}}$	$\bar{\pi}_{\text{sup}}$	$\bar{a}_{0,\text{sup}}$	Non-topo.	$\bar{\pi}-$TOHA	$\bar{a}_0-$TOHA	TOHA$\oplus\bar{\pi}$	TOHA$\oplus$NT
TQA Qw3B	57%	.976	.668	.670	.649	.892	.892	.892	.926	$+0.003^{\equiv}$	-0.018^{*}	$+0.003^{*\equiv}$	$+0.003^{*\equiv}$
TQA Qw1.5B	58%	.978	.646	.649	.628	.858	.857	.857	.917	$+0.003^{\equiv}$	-0.018^{*}	$+0.005^{*\equiv}$	$+0.002^{\equiv}$
TQA Phi-3	74%	.996	.751	.742	.715	.898	.901	.905	.936	-0.009^{*}	-0.036^{*}	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA TinyLl.	74%	.995	.666	.655	.618	.845	.843	.848	.885	-0.012^{*}	-0.049^{*}	$+0.002^{\equiv}$	$+0.001^{\equiv}$
TQA SmolLM	58%	.954	.701	.695	.595	.862	.860	.841	.899	$-0.007^{\equiv}$	-0.107^{*}	$+0.005^{*\equiv}$	$+0.002^{\equiv}$
TQA Mis7B	81%	.998	.751	.764	.766	.932	.934	.936	.959	$+0.013^{*}$	$+0.015^{*}$	$0.000^{\equiv}$	$+0.001^{\equiv}$
TQA Mis7B (chat)	80%	.998	.753	.762	.761	.934	.936	.937	.960	$+0.009^{*\equiv}$	$+0.007^{*\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$
TQA Qw7B	57%	.980	.705	.709	.544	.925	.927	.906	.950	$+0.004^{\equiv}$	-0.161^{*}	$+0.001^{\equiv}$	$+0.004^{*\equiv}$
HE Qw3B	68%	.987	.967	.982	.961	.999	.999	1.00	1.00	$+0.015^{*}$	$-0.007^{*\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
HE Qw1.5B	69%	.988	.970	.988	.972	.999	.999	.999	1.00	$+0.018^{*}$	$+0.002^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TrQA Qw3B	76%	.991	.728	.736	.742	.862	.863	.870	.870	$+0.009$	$+0.014^{*}$	$+0.002^{\equiv}$	$+0.012^{*}$
TrQA Qw1.5B	77%	.992	.712	.714	.708	.828	.827	.836	.873	$+0.002^{\equiv}$	-0.005	$+0.003^{\equiv}$	$+0.004^{\equiv}$
TrQA Phi-3	88%	.998	.724	.731	.733	.847	.851	.850	.883	$+0.007^{\equiv}$	$+0.008$	$0.000^{\equiv}$	$+0.001^{\equiv}$
TrQA TinyLl.	66%	.994	.714	.725	.693	.833	.826	.832	.842	$+0.011^{*}$	-0.021^{*}	$+0.007^{*\equiv}$	$+0.010^{*}$
TrQA Mis7B	73%	.992	.711	.704	.721	.821	.817	.822	.841	-0.007	$+0.010$	$+0.005^{*\equiv}$	$+0.009^{*\equiv}$
TrQA Qw7B	76%	.994	.754	.790	.617	.876	.880	.842	.881	$+0.036^{*}$	-0.137^{*}	$+0.001^{\equiv}$	$+0.012^{*}$

with token 0. The reduction is to max-prompt attention; that this matches the sink is a property of the model (Appendix G).

Topology has little effect on the first-order score. Stacking d onto the $\bar{\pi}$-score changes AUC by 0.000 to +0.007 (± 0.015 equivalent in 16/16; 95% interval > 0 in 5). Supervised probes on all heads show the $\bar{\pi}$-score differs from d by -0.007 to +0.004 (equivalent in 16/16); the response sink score differs by -0.034 to +0.008 (equivalent in 12). Exceptions are models without a first-token sink (Qwen2.5-7B and SmolLM-1.7B) and one on-policy setting. Discrete head selection is noisier: running on the $\bar{\pi}$-score shifts AUC by -0.012 to +0.036 (6 higher, 2 lower, 13/16 non-inferior at margin 0.015). On the sink score, it loses up to 0.049 AUC in sink models (0.107 to 0.161 in Qwen2.5-7B and SmolLM-1.7B): it selects heads prompt-coned on 60–95% of examples, whereas the sink is the top prompt token on 19–79% of responses (Table 14). Stacking TOHA onto non-topological features shifts AUC by 0.000 to +0.012.

Why "attention to $\langle s \rangle$" looked weak. Bazarova et al. (2026, Table 9) report that attention to the first token is a much weaker feature than their divergence. Proposition 1 identifies the quantity that d actually tracks: per head, the attention of the response tokens to their most-attended prompt token. In sink models, that token is the sink for every response token in 40–83% of head graphs (Table 14).

Benchmark artifacts. A TF-IDF model on the answer text alone attains 0.82 AUC on TruthfulQA and 0.98 on HaluEval, but only 0.59 to 0.74 on-policy: teacher-forced benchmarks reward recognizing how the dataset's answers were written (Hussain & Kantarcioglu, 2026; Janiak et al., 2025). We draw no conclusions from HaluEval, where lexical separability is near perfect.

6 Discussion

What attention-graph topology measures. In decoder-only models, Vietoris–Rips persistence of an attention graph, its stable summaries and TOHA's divergence depend beyond first-order attention on one scalar per graph: the coning defect, zero in most middle layers of every sink model and a bound on the deviation elsewhere (reporting guidelines in Appendix E).

Limitations. Our scales go 1.1B–7.62B; sink strength varies with scale and training (Gu et al., 2025), so the coned fraction is model-specific, though Theorem 1 is not. HalluZig's zigzag and learned graph models (Samaga et al., 2026; Frasca et al., 2026) are out of scope, as are sampling-based (semantic entropy; Farquhar et al., 2024) and latent-probe (CCS; Burns et al., 2023) baselines, which score quantities other than the attention graph. On RAGTruth's long retrieved contexts the sandwich still has no violations, but the exact case is rare: 11.8% of head graphs are prompt-coned against 57–88% on short-form, falling with context length, while the median per-head rank correlation between d and the $\bar{\pi}$-score holds at 0.955 (Appendix K). TOHA is therefore near-equivalent to a first-order statistic there rather than equal to one, and we do not reproduce its reported detection

accuracy on those benchmarks. On-policy labels use alias matching; an LLM-judge audit finds 5.9% to 11.4% disagreement without changing the orderings we claim (Appendix F).

Reproducibility Statement

All code, per-layer features for 4 settings and on-policy generations for TrQA Mis7B and TrQA Qw7B are at `https://anonymous.4open.science/r/ICLR-BE07/` (run `run_fix.sh`, then `run_toha.sh`); every other setting is reproducible via `extract.py`. Seeds, dtype, proofs: Appendices B, A.

AI Use Statement

We used generative AI tools to aid in proving mathematical claims and to aid and polish the writing of this paper.

References

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18(8): 1–35, 2017.

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

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.02732.

Jonathan Ariel Barmak and Elias Gabriel Minian. Strong homotopy types, nerves and collapses. *Discrete & Computational Geometry*, 47(2):301–328, 2012.

Ulrich Bauer. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, 2021.

Alexandra Bazarova, Andrei Volodichev, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, Andrey Savchenko, Serguei Barannikov, and Alexey Zaytsev. Hallucination detection in LLMs with topological divergence on attention graphs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL Volume 1: Long Papers)*, pp. 15449–15470, 2026. arXiv:2504.10063.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination detection in LLMs using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Jean-Daniel Boissonnat and Siddharth Pritam. Edge collapse and persistence of flag complexes. In *36th International Symposium on Computational Geometry (SoCG)*, volume 164 of *LIPIcs*, pp. 19:1–19:15, 2020.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16:77–102, 2015.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations (ICLR)*, 2023.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Y. Oudot. Proximity of persistence modules and their diagrams. In *Proceedings of the 25th Annual Symposium on Computational Geometry (SoCG)*, pp. 237–246, 2009.

Frédéric Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214, 2014.

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

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

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 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*, pp. 1419–1436, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, 2019.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Herbert Edelsbrunner and John L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

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

Fabrizio Frasca, Guy Bar-Shalom, Yftah Ziser, and Haggai Maron. Neural message-passing on attention graphs for hallucination detection. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.24770.

Alan M. Frieze. On the value of a random minimum spanning tree problem. *Discrete Applied Mathematics*, 10(1):47–56, 1985.

Yuri Gardinazzi, Karthik Viswanathan, Giada Panerai, Alessio Ansuini, Alberto Cazzaniga, and Matteo Biagetti. Persistent topological features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18811–18830, 2025. arXiv:2410.11042.

John C. Gower and Gavin J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 18(1):54–64, 1969.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT*, pp. 107–112, 2018.

Masanori Hino and Shu Kanazawa. Asymptotic behavior of lifetime sums for random simplicial complex processes. *Journal of the Mathematical Society of Japan*, 71(3):765–804, 2019.

Yasuaki Hiraoka and Tomoyuki Shirai. Minimum spanning acycle and lifetime of persistent homology in the Linial–Meshulam process. *Random Structures & Algorithms*, 51(2):315–340, 2017.

Khizar Hussain and Murat Kantarcioglu. PARALLAX: Separating genuine hallucination detection from benchmark construction artifacts. *arXiv preprint arXiv:2605.17028*, 2026.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

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

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Daniel Lütgehetmann, Dejan Govc, Jason P. Smith, and Ran Levi. Computing persistent homology of directed flag complexes. *Algorithms*, 13(1):19, 2020.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*, 2024.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024. arXiv:2401.00396.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025.

Ilan Perez and Raphael Reinauer. The topological BERT: Transforming attention into topology for natural language processing. *arXiv preprint arXiv:2206.15195*, 2022.

Shreyas N. Samaga, Gilberto Gonzalez Arroyo, and Tamal K. Dey. HalluZig: Hallucination detection using zigzag persistence. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026. arXiv:2601.01552.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

A PROOFS

Throughout, D is a symmetric $N \times N$ matrix with non-negative entries and zero diagonal, K_ϵ is the Vietoris–Rips complex at scale ϵ, $\mathrm{filt}_D(\sigma) = \max_{u,v\in\sigma} D_{uv}$ ($\mathrm{filt}_D(\{u\}) = 0$), and homology is taken over a field. Bars are half-open intervals $[b, d)$.

A.1 FACT 1

All vertices enter at 0. Processing edges in non-decreasing weight order, an edge either joins two components, in which case (elder rule) one 0D class dies at its weight and Kruskal's algorithm adds the edge to the spanning forest, or it closes a cycle, in which case no 0D class dies and Kruskal discards it. Hence the $N-1$ finite death times are the MST edge weights (Gower & Ross, 1969; Edelsbrunner & Harer, 2010).

A.2 THEOREM 1

Lemma 1 (Coned filtrations)**.** *If D is coned at s, then for every $\epsilon \geq 0$, $K_\epsilon = C_\epsilon \sqcup I_\epsilon$, where C_ϵ is a simplicial cone with apex s and I_ϵ is a set of isolated vertices. Consequently $\tilde{H}_k(K_\epsilon) = 0$ for every $k \geq 1$.*

Proof. Let $\sigma \in K_\epsilon$ with $s \notin \sigma$ and $|\sigma| \geq 2$. For each $u \in \sigma$ choose $v \in \sigma \setminus \{u\}$; coning gives $D_{su} \leq D_{uv} \leq \mathrm{filt}_D(\sigma) \leq \epsilon$. Therefore $\mathrm{filt}_D(\sigma \cup \{s\}) = \mathrm{filt}_D(\sigma)$ and $\sigma \cup \{s\} \in K_\epsilon$. Now let $u \neq s$ be a vertex with $D_{su} > \epsilon$. If u had an edge $\{u, v\} \in K_\epsilon$ then $D_{su} \leq D_{uv} \leq \epsilon$, a contradiction; so u is isolated. Let I_ϵ be these vertices and C_ϵ the subcomplex on the remaining vertices. Every vertex of C_ϵ is adjacent to s, and by the first part every simplex of C_ϵ extends by s, so C_ϵ is the cone over its link of s and is contractible. H_k of a disjoint union is the direct sum, and isolated points have no homology in degree ≥ 1. □

Proof of Theorem 1. Fix s and $\delta = \delta_s(D)$. Define $D'_{uv} = \max(D_{uv}, D_{su}, D_{sv})$ for $u, v \neq s$, $u \neq v$; $D'_{su} = D'_{us} = D_{su}$; $D'_{uu} = 0$.

Step 1: $D \leq D' \leq D + \delta$ and D' is coned at s. The first inequality is immediate. For $u, v \neq s$, $D'_{uv} - D_{uv} = [\max(D_{su}, D_{sv}) - D_{uv}]_+ \leq \delta$. Coning: $\max(D'_{su}, D'_{sv}) = \max(D_{su}, D_{sv}) \leq D'_{uv}$.

Step 2: interleaving. Since filtration values are maxima of entries, $\mathrm{filt}_D \leq \mathrm{filt}_{D'} \leq \mathrm{filt}_D + \delta$. Hence for every b, $K_b \subseteq K'_{b+\delta} \subseteq K_{b+\delta}$, where K' is the filtration of D'.

Step 3: (a). The persistence map $H_k(K_b) \to H_k(K_{b+\delta})$ is induced by the composite inclusion in Step 2 and therefore factors through $H_k(K'_{b+\delta})$, which is zero for $k \geq 1$ by Lemma 1. The rank of $H_k(K_b) \to H_k(K_{b'})$ equals the number of bars $[b_i, d_i)$ with $b_i \leq b$ and $d_i > b'$. A bar with $d_i - b_i > \delta$ would contribute to this rank at $b = b_i$, $b' = b_i + \delta$, a contradiction. So every bar has length at most δ, and summing gives the bound on P_k.

Step 4: (b). By Fact 1, $P_0(D) = w_D(T_D)$ for an MST T_D of D. The star $T_s = \{\{s, u\}\}_{u\neq s}$ is a spanning tree, so $P_0(D) \leq w_D(T_s) = S_s(D)$. For the lower bound, T_s is an MST of D', even with ties. Root any spanning tree T at s and let e_u be the edge from $u \neq s$ to its parent $p(u)$. If $p(u) = s$ then $D'_{e_u} = D'_{su}$; otherwise $D'_{e_u} = D'_{u\,p(u)} \geq \max(D'_{su}, D'_{s\,p(u)}) \geq D'_{su}$ because D' is coned. The map $u \mapsto e_u$ is a bijection onto the edges of T, so $w_{D'}(T) \geq \sum_{u\neq s} D'_{su} = w_{D'}(T_s)$ and $\max_{e\in T} D'_e \geq \max_u D'_{su}$. Hence $w_{D'}(T_{D'}) = w_{D'}(T_s) = S_s(D)$. The tree T_D spans s, so at least one of its $N-1$ edges is incident to s, where $D' = D$; each of the other at most $N-2$ edges satisfies $D'_e \leq D_e + \delta$. Therefore $P_0(D) = w_D(T_D) \geq w_{D'}(T_D) - (N-2)\delta \geq w_{D'}(T_{D'}) - (N-2)\delta = S_s(D) - (N-2)\delta$. The maximum finite 0D death is the bottleneck value $\min_T \max_{e\in T} w(e)$, which is attained by any MST; it is at most $\max_u D_{su}$ (the star is a spanning tree) and is 1-Lipschitz in the sup-norm, which gives the lower bound because the bottleneck of D' is $\max_u D_{su}$ (by the rooted-tree argument above).

Relation to known results. Lemma 1 is the classical observation that a flag complex in which one vertex is adjacent to every non-isolated vertex, and to every simplex, is a cone; this is the dominated-vertex mechanism behind strong collapses (Barmak & Minian, 2012) and edge collapses of flag filtrations (Boissonnat & Pritam, 2020). Step 3 is a one-sided instance of the interleaving argument

behind the stability of persistence (Cohen-Steiner et al., 2007; Chazal et al., 2009; 2014). The two-sided stability theorem would give bars of length at most 2δ, whereas the one-sided comparison $D \le D'$ gives δ, which is attained (Section 3).

Step 5: (c). If $\delta = 0$ then $D' = D$, and the claims follow from Lemma 1 and Step 4. □

Proof of Corollary 1. Let D' be as in the proof of Theorem 1. Since D' is coned at s, Lemma 1 and Step 4 give $\mathrm{dgm}_k(D') = \Sigma_s(D)_k$ for all k: no bars in degree $k \ge 1$, and 0D deaths equal to the star weights $D'_{su} = D_{su}$. For $1 \le j \le N-1$, the j-th smallest finite 0D death of a matrix E is $w_j(E) = \min\{\epsilon : \beta_0(K_\epsilon^E) \le N - j\}$. If $E \le E'$ entrywise then $K_\epsilon^{E'} \subseteq K_\epsilon^E$ on the same vertex set, so $\beta_0(K_\epsilon^E) \le \beta_0(K_\epsilon^{E'})$ and $w_j(E) \le w_j(E')$; and $w_j(E + c) = w_j(E) + c$ on off-diagonal entries. From $D \le D' \le D + \delta_s$ we get $w_j(D) \le D_{s(j)} \le w_j(D) + \delta_s$. The same inclusions give the Betti-0 sandwich, because $\beta_0(K'_\epsilon) = 1 + \#\{u : D_{su} > \epsilon\}$. Matching the j-th deaths shows that $d_B(\mathrm{dgm}_0(D), \Sigma_s(D)_0) \le \delta_s$. Matching them instead to $t_j = (D_{s(j)} - \delta_s/2)_+$ costs at most $\delta_s/2$ each: if $D_{s(j)} \ge \delta_s/2$ then $w_j \in [D_{s(j)} - \delta_s, D_{s(j)}] = [t_j - \delta_s/2, t_j + \delta_s/2]$, and otherwise $t_j = 0 \le w_j \le D_{s(j)} < \delta_s/2$. For W_1, all differences $D_{s(j)} - w_j$ are non-negative, so the cost of the sorted matching is $\sum_j (D_{s(j)} - w_j) = S_s(D) - P_0(D) \le (N-2)\delta_s$ by Theorem 1(b). In degree $k \ge 1$, every bar has length at most δ_s (Theorem 1(a)), so matching it to the diagonal costs at most $\delta_s/2$ in ℓ^∞; this gives $d_B(\mathrm{dgm}_k(D), \emptyset) \le \delta_s/2$ and the W_1 term $\frac{1}{2}\delta_s |\mathrm{dgm}_k(D)|$. The Lipschitz statements follow from the stability of landscapes in d_B (Bubenik, 2015) and of persistence images in W_1 (Adams et al., 2017). If $\delta_s = 0$ then $D = D'$. □

Proof of Proposition 1. Let $\tilde{D}$ equal D except that $\tilde{D}_{vv'} = 0$ for $v, v' \in P$. At scale 0, Kruskal's algorithm joins all of P into one component with $|P| - 1$ zero-length bars. Every later merge is a merge in the quotient graph Q on $\{p^*\} \cup R$ with $Q_{p^*u} = \min_{v \in P} \tilde{D}_{vu}$ and $Q_{uw} = D_{uw}$. So $\mathrm{MTopDiv}(R, P) = P_0(Q)$. For $u \in R$ and $v \in P$ we have $v < u$ and $D_{uv} = 1 - A_{uv}$, so $Q_{p^*u} = 1 - \pi_u$. For $w < u$ in R, $Q_{uw} = 1 - A_{uw}$, and $\max(Q_{p^*u}, Q_{p^*w}) - Q_{uw} = A_{uw} - \min(\pi_u, \pi_w)$, so $\delta_{p^*}(Q) = \delta_P$. Q has $|R| + 1$ vertices, so Theorem 1(b) gives $\sum_u (1 - \pi_u) - (|R| - 1)\delta_P \le P_0(Q) \le \sum_u (1 - \pi_u)$, with equality when $\delta_P = 0$ by part (c). Divide by $|R|$. Finally $\pi_u \ge A_{u0}$ because $0 \in P$. □

We checked both statements numerically. The fast MTop-Div (a dense Prim's algorithm on Q) agrees with ripser run on $\tilde{D}$ to $3 \cdot 10^{-8}$ on 400 random causal matrices. On these, and on 300 further random matrices checked for the elementwise death bound, the $(N-2)\delta$ sandwich, the $\delta/2$ shifted-star matching, and the bar bound in degrees 1 and 2, there are 0 violations (`sinktda/check_theory.py`). The tightened lower bounds also hold on real attention, with 0 violations on 774,356 layer graphs (15 settings) and 12,336,672 head graphs (11 settings) whose per-example features we retain (`sinktda/check_tight.py`); Table 4 shows how much of the slack real graphs use.

Proof of Corollary 2. For $u > v$, $A_{vu} = 0$, so $D_{uv} = 1 - A_{uv}$; in particular $D_{0u} = 1 - A_{u0}$. For $1 \le v < u$, $\max(D_{0u}, D_{0v}) - D_{uv} = A_{uv} - \min(A_{u0}, A_{v0})$. Substitute into Theorem 1. □

A.3 Linear growth of 1D persistence under i.i.d. weights

The usual reason for length normalizing 1D features is that there is growth in cycles with N. This is understood if weights are i.i.d.: 0D total persistence tends to $\zeta(3)$ (Frieze, 1985) and 1D total persistence of the random clique-complex process is $\Theta(N)$ (Hiraoka & Shirai, 2017; Hino & Kanazawa, 2019). Here is a simple bound that makes explicit the constant.

Proposition 2 (Linear growth under i.i.d. weights)**.** *If $\{D_{uv}\}_{u<v}$ are i.i.d.* Uniform$[0,1]$*, then* $\mathbb{E}[P_1(N)] \ge \int_0^1 \left[\binom{N}{2}p - N - \binom{N}{3}p^3\right]_+ dp = \frac{3N}{8} - \sqrt{3N} + O(1)$*. (There is also an $O(N)$ upper bound in Theorem 1.5 of Hino & Kanazawa, 2019.)*

Proof. With i.i.d. uniform weights, the edges with $D_{uv} \le p$ form an Erdős–Rényi graph $G(N, p)$ and K_p is its clique (flag) complex. Every 1D bar is contained in $[0, 1)$, because K_1 is the full simplex. Writing $\beta_1(p)$ for the first Betti number of K_p, the number of bars alive at p is $\beta_1(p)$, so $P_1 = \int_0^1 \beta_1(p)\, dp$. For any finite complex with f_i faces of dimension i, the strong Morse

Table 4: **How much of Theorem 1's slack real attention uses**, on every layer graph with $\delta_0 > 0$ (on coned graphs both bounds hold with equality). $u_0 = (S_0 - P_0)/((N-2)\delta_0) \in [0,1]$ is the share of the P_0 sandwich used; *vacuous*: share of these graphs with $(N-2)\delta_0 \geq S_0$, where the lower bound says nothing; u_1 = longest H_1 bar$/\delta_0 \in [0,1]$, over graphs with non-empty H_1. Settings whose per-example layer features we retain.

Setting	graphs with $\delta_0>0$	median δ_0	u_0: median	99%	max	vacuous	u_1: median	max
TQA Qw3B	22,248 (38%)	0.34	0.43	0.60	0.66	5.5%	0.16	0.974
TQA Qw1.5B	16,222 (35%)	0.19	0.05	0.65	0.69	0.9%	0.15	0.899
TQA Phi-3	6,062 (12%)	0.29	0.44	0.70	0.80	0.1%	0.08	0.969
TQA TinyLl.	22,948 (64%)	0.21	0.14	0.51	0.59	0.0%	0.15	0.997
TQA SmolLM	39,216 (100%)	0.84	0.69	0.84	0.88	0.0%	0.05	0.274
TQA Mis7B	938 (2%)	0.05	0.04	0.06	0.08	0.0%	0.41	0.655
TQA Mis7B (chat)	939 (2%)	0.05	0.03	0.06	0.07	0.0%	0.40	0.668
TQA Qw7B	45,752 (100%)	0.82	0.73	0.84	0.86	0.0%	0.02	0.186
HE Qw3B	34,038 (47%)	0.16	0.03	0.63	0.80	0.0%	0.11	1.000
HE Qw1.5B	10,501 (38%)	0.12	0.03	0.54	0.60	0.0%	0.18	0.997
TrQA Qw3B	26,656 (37%)	0.34	0.42	0.53	0.55	0.0%	0.17	1.000
TrQA Qw1.5B	27,661 (49%)	0.15	0.03	0.52	0.54	0.0%	0.16	1.000
TrQA Phi-3	8,237 (13%)	0.19	0.44	0.52	0.61	0.0%	0.04	0.999
TrQA TinyLl.	28,000 (64%)	0.22	0.13	0.40	0.44	0.0%	0.07	0.991
TrQA Qw7B	56,000 (100%)	0.84	0.71	0.81	0.84	0.0%	0.02	0.233

inequality at degree 2 reads $\beta_2 - \beta_1 + \beta_0 \leq f_2 - f_1 + f_0$ (Edelsbrunner & Harer, 2010), hence $\beta_1 \geq f_1 - f_0 - f_2 + \beta_0 + \beta_2 \geq f_1 - f_0 - f_2$. Since also $\beta_1 \geq 0$, Jensen's inequality gives $\mathbb{E}\beta_1(p) \geq [\mathbb{E}f_1 - f_0 - \mathbb{E}f_2]_+ = [\binom{N}{2}p - N - \binom{N}{3}p^3]_+$. Integrating over p gives the first inequality. With $p = x/\sqrt{N}$ and $dp = dx/\sqrt{N}$, the integrand divided by $\sqrt{N}$ is $N(x/2 - x^3/6) - \sqrt{N} + O(1)$ uniformly for $x \in [0, \sqrt{3}+1]$. It is positive on an interval whose endpoints are $O(N^{-1/2})$ from 0 and $\sqrt{3}$, and it is negative for larger x. Hence $\int_0^1 [\cdot]_+ dp = N\int_0^{\sqrt{3}} \left(\frac{x}{2} - \frac{x^3}{6}\right)dx - \sqrt{3N} + O(1) = \frac{3N}{8} - \sqrt{3N} + O(1)$. Numerically, $(\frac{3N}{8} - \text{bound})/\sqrt{N} = 1.62, 1.67, 1.70, 1.72$ at $N = 2^8, 2^{10}, 2^{12}, 2^{14}$, approaching $\sqrt{3} \approx 1.73$. The order $\Theta(N)$ itself is known (Hiraoka & Shirai, 2017; Hino & Kanazawa, 2019); the proposition only makes the constant explicit. □

Numerically, the integral equals 0.791, 3.92, 11.9, 30.2, 70.1 at $N = 16, 32, 64, 128, 256$, against simulated means 1.91, 5.83, 15.5, 36.2, 79.3 under i.i.d. Uniform$[0,1]$ weights (Figure 4, left). Over $N \in [16, 256]$ both grow faster than linearly (log-log slopes 1.55 and 1.31), a pre-asymptotic effect of the $-\sqrt{3N}$ term; both are $\Theta(N)$ as $N \to \infty$.

A.4 Directed flag complexes on causal attention

Proposition 3 (Directed flag collapse)**.** *Let A be lower-triangular with non-negative entries (self-loops are ignored). Let* $\mathrm{DFl}(A)$ *be the filtered directed flag complex (Lütgehetmann et al., 2020) of the digraph with an edge $i \to j$ for* every $j < i$ *(a transitive tournament) and weight $1 - A_{ij}$, and let* $\mathrm{VR}(D)$ *be as above with $D = 1 - \max(A, A^\top)$. These filtered complexes are isomorphic, so* $\mathrm{dgm}_k(\mathrm{DFl}(A)) = \mathrm{dgm}_k(\mathrm{VR}(D))$ *for all k. This remains true for any weighting $w(i \to j) = \phi(A_{ij})$ with $D_{ij} = \phi(\max(A_{ij}, A_{ji}))$ and ϕ non-increasing.*

Proof. A directed k-simplex is an ordered tuple $(v_0, \ldots, v_k)$ with an edge $v_a \to v_b$ for all $a < b$. Edges exist only from later to earlier tokens, so a vertex set S supports exactly one directed simplex, its elements in decreasing order; every other ordering uses a missing forward edge. (A missing edge is not a zero-weight edge; encoding it as weight 0 would admit all orderings, a common implementation error.) This bijection preserves faces. Its filtration value is $\max_{a<b}(1 - A_{v_a v_b})$, and for $u > v$, $D_{uv} = 1 - A_{uv}$, so the undirected value $\max_{u,v\in S} D_{uv}$ is the same number. □

We confirm the bijection and the filtration identity explicitly on 420 random causal matrices ($N \leq 8$, four families including tie-dense and underflowing matrices; 30,060 vertex sets; zero violations) and check that it does not hold for bidirectional attention, meaning that a set of size n has $n!$ orderings. The collapse comes from causality: for decoders, adding the directed flag complex brings nothing to

the symmetrized one. This statement does not cover directional features that are not coming from flag complexes, like Kushnareva et al. (2021) for bidirectional encoders.

B Implementation and protocol details

Extraction. Every setting runs in bfloat16 with eager attention and no quantization. The ones using the two 7B models (Mistral-7B on both benchmarks, its chat-format variant, and Qwen2.5-7B on both benchmarks) have been extracted on an NVIDIA RTX 5080 (CUDA), while others were extracted on Apple-silicon (MPS), where float16 produces non-finite values and degenerate generations for Qwen models and the extractor refuses non-finite logits. The paragraph on float32 below measures how much the platform itself moves a reported AUC. Model identifiers are in Table 10 (Hugging Face main revisions as of September 2026). For each example we run one forward pass with attentions and hidden states, average heads, and form D in float64. Ripser (Bauer, 2021) computes 0D/1D barcodes; since SciPy's sparse MST can omit near-zero-weight edges ($A_{u0} \approx 1$), we compute 0D MST summaries using a batched dense Prim's algorithm. The 0D bank is (max lifetime, total persistence) per layer; the 1D bank is (max lifetime, total persistence, birth of the longest bar, death of the longest bar) per layer. Only the 1D summaries are divided by a length: 1D and DEFLATED's 1D block by the sequence length N, ANSWER-ONLY's 1D block by $\max(\text{answer tokens}, 1)$. Everything else, including the 0D, DEFLATED 0D and SINK blocks, enters the classifier raw. That is deliberate: Theorem 1 pairs raw P_0 with raw S_0, and normalizing one of the two would compare banks that differ in their length content rather than in their topology. Both therefore carry the sequence length, which correlates with the label in some settings (Table 12); LEN (sequence and answer length alone) is scored as a separate bank in every setting for exactly that reason, and it is far below the topological banks everywhere except HaluEval (Table 23). LEGACY MSTPROXY, which appears in Tables 23 and 24, is the per-layer 0D total persistence on its own (one feature per layer): the MST-weight summary used by earlier attention-graph work, without the max-lifetime feature. δ_0 applies the closed form of Corollary 2; we brute force compute $\min_s \delta_s$ over all apexes. DEFLATED deletes token 0 and renormalizes each row over the remaining causal support. This is a first-token deletion, not an apex deletion: in the 2 models without a first-token sink (Qwen2.5-7B and SmolLM-1.7B) the apex is elsewhere, so their DEFLATED features remove an ordinary token and are excluded from the C4 numbers in §5.2. Every setting's per-layer apex is recorded (`delta_argmin`); AppendixL re-extracts 3 settings deflating that apex instead, which is the experiment C4 actually calls for. ANSWER-ONLY limits D to the answer tokens and does not renormalize. The Lookback ratio of head h is the mean over answer tokens t of $\bar{a}_{\text{ctx}}/(\bar{a}_{\text{ctx}} + \bar{a}_{\text{new}})$, where $\bar{a}$ is the mean attention from t to prompt tokens and to answer tokens up to t. The LLM-Check score of a head is the mean log diagonal attention (the log-determinant of the triangular attention kernel divided by N), summed over heads per layer. Log-probability features are the mean, minimum, and sum of answer-token log-probabilities, mean and maximum predictive entropy, and mean maximum softmax probability. Hidden states are mean-pooled over answer tokens at layer $\lfloor L/2 \rfloor$.

Prompts. For TruthfulQA we use each model's own format: the Qwen chat template with an explicit system turn "You are a helpful assistant.", `<|user|>...<|end|>` for Phi-3, `<|user|>...</s>` for TinyLlama, ChatML without a system turn for SmolLM, and `[INST]...[/INST]` for Mistral. HaluEval uses `Knowledge: ...\nQuestion: ...\nAnswer:` without a chat template. The boundary is computed by tokenizing the prefix up to but not including the final space, so that the space is attached to the first answer token when the full string is tokenized. When the chat template itself outputs the BOS token (e.g., Mistral's own template, Llama-3.1), it does not output a second one. Both the prefix and the full string are then tokenized in the same way, meaning the token ids of the prefix are a prefix of the full string's. This is verified for all models and templates without a GPU in `sinktda/check_tokenization.py`. HaluEval's raw format is the format that benchmark ships with, and it is what makes the prompt directly comparable across models, but the models we run on it are instruction-tuned, so it puts them outside the format they were tuned on and may change their sink behaviour; we keep it for comparability and draw no conclusion from HaluEval that we do not also draw from a chat-formatted benchmark (§5.3). The on-policy prompt is the model's chat template around "Answer the following question with a short phrase only, no explanation. Question: ...". The token at position 0 is the first token of the format (BOS for TinyLlama and Mistral, `<|user|>` for Phi-3, whose tokenizer adds no BOS, `<|im_start|>` for Qwen and SmolLM, and "Knowledge" for Qwen on HaluEval).

The theorem holds for any value, not necessarily a special token. We generate on-policy answers using the chat-formatted prompt with `add_special_tokens=False` and extract features with the tokenizer's default special tokens. They only differ for TinyLlama, where the input to feature extraction begins with a BOS token. Mistral runs with its hand-written `[INST]` string and with its own chat template. The two prefixes tokenize identically; the settings differ only in the template's end-of-sequence token. Banks that do not rely on the answer window agree to 0.011 AUC. The answer-restricted banks disagree more (answer-only PH by +0.071, deflated PH by +0.023), since that token becomes part of the answer subgraph and the deflated graph.

Sensitivity to inference precision. We also extracted five settings in float16, which lets us measure how much each bank depends on numerical precision (Table 5). All five are clean comparisons: same prompts, same prompt–response boundary, same protocol, precision the only difference. HaluEval Qwen2.5-3B was first extracted in float16 before a prompt–response boundary fix, which mixed precision with a one-token boundary shift; it has since been re-extracted in float16 under the current boundary, so its row isolates precision like the others and is included in every maximum quoted below. On the four clean settings, switching to bfloat16 alters the AUC of head-averaged topological banks by up to 0.058 (TQA Qw3B; at most 0.005 elsewhere), of sink features by up to 0.003, but of non-topological banks by at most 0.001. One more reason to report the precision along with any attention-topology result is that short 1D bars tend to be close to the scale of δ_0 and can be affected by small changes in attention.

Table 5: **AUC change from float16 to bfloat16** (same prompts and protocol). Every row isolates precision: same prompts, same prompt–response boundary, same protocol. The HaluEval Qwen2.5-3B row was re-extracted in float16 under the current boundary, its first float16 run having predated the boundary fix.

Setting	Sink	0D	1D	Defl.	$0D_h$	LogP	Lookb.	Hidden	Non-topo.
TQA Qw3B	-0.003	-0.018	$+0.058$	-0.001	-0.001	-0.001	0.000	0.000	-0.001
TQA Phi-3	-0.001	-0.001	-0.003	0.000	-0.001	0.000	0.000	0.000	0.000
TQA TinyLl.	0.000	0.000	-0.001	-0.001	0.000	0.000	0.000	0.000	0.000
TQA SmolLM	0.000	-0.001	-0.001	$+0.001$	$+0.002$	0.000	0.000	0.000	0.000
HE Qw3B	$+0.001$	-0.005	$+0.001$	$+0.002$	-0.014	0.000	-0.001	0.000	0.000

float32 against bfloat16. In the fp16/bf16 comparison above, we vary precision while keeping hardware fixed. Next, to isolate effects of precision versus platform, we repeated our extraction of 2 settings twice, once in bfloat16 and once in float32, on a single CUDA machine, and compared with the bfloat16 numbers the paper reports (which, for these two settings, were extracted on Apple-silicon MPS). The platform gap (bf16 CUDA against bf16 MPS, and so also torch and scikit-learn versions) is at most 0.005 AUC. The precision gap (float32 against bfloat16 on the same machine) is at most 0.007 (0.007 on the topological banks). Both are well within the ± 0.015 equivalence range, which reveals that the reported numbers transfer to other hardware and float32 does not change a conclusion (Table 6).

Table 6: **Platform and precision sensitivity** (ΔAUC). "bf16 CUDA $-$ bf16 MPS" isolates hardware and library versions at fixed precision; "fp32 $-$ bf16 (CUDA)" isolates precision at fixed hardware.

Setting	Contrast	Sink	0D	1D	Defl.	$0D_h$	LogP	Lookb.	Hidden	Non-topo.
TQA Qw1.5B	bf16 CUDA $-$ bf16 MPS	$+0.001$	$+0.005$	$+0.003$	$+0.003$	-0.002	0.000	$+0.001$	0.000	0.000
TQA Qw1.5B	fp32 $-$ bf16 (CUDA)	$+0.002$	$+0.004$	$+0.001$	-0.007	$+0.005$	$+0.003$	$+0.001$	-0.001	0.000
TQA TinyLl.	bf16 CUDA $-$ bf16 MPS	0.000	-0.001	$+0.001$	-0.001	$+0.001$	0.000	0.000	0.000	0.000
TQA TinyLl.	fp32 $-$ bf16 (CUDA)	0.000	$+0.001$	$+0.002$	$+0.003$	-0.001	0.000	0.000	0.000	0.000

On-policy benchmark. Table 7 presents the model generations produced and how they were evaluated.

Statistics. Bootstrap resamples are drawn over questions (both rows of a pair move together). TOST power at margin m is under a true difference of 0 and a normal sampling distribution with the bootstrap SE: $2\Phi(m/\mathrm{SE} - z_{0.95}) - 1$. Average repeated-CV predictions (seeds 42, 43, 44)

Table 7: **TriviaQA on-policy benchmark.** 2,000 validation questions sampled with seed 0; greedy decoding, first line of the answer, at most 24 new tokens. An answer is correct if any normalized gold alias (lower-cased, punctuation and articles removed) is contained as a whole-word span in the normalized answer. Empty answers are dropped.

Model	questions	accuracy	rows used	mean answer tokens
Qwen2.5-3B	2,000	42.5%	2,000	5.5
Qwen2.5-1.5B	2,000	31.4%	2,000	5.8
Phi-3-mini	2,000	56.2%	2,000	6.0
TinyLlama-1.1B	2,000	26.4%	2,000	18.9
Mistral-7B	2,000	66.5%	2,000	14.0
Qwen2.5-7B	2,000	51.4%	2,000	5.3

are used for bootstrapping (seed 0), i.e., the intervals report sampling of questions with a fixed, split-averaged predictor without refitting variance. Seed-to-seed SD of the pooled AUC is given separately (Appendix M). It is at most 0.008, with 24 out of the bank–setting cells exceeding 0.004. The C rule is a function of bank width alone: banks of at most 64 dimensions are fit at the fixed $C = 1$, wider banks at a C chosen from $\{0.01, 0.1, 1\}$ by inner grouped 3-fold CV on the training fold. Because the rule depends only on width, the same bank always receives the same regularization treatment in every comparison it enters. That grid is narrow for the widest banks, and an under-regularized baseline would understate exactly the comparisons the paper leans on, so we repeated every reported bank in each of the 15 settings whose features we retain with the wider grid $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$, applied to every bank regardless of width (Table 9). The non-topological banks, which are the ones the narrow grid could have disadvantaged, gain at most 0.015 (hidden-state probe 0.010, full non-topological bank 0.006), so they were not materially under-regularized. The topological banks gain more, up to 0.061 (TQA Mis7B Ans.); the wider grid therefore favours the topological side of every comparison, and the conclusions below are the ones that survive that. Answer-only PH gains the most and still overtakes the non-topological bank in 0 of the settings where that bank is available, and across all settings the four dominance orderings the paper claims (NONTOPO $\geq$ 0D, NONTOPO $\geq$ DEFL, HIDDEN $\geq$ 0D, NONTOPO $\geq$ PH-0D) reverse 0 times. The reduction's own comparison is a difference of two banks rather than either one, and AUC(0D) $-$ AUC(SINK) moves by at most 0.023, in the direction of the sink. The conclusions are not an artefact of the grid. Late fusion trains the second-level model on first-level out-of-fold scores as in regular stacking. Training rows' first-level scores are obtained by training models which have seen the corresponding test questions, which is a weak, standard optimism that both arms of the comparison share. Nested stacking removes it: within each outer training fold, every first-level model is refit on inner grouped 5-fold splits of that fold only, the second-level model is trained on those inner out-of-fold scores, and it is applied to the test fold's scores from a first-level model fit on the whole training fold. No input to the stacker then depends on a test label. Table 8 reports it on the 11 settings whose raw features we retain (HE Qw1.5B, HE Qw3B, TQA Phi-3, TQA Qw1.5B, TQA Qw3B, TQA SmolLM, TQA TinyLl., TrQA Phi-3, TrQA Qw1.5B, TrQA Qw3B and TrQA TinyLl.); it moves no stacked comparison by more than 0.002 AUC.

C ADDITIONAL ANALYSES

Models. Table 10 lists the model architectures. All models use grouped or multi-head attention; we average attention over query heads.

Table 10: **Models.**

Model	params	layers	heads	KV heads	hidden
`Qwen/Qwen2.5-3B-Instruct`	3.09B	36	16	2	2048
`Qwen/Qwen2.5-1.5B-Instruct`	1.54B	28	12	2	1536
`microsoft/Phi-3-mini-4k-instruct`	3.82B	32	32	32	3072
`TinyLlama/TinyLlama-1.1B-Chat-v1.0`	1.10B	22	32	4	2048
`HuggingFaceTB/SmolLM-1.7B-Instruct`	1.71B	24	32	32	2048
`mistralai/Mistral-7B-Instruct-v0.2`	7.24B	32	32	8	4096
`Qwen/Qwen2.5-7B-Instruct`	7.62B	28	28	4	3584

Table 8: **Late fusion with nested stacking** (ΔAUC; first-level models refit inside every outer training fold). NT: all non-topological features; δ_h: per-head coning defect; TOHA: supervised probe on TOHA's per-head scores; $\bar{\pi}$: the same probe on the $\bar{\pi}$-score. $^{\equiv}$: equivalent within ± 0.015; *: 95% interval excludes 0; both marks can apply.

Setting	0D⊕NT	Defl.⊕NT	δ_h⊕NT	TOHA⊕NT	TOHA⊕$\bar{\pi}$	δ_h⊕Sink$_h$
TQA Qw3B	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.002^{\equiv}$	$+0.004^{*\equiv}$	$+0.003^{*\equiv}$	$+0.018^{*}$
TQA Qw1.5B	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.002^{\equiv}$	$+0.005^{*\equiv}$	$+0.010^{*\equiv}$
TQA Phi-3	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.010^{*\equiv}$
TQA TinyLl.	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.002^{\equiv}$	$+0.006^{*\equiv}$
TQA SmolLM	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.002^{\equiv}$	$+0.005^{*\equiv}$	$+0.019^{*}$
HE Qw3B	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.018^{*}$
HE Qw1.5B	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.025^{*}$
TrQA Qw3B	$+0.002^{\equiv}$	$+0.001^{\equiv}$	$+0.004^{*\equiv}$	$+0.013^{*}$	$+0.002^{\equiv}$	$+0.013^{*}$
TrQA Qw1.5B	$+0.001^{\equiv}$	$0.000^{\equiv}$	$+0.004^{\equiv}$	$+0.004^{\equiv}$	$+0.002^{\equiv}$	$+0.027^{*}$
TrQA Phi-3	$+0.001^{\equiv}$	$+0.001^{\equiv}$	$+0.002^{\equiv}$	$+0.001^{\equiv}$	$0.000^{\equiv}$	$+0.010^{*\equiv}$
TrQA TinyLl.	$+0.001^{\equiv}$	$+0.002^{\equiv}$	$+0.007^{*\equiv}$	$+0.010^{*}$	$+0.006^{*\equiv}$	$+0.010^{*}$

Table 9: **A wider regularization grid** (ΔAUC, wider grid minus reported). Every bank of every setting whose features we retain is refit with C tuned over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}\}$ by the same inner grouped 3-fold CV, including the banks the reported rule fits at a fixed $C = 1$. Same seeds, splitter and scaler otherwise. 16 banks per setting are compared; the 11 of Table 2, plus answer-only PH (*Ans.*), are shown. 4 settings kept only their per-layer features, so their hidden-state and per-head banks cannot be rebuilt and are shown as dashes.

Setting	Sink	0D	1D	Defl.	Ans.	Sink$_h$	0D$_h$	Lookb.	LogP	Hidden	Non-topo.
TQA Qw3B	+0.027	+0.004	0.000	+0.003	+0.014	0.000	−0.005	0.000	0.000	−0.002	0.000
TQA Qw1.5B	+0.015	+0.004	0.000	+0.011	+0.038	0.000	0.000	0.000	−0.001	0.000	0.000
TQA Phi-3	0.000	+0.001	−0.005	+0.001	+0.046	0.000	0.000	0.000	0.000	0.000	0.000
TQA TinyLl.	+0.003	+0.002	0.000	+0.003	+0.046	0.000	0.000	0.000	0.000	−0.001	−0.001
TQA SmolLM	+0.007	+0.001	0.000	+0.003	+0.038	0.000	0.000	0.000	−0.001	0.000	0.000
TQA Mis7B	+0.001	+0.001	0.000	0.000	+0.061	–	–	–	0.000	–	–
TQA Mis7B (chat)	+0.001	+0.001	0.000	+0.003	+0.024	–	–	–	0.000	–	–
TQA Qw7B	0.000	+0.004	0.000	+0.005	+0.018	–	–	–	0.000	–	–
HE Qw3B	+0.004	+0.008	−0.001	+0.018	+0.001	0.000	0.000	−0.001	0.000	0.000	0.000
HE Qw1.5B	+0.002	+0.004	0.000	0.000	−0.001	0.000	0.000	0.000	+0.003	0.000	0.000
TrQA Qw3B	+0.009	+0.001	−0.001	+0.021	0.000	0.000	+0.002	0.000	0.000	+0.009	+0.006
TrQA Qw1.5B	+0.003	0.000	−0.005	+0.006	+0.020	0.000	−0.002	0.000	−0.001	+0.008	+0.002
TrQA Phi-3	+0.004	+0.005	−0.011	+0.008	+0.001	0.000	+0.010	0.000	−0.007	+0.004	+0.002
TrQA TinyLl.	+0.008	+0.007	0.000	+0.005	+0.043	0.000	0.000	0.000	0.000	+0.003	0.000
TrQA Qw7B	0.000	−0.001	−0.005	+0.001	0.000	–	–	–	0.000	–	–

Length scaling of 1D persistence on real attention. Table 11 regresses log total 1D persistence (summed over layers) on $\log N$ with a class fixed effect, on rows with $P_1 > 0$, so that the exponent is not driven by a length difference between correct and hallucinated answers. Exponents range from 0.03 to 1.39, against 1.31 for i.i.d. uniform weights and 1.27 for random causal attention without a sink over the same range of N. 12 of the 13 fitted settings grow significantly more slowly than the i.i.d. reference (not HE Qw3B). The exponent exceeds 1 in HE Qw1.5B and HE Qw3B, so growth on real attention is not uniformly sub-linear. TQA Mis7B and TQA Mis7B (chat) have too few rows with $P_1 > 0$ to fit (first column): almost all of their graphs are coned (Table 1), so 1D persistence is zero for almost every example whatever its length. In the other settings almost every example has some 1D bar summed over layers, because those models have some layers that are not coned; within an example, the coned layers contribute nothing to P_1 whatever the length (Table 1).

Table 11: **Length scaling of 1D persistence.** OLS of $\log P_1$ on $\log N$ and the class label, fit on the rows with $P_1 > 0$; the first column is the share of rows that are retained, and the rows with $P_1 = 0$ are dropped from the fit. Rows need per-example layer features, which we retain for 15 of the 16 settings (not TrQA Mis7B).

Setting	rows with $P_1>0$	$\hat{\alpha}$ (class FE)	95% CI	R^2	Spearman(P_1, N)
TQA Qw3B	100.0%	0.25	[0.23, 0.27]	0.27	0.57
TQA Qw1.5B	100.0%	0.57	[0.52, 0.61]	0.29	0.51
TQA Phi-3	99.9%	0.93	[0.82, 1.04]	0.15	0.39
TQA TinyLl.	100.0%	0.32	[0.28, 0.36]	0.14	0.28
TQA SmolLM	100.0%	0.61	[0.54, 0.69]	0.13	0.36
TQA Mis7B	0.2%	–	–	–	–
TQA Mis7B (chat)	0.2%	–	–	–	–
TQA Qw7B	100.0%	0.59	[0.55, 0.62]	0.41	0.63
HE Qw3B	100.0%	1.39	[1.31, 1.47]	0.36	0.60
HE Qw1.5B	100.0%	1.10	[1.07, 1.13]	0.84	0.91
TrQA Qw3B	100.0%	0.72	[0.68, 0.75]	0.46	0.68
TrQA Qw1.5B	100.0%	0.75	[0.72, 0.78]	0.59	0.68
TrQA Phi-3	100.0%	0.70	[0.65, 0.75]	0.29	0.49
TrQA TinyLl.	100.0%	0.03	[0.02, 0.03]	0.03	0.15
TrQA Qw7B	100.0%	0.84	[0.81, 0.87]	0.57	0.65
i.i.d. uniform weights (simulation)	100%	1.31	–	–	–
random causal attention, no sink (simulation)	100%	1.27	–	–	–

Sequence lengths. Table 12 gives class-conditional lengths. TruthfulQA is length-balanced by construction; HaluEval's hallucinated answers are systematically longer; on-policy, incorrect answers differ in length from correct ones because they are the model's own.

Table 12: **Class-conditional sequence length** (mean $\pm$ SD, full sequence) and mean answer length. Rows need per-example layer features, which we retain for 15 of the 16 settings (not TrQA Mis7B).

Setting	length N (correct)	length N (hallucinated)	answer tokens (c./h.)	point-biserial r	Welch p
TQA Qw3B	44.4 ± 7.9	44.0 ± 8.0	12.6 / 12.1	-0.030	0.22
TQA Qw1.5B	44.4 ± 7.9	44.0 ± 8.0	12.6 / 12.1	-0.030	0.22
TQA Phi-3	29.2 ± 8.9	28.7 ± 8.8	11.8 / 11.4	-0.024	0.32
TQA TinyLl.	42.5 ± 9.0	42.1 ± 8.9	12.1 / 11.7	-0.024	0.32
TQA SmolLM	32.6 ± 8.2	32.1 ± 8.2	10.6 / 10.2	-0.026	0.29
TQA Mis7B	33.3 ± 8.7	32.8 ± 8.6	11.4 / 10.9	-0.029	0.25
TQA Mis7B (chat)	34.3 ± 8.7	33.8 ± 8.6	12.4 / 11.9	-0.029	0.25
TQA Qw7B	44.4 ± 7.9	44.0 ± 8.0	12.6 / 12.1	-0.030	0.22
HE Qw3B	121.2 ± 39.7	132.7 ± 41.2	3.8 / 15.4	+0.141	2.1×10^{-10}
HE Qw1.5B	122.2 ± 39.5	133.9 ± 41.3	3.6 / 15.3	+0.144	5.1×10^{-6}
TrQA Qw3B	68.0 ± 10.4	69.0 ± 9.7	5.3 / 5.7	+0.052	0.02
TrQA Qw1.5B	67.5 ± 10.3	69.4 ± 10.0	5.4 / 5.9	+0.086	1.3×10^{-4}
TrQA Phi-3	45.6 ± 11.3	46.9 ± 11.2	5.6 / 6.5	+0.057	0.01
TrQA TinyLl.	67.3 ± 12.5	73.7 ± 14.9	17.0 / 19.6	+0.196	1.8×10^{-21}
TrQA Qw7B	67.4 ± 9.6	69.2 ± 10.2	5.0 / 5.5	+0.092	3.8×10^{-5}

Margin sensitivity. Table 13 reports TOST p-values at four equivalence margins for the four comparisons of Figure 5; bold marks $p < 0.05$ (equivalence).

Table 13: **TOST p-values by margin m.** NT: all non-topological features.

Setting	Sink vs. 0D				0D | Sink				1D | 0D				0D | NT			
m	.010	.015	.020	.025	.010	.015	.020	.025	.010	.015	.020	.025	.010	.015	.020	.025
TQA Qw3B	**.029**	**.001**	**<.001**	**<.001**	**.040**	**.001**	**<.001**	**<.001**	1.000	.995	.972	.882	**<.001**	**<.001**	**<.001**	**<.001**
TQA Qw1.5B	.109	**.016**	**<.001**	**<.001**	.692	.276	**.040**	**.002**	.311	**.034**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TQA Phi-3	.138	**.013**	**<.001**	**<.001**	.112	**.003**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TQA TinyLl.	.520	.234	.064	**.012**	.962	.809	.483	.169	.210	**.021**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TQA SmolLM	.939	.865	.748	.602	1.000	.996	.960	.805	.179	**.036**	**.003**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TQA Mis7B	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TQA Mis7B (chat)	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TQA Qw7B	1.000	1.000	1.000	1.000	1.000	.999	.964	.719	.962	.836	.578	.281	**<.001**	**<.001**	**<.001**	**<.001**
HE Qw3B	.910	.676	.351	.110	1.000	1.000	.982	.804	.906	.551	.148	**.012**	**<.001**	**<.001**	**<.001**	**<.001**
HE Qw1.5B	.993	.970	.896	.755	.679	.263	**.041**	**.002**	.980	.918	.786	.569	**<.001**	**<.001**	**<.001**	**<.001**
TrQA Qw3B	.792	.522	.239	.069	.982	.866	.561	.219	.106	**.015**	**.002**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TrQA Qw1.5B	.431	.176	**.050**	**.010**	.528	.211	**.044**	**.005**	.079	**.011**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TrQA Phi-3	**.047**	**.002**	**<.001**	**<.001**	.114	**.005**	**<.001**	**<.001**	.086	**.003**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TrQA TinyLl.	.091	**.021**	**.003**	**<.001**	.071	**.005**	**<.001**	**<.001**	.083	**.013**	**.002**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**
TrQA Mis7B	.111	**.023**	**.003**	**<.001**	.052	**.004**	**<.001**	**<.001**	.315	.096	**.017**	**.002**	**<.001**	**<.001**	**<.001**	**<.001**
TrQA Qw7B	.993	.982	.962	.924	.959	.852	.606	.314	.054	**.007**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**	**<.001**

Per-layer signal. Figure 2 shows single-feature AUCs by depth. P_0 and the star weight trace nearly identical curves in every model with a sink, while P_1 carries signal only in the layers that are not coned. These AUCs are orientation-free ($\max(\text{AUC}, 1 - \text{AUC})$) and therefore slightly optimistic; they are descriptive, not a detector.

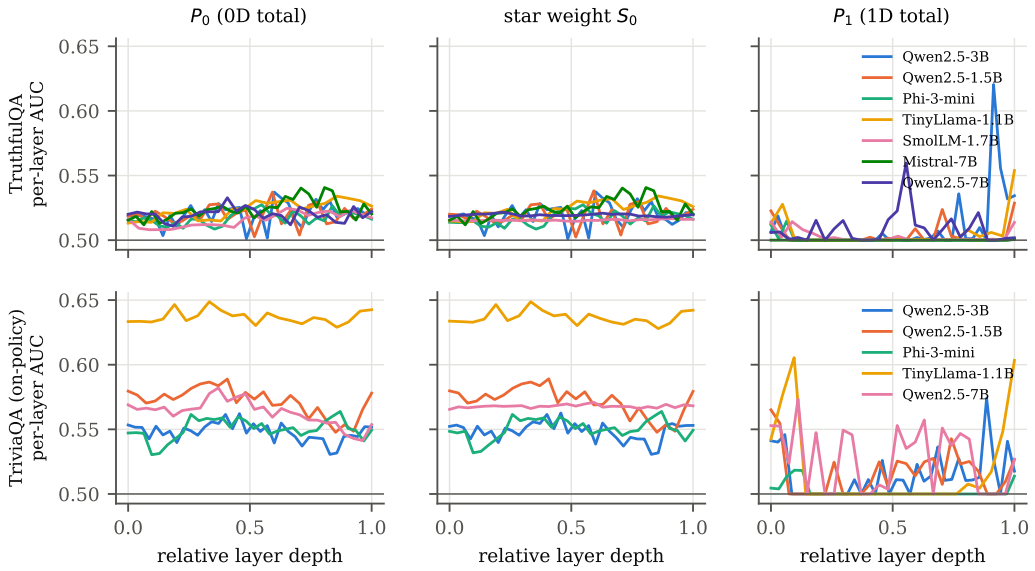


Figure 2: **Single-feature AUC by layer.** Curves need per-example layer features, which we retain for 15 of the 16 settings (not TrQA Mis7B). Mistral-7B is shown in its `[INST]` format only.

Worked example. Figure 3 shows TruthfulQA question 67 with its incorrect answer ("Yes, vampires are real") under TinyLlama-1.1B (N=29). In layer 5 the model sends 0.85 of its attention to token 0 on average, $\delta_0 = 0.000$, and accordingly $P_0 = 4.225$ equals the star weight $S_0 = 4.225$ and there are 0 H_1 bars. In layer 21 sink attention drops to 0.53 and $\delta_0 = 0.281$; now $P_0 = 12.287$ falls below $S_0 = 13.043$ (within the bound $(N-2)\delta_0 = 7.59$), and 10 H_1 bars appear, the longest of length $0.172 \leq \delta_0$. The theorem bounds loops but does not force them: in layers 2 and 11–18 of the same example δ_0 lies between 0.11 and 0.52, yet H_1 is empty.

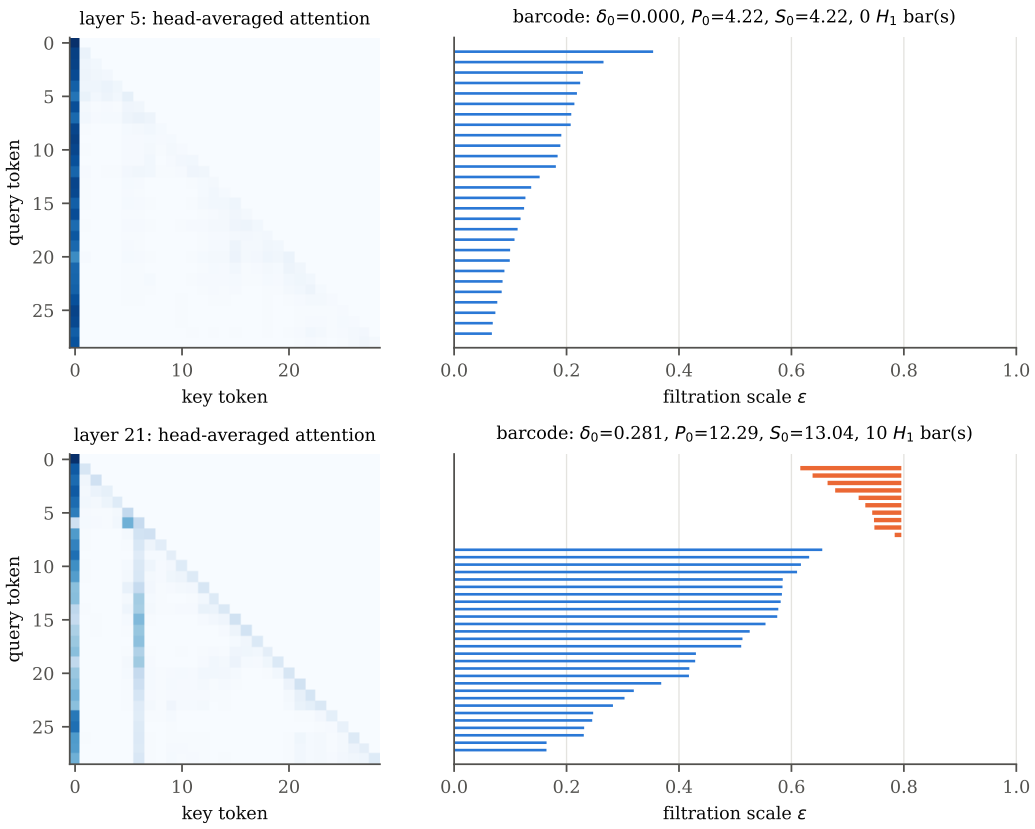


Figure 3: **A coned and a non-coned layer of the same example** (TinyLlama-1.1B, TruthfulQA). Left: head-averaged attention; the bright first column is the sink. Right: 0D bars (blue) and 1D bars (orange) of the Vietoris–Rips filtration of $D = 1 - \max(A, A^\top)$.

D ONE-DIMENSIONAL PERSISTENCE

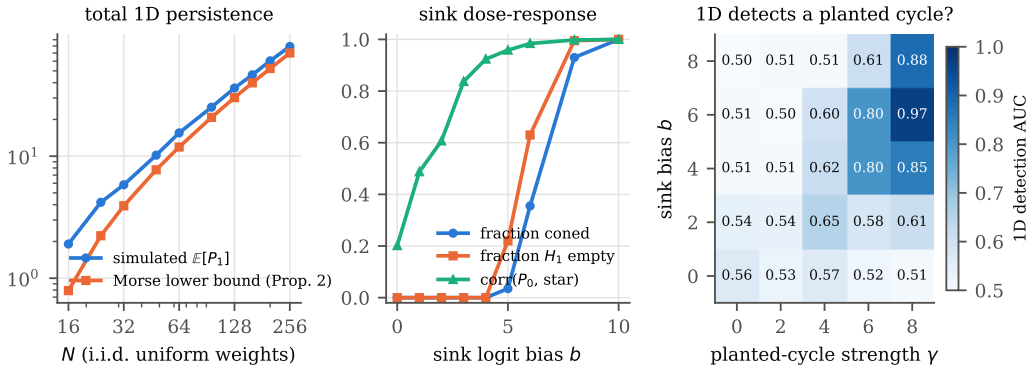


Figure 4: **Synthetic checks.** Left: Lower bound of Proposition 2 compared to simulated mean total 1D persistence under i.i.d. Uniform$[0, 1]$ weights (20 draws per $N \leq 128$, 8 above). Middle: random causal attention ($N = 40$, 200 draws per point) with a logit bias b on token 0: graphs become exactly coned, H_1 empties, and P_0 becomes a function of the star weight, growing as the sink becomes stronger. Right: AUC of 1D total persistence for detecting a planted 6-cycle of logit strength γ (150 draws per class). Without a sink ($b = 0$) the planted cycle is lost among background cycles; under a strong sink ($b = 8$) it is invisible unless it is strong enough to break the cone.

Under i.i.d. weights, 1D persistence is thus growing faster than linearly in the lengths relevant for us and linearly asymptotically (Proposition 2; Figure 4, left). For real attention, almost no 1D bars exist (Table 1) and they appear only on graphs with a non-zero coning defect. Figure 4 (middle) simulates this change in a controlled manner: as we increase the logit bias of the sink, random causal attention goes from looking i.i.d.-like to doing exact coning, while H_1 empties out in lockstep.

This explains a common finding in practice, that 1D features add little to 0D. At full sample size, adding length-normalized 1D to 0D changes AUC by -0.008 to +0.033, which is equivalent within ± 0.015 in 9 of 13 sink settings; the exceptions are HE Qw1.5B, HE Qw3B, TQA Qw3B and TrQA Mis7B; of these, adding 1D has a significant positive effect on AUC in HE Qw1.5B, HE Qw3B and TQA Qw3B; the other exception has a negative point estimate but an interval that is too wide to be certain about equality. Note that the improvement seen in TruthfulQA Qwen2.5-3B is also the most sensitive result in terms of inference precision (Appendix B). We want to emphasize that this is not because loops are generally uninformative. For instance, in a strongly sinked model, if a loop exists, there must be some token paying more attention to another token than one of those two tokens is paying to the sink, and the duration of such a loop will be limited by how much extra attention one of them pays to the other. See Figure 4 (right). H_1 can readily find a planted 6-cycle under a moderate sink ($b = 6, \gamma = 8$: AUC 0.97). But a strong sink ($b = 8, \gamma \leq 4$: AUC $\leq$0.51) will mask it unless it is strong enough to shatter the cone. Without any sink ($b = 0$) H_1 will also fail (AUC $\leq$0.57), since the random attention will already have plenty of background cycles. 1D persistence is useful only in between.

E Recommendations for reporting attention-graph topology

(i) Compare a topological detector against the first-order statistic named by the reduction, and report sink attention and δ_0 (or the relevant defect, such as δ_P for TOHA). (ii) Compare topological detectors against per-head non-topological statistics (sink mass, entropy, Lookback ratios) at the same resolution, as well as against a hidden-state probe. (iii) Evaluate on the model's own generations rather than using teacher-forced benchmarks. These mix detection and recognition of dataset authorship. (iv) If we want to go beyond the sink, either remove the sink first (deflation) or consider only answer tokens. These reveal real non-sink geometry, but in our settings it was not new information. Deflate the token that is actually the apex, which need not be token 0 (§5.1). (v) Report the inference precision and the hardware. Short 1D bars sit at the scale of δ_0, and switching float16 to bfloat16 moves the AUC of 1D banks by up to 0.058 while leaving non-topological banks within 0.001 (Appendix B).

F Validation of the TOHA implementation and the on-policy labels

Against the authors' released code. Our per-head score d is our own implementation, so we check it against the routine the TOHA authors release (Bazarova et al., 2026; `src/methods/mtopdiv/utils.py`), which builds the same distance matrix and calls ripser rather than our dense Prim's algorithm. Over 21,696 head graphs from 2 settings the two agree to 1.2×10^{-8} (Table 15). So the reduction in Proposition 1 is not an artefact of how we compute d.

Per-setting diagnostics. Table 14 gives, for every setting, the quantities behind §5.4: the identity check, prompt-coning, whether token 0 is every response token's most-attended prompt token (*sink top*), the median per-head rank correlations of d with the $\bar{\pi}$-score and with the response sink score $1 - \bar{a}_0$, and the same coning and sink shares on the heads that TOHA's selection picks.

LLM-judge audit of the on-policy labels. The on-policy label is an alias string match, which mislabels paraphrases and hedged answers. We had Qwen2.5-7B-Instruct grade all 2000 answers of each of 2 on-policy settings from the question, the gold aliases and the answer. Disagreement is 5.9% to 11.4% (Table 16). On TrQA Mis7B, graded by a different model, it is mostly the string match being too lax (133 of the disagreements, against 95 the other way); on TrQA Qw7B the two directions are closer to balanced (35 against 84). Re-scoring the same out-of-fold predictions on those rows under the judge's labels shifts every bank by at most 0.027 AUC and leaves the dominance orderings the paper relies on intact (NONTOPO $\geq$ 0D, NONTOPO $\geq$ DEFL, HIDDEN $\geq$ 0D, NONTOPO $\geq$ PH-0D),

Table 14: **TOHA diagnostics.** *viol.*: violations of Proposition 1; *max gap*: largest $|d - \bar{\pi}\text{-score}|$ on prompt-coned heads; ρ: median over heads of the Spearman correlation across examples; *sel.*: over the heads TOHA's Algorithm 1 selects in each training fold (mean number in the last column).

Setting	head graphs	viol.	$\delta_P{=}0$	max gap	sink top	$\rho(d, \bar{\pi})$	$\rho(d, 1{-}\bar{a}_0)$	sel. $\delta_P{=}0$	sel. sink top	sel. heads
TQA Qw3B	941,184	0	57%	$6 \cdot 10^{-8}$	41%	0.976	+0.888	80%	19%	4.4
TQA Qw1.5B	549,024	0	58%	$6 \cdot 10^{-8}$	45%	0.978	+0.893	74%	34%	8.6
TQA Phi-3	1,673,216	0	74%	0	71%	0.996	+0.985	74%	69%	3.5
TQA TinyLl.	1,150,336	0	74%	0	63%	0.995	+0.967	83%	44%	4.5
TQA SmolLM	1,254,912	0	58%	0	0%	0.954	-0.087	76%	0%	6.6
TQA Mis7B	1,673,216	0	81%	0	83%	0.998	+0.992	64%	72%	3.7
TQA Mis7B (chat)	1,673,216	0	80%	0	83%	0.998	+0.992	60%	77%	3.0
TQA Qw7B	1,281,056	0	57%	0	0%	0.980	-0.110	66%	0%	8.1
HE Qw3B	1,152,000	0	68%	$6 \cdot 10^{-8}$	42%	0.987	+0.777	70%	28%	6.7
HE Qw1.5B	336,000	0	69%	$6 \cdot 10^{-8}$	46%	0.988	+0.842	71%	59%	9.8
TrQA Qw3B	1,152,000	0	76%	$6 \cdot 10^{-8}$	40%	0.991	+0.868	89%	70%	8.1
TrQA Qw1.5B	672,000	0	77%	$6 \cdot 10^{-8}$	42%	0.992	+0.894	88%	51%	8.6
TrQA Phi-3	2,048,000	0	88%	0	75%	0.998	+0.987	95%	79%	9.1
TrQA TinyLl.	1,408,000	0	66%	0	42%	0.994	+0.961	92%	62%	9.4
TrQA Mis7B	2,048,000	0	73%	0	67%	0.992	+0.971	63%	33%	4.6
TrQA Qw7B	1,568,000	0	76%	$6 \cdot 10^{-8}$	0%	0.994	-0.164	93%	0%	9.7

Table 15: **Our d against the authors' MTop-Div code.** Δ is the per-head difference; both are divided by $|R|$.

Setting	head graphs	max $\lvert\Delta\rvert$	mean $\lvert\Delta\rvert$	Pearson r
TQA Mis7B	12,288	0	0	1.00000000
TQA Qw7B	9,408	1.2×10^{-8}	1.1×10^{-11}	1.00000000

so the conclusions do not rest on the labeling. The audit now covers every non-empty generation in both settings, so it is no longer a subsample; the standard error of an AUC on $n = 2000$ rows is about 0.011, and near-ties closer than that between other banks can still reorder. Only the four orderings above are claimed. The judge is itself a language model, not a human annotator, and TrQA Qw7B is graded by the same model that produced its answers, so its 5.9% should be read as a floor rather than an estimate.

Table 16: **LLM-judge audit of the on-policy labels.** "match 0/judge 1": the string match calls it hallucinated and the judge does not. Last column: largest change in any bank's AUC on the audited rows when the judge's labels replace the string match.

Setting	judged	disagree	match 0/judge 1	match 1/judge 0	max $\lvert\Delta\text{AUC}\rvert$
TrQA Mis7B	2,000	11.4%	95	133	0.027
TrQA Qw7B	2,000	5.9%	84	35	0.014

G INTERVENING ON THE SINK

Proposition 1 states that TOHA should correlate more closely with the response sink score the stronger the sink. We can check this directly by intervening on the sink itself. For Qwen2.5-1.5B and TinyLlama on TruthfulQA, we add a constant $b \in \{-2, 0, 2, 4\}$ to the first token's attention logit in every head and recompute all per-head quantities (Table 17). The intervention does not affect the identity (no violations in any run). It affects the structure the proposition says matters. Weakening the sink ($b = -2$) reduces the fraction of prompt-coned head graphs from 58% to 37% (Qwen) and from 74% to 63% (TinyLlama), the fraction where the sink is every response token's top prompt token from 45% to 21% and from 63% to 21%, and the median rank correlation between d and the response sink score from 0.893 to 0.605 and from 0.967 to 0.406. Strengthening it ($b = 4$) increases the correlation to 0.962 and 0.996. In TinyLlama all three numbers increase monotonically with b. In Qwen they peak at $b = 2$ and drop a little at $b = 4$. TOHA's own AUC hardly changes (0.65 to 0.67 and 0.65 to 0.70). The gap to the sink-score version is not monotone in b, showing that our intervention affects TOHA's accuracy through the attention graph rather than being a dose-response.

Table 17: **Sink-bias intervention** (teacher-forced TruthfulQA; labels unchanged). A constant b is added to the attention logit of key 0 for every query after token 0, in every head and layer. $\delta_P{=}0$: share of prompt-coned head graphs. Sink top: share of head graphs in which token 0 is every response token's most-attended prompt token. ρ: median per-head rank correlation between TOHA's d and the response sink score. Remaining columns: AUCs as in Table 3.

Model	b	$\delta_P{=}0$	sink top	$\rho(d, 1-\bar{a}_0)$	TOHA	TOHA$[\bar{a}_0]$	TOHA$_{\text{sup}}$	$\bar{a}_{0,\text{sup}}$
Qwen2.5-1.5B	-2	37%	21%	0.605	0.667	0.624	0.843	0.849
Qwen2.5-1.5B	0	58%	45%	0.893	0.646	0.628	0.858	0.857
Qwen2.5-1.5B	$+2$	72%	60%	0.970	0.667	0.637	0.857	0.857
Qwen2.5-1.5B	$+4$	69%	56%	0.962	0.647	0.640	0.839	0.838
TinyLlama-1.1B	-2	63%	21%	0.406	0.698	0.626	0.845	0.834
TinyLlama-1.1B	0	74%	63%	0.967	0.666	0.618	0.845	0.848
TinyLlama-1.1B	$+2$	81%	75%	0.992	0.652	0.610	0.848	0.848
TinyLlama-1.1B	$+4$	86%	83%	0.996	0.665	0.616	0.844	0.842

H MISTRAL-7B

Mistral-7B-Instruct-v0.2 is the most sink-dominated model in this study, and the reduction is correspondingly tightest there. In its `[INST]` format it puts 0.75 of its attention on token 0 — more than any smaller model, and 98.2% of its 52,288 layer graphs are exactly coned ($\delta_0 = 0$), with 0 violations of any bound in Theorem 1.

Its 1D persistence is exactly zero on 99.99% of its graphs, and $\delta_0 = 0$ on 98.2%, so near-universal coning is what empties its 1D bank: adding 1D to 0D changes AUC by 0.000 (Table 22). The length-normalized rank correlation between $P_0/(N{-}1)$ and $1 - \bar{m}$ has median 1.000 over layers and never drops below 0.994, so on this model 0D total persistence is, up to rank, an affine function of sink attention in every layer. Per head, 81% of its graphs are exactly prompt-coned. Its median $\rho(d, \bar{\pi})$ is 0.998, and Proposition 1 has 0 violations.

I WHAT THE THEORY PREDICTS FOR ACCURACY-LEVEL COMPARISONS, AND COST

Four regularities that an accuracy-only comparison of attention-graph PH would observe follow from the theory: (i) 0D persistence is an MST statistic (Fact 1); (ii) 1D adds little beyond 0D (Table 22, column "1D | 0D"; Theorem 1(a)); (iii) where coning is near-universal, the 1D bank is almost all zeros, e.g. 99.99% of Mistral-7B's graphs in its `[INST]` format (Appendix H); (iv) HaluEval answers can be separated from text alone (§5.3).

Cost. Median per-matrix wall-clock time on real attention from TinyLlama-1.1B-Chat-v1.0 (sink features / δ_0 / dense MST / ripser 0D / ripser 0D+1D) is 0.001/0.02/0.10/0.02/0.02 ms at $N{=}32$; 0.001/0.03/0.21/0.03/0.03 ms at $N{=}64$; 0.001/0.07/0.42/0.07/0.06 ms at $N{=}128$; 0.001/0.25/0.90/0.18/0.18 ms at $N{=}256$. On i.i.d. random matrices of the same size, ripser 0D+1D takes 0.171 ms ($N{=}32$), 3.46 ms ($N{=}64$), 115 ms ($N{=}128$), 4,631 ms ($N{=}256$). On real attention the same computation is four orders of magnitude cheaper at $N{=}256$, consistent with Theorem 1: sink-dominated graphs have few, short 1D bars to compute. Sink features, read off one column, are one to two orders of magnitude cheaper still.

J THEOREM 1 ON A LONG RETRIEVED CONTEXT

Appendix K asks what a retrieved passage does to the *prompt-level* defect that governs TOHA. This asks the same of the whole-graph defect δ_0 that governs Theorem 1. It needs no hallucination labels, because the theorem is a statement about the attention graph: we run the per-layer path on 396 RAGTruth responses with Qwen2.5-1.5B at sequence lengths up to about 2,000 tokens and report the coned share, δ_0 and the length-normalized rank correlation, exactly as Table 1 does for the short-form settings. Only 0D persistence is computed (`--maxdim 0`); a dense Vietoris–Rips filtration to maximum homology dimension 1 at that length is not affordable, so this appendix makes no claim about H_1.

Coning thins out, and much less than the prompt-level defect did. Across 11,088 layer graphs, 30.1% are exactly coned, against 50.6–64.5% for the same model on the three short-form benchmarks. The decline is monotone in length, from 39.7% on the shortest contexts to 22.9% on the longest, and the median δ_0 rises from 0.042 to 0.091 (Table 18). Mean attention to token 0 falls to 0.361, against 0.45–0.53 on short-form: attention does spread over a retrieved passage, as one would expect. But the defect stays small in absolute terms, a median of 0.069 against the 0.34 that the prompt-level defect reaches on the same corpus.

The reduction itself barely moves. The median per-layer rank correlation between $P_0/(N-1)$ and $1-\bar{m}$ is 0.9999 overall and 0.9997 to 0.9999 across the 4 length buckets, against 0.9997–1.0000 on short-form. Length-normalized 0D total persistence is therefore still, up to rank, one minus the mean attention paid to the first token, on contexts an order of magnitude longer than anything in the body. The median hides a tail: individual layers fall as low as -0.47 in the longest bucket, so the statement is about the typical layer and not about every one of them.

Read together with Appendix K. The two defects behave differently on the same corpus. δ_P, which sits between TOHA and its first-order counterpart, grows until the exact case is rare (11.8% of head graphs prompt-coned). δ_0, which sits between the whole barcode and the star, grows far less, and the correlation the reduction actually rests on is essentially unchanged. A long prompt dilutes the prompt–response interface that TOHA measures more than it dilutes the sink itself.

Table 18: **Theorem 1 on RAGTruth** (Qwen2.5-1.5B, 396 responses, 0D only). *coned*: share of layer graphs with $\delta_0 = 0$. $\rho_{/N}$: median over layers of the Spearman correlation across examples between $P_0/(N-1)$ and $1-\bar{m}$. Buckets are by sequence length in tokens; buckets with fewer than 20 examples are omitted.

Seq. len.	examples	layer graphs	coned	median δ_0	sink attn.	$\rho_{/N}$
all	396	11,088	30.1%	0.069	0.361	0.9999
0–512	105	2,940	39.7%	0.042	0.375	0.9999
512–768	130	3,640	29.8%	0.069	0.359	0.9999
768–1024	81	2,268	26.3%	0.076	0.361	0.9998
1024–1536	67	1,876	22.9%	0.091	0.348	0.9997

K TOHA ON A LONG RETRIEVED CONTEXT

Every benchmark in the body puts tens of tokens in the prompt. TOHA was built for retrieval-augmented generation, where the prompt is a retrieved passage, and Proposition 1's error term is a prompt-level coning defect δ_P: a long prompt gives it more room to grow. We therefore run the per-head extractor on RAGTruth (Niu et al., 2024), which pairs real retrieved contexts with responses annotated for hallucinated spans, using Qwen2.5-1.5B on 1,000 responses spanning all three task types (prompts up to 2,300 tokens). Only the per-head path is run: d, the $\bar{\pi}$-score, δ_P and the prompt-coned share are closed forms or dense Prim's, with no Vietoris–Rips filtration to compute at that length.

The bound holds; the equality does not. On 336,000 head graphs the sandwich has 0 violations, and on the heads where δ_P vanishes d still equals the $\bar{\pi}$-score to 6×10^{-8}. But those heads are now the minority: 11.8% of head graphs are exactly prompt-coned, against 57–88% on the short-form settings, and the median δ_P is 0.34. Bucketing by sequence length makes it monotone: the coned share falls from 18.0% on the shortest contexts to 7.1% on the longest, while the median δ_P rises from 0.27 to 0.39 (Table 19). The exact identity is a short-prompt phenomenon.

The ordering survives anyway. What does not degrade is how tightly d tracks the statistic it reduces to. The median per-head rank correlation between d and the $\bar{\pi}$-score is 0.955 overall and 0.951 to 0.954 across the 5 length buckets, flat where the coned share falls by more than half; the same model gives 0.978 on short-form TruthfulQA. With the response sink score it is 0.917. So

on long retrieved contexts TOHA is rarely *equal* to a first-order attention statistic and remains very nearly a monotone function of one. A detector that ranks examples, which is what an AUC measures, is therefore still reproduced by the first-order score; a claim that the two are identical is not. We report this as a boundary on Proposition 1, not as a failure of it: the theorem bounds d by δ_P and δ_P is simply larger here.

Table 19: **TOHA on RAGTruth** (Qwen2.5-1.5B, 1,000 responses). *viol.*: violations of the Proposition 1 sandwich. *coned*: share of head graphs with $\delta_P = 0$, on which d equals the $\bar{\pi}$-score exactly. ρ: median over heads of the Spearman correlation across examples. Buckets are by total sequence length in tokens.

Seq. len.	examples	head graphs	viol.	coned	median δ_P	$\rho(d, \bar{\pi})$	$\rho(d, 1 - \bar{a}_0)$
all	1,000	336,000	0	11.8%	0.338	0.955	0.917
0–512	254	85,344	0	18.0%	0.272	0.953	0.917
512–768	294	98,784	0	12.1%	0.331	0.951	0.911
768–1024	233	78,288	0	9.0%	0.369	0.953	0.926
1024–1536	181	60,816	0	7.6%	0.387	0.952	0.921
1536+	38	12,768	0	7.1%	0.391	0.954	0.912

L DEFLATING THE APEX INSTEAD OF TOKEN 0

Prediction C4 says that removing the sink should expose whatever non-sink topology exists, and DEFLATED tests it by deleting token 0. That is the right vertex only when token 0 is the apex. Across the 3 settings re-extracted here (TQA Qw3B, TQA Qw7B and TQA SmolLM), token 0 is the per-layer apex $\arg\min_s \delta_s$ in 0% to 92% of (example, layer) graphs: 92% in the sink-dominated control and 0% in the model without a first-token sink, where deflation had therefore been deleting an ordinary token every time. We re-extract with `--apex-deflation`, which deletes each layer's own apex, and evaluate into a separate results directory so the reported tables are untouched.

The correction matters, and it matters where the theory says. Against SINK, apex deflation adds +0.026 to +0.069 AUC where token-0 deflation adds +0.028 to +0.043. The two agree in the control, where they are usually the same deletion (1 of 3 settings equivalent within ± 0.015, overall range -0.010 to +0.026), and differ most in the model without a first-token sink, where deflating the apex is worth +0.026 AUC over deflating token 0 (TQA SmolLM, 95% interval above zero). A first-token deletion therefore understates non-sink structure in exactly the models whose apex is not token 0, which is where C4 needed testing.

It does not change the conclusion. Stacked onto the non-topological probes, correctly apex-deflated PH changes AUC by -0.002 to +0.001, equivalent within ± 0.015 in 3 of 3 settings. Fixing the deflation makes it a better topological feature and leaves it just as redundant against cheap first-order probes. The apex is chosen per layer from the same graph the features are computed on, so this is a label-free but data-dependent choice of vertex: read it as an intervention that tests C4, not as a detector competing on AUC.

M FULL RESULTS AND COMPLETE TABLES

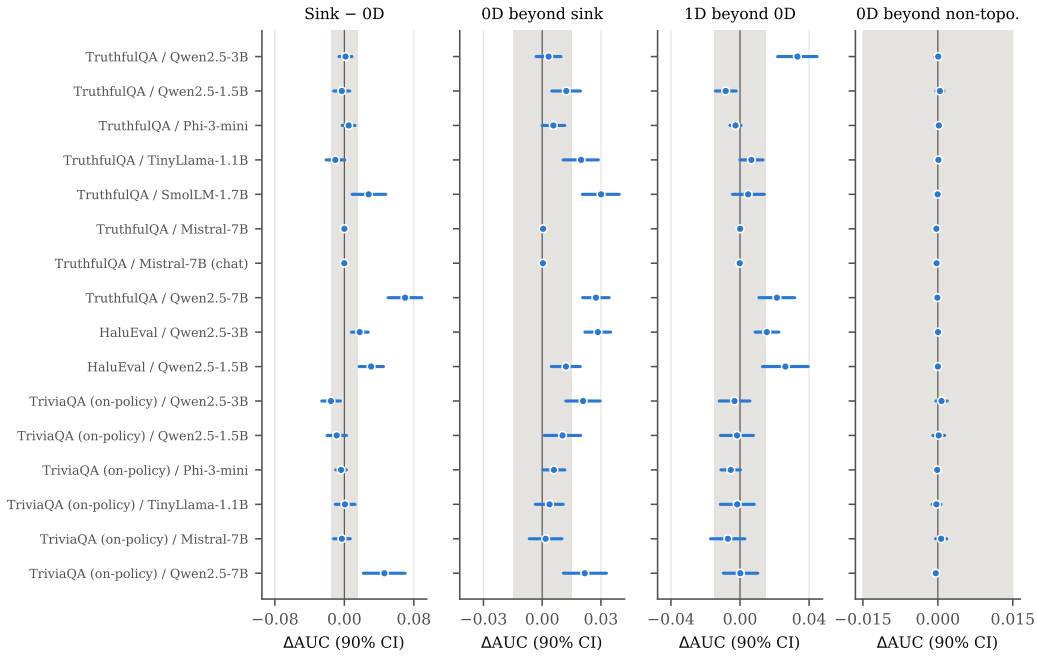


Figure 5: **Four comparisons across all settings** (ΔAUC with 90% cluster-bootstrap intervals; shaded band: ± 0.015). Sink features nearly reproduce 0D; 0D adds little beyond them; 1D adds little beyond 0D except in 4 settings; and 0D adds nothing beyond non-topological probes under early fusion (late fusion in Table 22).

Table 20: **Where 0D adds to the sink** (§5.2). Layers are split without labels into coned ($\geq 90\%$ of graphs with $\delta_0 = 0$) and open ($< 50\%$); Δ is the AUC change from adding that subset's per-layer 0D features to SINK (early fusion), with its 90% interval. Marks as in Table 22. Settings whose per-example layer features we retain.

Setting	coned layers	Δ ($0D_{coned}$ |Sink)	open layers	Δ ($0D_{open}$ |Sink)
TQA Qw3B	21	$-0.001^{\equiv}$ [-0.003, +0.000]	13	$+0.004^{\equiv}$ [-0.002, +0.010]
TQA Qw1.5B	14	$0.000^{\equiv}$ [-0.001, +0.001]	10	$+0.010^{*}$ [+0.003, +0.018]
TQA Phi-3	26	$+0.001^{\equiv}$ [-0.000, +0.002]	3	$+0.006^{\equiv}$ [-0.000, +0.012]
TQA TinyLl.	8	$+0.001^{\equiv}$ [-0.000, +0.002]	14	$+0.019^{*}$ [+0.010, +0.028]
TQA SmolLM	0	–	24	$+0.030^{*}$ [+0.021, +0.039]
TQA Mis7B	31	$+0.001^{\equiv}$ [-0.000, +0.002]	1	$0.000^{\equiv}$ [-0.001, +0.001]
TQA Mis7B (chat)	31	$0.000^{\equiv}$ [-0.001, +0.002]	1	$0.000^{\equiv}$ [-0.001, +0.000]
TQA Qw7B	0	–	28	$+0.027^{*}$ [+0.021, +0.034]
HE Qw3B	14	$+0.001^{\equiv}$ [-0.000, +0.002]	19	$+0.027^{*}$ [+0.020, +0.033]
HE Qw1.5B	12	$-0.001^{\equiv}$ [-0.003, +0.001]	10	$+0.007^{\equiv}$ [+0.001, +0.014]
TrQA Qw3B	21	$+0.004^{\equiv}$ [-0.000, +0.008]	13	$+0.014^{*}$ [+0.006, +0.021]
TrQA Qw1.5B	12	$-0.001^{\equiv}$ [-0.003, +0.001]	13	$+0.012^{*}$ [+0.003, +0.021]
TrQA Phi-3	26	$+0.001^{\equiv}$ [-0.000, +0.003]	3	$+0.004^{\equiv}$ [-0.001, +0.009]
TrQA TinyLl.	8	$+0.001^{*\equiv}$ [+0.000, +0.002]	14	$+0.003^{\equiv}$ [-0.005, +0.010]
TrQA Qw7B	0	–	28	$+0.022^{*}$ [+0.011, +0.033]

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TQA Qw3B	LEN	2	0.529	0.0003	0.567
TQA Qw3B	LEX	–	0.820	0.0014	0.829
TQA Qw3B	LOGPROB	6	0.630	0.0006	0.694
TQA Qw3B	LLMCHECK	36	0.693	0.0011	0.764
TQA Qw3B	ROWSTAT	72	0.750	0.0003	0.823
TQA Qw3B	SINK	72	0.737	0.0022	0.767
TQA Qw3B	0D	72	0.736	0.0013	0.778
TQA Qw3B	LEGACY MSTPROXY	36	0.713	0.0008	0.788
TQA Qw3B	1D	144	0.660	0.0001	0.674
TQA Qw3B	DEFL	216	0.715	0.0028	0.753
TQA Qw3B	ANS	216	0.748	0.0017	0.749
TQA Qw3B	PH SINK	576	0.851	0.0015	0.870
TQA Qw3B	PH 0DTOT	576	0.842	0.0011	0.869
TQA Qw3B	PH 0D	1152	0.853	0.0013	0.862
TQA Qw3B	PH ENT	576	0.857	0.0013	0.901
TQA Qw3B	LOOKBACK	576	0.897	0.0022	0.908
TQA Qw3B	HIDDEN	2048	0.913	0.0008	0.918
TQA Qw3B	NONTOPO	2738	0.926	0.0006	0.936
TQA Qw3B	0D|1D	216	0.769	0.0004	0.791
TQA Qw3B	0D|SINK	144	0.741	0.0015	0.776
TQA Qw3B	HIDDEN|0D	2120	0.913	0.0007	0.920
TQA Qw3B	LEX|0D	–	0.840	0.0015	0.848
TQA Qw3B	LEX|DEFL	–	0.830	0.0004	0.854
TQA Qw3B	LEX|SINK	–	0.843	0.0020	0.851
TQA Qw3B	LOGPROB|0D	78	0.776	0.0022	0.810
TQA Qw3B	NONTOPO|0D	2810	0.926	0.0004	0.936
TQA Qw3B	NONTOPO|DEFL	2954	0.928	0.0006	0.933
TQA Qw3B	PH SINK|PH 0DTOT	1152	0.860	0.0022	0.892
TQA Qw3B	SINK|1D	216	0.766	0.0007	0.787
TQA Qw3B	SINK|ANS	288	0.785	0.0031	0.797
TQA Qw3B	SINK|DEFL	288	0.766	0.0024	0.799
TQA Qw1.5B	LEN	2	0.529	0.0003	0.567
TQA Qw1.5B	LEX	–	0.820	0.0014	0.829
TQA Qw1.5B	LOGPROB	6	0.638	0.0007	0.682
TQA Qw1.5B	LLMCHECK	28	0.666	0.0005	0.743
TQA Qw1.5B	ROWSTAT	56	0.738	0.0014	0.815
TQA Qw1.5B	SINK	56	0.723	0.0019	0.771
TQA Qw1.5B	0D	56	0.726	0.0007	0.767
TQA Qw1.5B	LEGACY MSTPROXY	28	0.698	0.0010	0.786
TQA Qw1.5B	1D	112	0.586	0.0014	0.644
TQA Qw1.5B	DEFL	168	0.703	0.0042	0.755
TQA Qw1.5B	ANS	168	0.697	0.0040	0.700
TQA Qw1.5B	PH SINK	336	0.830	0.0008	0.851
TQA Qw1.5B	PH 0DTOT	336	0.819	0.0011	0.838
TQA Qw1.5B	PH 0D	672	0.838	0.0012	0.847

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TQA Qw1.5B	PH ENT	336	0.833	0.0025	0.852
TQA Qw1.5B	LOOKBACK	336	0.851	0.0020	0.856
TQA Qw1.5B	HIDDEN	1536	0.916	0.0012	0.923
TQA Qw1.5B	NONTOPO	1962	0.917	0.0013	0.928
TQA Qw1.5B	0D|1D	168	0.718	0.0013	0.767
TQA Qw1.5B	0D|SINK	112	0.735	0.0012	0.766
TQA Qw1.5B	HIDDEN|0D	1592	0.916	0.0014	0.924
TQA Qw1.5B	LEX|0D	–	0.835	0.0017	0.841
TQA Qw1.5B	LEX|DEFL	–	0.825	0.0019	0.836
TQA Qw1.5B	LEX|SINK	–	0.833	0.0008	0.846
TQA Qw1.5B	LOGPROB|0D	62	0.740	0.0013	0.797
TQA Qw1.5B	NONTOPO|0D	2018	0.917	0.0014	0.927
TQA Qw1.5B	NONTOPO|DEFL	2130	0.917	0.0014	0.923
TQA Qw1.5B	PH SINK|PH 0DTOT	672	0.841	0.0013	0.867
TQA Qw1.5B	SINK|1D	168	0.715	0.0017	0.769
TQA Qw1.5B	SINK|ANS	224	0.761	0.0027	0.753
TQA Qw1.5B	SINK|DEFL	224	0.757	0.0028	0.802
TQA Phi-3	LEN	2	0.522	0.0004	0.562
TQA Phi-3	LEX	–	0.820	0.0014	0.829
TQA Phi-3	LOGPROB	6	0.578	0.0016	0.597
TQA Phi-3	LLMCHECK	32	0.653	0.0008	0.755
TQA Phi-3	ROWSTAT	64	0.661	0.0011	0.738
TQA Phi-3	SINK	64	0.651	0.0006	0.700
TQA Phi-3	0D	64	0.646	0.0011	0.709
TQA Phi-3	LEGACY MSTPROXY	32	0.613	0.0008	0.694
TQA Phi-3	1D	128	0.526	0.0030	0.551
TQA Phi-3	DEFL	192	0.639	0.0011	0.684
TQA Phi-3	ANS	192	0.629	0.0042	0.651
TQA Phi-3	PH SINK	1024	0.871	0.0008	0.901
TQA Phi-3	PH 0DTOT	1024	0.845	0.0014	0.882
TQA Phi-3	PH 0D	2048	0.868	0.0007	0.881
TQA Phi-3	PH ENT	1024	0.864	0.0010	0.895
TQA Phi-3	LOOKBACK	1024	0.892	0.0004	0.902
TQA Phi-3	HIDDEN	3072	0.933	0.0007	0.945
TQA Phi-3	NONTOPO	4198	0.936	0.0006	0.942
TQA Phi-3	0D|1D	192	0.643	0.0014	0.705
TQA Phi-3	0D|SINK	128	0.657	0.0007	0.704
TQA Phi-3	HIDDEN|0D	3136	0.933	0.0008	0.944
TQA Phi-3	LEX|0D	–	0.821	0.0003	0.842
TQA Phi-3	LEX|DEFL	–	0.802	0.0001	0.827
TQA Phi-3	LEX|SINK	–	0.820	0.0007	0.834
TQA Phi-3	LOGPROB|0D	70	0.655	0.0011	0.694
TQA Phi-3	NONTOPO|0D	4262	0.936	0.0006	0.942
TQA Phi-3	NONTOPO|DEFL	4390	0.937	0.0007	0.939
TQA Phi-3	PH SINK|PH 0DTOT	2048	0.871	0.0007	0.905
TQA Phi-3	SINK|1D	192	0.651	0.0011	0.696
TQA Phi-3	SINK|ANS	256	0.675	0.0015	0.698
TQA Phi-3	SINK|DEFL	256	0.684	0.0023	0.718
TQA TinyLl.	LEN	2	0.523	0.0005	0.556
TQA TinyLl.	LEX	–	0.820	0.0014	0.829
TQA TinyLl.	LOGPROB	6	0.637	0.0010	0.645
TQA TinyLl.	LLMCHECK	22	0.614	0.0005	0.705
TQA TinyLl.	ROWSTAT	44	0.661	0.0015	0.749
TQA TinyLl.	SINK	44	0.643	0.0006	0.718
TQA TinyLl.	0D	44	0.654	0.0011	0.729

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TQA TinyLl.	LEGACY MSTPROXY	22	0.627	0.0009	0.741
TQA TinyLl.	1D	88	0.592	0.0006	0.673
TQA TinyLl.	DEFL	132	0.651	0.0019	0.705
TQA TinyLl.	ANS	132	0.644	0.0049	0.678
TQA TinyLl.	PH SINK	704	0.808	0.0015	0.841
TQA TinyLl.	PH 0DTOT	704	0.804	0.0021	0.830
TQA TinyLl.	PH 0D	1408	0.806	0.0015	0.820
TQA TinyLl.	PH ENT	704	0.817	0.0008	0.838
TQA TinyLl.	LOOKBACK	704	0.850	0.0028	0.846
TQA TinyLl.	HIDDEN	2048	0.883	0.0013	0.879
TQA TinyLl.	NONTOPO	2824	0.885	0.0011	0.887
TQA TinyLl.	0D|1D	132	0.660	0.0010	0.744
TQA TinyLl.	0D|SINK	88	0.663	0.0004	0.728
TQA TinyLl.	HIDDEN|0D	2092	0.883	0.0014	0.880
TQA TinyLl.	LEX|0D	–	0.830	0.0012	0.836
TQA TinyLl.	LEX|DEFL	–	0.826	0.0012	0.846
TQA TinyLl.	LEX|SINK	–	0.831	0.0011	0.832
TQA TinyLl.	LOGPROB|0D	50	0.683	0.0010	0.711
TQA TinyLl.	NONTOPO|0D	2868	0.885	0.0011	0.889
TQA TinyLl.	NONTOPO|DEFL	2956	0.884	0.0010	0.885
TQA TinyLl.	PH SINK|PH 0DTOT	1408	0.810	0.0031	0.841
TQA TinyLl.	SINK|1D	132	0.651	0.0016	0.729
TQA TinyLl.	SINK|ANS	176	0.685	0.0041	0.720
TQA TinyLl.	SINK|DEFL	176	0.688	0.0015	0.749
TQA SmolLM	LEN	2	0.524	0.0003	0.554
TQA SmolLM	LEX	–	0.820	0.0014	0.829
TQA SmolLM	LOGPROB	6	0.588	0.0010	0.592
TQA SmolLM	LLMCHECK	24	0.604	0.0004	0.684
TQA SmolLM	ROWSTAT	48	0.673	0.0014	0.741
TQA SmolLM	SINK	48	0.611	0.0019	0.628
TQA SmolLM	0D	48	0.583	0.0029	0.656
TQA SmolLM	LEGACY MSTPROXY	24	0.585	0.0011	0.674
TQA SmolLM	1D	96	0.543	0.0015	0.581
TQA SmolLM	DEFL	144	0.594	0.0023	0.650
TQA SmolLM	ANS	144	0.628	0.0041	0.630
TQA SmolLM	PH SINK	768	0.720	0.0008	0.775
TQA SmolLM	PH 0DTOT	768	0.819	0.0019	0.845
TQA SmolLM	PH 0D	1536	0.800	0.0012	0.818
TQA SmolLM	PH ENT	768	0.834	0.0007	0.841
TQA SmolLM	LOOKBACK	768	0.863	0.0017	0.864
TQA SmolLM	HIDDEN	2048	0.894	0.0022	0.892
TQA SmolLM	NONTOPO	2894	0.899	0.0015	0.903
TQA SmolLM	0D|1D	144	0.588	0.0019	0.643
TQA SmolLM	0D|SINK	96	0.641	0.0026	0.662
TQA SmolLM	HIDDEN|0D	2096	0.894	0.0020	0.891
TQA SmolLM	LEX|0D	–	0.827	0.0008	0.838
TQA SmolLM	LEX|DEFL	–	0.822	0.0001	0.845
TQA SmolLM	LEX|SINK	–	0.819	0.0015	0.826
TQA SmolLM	LOGPROB|0D	54	0.613	0.0009	0.630
TQA SmolLM	NONTOPO|0D	2942	0.899	0.0013	0.902
TQA SmolLM	NONTOPO|DEFL	3038	0.899	0.0012	0.900
TQA SmolLM	PH SINK|PH 0DTOT	1536	0.826	0.0023	0.854
TQA SmolLM	SINK|1D	144	0.629	0.0013	0.650
TQA SmolLM	SINK|ANS	192	0.646	0.0021	0.651
TQA SmolLM	SINK|DEFL	192	0.652	0.0022	0.681

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TQA Mis7B	LEN	2	0.525	0.0002	0.571
TQA Mis7B	LEX	–	0.820	0.0008	0.828
TQA Mis7B	LOGPROB	6	0.641	0.0004	0.652
TQA Mis7B	LLMCHECK	32	0.652	0.0002	0.770
TQA Mis7B	ROWSTAT	64	0.694	0.0009	0.821
TQA Mis7B	SINK	64	0.685	0.0008	0.755
TQA Mis7B	0D	64	0.684	0.0008	0.755
TQA Mis7B	LEGACY MSTPROXY	32	0.646	0.0009	0.769
TQA Mis7B	1D	128	0.501	0.0000	0.499
TQA Mis7B	DEFL	192	0.710	0.0015	0.756
TQA Mis7B	ANS	192	0.677	0.0043	0.704
TQA Mis7B	PH SINK	1024	0.903	0.0009	0.938
TQA Mis7B	PH 0DTOT	1024	0.887	0.0020	0.928
TQA Mis7B	PH 0D	2048	0.893	0.0013	0.908
TQA Mis7B	PH ENT	1024	0.897	0.0010	0.929
TQA Mis7B	LOOKBACK	1024	0.933	0.0008	0.945
TQA Mis7B	HIDDEN	4096	0.956	0.0005	0.957
TQA Mis7B	NONTOPO	5222	0.959	0.0006	0.967
TQA Mis7B	0D|1D	192	0.685	0.0007	0.760
TQA Mis7B	0D|SINK	128	0.685	0.0007	0.754
TQA Mis7B	HIDDEN|0D	4160	0.956	0.0005	0.956
TQA Mis7B	LEX|0D	–	0.834	0.0013	0.859
TQA Mis7B	LEX|DEFL	–	0.823	0.0013	0.856
TQA Mis7B	LEX|SINK	–	0.833	0.0015	0.859
TQA Mis7B	LOGPROB|0D	70	0.722	0.0012	0.775
TQA Mis7B	NONTOPO|0D	5286	0.958	0.0006	0.967
TQA Mis7B	NONTOPO|DEFL	5414	0.959	0.0006	0.966
TQA Mis7B	PH SINK|PH 0DTOT	2048	0.903	0.0013	0.936
TQA Mis7B	SINK|1D	192	0.685	0.0008	0.758
TQA Mis7B	SINK|ANS	256	0.727	0.0023	0.763
TQA Mis7B	SINK|DEFL	256	0.740	0.0004	0.800
TQA Mis7B (chat)	LEN	2	0.525	0.0002	0.571
TQA Mis7B (chat)	LEX	–	0.820	0.0008	0.828
TQA Mis7B (chat)	LOGPROB	6	0.651	0.0014	0.668
TQA Mis7B (chat)	LLMCHECK	32	0.653	0.0001	0.769
TQA Mis7B (chat)	ROWSTAT	64	0.700	0.0008	0.823
TQA Mis7B (chat)	SINK	64	0.688	0.0010	0.754
TQA Mis7B (chat)	0D	64	0.688	0.0010	0.755
TQA Mis7B (chat)	LEGACY MSTPROXY	32	0.652	0.0009	0.783
TQA Mis7B (chat)	1D	128	0.501	0.0000	0.499
TQA Mis7B (chat)	DEFL	192	0.733	0.0013	0.781
TQA Mis7B (chat)	ANS	192	0.748	0.0012	0.767
TQA Mis7B (chat)	PH SINK	1024	0.907	0.0011	0.934
TQA Mis7B (chat)	PH 0DTOT	1024	0.890	0.0021	0.931
TQA Mis7B (chat)	PH 0D	2048	0.900	0.0017	0.919
TQA Mis7B (chat)	PH ENT	1024	0.901	0.0010	0.928
TQA Mis7B (chat)	LOOKBACK	1024	0.934	0.0008	0.942
TQA Mis7B (chat)	HIDDEN	4096	0.958	0.0004	0.957
TQA Mis7B (chat)	NONTOPO	5222	0.960	0.0006	0.965
TQA Mis7B (chat)	0D|1D	192	0.688	0.0009	0.754
TQA Mis7B (chat)	0D|SINK	128	0.689	0.0008	0.755
TQA Mis7B (chat)	HIDDEN|0D	4160	0.958	0.0004	0.957
TQA Mis7B (chat)	LEX|0D	–	0.834	0.0013	0.860
TQA Mis7B (chat)	LEX|DEFL	–	0.835	0.0020	0.864
TQA Mis7B (chat)	LEX|SINK	–	0.834	0.0015	0.859

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TQA Mis7B (chat)	LOGPROB|0D	70	0.733	0.0021	0.798
TQA Mis7B (chat)	NONTOPO|0D	5286	0.960	0.0005	0.965
TQA Mis7B (chat)	NONTOPO|DEFL	5414	0.961	0.0006	0.966
TQA Mis7B (chat)	PH SINK|PH 0DTOT	2048	0.907	0.0012	0.938
TQA Mis7B (chat)	SINK|1D	192	0.688	0.0009	0.756
TQA Mis7B (chat)	SINK|ANS	256	0.767	0.0028	0.792
TQA Mis7B (chat)	SINK|DEFL	256	0.756	0.0018	0.803
TQA Qw7B	LEN	2	0.529	0.0003	0.567
TQA Qw7B	LEX	–	0.820	0.0008	0.828
TQA Qw7B	LOGPROB	6	0.674	0.0007	0.717
TQA Qw7B	LLMCHECK	28	0.667	0.0011	0.729
TQA Qw7B	ROWSTAT	56	0.736	0.0008	0.807
TQA Qw7B	SINK	56	0.749	0.0001	0.761
TQA Qw7B	0D	56	0.679	0.0010	0.774
TQA Qw7B	LEGACY MSTPROXY	28	0.678	0.0011	0.785
TQA Qw7B	1D	112	0.610	0.0005	0.634
TQA Qw7B	DEFL	168	0.697	0.0009	0.748
TQA Qw7B	ANS	168	0.741	0.0019	0.749
TQA Qw7B	PH SINK	784	0.821	0.0025	0.843
TQA Qw7B	PH 0DTOT	784	0.872	0.0016	0.909
TQA Qw7B	PH 0D	1568	0.850	0.0022	0.867
TQA Qw7B	PH ENT	784	0.887	0.0003	0.901
TQA Qw7B	LOOKBACK	784	0.922	0.0004	0.923
TQA Qw7B	HIDDEN	3584	0.933	0.0008	0.934
TQA Qw7B	NONTOPO	4458	0.950	0.0013	0.950
TQA Qw7B	0D|1D	168	0.701	0.0002	0.756
TQA Qw7B	0D|SINK	112	0.777	0.0004	0.794
TQA Qw7B	HIDDEN|0D	3640	0.933	0.0008	0.933
TQA Qw7B	LEX|0D	–	0.834	0.0008	0.858
TQA Qw7B	LEX|DEFL	–	0.828	0.0014	0.844
TQA Qw7B	LEX|SINK	–	0.848	0.0001	0.849
TQA Qw7B	LOGPROB|0D	62	0.727	0.0002	0.792
TQA Qw7B	NONTOPO|0D	4514	0.949	0.0013	0.949
TQA Qw7B	NONTOPO|DEFL	4626	0.949	0.0012	0.952
TQA Qw7B	PH SINK|PH 0DTOT	1568	0.881	0.0005	0.920
TQA Qw7B	SINK|1D	168	0.767	0.0023	0.780
TQA Qw7B	SINK|ANS	224	0.794	0.0010	0.802
TQA Qw7B	SINK|DEFL	224	0.788	0.0006	0.797
HE Qw3B	LEN	2	0.962	0.0001	0.970
HE Qw3B	LEX	–	0.981	0.0004	0.977
HE Qw3B	LOGPROB	6	0.993	0.0001	0.991
HE Qw3B	LLMCHECK	36	0.745	0.0004	0.956
HE Qw3B	ROWSTAT	72	0.829	0.0002	0.976
HE Qw3B	SINK	72	0.829	0.0008	0.930
HE Qw3B	0D	72	0.811	0.0006	0.927
HE Qw3B	LEGACY MSTPROXY	36	0.731	0.0002	0.960
HE Qw3B	1D	144	0.710	0.0013	0.891
HE Qw3B	DEFL	216	0.834	0.0009	0.959
HE Qw3B	ANS	216	0.978	0.0009	0.980
HE Qw3B	PH SINK	576	0.958	0.0007	0.999
HE Qw3B	PH 0DTOT	576	0.958	0.0005	0.995
HE Qw3B	PH 0D	1152	0.971	0.0006	0.988
HE Qw3B	PH ENT	576	0.964	0.0005	0.993
HE Qw3B	LOOKBACK	576	0.998	0.0001	0.998
HE Qw3B	HIDDEN	2048	1.000	0.0000	1.000

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
HE Qw3B	NONTOPO	2738	1.000	0.0002	1.000
HE Qw3B	0D|1D	216	0.826	0.0013	0.945
HE Qw3B	0D|SINK	144	0.857	0.0010	0.948
HE Qw3B	HIDDEN|0D	2120	1.000	0.0000	1.000
HE Qw3B	LEX|0D	–	0.979	0.0001	0.983
HE Qw3B	LEX|DEFL	–	0.971	0.0003	0.975
HE Qw3B	LEX|SINK	–	0.980	0.0003	0.981
HE Qw3B	LOGPROB|0D	78	0.993	0.0003	0.992
HE Qw3B	NONTOPO|0D	2810	1.000	0.0002	1.000
HE Qw3B	NONTOPO|DEFL	2954	1.000	0.0002	1.000
HE Qw3B	PH SINK|PH 0DTOT	1152	0.962	0.0006	0.997
HE Qw3B	SINK|1D	216	0.843	0.0011	0.938
HE Qw3B	SINK|ANS	288	0.978	0.0003	0.982
HE Qw3B	SINK|DEFL	288	0.887	0.0020	0.957
HE Qw1.5B	LEN	2	0.967	0.0002	0.976
HE Qw1.5B	LEX	–	0.979	0.0004	0.980
HE Qw1.5B	LOGPROB	6	0.995	0.0002	0.988
HE Qw1.5B	LLMCHECK	28	0.791	0.0005	0.978
HE Qw1.5B	ROWSTAT	56	0.788	0.0017	0.984
HE Qw1.5B	SINK	56	0.820	0.0015	0.934
HE Qw1.5B	0D	56	0.789	0.0014	0.918
HE Qw1.5B	LEGACY MSTPROXY	28	0.719	0.0003	0.976
HE Qw1.5B	1D	112	0.747	0.0006	0.878
HE Qw1.5B	DEFL	168	0.825	0.0021	0.952
HE Qw1.5B	ANS	168	0.972	0.0006	0.978
HE Qw1.5B	PH SINK	336	0.922	0.0007	0.998
HE Qw1.5B	PH 0DTOT	336	0.921	0.0007	0.996
HE Qw1.5B	PH 0D	672	0.949	0.0027	0.974
HE Qw1.5B	PH ENT	336	0.933	0.0011	0.996
HE Qw1.5B	LOOKBACK	336	0.999	0.0002	0.998
HE Qw1.5B	HIDDEN	1536	1.000	0.0001	1.000
HE Qw1.5B	NONTOPO	1962	1.000	0.0002	1.000
HE Qw1.5B	0D|1D	168	0.815	0.0032	0.922
HE Qw1.5B	0D|SINK	112	0.832	0.0015	0.924
HE Qw1.5B	HIDDEN|0D	1592	1.000	0.0001	1.000
HE Qw1.5B	LEX|0D	–	0.971	0.0004	0.986
HE Qw1.5B	LEX|DEFL	–	0.956	0.0002	0.986
HE Qw1.5B	LEX|SINK	–	0.968	0.0005	0.984
HE Qw1.5B	LOGPROB|0D	62	0.994	0.0002	0.990
HE Qw1.5B	NONTOPO|0D	2018	1.000	0.0000	1.000
HE Qw1.5B	NONTOPO|DEFL	2130	1.000	0.0001	1.000
HE Qw1.5B	PH SINK|PH 0DTOT	672	0.931	0.0024	0.998
HE Qw1.5B	SINK|1D	168	0.826	0.0035	0.940
HE Qw1.5B	SINK|ANS	224	0.975	0.0015	0.990
HE Qw1.5B	SINK|DEFL	224	0.872	0.0028	0.962
TrQA Qw3B	LEN	2	0.569	0.0014	–
TrQA Qw3B	LEX	–	0.596	0.0039	–
TrQA Qw3B	LOGPROB	6	0.839	0.0002	–
TrQA Qw3B	LLMCHECK	36	0.688	0.0012	–
TrQA Qw3B	ROWSTAT	72	0.724	0.0046	–
TrQA Qw3B	SINK	72	0.717	0.0038	–
TrQA Qw3B	0D	72	0.732	0.0038	–
TrQA Qw3B	LEGACY MSTPROXY	36	0.691	0.0033	–
TrQA Qw3B	1D	144	0.601	0.0031	–
TrQA Qw3B	DEFL	216	0.684	0.0043	–

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TrQA Qw3B	ANS	216	0.719	0.0020	–
TrQA Qw3B	PH SINK	576	0.820	0.0021	–
TrQA Qw3B	PH 0DTOT	576	0.807	0.0012	–
TrQA Qw3B	PH 0D	1152	0.807	0.0017	–
TrQA Qw3B	PH ENT	576	0.810	0.0007	–
TrQA Qw3B	LOOKBACK	576	0.861	0.0013	–
TrQA Qw3B	HIDDEN	2048	0.841	0.0042	–
TrQA Qw3B	NONTOPO	2738	0.870	0.0028	–
TrQA Qw3B	0D|1D	216	0.729	0.0022	–
TrQA Qw3B	0D|SINK	144	0.738	0.0050	–
TrQA Qw3B	HIDDEN|0D	2120	0.843	0.0040	–
TrQA Qw3B	LEX|0D	–	0.733	0.0011	–
TrQA Qw3B	LEX|DEFL	–	0.690	0.0031	–
TrQA Qw3B	LEX|SINK	–	0.724	0.0005	–
TrQA Qw3B	LOGPROB|0D	78	0.843	0.0022	–
TrQA Qw3B	NONTOPO|0D	2810	0.870	0.0027	–
TrQA Qw3B	NONTOPO|DEFL	2954	0.870	0.0029	–
TrQA Qw3B	PH SINK|PH 0DTOT	1152	0.824	0.0017	–
TrQA Qw3B	SINK|1D	216	0.718	0.0035	–
TrQA Qw3B	SINK|ANS	288	0.752	0.0010	–
TrQA Qw3B	SINK|DEFL	288	0.719	0.0038	–
TrQA Qw1.5B	LEN	2	0.588	0.0012	–
TrQA Qw1.5B	LEX	–	0.609	0.0033	–
TrQA Qw1.5B	LOGPROB	6	0.823	0.0006	–
TrQA Qw1.5B	LLMCHECK	28	0.676	0.0029	–
TrQA Qw1.5B	ROWSTAT	56	0.720	0.0008	–
TrQA Qw1.5B	SINK	56	0.707	0.0000	–
TrQA Qw1.5B	0D	56	0.716	0.0004	–
TrQA Qw1.5B	LEGACY MSTPROXY	28	0.685	0.0014	–
TrQA Qw1.5B	1D	112	0.574	0.0024	–
TrQA Qw1.5B	DEFL	168	0.655	0.0016	–
TrQA Qw1.5B	ANS	168	0.675	0.0035	–
TrQA Qw1.5B	PH SINK	336	0.767	0.0034	–
TrQA Qw1.5B	PH 0DTOT	336	0.759	0.0051	–
TrQA Qw1.5B	PH 0D	672	0.768	0.0013	–
TrQA Qw1.5B	PH ENT	336	0.768	0.0025	–
TrQA Qw1.5B	LOOKBACK	336	0.824	0.0017	–
TrQA Qw1.5B	HIDDEN	1536	0.855	0.0031	–
TrQA Qw1.5B	NONTOPO	1962	0.873	0.0023	–
TrQA Qw1.5B	0D|1D	168	0.714	0.0015	–
TrQA Qw1.5B	0D|SINK	112	0.717	0.0006	–
TrQA Qw1.5B	HIDDEN|0D	1592	0.855	0.0032	–
TrQA Qw1.5B	LEX|0D	–	0.721	0.0031	–
TrQA Qw1.5B	LEX|DEFL	–	0.672	0.0047	–
TrQA Qw1.5B	LEX|SINK	–	0.715	0.0036	–
TrQA Qw1.5B	LOGPROB|0D	62	0.828	0.0007	–
TrQA Qw1.5B	NONTOPO|0D	2018	0.873	0.0025	–
TrQA Qw1.5B	NONTOPO|DEFL	2130	0.872	0.0024	–
TrQA Qw1.5B	PH SINK|PH 0DTOT	672	0.772	0.0022	–
TrQA Qw1.5B	SINK|1D	168	0.709	0.0028	–
TrQA Qw1.5B	SINK|ANS	224	0.726	0.0059	–
TrQA Qw1.5B	SINK|DEFL	224	0.719	0.0023	–
TrQA Phi-3	LEN	2	0.581	0.0022	–
TrQA Phi-3	LEX	–	0.635	0.0083	–
TrQA Phi-3	LOGPROB	6	0.877	0.0002	–

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TrQA Phi-3	LLMCHECK	32	0.676	0.0018	–
TrQA Phi-3	ROWSTAT	64	0.728	0.0006	–
TrQA Phi-3	SINK	64	0.707	0.0018	–
TrQA Phi-3	0D	64	0.710	0.0022	–
TrQA Phi-3	LEGACY MSTPROXY	32	0.669	0.0002	–
TrQA Phi-3	1D	128	0.546	0.0031	–
TrQA Phi-3	DEFL	192	0.698	0.0021	–
TrQA Phi-3	ANS	192	0.710	0.0037	–
TrQA Phi-3	PH SINK	1024	0.826	0.0010	–
TrQA Phi-3	PH 0DTOT	1024	0.810	0.0016	–
TrQA Phi-3	PH 0D	2048	0.807	0.0024	–
TrQA Phi-3	PH ENT	1024	0.817	0.0021	–
TrQA Phi-3	LOOKBACK	1024	0.840	0.0016	–
TrQA Phi-3	HIDDEN	3072	0.867	0.0035	–
TrQA Phi-3	NONTOPO	4198	0.883	0.0030	–
TrQA Phi-3	0D|1D	192	0.705	0.0011	–
TrQA Phi-3	0D|SINK	128	0.713	0.0024	–
TrQA Phi-3	HIDDEN|0D	3136	0.867	0.0037	–
TrQA Phi-3	LEX|0D	–	0.721	0.0031	–
TrQA Phi-3	LEX|DEFL	–	0.712	0.0022	–
TrQA Phi-3	LEX|SINK	–	0.718	0.0044	–
TrQA Phi-3	LOGPROB|0D	70	0.878	0.0013	–
TrQA Phi-3	NONTOPO|0D	4262	0.883	0.0030	–
TrQA Phi-3	NONTOPO|DEFL	4390	0.885	0.0030	–
TrQA Phi-3	PH SINK|PH 0DTOT	2048	0.828	0.0009	–
TrQA Phi-3	SINK|1D	192	0.699	0.0012	–
TrQA Phi-3	SINK|ANS	256	0.736	0.0046	–
TrQA Phi-3	SINK|DEFL	256	0.736	0.0016	–
TrQA TinyLl.	LEN	2	0.631	0.0005	–
TrQA TinyLl.	LEX	–	0.740	0.0003	–
TrQA TinyLl.	LOGPROB	6	0.724	0.0006	–
TrQA TinyLl.	LLMCHECK	22	0.688	0.0010	–
TrQA TinyLl.	ROWSTAT	44	0.752	0.0001	–
TrQA TinyLl.	SINK	44	0.720	0.0020	–
TrQA TinyLl.	0D	44	0.720	0.0020	–
TrQA TinyLl.	LEGACY MSTPROXY	22	0.729	0.0006	–
TrQA TinyLl.	1D	88	0.659	0.0007	–
TrQA TinyLl.	DEFL	132	0.726	0.0024	–
TrQA TinyLl.	ANS	132	0.701	0.0021	–
TrQA TinyLl.	PH SINK	704	0.814	0.0016	–
TrQA TinyLl.	PH 0DTOT	704	0.792	0.0018	–
TrQA TinyLl.	PH 0D	1408	0.775	0.0027	–
TrQA TinyLl.	PH ENT	704	0.807	0.0003	–
TrQA TinyLl.	LOOKBACK	704	0.823	0.0024	–
TrQA TinyLl.	HIDDEN	2048	0.827	0.0028	–
TrQA TinyLl.	NONTOPO	2824	0.842	0.0031	–
TrQA TinyLl.	0D|1D	132	0.718	0.0002	–
TrQA TinyLl.	0D|SINK	88	0.724	0.0015	–
TrQA TinyLl.	HIDDEN|0D	2092	0.828	0.0027	–
TrQA TinyLl.	LEX|0D	–	0.764	0.0010	–
TrQA TinyLl.	LEX|DEFL	–	0.757	0.0032	–
TrQA TinyLl.	LEX|SINK	–	0.762	0.0021	–
TrQA TinyLl.	LOGPROB|0D	50	0.762	0.0021	–
TrQA TinyLl.	NONTOPO|0D	2868	0.842	0.0029	–
TrQA TinyLl.	NONTOPO|DEFL	2956	0.842	0.0032	–

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TrQA TinyLl.	PH SINK|PH 0DTOT	1408	0.813	0.0016	–
TrQA TinyLl.	SINK|1D	132	0.719	0.0013	–
TrQA TinyLl.	SINK|ANS	176	0.725	0.0030	–
TrQA TinyLl.	SINK|DEFL	176	0.734	0.0032	–
TrQA Mis7B	LEN	2	0.635	0.0017	–
TrQA Mis7B	LEX	–	0.677	0.0050	–
TrQA Mis7B	LOGPROB	6	0.728	0.0007	–
TrQA Mis7B	LLMCHECK	32	0.686	0.0019	–
TrQA Mis7B	ROWSTAT	64	0.731	0.0016	–
TrQA Mis7B	SINK	64	0.696	0.0017	–
TrQA Mis7B	0D	64	0.699	0.0003	–
TrQA Mis7B	LEGACY MSTPROXY	32	0.693	0.0017	–
TrQA Mis7B	1D	128	0.611	0.0028	–
TrQA Mis7B	DEFL	192	0.721	0.0024	–
TrQA Mis7B	ANS	192	0.662	0.0027	–
TrQA Mis7B	PH SINK	1024	0.812	0.0024	–
TrQA Mis7B	PH 0DTOT	1024	0.808	0.0037	–
TrQA Mis7B	PH 0D	2048	0.775	0.0013	–
TrQA Mis7B	PH ENT	1024	0.812	0.0027	–
TrQA Mis7B	LOOKBACK	1024	0.829	0.0041	–
TrQA Mis7B	HIDDEN	4096	0.833	0.0008	–
TrQA Mis7B	NONTOPO	5222	0.841	0.0011	–
TrQA Mis7B	0D|1D	192	0.692	0.0035	–
TrQA Mis7B	0D|SINK	128	0.698	0.0018	–
TrQA Mis7B	HIDDEN|0D	4160	0.834	0.0011	–
TrQA Mis7B	LEX|0D	–	0.723	0.0030	–
TrQA Mis7B	LEX|DEFL	–	0.744	0.0057	–
TrQA Mis7B	LEX|SINK	–	0.721	0.0027	–
TrQA Mis7B	LOGPROB|0D	70	0.755	0.0016	–
TrQA Mis7B	NONTOPO|0D	5286	0.842	0.0015	–
TrQA Mis7B	NONTOPO|DEFL	5414	0.841	0.0010	–
TrQA Mis7B	PH SINK|PH 0DTOT	2048	0.814	0.0022	–
TrQA Mis7B	SINK|1D	192	0.686	0.0042	–
TrQA Mis7B	SINK|ANS	256	0.693	0.0048	–
TrQA Mis7B	SINK|DEFL	256	0.739	0.0019	–
TrQA Qw7B	LEN	2	0.595	0.0004	–
TrQA Qw7B	LEX	–	0.594	0.0029	–
TrQA Qw7B	LOGPROB	6	0.847	0.0003	–
TrQA Qw7B	LLMCHECK	28	0.725	0.0015	–
TrQA Qw7B	ROWSTAT	56	0.734	0.0016	–
TrQA Qw7B	SINK	56	0.714	0.0023	–
TrQA Qw7B	0D	56	0.668	0.0006	–
TrQA Qw7B	LEGACY MSTPROXY	28	0.672	0.0005	–
TrQA Qw7B	1D	112	0.582	0.0022	–
TrQA Qw7B	DEFL	168	0.672	0.0016	–
TrQA Qw7B	ANS	168	0.755	0.0026	–
TrQA Qw7B	PH SINK	784	0.801	0.0009	–
TrQA Qw7B	PH 0DTOT	784	0.822	0.0011	–
TrQA Qw7B	PH 0D	1568	0.798	0.0029	–
TrQA Qw7B	PH ENT	784	0.841	0.0021	–
TrQA Qw7B	LOOKBACK	784	0.879	0.0010	–
TrQA Qw7B	HIDDEN	3584	0.854	0.0007	–
TrQA Qw7B	NONTOPO	4458	0.881	0.0008	–
TrQA Qw7B	0D|1D	168	0.668	0.0021	–
TrQA Qw7B	0D|SINK	112	0.735	0.0008	–

Table 23: **Complete out-of-fold ROC-AUC results across all settings and feature banks.** LEGACY MSTPROXY is per-layer 0D total persistence alone, the MST-weight summary used by earlier attention-graph work (Appendix B); X|Y is early fusion, as in the main text. A dash in *dim* means the bank has no fixed width: LEX refits a TF-IDF vocabulary inside every training fold. A dash in *pair acc.* means the benchmark is not paired: the on-policy TriviaQA settings have one row per question, so there is no grounded/hallucinated pair to score.

Setting	Bank	dim	AUC	seed SD	pair acc.
TrQA Qw7B	HIDDEN|0D	3640	0.853	0.0008	–
TrQA Qw7B	LEX|0D	–	0.678	0.0043	–
TrQA Qw7B	LEX|DEFL	–	0.678	0.0022	–
TrQA Qw7B	LEX|SINK	–	0.728	0.0007	–
TrQA Qw7B	LOGPROB|0D	62	0.842	0.0012	–
TrQA Qw7B	NONTOPO|0D	4514	0.881	0.0007	–
TrQA Qw7B	NONTOPO|DEFL	4626	0.881	0.0007	–
TrQA Qw7B	PH SINK|PH 0DTOT	1568	0.832	0.0010	–
TrQA Qw7B	SINK|1D	168	0.718	0.0023	–
TrQA Qw7B	SINK|ANS	224	0.781	0.0035	–
TrQA Qw7B	SINK|DEFL	224	0.734	0.0021	–

Table 24: **Complete paired difference tests and equivalence tests across all settings and comparisons.** Bank names follow the main text: X|Y is early fusion (the concatenated banks) and X⊕Y is late fusion (the two banks' out-of-fold logits stacked in a second-level model), so X|Y and X⊕Y are two different models over the same features and need not agree. LEGACY MSTPROXY is per-layer 0D total persistence alone (Appendix B).

Setting	Test (A → B)	AUC_A	AUC_B	Δ	90% CI	$\equiv_{.015}$	power
TQA Qw3B	0D→SINK	0.736	0.737	+0.001	[-0.006, +0.009]	yes	0.90
TQA Qw3B	0D→LEGACY MSTPROXY	0.736	0.713	-0.023	[-0.034, -0.012]	no	0.42
TQA Qw3B	SINK→0D|SINK	0.737	0.741	+0.003	[-0.003, +0.010]	yes	0.97
TQA Qw3B	SINK→SINK|DEFL	0.737	0.766	+0.029	[+0.017, +0.041]	no	0.30
TQA Qw3B	0D→0D|1D	0.736	0.769	+0.033	[+0.022, +0.045]	no	0.38
TQA Qw3B	SINK→SINK|1D	0.737	0.766	+0.028	[+0.017, +0.039]	no	0.46
TQA Qw3B	SINK→SINK|ANS	0.737	0.785	+0.047	[+0.033, +0.062]	no	0.03
TQA Qw3B	PH 0DTOT→PH SINK	0.842	0.851	+0.009	[+0.002, +0.016]	no	0.93
TQA Qw3B	PH SINK→PH SINK|PH 0DTOT	0.851	0.860	+0.009	[+0.005, +0.012]	yes	1.00
TQA Qw3B	PH 0D→PH ENT	0.853	0.857	+0.004	[-0.006, +0.014]	yes	0.57
TQA Qw3B	0D→LOGPROB	0.736	0.630	-0.106	[-0.129, -0.083]	no	0.00
TQA Qw3B	0D→HIDDEN	0.736	0.913	+0.177	[+0.163, +0.192]	no	0.03
TQA Qw3B	0D→LOOKBACK	0.736	0.897	+0.161	[+0.146, +0.175]	no	0.02
TQA Qw3B	0D→LLMCHECK	0.736	0.693	-0.043	[-0.059, -0.026]	no	0.00
TQA Qw3B	0D→LEX	0.736	0.820	+0.084	[+0.066, +0.102]	no	0.00
TQA Qw3B	0D→ROWSTAT	0.736	0.750	+0.014	[+0.002, +0.026]	no	0.30
TQA Qw3B	0D→LEN	0.736	0.529	-0.207	[-0.229, -0.185]	no	0.00
TQA Qw3B	NONTOPO→NONTOPO|0D	0.926	0.926	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw3B	NONTOPO→NONTOPO|DEFL	0.926	0.928	+0.001	[-0.000, +0.003]	yes	1.00
TQA Qw3B	LEX→LEX|0D	0.820	0.840	+0.020	[+0.011, +0.030]	no	0.63
TQA Qw3B	LEX→LEX|SINK	0.820	0.843	+0.022	[+0.013, +0.033]	no	0.59
TQA Qw3B	LEX→LEX|DEFL	0.820	0.830	+0.010	[-0.003, +0.023]	no	0.19
TQA Qw3B	LOGPROB→LOGPROB|0D	0.630	0.776	+0.146	[+0.127, +0.165]	no	0.00
TQA Qw3B	HIDDEN→HIDDEN|0D	0.913	0.913	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw3B	NONTOPO→NONTOPO⊕0D	0.926	0.926	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw3B	NONTOPO→NONTOPO⊕DEFL	0.926	0.926	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw3B	NONTOPO→NONTOPO⊕SINK	0.926	0.926	+0.000	[-0.000, +0.001]	yes	1.00
TQA Qw3B	HIDDEN→HIDDEN⊕0D	0.913	0.914	+0.001	[-0.000, +0.002]	yes	1.00
TQA Qw3B	LOGPROB→LOGPROB⊕0D	0.630	0.757	+0.128	[+0.109, +0.147]	no	0.00
TQA Qw3B	SINK→SINK⊕0D	0.737	0.739	+0.002	[-0.001, +0.006]	yes	1.00
TQA Qw1.5B	0D→SINK	0.726	0.723	-0.003	[-0.012, +0.006]	yes	0.68
TQA Qw1.5B	0D→LEGACY MSTPROXY	0.726	0.698	-0.028	[-0.040, -0.016]	no	0.36
TQA Qw1.5B	SINK→0D|SINK	0.723	0.735	+0.012	[+0.005, +0.020]	no	0.91
TQA Qw1.5B	SINK→SINK|DEFL	0.723	0.757	+0.034	[+0.022, +0.045]	no	0.35
TQA Qw1.5B	0D→0D|1D	0.726	0.718	-0.008	[-0.014, -0.002]	yes	0.99
TQA Qw1.5B	SINK→SINK|1D	0.723	0.715	-0.008	[-0.014, -0.002]	yes	0.99
TQA Qw1.5B	SINK→SINK|ANS	0.723	0.761	+0.038	[+0.024, +0.051]	no	0.11
TQA Qw1.5B	PH 0DTOT→PH SINK	0.819	0.830	+0.011	[+0.003, +0.018]	no	0.90
TQA Qw1.5B	PH SINK→PH SINK|PH 0DTOT	0.830	0.841	+0.011	[+0.007, +0.014]	yes	1.00
TQA Qw1.5B	PH 0D→PH ENT	0.838	0.833	-0.004	[-0.016, +0.007]	no	0.44
TQA Qw1.5B	0D→LOGPROB	0.726	0.638	-0.088	[-0.108, -0.067]	no	0.00
TQA Qw1.5B	0D→HIDDEN	0.726	0.916	+0.190	[+0.175, +0.204]	no	0.05
TQA Qw1.5B	0D→LOOKBACK	0.726	0.851	+0.125	[+0.109, +0.140]	no	0.00
TQA Qw1.5B	0D→LLMCHECK	0.726	0.666	-0.060	[-0.076, -0.044]	no	0.00
TQA Qw1.5B	0D→LEX	0.726	0.820	+0.094	[+0.076, +0.112]	no	0.00
TQA Qw1.5B	0D→ROWSTAT	0.726	0.738	+0.012	[-0.000, +0.025]	no	0.27
TQA Qw1.5B	0D→LEN	0.726	0.529	-0.197	[-0.218, -0.176]	no	0.00
TQA Qw1.5B	NONTOPO→NONTOPO|0D	0.917	0.917	+0.000	[-0.000, +0.001]	yes	1.00
TQA Qw1.5B	NONTOPO→NONTOPO|DEFL	0.917	0.917	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw1.5B	LEX→LEX|0D	0.820	0.835	+0.015	[+0.006, +0.024]	no	0.71
TQA Qw1.5B	LEX→LEX|SINK	0.820	0.833	+0.013	[+0.004, +0.022]	no	0.73
TQA Qw1.5B	LEX→LEX|DEFL	0.820	0.825	+0.005	[-0.006, +0.016]	no	0.45

Table 24: **Complete paired difference tests and equivalence tests across all settings and comparisons.** Bank names follow the main text: X|Y is early fusion (the concatenated banks) and X⊕Y is late fusion (the two banks' out-of-fold logits stacked in a second-level model), so X|Y and X⊕Y are two different models over the same features and need not agree. LEGACY MSTPROXY is per-layer 0D total persistence alone (Appendix B).

Setting	Test (A → B)	AUC$_A$	AUC$_B$	Δ	90% CI	$\equiv_{.015}$	power
TQA Qw1.5B	LOGPROB→LOGPROB|0D	0.638	0.740	+0.101	[+0.086, +0.116]	no	0.00
TQA Qw1.5B	HIDDEN→HIDDEN|0D	0.916	0.916	+0.001	[-0.000, +0.002]	yes	1.00
TQA Qw1.5B	NONTOPO→NONTOPO⊕0D	0.916	0.916	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw1.5B	NONTOPO→NONTOPO⊕DEFL	0.916	0.916	-0.000	[-0.001, +0.001]	yes	1.00
TQA Qw1.5B	NONTOPO→NONTOPO⊕SINK	0.916	0.916	-0.000	[-0.001, +0.000]	yes	1.00
TQA Qw1.5B	HIDDEN→HIDDEN⊕0D	0.915	0.915	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw1.5B	LOGPROB→LOGPROB⊕0D	0.638	0.737	+0.100	[+0.084, +0.115]	no	0.00
TQA Qw1.5B	SINK→SINK⊕0D	0.722	0.730	+0.007	[+0.002, +0.012]	yes	1.00
TQA Phi-3	0D→SINK	0.646	0.651	+0.005	[-0.003, +0.012]	yes	0.90
TQA Phi-3	0D→LEGACY MSTPROXY	0.646	0.613	-0.033	[-0.046, -0.020]	no	0.20
TQA Phi-3	SINK→0D|SINK	0.651	0.657	+0.006	[-0.000, +0.012]	yes	0.99
TQA Phi-3	SINK→SINK|DEFL	0.651	0.684	+0.033	[+0.020, +0.048]	no	0.10
TQA Phi-3	0D→0D|1D	0.646	0.643	-0.003	[-0.006, +0.001]	yes	1.00
TQA Phi-3	SINK→SINK|1D	0.651	0.651	+0.000	[-0.004, +0.005]	yes	1.00
TQA Phi-3	SINK→SINK|ANS	0.651	0.675	+0.024	[+0.007, +0.039]	no	0.00
TQA Phi-3	PH 0DTOT→PH SINK	0.845	0.871	+0.026	[+0.020, +0.031]	no	1.00
TQA Phi-3	PH SINK→PH SINK|PH 0DTOT	0.871	0.871	-0.000	[-0.001, +0.001]	yes	1.00
TQA Phi-3	PH 0D→PH ENT	0.868	0.864	-0.004	[-0.012, +0.004]	yes	0.87
TQA Phi-3	0D→LOGPROB	0.646	0.578	-0.068	[-0.092, -0.044]	no	0.00
TQA Phi-3	0D→HIDDEN	0.646	0.933	+0.287	[+0.272, +0.302]	no	0.00
TQA Phi-3	0D→LOOKBACK	0.646	0.892	+0.246	[+0.230, +0.262]	no	0.00
TQA Phi-3	0D→LLMCHECK	0.646	0.653	+0.007	[-0.010, +0.024]	no	0.00
TQA Phi-3	0D→LEX	0.646	0.820	+0.174	[+0.154, +0.193]	no	0.00
TQA Phi-3	0D→ROWSTAT	0.646	0.661	+0.015	[+0.002, +0.029]	no	0.12
TQA Phi-3	0D→LEN	0.646	0.522	-0.124	[-0.145, -0.102]	no	0.00
TQA Phi-3	NONTOPO→NONTOPO|0D	0.936	0.936	+0.000	[-0.000, +0.001]	yes	1.00
TQA Phi-3	NONTOPO→NONTOPO|DEFL	0.936	0.937	+0.001	[-0.000, +0.002]	yes	1.00
TQA Phi-3	LEX→LEX|0D	0.820	0.821	+0.001	[-0.007, +0.009]	yes	0.85
TQA Phi-3	LEX→LEX|SINK	0.820	0.820	-0.000	[-0.008, +0.007]	yes	0.89
TQA Phi-3	LEX→LEX|DEFL	0.820	0.802	-0.018	[-0.028, -0.007]	no	0.52
TQA Phi-3	LOGPROB→LOGPROB|0D	0.578	0.655	+0.077	[+0.058, +0.096]	no	0.00
TQA Phi-3	HIDDEN→HIDDEN|0D	0.933	0.933	+0.000	[-0.001, +0.001]	yes	1.00
TQA Phi-3	NONTOPO→NONTOPO⊕0D	0.936	0.936	-0.000	[-0.000, -0.000]	yes	1.00
TQA Phi-3	NONTOPO→NONTOPO⊕DEFL	0.936	0.936	-0.000	[-0.000, -0.000]	yes	1.00
TQA Phi-3	NONTOPO→NONTOPO⊕SINK	0.936	0.936	-0.000	[-0.000, +0.000]	yes	1.00
TQA Phi-3	HIDDEN→HIDDEN⊕0D	0.932	0.932	-0.000	[-0.000, -0.000]	yes	1.00
TQA Phi-3	LOGPROB→LOGPROB⊕0D	0.577	0.650	+0.073	[+0.055, +0.091]	no	0.00
TQA Phi-3	SINK→SINK⊕0D	0.650	0.651	+0.000	[-0.003, +0.003]	yes	1.00
TQA TinyLl.	0D→SINK	0.654	0.643	-0.010	[-0.021, +0.000]	no	0.51
TQA TinyLl.	0D→LEGACY MSTPROXY	0.654	0.627	-0.027	[-0.037, -0.016]	no	0.46
TQA TinyLl.	SINK→0D|SINK	0.643	0.663	+0.020	[+0.011, +0.029]	no	0.73
TQA TinyLl.	SINK→SINK|DEFL	0.643	0.688	+0.045	[+0.032, +0.057]	no	0.28
TQA TinyLl.	0D→0D|1D	0.654	0.660	+0.007	[-0.000, +0.013]	yes	0.95
TQA TinyLl.	SINK→SINK|1D	0.643	0.651	+0.008	[+0.001, +0.015]	no	0.93
TQA TinyLl.	SINK→SINK|ANS	0.643	0.685	+0.041	[+0.025, +0.058]	no	0.00
TQA TinyLl.	PH 0DTOT→PH SINK	0.804	0.808	+0.004	[-0.002, +0.011]	yes	0.97
TQA TinyLl.	PH SINK→PH SINK|PH 0DTOT	0.808	0.810	+0.003	[+0.000, +0.005]	yes	1.00
TQA TinyLl.	PH 0D→PH ENT	0.806	0.817	+0.011	[+0.000, +0.022]	no	0.50
TQA TinyLl.	0D→LOGPROB	0.654	0.637	-0.017	[-0.039, +0.004]	no	0.00
TQA TinyLl.	0D→HIDDEN	0.654	0.883	+0.229	[+0.214, +0.244]	no	0.00
TQA TinyLl.	0D→LOOKBACK	0.654	0.850	+0.197	[+0.182, +0.212]	no	0.00
TQA TinyLl.	0D→LLMCHECK	0.654	0.614	-0.039	[-0.055, -0.024]	no	0.00
TQA TinyLl.	0D→LEX	0.654	0.820	+0.166	[+0.149, +0.184]	no	0.00
TQA TinyLl.	0D→ROWSTAT	0.654	0.661	+0.007	[-0.004, +0.019]	no	0.37
TQA TinyLl.	0D→LEN	0.654	0.523	-0.131	[-0.152, -0.110]	no	0.00
TQA TinyLl.	NONTOPO→NONTOPO|0D	0.885	0.885	+0.000	[-0.000, +0.000]	yes	1.00
TQA TinyLl.	NONTOPO→NONTOPO|DEFL	0.885	0.884	-0.001	[-0.002, +0.000]	yes	1.00
TQA TinyLl.	LEX→LEX|0D	0.820	0.830	+0.010	[+0.003, +0.016]	no	0.97
TQA TinyLl.	LEX→LEX|SINK	0.820	0.831	+0.011	[+0.005, +0.017]	no	0.98
TQA TinyLl.	LEX→LEX|DEFL	0.820	0.826	+0.005	[-0.003, +0.014]	yes	0.75
TQA TinyLl.	LOGPROB→LOGPROB|0D	0.637	0.683	+0.046	[+0.033, +0.060]	no	0.13
TQA TinyLl.	HIDDEN→HIDDEN|0D	0.883	0.883	-0.000	[-0.001, +0.000]	yes	1.00
TQA TinyLl.	NONTOPO→NONTOPO⊕0D	0.884	0.884	-0.000	[-0.000, -0.000]	yes	1.00
TQA TinyLl.	NONTOPO→NONTOPO⊕DEFL	0.884	0.884	-0.000	[-0.001, +0.000]	yes	1.00
TQA TinyLl.	NONTOPO→NONTOPO⊕SINK	0.884	0.884	-0.000	[-0.000, -0.000]	yes	1.00
TQA TinyLl.	HIDDEN→HIDDEN⊕0D	0.883	0.882	-0.000	[-0.000, +0.000]	yes	1.00
TQA TinyLl.	LOGPROB→LOGPROB⊕0D	0.636	0.678	+0.043	[+0.030, +0.055]	no	0.25
TQA TinyLl.	SINK→SINK⊕0D	0.643	0.657	+0.014	[+0.007, +0.021]	no	0.96
TQA SmolLM	0D→SINK	0.583	0.611	+0.028	[+0.009, +0.048]	no	0.00
TQA SmolLM	0D→LEGACY MSTPROXY	0.583	0.585	+0.002	[-0.003, +0.006]	yes	1.00
TQA SmolLM	SINK→0D|SINK	0.611	0.641	+0.030	[+0.021, +0.039]	no	0.66
TQA SmolLM	SINK→SINK|DEFL	0.611	0.652	+0.041	[+0.030, +0.053]	no	0.34
TQA SmolLM	0D→0D|1D	0.583	0.588	+0.005	[-0.004, +0.014]	yes	0.68
TQA SmolLM	SINK→SINK|1D	0.611	0.629	+0.018	[+0.008, +0.028]	no	0.62
TQA SmolLM	SINK→SINK|ANS	0.611	0.646	+0.036	[+0.018, +0.053]	no	0.00
TQA SmolLM	PH 0DTOT→PH SINK	0.819	0.720	-0.099	[-0.113, -0.086]	no	0.15
TQA SmolLM	PH SINK→PH SINK|PH 0DTOT	0.720	0.826	+0.106	[+0.094, +0.118]	no	0.29
TQA SmolLM	PH 0D→PH ENT	0.800	0.834	+0.034	[+0.024, +0.044]	no	0.60
TQA SmolLM	0D→LOGPROB	0.583	0.588	+0.005	[-0.017, +0.027]	no	0.00
TQA SmolLM	0D→HIDDEN	0.583	0.894	+0.311	[+0.296, +0.326]	no	0.01
TQA SmolLM	0D→LOOKBACK	0.583	0.863	+0.280	[+0.264, +0.295]	no	0.00
TQA SmolLM	0D→LLMCHECK	0.583	0.604	+0.021	[+0.006, +0.037]	no	0.00
TQA SmolLM	0D→LEX	0.583	0.820	+0.237	[+0.218, +0.256]	no	0.00
TQA SmolLM	0D→ROWSTAT	0.583	0.673	+0.090	[+0.079, +0.102]	no	0.37
TQA SmolLM	0D→LEN	0.583	0.524	-0.059	[-0.074, -0.043]	no	0.00
TQA SmolLM	NONTOPO→NONTOPO|0D	0.899	0.899	-0.000	[-0.001, +0.001]	yes	1.00

Table 24: **Complete paired difference tests and equivalence tests across all settings and comparisons.** Bank names follow the main text: X|Y is early fusion (the concatenated banks) and X⊕Y is late fusion (the two banks' out-of-fold logits stacked in a second-level model), so X|Y and X⊕Y are two different models over the same features and need not agree. LEGACY MSTPROXY is per-layer 0D total persistence alone (Appendix B).

Setting	Test (A → B)	AUC_A	AUC_B	Δ	90% CI	$\equiv_{.015}$	power
TQA SmolLM	NONTOPO→NONTOPO|DEFL	0.899	0.899	-0.001	[-0.002, +0.000]	yes	1.00
TQA SmolLM	LEX→LEX|0D	0.820	0.827	+0.007	[+0.001, +0.012]	yes	1.00
TQA SmolLM	LEX→LEX|SINK	0.820	0.819	-0.001	[-0.007, +0.004]	yes	1.00
TQA SmolLM	LEX→LEX|DEFL	0.820	0.822	+0.002	[-0.005, +0.009]	yes	0.91
TQA SmolLM	LOGPROB→LOGPROB|0D	0.588	0.613	+0.025	[+0.015, +0.035]	no	0.59
TQA SmolLM	HIDDEN→HIDDEN|0D	0.894	0.894	+0.000	[-0.001, +0.001]	yes	1.00
TQA SmolLM	NONTOPO→NONTOPO⊕0D	0.899	0.899	-0.000	[-0.000, -0.000]	yes	1.00
TQA SmolLM	NONTOPO→NONTOPO⊕DEFL	0.899	0.899	-0.000	[-0.000, -0.000]	yes	1.00
TQA SmolLM	NONTOPO→NONTOPO⊕SINK	0.899	0.899	-0.000	[-0.000, +0.000]	yes	1.00
TQA SmolLM	HIDDEN→HIDDEN⊕0D	0.893	0.893	-0.000	[-0.001, -0.000]	yes	1.00
TQA SmolLM	LOGPROB→LOGPROB⊕0D	0.588	0.607	+0.019	[+0.010, +0.029]	no	0.67
TQA SmolLM	SINK→SINK⊕0D	0.610	0.633	+0.022	[+0.014, +0.031]	no	0.81
TQA Mis7B	0D→SINK	0.684	0.685	+0.000	[-0.001, +0.001]	yes	1.00
TQA Mis7B	0D→LEGACY MSTPROXY	0.684	0.646	-0.038	[-0.051, -0.026]	no	0.24
TQA Mis7B	SINK→0D|SINK	0.685	0.685	+0.000	[-0.001, +0.002]	yes	1.00
TQA Mis7B	SINK→SINK|DEFL	0.685	0.740	+0.055	[+0.040, +0.070]	no	0.00
TQA Mis7B	0D→0D|1D	0.684	0.685	+0.000	[-0.000, +0.001]	yes	1.00
TQA Mis7B	SINK→SINK|1D	0.685	0.685	+0.000	[-0.000, +0.001]	yes	1.00
TQA Mis7B	SINK→SINK|ANS	0.685	0.727	+0.043	[+0.027, +0.058]	no	0.00
TQA Mis7B	PH 0DTOT→PH SINK	0.887	0.903	+0.016	[+0.011, +0.020]	no	1.00
TQA Mis7B	PH SINK→PH SINK|PH 0DTOT	0.903	0.903	+0.001	[-0.000, +0.002]	yes	1.00
TQA Mis7B	PH 0D→PH ENT	0.893	0.897	+0.004	[-0.003, +0.011]	yes	0.94
TQA Mis7B	0D→LOGPROB	0.684	0.641	-0.044	[-0.067, -0.020]	no	0.00
TQA Mis7B	0D→HIDDEN	0.684	0.956	+0.271	[+0.256, +0.286]	no	0.00
TQA Mis7B	0D→LOOKBACK	0.684	0.933	+0.248	[+0.233, +0.263]	no	0.00
TQA Mis7B	0D→LLMCHECK	0.684	0.652	-0.032	[-0.048, -0.016]	no	0.00
TQA Mis7B	0D→LEX	0.684	0.820	+0.136	[+0.116, +0.155]	no	0.00
TQA Mis7B	0D→ROWSTAT	0.684	0.694	+0.010	[-0.004, +0.024]	no	0.08
TQA Mis7B	0D→LEN	0.684	0.525	-0.159	[-0.179, -0.140]	no	0.00
TQA Mis7B	NONTOPO→NONTOPO|0D	0.959	0.958	-0.000	[-0.000, -0.000]	yes	1.00
TQA Mis7B	NONTOPO→NONTOPO|DEFL	0.959	0.959	+0.000	[-0.000, +0.001]	yes	1.00
TQA Mis7B	LEX→LEX|0D	0.820	0.834	+0.013	[+0.004, +0.023]	no	0.66
TQA Mis7B	LEX→LEX|SINK	0.820	0.833	+0.013	[+0.004, +0.023]	no	0.66
TQA Mis7B	LEX→LEX|DEFL	0.820	0.823	+0.003	[-0.009, +0.015]	no	0.29
TQA Mis7B	LOGPROB→LOGPROB|0D	0.641	0.722	+0.081	[+0.066, +0.097]	no	0.00
TQA Mis7B	HIDDEN→HIDDEN|0D	0.956	0.956	+0.000	[-0.000, +0.001]	yes	1.00
TQA Mis7B	NONTOPO→NONTOPO⊕0D	0.958	0.958	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B	NONTOPO→NONTOPO⊕DEFL	0.958	0.958	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B	NONTOPO→NONTOPO⊕SINK	0.958	0.958	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B	HIDDEN→HIDDEN⊕0D	0.955	0.955	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B	LOGPROB→LOGPROB⊕0D	0.640	0.719	+0.079	[+0.064, +0.094]	no	0.00
TQA Mis7B	SINK→SINK⊕0D	0.684	0.684	+0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B (chat)	0D→SINK	0.688	0.688	-0.000	[-0.001, +0.000]	yes	1.00
TQA Mis7B (chat)	0D→LEGACY MSTPROXY	0.688	0.652	-0.037	[-0.049, -0.024]	no	0.23
TQA Mis7B (chat)	SINK→0D|SINK	0.688	0.689	+0.000	[-0.001, +0.001]	yes	1.00
TQA Mis7B (chat)	SINK→SINK|DEFL	0.688	0.756	+0.067	[+0.052, +0.082]	no	0.00
TQA Mis7B (chat)	0D→0D|1D	0.688	0.688	-0.000	[-0.001, +0.000]	yes	1.00
TQA Mis7B (chat)	SINK→SINK|1D	0.688	0.688	-0.000	[-0.001, +0.000]	yes	1.00
TQA Mis7B (chat)	SINK→SINK|ANS	0.688	0.767	+0.079	[+0.061, +0.096]	no	0.00
TQA Mis7B (chat)	PH 0DTOT→PH SINK	0.890	0.907	+0.016	[+0.012, +0.021]	no	1.00
TQA Mis7B (chat)	PH SINK→PH SINK|PH 0DTOT	0.907	0.907	+0.001	[-0.001, +0.002]	yes	1.00
TQA Mis7B (chat)	PH 0D→PH ENT	0.900	0.901	+0.001	[-0.006, +0.008]	yes	0.95
TQA Mis7B (chat)	0D→LOGPROB	0.688	0.651	-0.037	[-0.062, -0.013]	no	0.00
TQA Mis7B (chat)	0D→HIDDEN	0.688	0.958	+0.270	[+0.255, +0.284]	no	0.03
TQA Mis7B (chat)	0D→LOOKBACK	0.688	0.934	+0.246	[+0.231, +0.260]	no	0.03
TQA Mis7B (chat)	0D→LLMCHECK	0.688	0.653	-0.036	[-0.052, -0.019]	no	0.00
TQA Mis7B (chat)	0D→LEX	0.688	0.820	+0.132	[+0.112, +0.151]	no	0.00
TQA Mis7B (chat)	0D→ROWSTAT	0.688	0.700	+0.011	[-0.003, +0.025]	no	0.09
TQA Mis7B (chat)	0D→LEN	0.688	0.525	-0.163	[-0.184, -0.143]	no	0.00
TQA Mis7B (chat)	NONTOPO→NONTOPO|0D	0.960	0.960	-0.000	[-0.000, -0.000]	yes	1.00
TQA Mis7B (chat)	NONTOPO→NONTOPO|DEFL	0.960	0.961	+0.000	[-0.000, +0.001]	yes	1.00
TQA Mis7B (chat)	LEX→LEX|0D	0.820	0.834	+0.014	[+0.004, +0.023]	no	0.65
TQA Mis7B (chat)	LEX→LEX|SINK	0.820	0.834	+0.014	[+0.004, +0.023]	no	0.65
TQA Mis7B (chat)	LEX→LEX|DEFL	0.820	0.835	+0.015	[+0.002, +0.027]	no	0.25
TQA Mis7B (chat)	LOGPROB→LOGPROB|0D	0.651	0.733	+0.082	[+0.067, +0.097]	no	0.01
TQA Mis7B (chat)	HIDDEN→HIDDEN|0D	0.958	0.958	-0.000	[-0.001, +0.000]	yes	1.00
TQA Mis7B (chat)	NONTOPO→NONTOPO⊕0D	0.960	0.960	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B (chat)	NONTOPO→NONTOPO⊕DEFL	0.960	0.960	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B (chat)	NONTOPO→NONTOPO⊕SINK	0.960	0.960	-0.000	[-0.000, +0.000]	yes	1.00
TQA Mis7B (chat)	HIDDEN→HIDDEN⊕0D	0.958	0.958	-0.000	[-0.000, -0.000]	yes	1.00
TQA Mis7B (chat)	LOGPROB→LOGPROB⊕0D	0.651	0.728	+0.077	[+0.062, +0.092]	no	0.00
TQA Mis7B (chat)	SINK→SINK⊕0D	0.688	0.688	+0.000	[-0.000, +0.000]	yes	1.00
TQA Qw7B	0D→SINK	0.679	0.749	+0.070	[+0.050, +0.089]	no	0.00
TQA Qw7B	0D→LEGACY MSTPROXY	0.679	0.678	-0.001	[-0.007, +0.004]	yes	0.99
TQA Qw7B	SINK→0D|SINK	0.749	0.777	+0.027	[+0.021, +0.034]	no	0.95
TQA Qw7B	SINK→SINK|DEFL	0.749	0.788	+0.038	[+0.029, +0.048]	no	0.64
TQA Qw7B	0D→0D|1D	0.679	0.701	+0.021	[+0.011, +0.032]	no	0.52
TQA Qw7B	SINK→SINK|1D	0.749	0.767	+0.017	[+0.010, +0.025]	no	0.89
TQA Qw7B	SINK→SINK|ANS	0.749	0.794	+0.045	[+0.030, +0.060]	no	0.02
TQA Qw7B	PH 0DTOT→PH SINK	0.872	0.821	-0.051	[-0.062, -0.040]	no	0.45
TQA Qw7B	PH SINK→PH SINK|PH 0DTOT	0.821	0.881	+0.060	[+0.051, +0.069]	no	0.71
TQA Qw7B	PH 0D→PH ENT	0.850	0.887	+0.037	[+0.027, +0.048]	no	0.52
TQA Qw7B	0D→LOGPROB	0.679	0.674	-0.006	[-0.028, +0.017]	no	0.00
TQA Qw7B	0D→HIDDEN	0.679	0.933	+0.254	[+0.240, +0.268]	no	0.09
TQA Qw7B	0D→LOOKBACK	0.679	0.922	+0.242	[+0.229, +0.256]	no	0.17
TQA Qw7B	0D→LLMCHECK	0.679	0.667	-0.012	[-0.027, +0.004]	no	0.00

Table 24: **Complete paired difference tests and equivalence tests across all settings and comparisons.** Bank names follow the main text: X|Y is early fusion (the concatenated banks) and X⊕Y is late fusion (the two banks' out-of-fold logits stacked in a second-level model), so X|Y and X⊕Y are two different models over the same features and need not agree. LEGACY MSTPROXY is per-layer 0D total persistence alone (Appendix B).

Setting	Test (A → B)	AUC$_A$	AUC$_B$	Δ	90% CI	$\equiv_{.015}$	power
TQA Qw7B	0D→LEX	0.679	0.820	+0.141	[+0.122, +0.159]	no	0.00
TQA Qw7B	0D→ROWSTAT	0.679	0.736	+0.057	[+0.048, +0.066]	no	0.74
TQA Qw7B	0D→LEN	0.679	0.529	-0.151	[-0.171, -0.130]	no	0.00
TQA Qw7B	NONTOPO→NONTOPO|0D	0.950	0.949	-0.000	[-0.000, +0.000]	yes	1.00
TQA Qw7B	NONTOPO→NONTOPO|DEFL	0.950	0.949	-0.000	[-0.001, +0.000]	yes	1.00
TQA Qw7B	LEX→LEX|0D	0.820	0.834	+0.014	[+0.007, +0.021]	no	0.94
TQA Qw7B	LEX→LEX|SINK	0.820	0.848	+0.028	[+0.016, +0.040]	no	0.29
TQA Qw7B	LEX→LEX|DEFL	0.820	0.828	+0.008	[-0.001, +0.016]	no	0.78
TQA Qw7B	LOGPROB→LOGPROB|0D	0.674	0.727	+0.053	[+0.042, +0.065]	no	0.40
TQA Qw7B	HIDDEN→HIDDEN|0D	0.933	0.933	-0.000	[-0.001, +0.000]	yes	1.00
TQA Qw7B	NONTOPO→NONTOPO⊕0D	0.949	0.949	-0.000	[-0.000, +0.000]	yes	1.00
TQA Qw7B	NONTOPO→NONTOPO⊕DEFL	0.949	0.949	+0.000	[-0.000, +0.000]	yes	1.00
TQA Qw7B	NONTOPO→NONTOPO⊕SINK	0.949	0.949	+0.000	[-0.001, +0.001]	yes	1.00
TQA Qw7B	HIDDEN→HIDDEN⊕0D	0.933	0.933	+0.000	[-0.000, +0.001]	yes	1.00
TQA Qw7B	LOGPROB→LOGPROB⊕0D	0.673	0.725	+0.052	[+0.039, +0.064]	no	0.23
TQA Qw7B	SINK→SINK⊕0D	0.749	0.775	+0.026	[+0.018, +0.034]	no	0.88
HE Qw3B	0D→SINK	0.811	0.829	+0.018	[+0.008, +0.028]	no	0.64
HE Qw3B	0D→LEGACY MSTPROXY	0.811	0.731	-0.080	[-0.091, -0.069]	no	0.46
HE Qw3B	SINK→0D|SINK	0.829	0.857	+0.028	[+0.022, +0.035]	no	0.97
HE Qw3B	SINK→SINK|DEFL	0.829	0.887	+0.058	[+0.049, +0.067]	no	0.73
HE Qw3B	0D→0D|1D	0.811	0.826	+0.015	[+0.009, +0.023]	no	0.94
HE Qw3B	SINK→SINK|1D	0.829	0.843	+0.014	[+0.007, +0.021]	no	0.97
HE Qw3B	SINK→SINK|ANS	0.829	0.978	+0.150	[+0.139, +0.161]	no	0.44
HE Qw3B	PH 0DTOT→PH SINK	0.958	0.958	+0.000	[-0.005, +0.006]	yes	1.00
HE Qw3B	PH SINK→PH SINK|PH 0DTOT	0.958	0.962	+0.004	[+0.003, +0.005]	yes	1.00
HE Qw3B	PH 0D→PH ENT	0.971	0.964	-0.006	[-0.012, -0.000]	yes	0.99
HE Qw3B	0D→LOGPROB	0.811	0.993	+0.182	[+0.170, +0.194]	no	0.33
HE Qw3B	0D→HIDDEN	0.811	1.000	+0.189	[+0.177, +0.201]	no	0.32
HE Qw3B	0D→LOOKBACK	0.811	0.998	+0.187	[+0.175, +0.199]	no	0.32
HE Qw3B	0D→LLMCHECK	0.811	0.745	-0.066	[-0.078, -0.054]	no	0.27
HE Qw3B	0D→LEX	0.811	0.981	+0.170	[+0.159, +0.182]	no	0.37
HE Qw3B	0D→ROWSTAT	0.811	0.829	+0.018	[+0.007, +0.030]	no	0.35
HE Qw3B	0D→LEN	0.811	0.962	+0.151	[+0.140, +0.163]	no	0.36
HE Qw3B	NONTOPO→NONTOPO|0D	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	NONTOPO→NONTOPO|DEFL	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	LEX→LEX|0D	0.981	0.979	-0.002	[-0.005, +0.000]	yes	1.00
HE Qw3B	LEX→LEX|SINK	0.981	0.980	-0.001	[-0.003, +0.001]	yes	1.00
HE Qw3B	LEX→LEX|DEFL	0.981	0.971	-0.010	[-0.014, -0.007]	yes	1.00
HE Qw3B	LOGPROB→LOGPROB|0D	0.993	0.993	+0.000	[-0.001, +0.001]	yes	1.00
HE Qw3B	HIDDEN→HIDDEN|0D	1.000	1.000	-0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	NONTOPO→NONTOPO⊕0D	1.000	1.000	-0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	NONTOPO→NONTOPO⊕DEFL	1.000	1.000	-0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	NONTOPO→NONTOPO⊕SINK	1.000	1.000	-0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	HIDDEN→HIDDEN⊕0D	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw3B	LOGPROB→LOGPROB⊕0D	0.993	0.993	+0.001	[+0.000, +0.001]	yes	1.00
HE Qw3B	SINK→SINK⊕0D	0.828	0.839	+0.011	[+0.007, +0.015]	yes	1.00
HE Qw1.5B	0D→SINK	0.789	0.820	+0.031	[+0.017, +0.045]	no	0.08
HE Qw1.5B	0D→LEGACY MSTPROXY	0.789	0.719	-0.070	[-0.086, -0.053]	no	0.00
HE Qw1.5B	SINK→0D|SINK	0.820	0.832	+0.012	[+0.005, +0.019]	no	0.91
HE Qw1.5B	SINK→SINK|DEFL	0.820	0.872	+0.052	[+0.038, +0.065]	no	0.17
HE Qw1.5B	0D→0D|1D	0.789	0.815	+0.026	[+0.013, +0.040]	no	0.17
HE Qw1.5B	SINK→SINK|1D	0.820	0.826	+0.006	[-0.005, +0.017]	no	0.45
HE Qw1.5B	SINK→SINK|ANS	0.820	0.975	+0.155	[+0.138, +0.172]	no	0.00
HE Qw1.5B	PH 0DTOT→PH SINK	0.921	0.922	+0.002	[-0.009, +0.013]	yes	0.46
HE Qw1.5B	PH SINK→PH SINK|PH 0DTOT	0.922	0.931	+0.008	[+0.006, +0.011]	yes	1.00
HE Qw1.5B	PH 0D→PH ENT	0.949	0.933	-0.016	[-0.028, -0.004]	no	0.33
HE Qw1.5B	0D→LOGPROB	0.789	0.995	+0.205	[+0.188, +0.222]	no	0.00
HE Qw1.5B	0D→HIDDEN	0.789	1.000	+0.210	[+0.193, +0.227]	no	0.00
HE Qw1.5B	0D→LOOKBACK	0.789	0.999	+0.210	[+0.193, +0.226]	no	0.00
HE Qw1.5B	0D→LLMCHECK	0.789	0.791	+0.002	[-0.017, +0.021]	no	0.00
HE Qw1.5B	0D→LEX	0.789	0.979	+0.190	[+0.172, +0.207]	no	0.00
HE Qw1.5B	0D→ROWSTAT	0.789	0.788	-0.001	[-0.018, +0.016]	no	0.00
HE Qw1.5B	0D→LEN	0.789	0.967	+0.178	[+0.160, +0.196]	no	0.00
HE Qw1.5B	NONTOPO→NONTOPO|0D	1.000	1.000	+0.000	[+0.000, +0.000]	yes	1.00
HE Qw1.5B	NONTOPO→NONTOPO|DEFL	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw1.5B	LEX→LEX|0D	0.979	0.971	-0.008	[-0.014, -0.002]	yes	0.99
HE Qw1.5B	LEX→LEX|SINK	0.979	0.968	-0.011	[-0.017, -0.005]	no	0.98
HE Qw1.5B	LEX→LEX|DEFL	0.979	0.956	-0.022	[-0.030, -0.016]	no	0.94
HE Qw1.5B	LOGPROB→LOGPROB|0D	0.995	0.994	-0.001	[-0.003, +0.001]	yes	1.00
HE Qw1.5B	HIDDEN→HIDDEN|0D	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw1.5B	NONTOPO→NONTOPO⊕0D	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw1.5B	NONTOPO→NONTOPO⊕DEFL	1.000	1.000	-0.000	[-0.000, +0.000]	yes	1.00
HE Qw1.5B	NONTOPO→NONTOPO⊕SINK	1.000	1.000	+0.000	[-0.000, +0.000]	yes	1.00
HE Qw1.5B	HIDDEN→HIDDEN⊕0D	1.000	1.000	-0.000	[-0.000, +0.000]	yes	1.00
HE Qw1.5B	LOGPROB→LOGPROB⊕0D	0.994	0.994	-0.000	[-0.001, +0.001]	yes	1.00
HE Qw1.5B	SINK→SINK⊕0D	0.819	0.823	+0.004	[-0.000, +0.009]	yes	1.00
TrQA Qw3B	0D→SINK	0.732	0.717	-0.015	[-0.026, -0.004]	no	0.48
TrQA Qw3B	0D→LEGACY MSTPROXY	0.732	0.691	-0.041	[-0.057, -0.025]	no	0.00
TrQA Qw3B	SINK→0D|SINK	0.717	0.738	+0.021	[+0.012, +0.030]	no	0.76
TrQA Qw3B	SINK→SINK|DEFL	0.717	0.719	+0.002	[-0.012, +0.015]	no	0.13
TrQA Qw3B	0D→0D|1D	0.732	0.729	-0.003	[-0.012, +0.006]	yes	0.74
TrQA Qw3B	SINK→SINK|1D	0.717	0.718	+0.001	[-0.009, +0.010]	yes	0.65
TrQA Qw3B	SINK→SINK|ANS	0.717	0.752	+0.035	[+0.019, +0.051]	no	0.00
TrQA Qw3B	PH 0DTOT→PH SINK	0.807	0.820	+0.012	[+0.004, +0.021]	no	0.82
TrQA Qw3B	PH SINK→PH SINK|PH 0DTOT	0.820	0.824	+0.004	[+0.002, +0.006]	yes	1.00
TrQA Qw3B	PH 0D→PH ENT	0.807	0.810	+0.003	[-0.011, +0.017]	no	0.08

Table 24: **Complete paired difference tests and equivalence tests across all settings and comparisons.** Bank names follow the main text: X|Y is early fusion (the concatenated banks) and X⊕Y is late fusion (the two banks' out-of-fold logits stacked in a second-level model), so X|Y and X⊕Y are two different models over the same features and need not agree. LEGACY MSTPROXY is per-layer 0D total persistence alone (Appendix B).

Setting	Test (A → B)	AUC_A	AUC_B	Δ	90% CI	$\equiv_{.015}$	power
TrQA Qw3B	0D→LOGPROB	0.732	0.839	+0.107	[+0.088, +0.125]	no	0.00
TrQA Qw3B	0D→HIDDEN	0.732	0.841	+0.109	[+0.090, +0.129]	no	0.00
TrQA Qw3B	0D→LOOKBACK	0.732	0.861	+0.129	[+0.111, +0.147]	no	0.00
TrQA Qw3B	0D→LLMCHECK	0.732	0.688	-0.044	[-0.065, -0.023]	no	0.00
TrQA Qw3B	0D→LEX	0.732	0.596	-0.137	[-0.163, -0.110]	no	0.00
TrQA Qw3B	0D→ROWSTAT	0.732	0.724	-0.009	[-0.025, +0.008]	no	0.00
TrQA Qw3B	0D→LEN	0.732	0.569	-0.163	[-0.189, -0.137]	no	0.00
TrQA Qw3B	NONTOPO→NONTOPO|0D	0.870	0.870	+0.001	[-0.000, +0.002]	yes	1.00
TrQA Qw3B	NONTOPO→NONTOPO|DEFL	0.870	0.870	+0.000	[-0.001, +0.002]	yes	1.00
TrQA Qw3B	LEX→LEX|0D	0.596	0.733	+0.137	[+0.117, +0.158]	no	0.00
TrQA Qw3B	LEX→LEX|SINK	0.596	0.724	+0.128	[+0.108, +0.149]	no	0.00
TrQA Qw3B	LEX→LEX|DEFL	0.596	0.690	+0.095	[+0.073, +0.116]	no	0.00
TrQA Qw3B	LOGPROB→LOGPROB|0D	0.839	0.843	+0.004	[-0.002, +0.011]	yes	0.95
TrQA Qw3B	HIDDEN→HIDDEN|0D	0.841	0.843	+0.002	[+0.000, +0.004]	yes	1.00
TrQA Qw3B	NONTOPO→NONTOPO⊕0D	0.869	0.871	+0.002	[-0.000, +0.004]	yes	1.00
TrQA Qw3B	NONTOPO→NONTOPO⊕DEFL	0.869	0.870	+0.001	[-0.001, +0.002]	yes	1.00
TrQA Qw3B	NONTOPO→NONTOPO⊕SINK	0.869	0.871	+0.001	[-0.001, +0.004]	yes	1.00
TrQA Qw3B	HIDDEN→HIDDEN⊕0D	0.841	0.849	+0.008	[+0.003, +0.012]	yes	1.00
TrQA Qw3B	LOGPROB→LOGPROB⊕0D	0.838	0.844	+0.006	[+0.002, +0.010]	yes	1.00
TrQA Qw3B	SINK→SINK⊕0D	0.716	0.733	+0.017	[+0.010, +0.025]	no	0.89
TrQA Qw1.5B	0D→SINK	0.716	0.707	-0.009	[-0.020, +0.002]	no	0.43
TrQA Qw1.5B	0D→LEGACY MSTPROXY	0.716	0.685	-0.031	[-0.046, -0.015]	no	0.00
TrQA Qw1.5B	SINK→0D|SINK	0.707	0.717	+0.010	[+0.001, +0.020]	no	0.68
TrQA Qw1.5B	SINK→SINK|DEFL	0.707	0.719	+0.012	[-0.003, +0.027]	no	0.00
TrQA Qw1.5B	0D→0D|1D	0.716	0.714	-0.002	[-0.011, +0.008]	yes	0.65
TrQA Qw1.5B	SINK→SINK|1D	0.707	0.709	+0.002	[-0.008, +0.013]	yes	0.49
TrQA Qw1.5B	SINK→SINK|ANS	0.707	0.726	+0.019	[+0.001, +0.036]	no	0.00
TrQA Qw1.5B	PH 0DTOT→PH SINK	0.759	0.767	+0.008	[-0.003, +0.018]	no	0.53
TrQA Qw1.5B	PH SINK→PH SINK|PH 0DTOT	0.767	0.772	+0.004	[+0.001, +0.008]	yes	1.00
TrQA Qw1.5B	PH 0D→PH ENT	0.768	0.768	-0.000	[-0.017, +0.017]	no	0.00
TrQA Qw1.5B	0D→LOGPROB	0.716	0.823	+0.107	[+0.085, +0.129]	no	0.00
TrQA Qw1.5B	0D→HIDDEN	0.716	0.855	+0.140	[+0.119, +0.160]	no	0.00
TrQA Qw1.5B	0D→LOOKBACK	0.716	0.824	+0.108	[+0.088, +0.129]	no	0.00
TrQA Qw1.5B	0D→LLMCHECK	0.716	0.676	-0.039	[-0.062, -0.016]	no	0.00
TrQA Qw1.5B	0D→LEX	0.716	0.609	-0.106	[-0.136, -0.078]	no	0.00
TrQA Qw1.5B	0D→ROWSTAT	0.716	0.720	+0.005	[-0.011, +0.020]	no	0.00
TrQA Qw1.5B	0D→LEN	0.716	0.588	-0.128	[-0.154, -0.102]	no	0.00
TrQA Qw1.5B	NONTOPO→NONTOPO|0D	0.873	0.873	+0.000	[-0.001, +0.001]	yes	1.00
TrQA Qw1.5B	NONTOPO→NONTOPO|DEFL	0.873	0.872	-0.001	[-0.002, +0.001]	yes	1.00
TrQA Qw1.5B	LEX→LEX|0D	0.609	0.721	+0.112	[+0.091, +0.134]	no	0.00
TrQA Qw1.5B	LEX→LEX|SINK	0.609	0.715	+0.106	[+0.085, +0.127]	no	0.00
TrQA Qw1.5B	LEX→LEX|DEFL	0.609	0.672	+0.063	[+0.040, +0.085]	no	0.00
TrQA Qw1.5B	LOGPROB→LOGPROB|0D	0.823	0.828	+0.006	[-0.001, +0.013]	yes	0.92
TrQA Qw1.5B	HIDDEN→HIDDEN|0D	0.855	0.855	+0.000	[-0.001, +0.002]	yes	1.00
TrQA Qw1.5B	NONTOPO→NONTOPO⊕0D	0.872	0.873	+0.000	[-0.001, +0.002]	yes	1.00
TrQA Qw1.5B	NONTOPO→NONTOPO⊕DEFL	0.872	0.873	+0.000	[-0.001, +0.002]	yes	1.00
TrQA Qw1.5B	NONTOPO→NONTOPO⊕SINK	0.872	0.873	+0.001	[-0.001, +0.002]	yes	1.00
TrQA Qw1.5B	HIDDEN→HIDDEN⊕0D	0.855	0.857	+0.002	[-0.000, +0.004]	yes	1.00
TrQA Qw1.5B	LOGPROB→LOGPROB⊕0D	0.822	0.828	+0.006	[+0.001, +0.010]	yes	1.00
TrQA Qw1.5B	SINK→SINK⊕0D	0.706	0.717	+0.011	[+0.003, +0.018]	no	0.93
TrQA Phi-3	0D→SINK	0.710	0.707	-0.004	[-0.010, +0.002]	yes	0.98
TrQA Phi-3	0D→LEGACY MSTPROXY	0.710	0.669	-0.041	[-0.057, -0.025]	no	0.00
TrQA Phi-3	SINK→0D|SINK	0.707	0.713	+0.006	[+0.000, +0.012]	yes	0.99
TrQA Phi-3	SINK→SINK|DEFL	0.707	0.736	+0.029	[+0.013, +0.045]	no	0.00
TrQA Phi-3	0D→0D|1D	0.710	0.705	-0.005	[-0.011, +0.000]	yes	0.99
TrQA Phi-3	SINK→SINK|1D	0.707	0.699	-0.008	[-0.013, -0.002]	yes	1.00
TrQA Phi-3	SINK→SINK|ANS	0.707	0.736	+0.029	[+0.012, +0.046]	no	0.00
TrQA Phi-3	PH 0DTOT→PH SINK	0.810	0.826	+0.016	[+0.009, +0.022]	no	0.96
TrQA Phi-3	PH SINK→PH SINK|PH 0DTOT	0.826	0.828	+0.002	[-0.001, +0.004]	yes	1.00
TrQA Phi-3	PH 0D→PH ENT	0.807	0.817	+0.010	[-0.002, +0.022]	no	0.31
TrQA Phi-3	0D→LOGPROB	0.710	0.877	+0.167	[+0.148, +0.186]	no	0.00
TrQA Phi-3	0D→HIDDEN	0.710	0.867	+0.157	[+0.137, +0.176]	no	0.00
TrQA Phi-3	0D→LOOKBACK	0.710	0.840	+0.129	[+0.109, +0.150]	no	0.00
TrQA Phi-3	0D→LLMCHECK	0.710	0.676	-0.034	[-0.055, -0.014]	no	0.00
TrQA Phi-3	0D→LEX	0.710	0.635	-0.076	[-0.102, -0.050]	no	0.00
TrQA Phi-3	0D→ROWSTAT	0.710	0.728	+0.018	[+0.002, +0.034]	no	0.00
TrQA Phi-3	0D→LEN	0.710	0.581	-0.130	[-0.156, -0.104]	no	0.00
TrQA Phi-3	NONTOPO→NONTOPO|0D	0.883	0.883	-0.000	[-0.001, +0.000]	yes	1.00
TrQA Phi-3	NONTOPO→NONTOPO|DEFL	0.883	0.885	+0.002	[+0.000, +0.004]	yes	1.00
TrQA Phi-3	LEX→LEX|0D	0.635	0.721	+0.086	[+0.069, +0.104]	no	0.00
TrQA Phi-3	LEX→LEX|SINK	0.635	0.718	+0.083	[+0.066, +0.101]	no	0.00
TrQA Phi-3	LEX→LEX|DEFL	0.635	0.712	+0.077	[+0.057, +0.098]	no	0.00
TrQA Phi-3	LOGPROB→LOGPROB|0D	0.877	0.878	+0.000	[-0.002, +0.002]	yes	1.00
TrQA Phi-3	HIDDEN→HIDDEN|0D	0.867	0.867	-0.000	[-0.001, +0.000]	yes	1.00
TrQA Phi-3	NONTOPO→NONTOPO⊕0D	0.883	0.884	+0.000	[-0.001, +0.002]	yes	1.00
TrQA Phi-3	NONTOPO→NONTOPO⊕DEFL	0.883	0.884	+0.001	[-0.001, +0.003]	yes	1.00
TrQA Phi-3	NONTOPO→NONTOPO⊕SINK	0.883	0.883	+0.000	[-0.001, +0.001]	yes	1.00
TrQA Phi-3	HIDDEN→HIDDEN⊕0D	0.867	0.869	+0.002	[-0.000, +0.004]	yes	1.00
TrQA Phi-3	LOGPROB→LOGPROB⊕0D	0.877	0.878	+0.001	[-0.001, +0.003]	yes	1.00
TrQA Phi-3	SINK→SINK⊕0D	0.706	0.710	+0.004	[-0.001, +0.009]	yes	1.00
TrQA TinyLl.	0D→SINK	0.720	0.720	+0.001	[-0.010, +0.012]	yes	0.40
TrQA TinyLl.	0D→LEGACY MSTPROXY	0.720	0.729	+0.010	[-0.000, +0.020]	no	0.55
TrQA TinyLl.	SINK→0D|SINK	0.720	0.724	+0.004	[-0.003, +0.011]	yes	0.93
TrQA TinyLl.	SINK→SINK|DEFL	0.720	0.734	+0.014	[-0.000, +0.028]	no	0.10
TrQA TinyLl.	0D→0D|1D	0.720	0.718	-0.002	[-0.011, +0.008]	yes	0.60
TrQA TinyLl.	SINK→SINK|1D	0.720	0.719	-0.001	[-0.010, +0.008]	yes	0.73

Table 24: **Complete paired difference tests and equivalence tests across all settings and comparisons.** Bank names follow the main text: X|Y is early fusion (the concatenated banks) and X⊕Y is late fusion (the two banks' out-of-fold logits stacked in a second-level model), so X|Y and X⊕Y are two different models over the same features and need not agree. LEGACY MSTPROXY is per-layer 0D total persistence alone (Appendix B).

Setting	Test (A → B)	AUC$_A$	AUC$_B$	Δ	90% CI	$\equiv_{.015}$	power
TrQA TinyLl.	SINK→SINK|ANS	0.720	0.725	+0.005	[-0.011, +0.021]	no	0.00
TrQA TinyLl.	PH 0DTOT→PH SINK	0.792	0.814	+0.022	[+0.015, +0.029]	no	0.94
TrQA TinyLl.	PH SINK→PH SINK|PH 0DTOT	0.814	0.813	-0.001	[-0.002, +0.000]	yes	1.00
TrQA TinyLl.	PH 0D→PH ENT	0.775	0.807	+0.032	[+0.017, +0.047]	no	0.01
TrQA TinyLl.	0D→LOGPROB	0.720	0.724	+0.004	[-0.019, +0.027]	no	0.00
TrQA TinyLl.	0D→HIDDEN	0.720	0.827	+0.108	[+0.088, +0.128]	no	0.00
TrQA TinyLl.	0D→LOOKBACK	0.720	0.823	+0.104	[+0.087, +0.121]	no	0.00
TrQA TinyLl.	0D→LLMCHECK	0.720	0.688	-0.032	[-0.050, -0.013]	no	0.00
TrQA TinyLl.	0D→LEX	0.720	0.740	+0.021	[-0.003, +0.044]	no	0.00
TrQA TinyLl.	0D→ROWSTAT	0.720	0.752	+0.032	[+0.020, +0.045]	no	0.26
TrQA TinyLl.	0D→LEN	0.720	0.631	-0.089	[-0.109, -0.068]	no	0.00
TrQA TinyLl.	NONTOPO→NONTOPO|0D	0.842	0.842	-0.000	[-0.001, +0.001]	yes	1.00
TrQA TinyLl.	NONTOPO→NONTOPO|DEFL	0.842	0.842	+0.000	[-0.003, +0.003]	yes	1.00
TrQA TinyLl.	LEX→LEX|0D	0.740	0.764	+0.024	[+0.012, +0.036]	no	0.33
TrQA TinyLl.	LEX→LEX|SINK	0.740	0.762	+0.021	[+0.010, +0.032]	no	0.45
TrQA TinyLl.	LEX→LEX|DEFL	0.740	0.757	+0.016	[+0.002, +0.031]	no	0.04
TrQA TinyLl.	LOGPROB→LOGPROB|0D	0.724	0.762	+0.038	[+0.025, +0.052]	no	0.17
TrQA TinyLl.	HIDDEN→HIDDEN|0D	0.827	0.828	+0.001	[-0.001, +0.002]	yes	1.00
TrQA TinyLl.	NONTOPO→NONTOPO⊕0D	0.842	0.842	+0.000	[-0.001, +0.002]	yes	1.00
TrQA TinyLl.	NONTOPO→NONTOPO⊕DEFL	0.842	0.843	+0.001	[-0.001, +0.004]	yes	1.00
TrQA TinyLl.	NONTOPO→NONTOPO⊕SINK	0.842	0.842	+0.001	[-0.001, +0.002]	yes	1.00
TrQA TinyLl.	HIDDEN→HIDDEN⊕0D	0.827	0.828	+0.001	[-0.001, +0.004]	yes	1.00
TrQA TinyLl.	LOGPROB→LOGPROB⊕0D	0.723	0.755	+0.031	[+0.020, +0.043]	no	0.40
TrQA TinyLl.	SINK→SINK⊕0D	0.720	0.725	+0.005	[+0.001, +0.009]	yes	1.00
TrQA Mis7B	0D→SINK	0.699	0.696	-0.003	[-0.013, +0.007]	yes	0.62
TrQA Mis7B	0D→LEGACY MSTPROXY	0.699	0.693	-0.006	[-0.019, +0.007]	no	0.21
TrQA Mis7B	SINK→0D|SINK	0.696	0.698	+0.002	[-0.007, +0.010]	yes	0.80
TrQA Mis7B	SINK→SINK|DEFL	0.696	0.739	+0.043	[+0.026, +0.060]	no	0.00
TrQA Mis7B	0D→0D|1D	0.699	0.692	-0.007	[-0.017, +0.003]	no	0.59
TrQA Mis7B	SINK→SINK|1D	0.696	0.686	-0.010	[-0.020, +0.001]	no	0.52
TrQA Mis7B	SINK→SINK|ANS	0.696	0.693	-0.004	[-0.020, +0.014]	no	0.00
TrQA Mis7B	PH 0DTOT→PH SINK	0.808	0.812	+0.004	[-0.004, +0.011]	yes	0.88
TrQA Mis7B	PH SINK→PH SINK|PH 0DTOT	0.812	0.814	+0.002	[+0.000, +0.004]	yes	1.00
TrQA Mis7B	PH 0D→PH ENT	0.775	0.812	+0.037	[+0.023, +0.051]	no	0.09
TrQA Mis7B	0D→LOGPROB	0.699	0.728	+0.029	[+0.005, +0.053]	no	0.00
TrQA Mis7B	0D→HIDDEN	0.699	0.833	+0.134	[+0.113, +0.155]	no	0.00
TrQA Mis7B	0D→LOOKBACK	0.699	0.829	+0.130	[+0.110, +0.150]	no	0.00
TrQA Mis7B	0D→LLMCHECK	0.699	0.686	-0.013	[-0.034, +0.008]	no	0.00
TrQA Mis7B	0D→LEX	0.699	0.677	-0.023	[-0.047, +0.002]	no	0.00
TrQA Mis7B	0D→ROWSTAT	0.699	0.731	+0.032	[+0.017, +0.047]	no	0.03
TrQA Mis7B	0D→LEN	0.699	0.635	-0.064	[-0.086, -0.043]	no	0.00
TrQA Mis7B	NONTOPO→NONTOPO|0D	0.841	0.842	+0.001	[-0.000, +0.002]	yes	1.00
TrQA Mis7B	NONTOPO→NONTOPO|DEFL	0.841	0.841	-0.000	[-0.001, +0.001]	yes	1.00
TrQA Mis7B	LEX→LEX|0D	0.677	0.723	+0.046	[+0.030, +0.062]	no	0.00
TrQA Mis7B	LEX→LEX|SINK	0.677	0.721	+0.045	[+0.029, +0.060]	no	0.00
TrQA Mis7B	LEX→LEX|DEFL	0.677	0.744	+0.068	[+0.049, +0.087]	no	0.00
TrQA Mis7B	LOGPROB→LOGPROB|0D	0.728	0.755	+0.027	[+0.012, +0.043]	no	0.00
TrQA Mis7B	HIDDEN→HIDDEN|0D	0.833	0.834	+0.000	[-0.001, +0.002]	yes	1.00
TrQA Mis7B	NONTOPO→NONTOPO⊕0D	0.841	0.843	+0.003	[+0.000, +0.005]	yes	1.00
TrQA Mis7B	NONTOPO→NONTOPO⊕DEFL	0.841	0.846	+0.005	[+0.002, +0.008]	yes	1.00
TrQA Mis7B	NONTOPO→NONTOPO⊕SINK	0.841	0.843	+0.002	[-0.000, +0.005]	yes	1.00
TrQA Mis7B	HIDDEN→HIDDEN⊕0D	0.833	0.836	+0.003	[+0.000, +0.006]	yes	1.00
TrQA Mis7B	LOGPROB→LOGPROB⊕0D	0.727	0.755	+0.028	[+0.014, +0.042]	no	0.11
TrQA Mis7B	SINK→SINK⊕0D	0.696	0.700	+0.005	[-0.001, +0.011]	yes	0.98
TrQA Qw7B	0D→SINK	0.668	0.714	+0.046	[+0.022, +0.070]	no	0.00
TrQA Qw7B	0D→LEGACY MSTPROXY	0.668	0.672	+0.004	[-0.002, +0.010]	yes	0.98
TrQA Qw7B	SINK→0D|SINK	0.714	0.735	+0.022	[+0.011, +0.033]	no	0.46
TrQA Qw7B	SINK→SINK|DEFL	0.714	0.734	+0.020	[+0.008, +0.033]	no	0.27
TrQA Qw7B	0D→0D|1D	0.668	0.668	+0.000	[-0.010, +0.010]	yes	0.59
TrQA Qw7B	SINK→SINK|1D	0.714	0.718	+0.004	[-0.004, +0.013]	yes	0.73
TrQA Qw7B	SINK→SINK|ANS	0.714	0.781	+0.067	[+0.051, +0.083]	no	0.00
TrQA Qw7B	PH 0DTOT→PH SINK	0.822	0.801	-0.020	[-0.035, -0.006]	no	0.07
TrQA Qw7B	PH SINK→PH SINK|PH 0DTOT	0.801	0.832	+0.031	[+0.024, +0.039]	no	0.91
TrQA Qw7B	PH 0D→PH ENT	0.798	0.841	+0.044	[+0.032, +0.056]	no	0.33
TrQA Qw7B	0D→LOGPROB	0.668	0.847	+0.180	[+0.158, +0.201]	no	0.00
TrQA Qw7B	0D→HIDDEN	0.668	0.854	+0.186	[+0.165, +0.207]	no	0.00
TrQA Qw7B	0D→LOOKBACK	0.668	0.879	+0.211	[+0.191, +0.231]	no	0.00
TrQA Qw7B	0D→LLMCHECK	0.668	0.725	+0.058	[+0.038, +0.078]	no	0.00
TrQA Qw7B	0D→LEX	0.668	0.594	-0.074	[-0.101, -0.046]	no	0.00
TrQA Qw7B	0D→ROWSTAT	0.668	0.734	+0.067	[+0.053, +0.081]	no	0.10
TrQA Qw7B	0D→LEN	0.668	0.595	-0.073	[-0.097, -0.049]	no	0.00
TrQA Qw7B	NONTOPO→NONTOPO|0D	0.881	0.881	-0.000	[-0.001, +0.000]	yes	1.00
TrQA Qw7B	NONTOPO→NONTOPO|DEFL	0.881	0.881	+0.000	[-0.001, +0.001]	yes	1.00
TrQA Qw7B	LEX→LEX|0D	0.594	0.678	+0.085	[+0.067, +0.102]	no	0.00
TrQA Qw7B	LEX→LEX|SINK	0.594	0.728	+0.134	[+0.113, +0.154]	no	0.00
TrQA Qw7B	LEX→LEX|DEFL	0.594	0.678	+0.085	[+0.066, +0.103]	no	0.00
TrQA Qw7B	LOGPROB→LOGPROB|0D	0.847	0.842	-0.005	[-0.010, +0.001]	yes	1.00
TrQA Qw7B	HIDDEN→HIDDEN|0D	0.854	0.853	-0.001	[-0.001, -0.000]	yes	1.00
TrQA Qw7B	NONTOPO→NONTOPO⊕0D	0.881	0.881	-0.000	[-0.001, +0.001]	yes	1.00
TrQA Qw7B	NONTOPO→NONTOPO⊕DEFL	0.881	0.881	-0.000	[-0.001, +0.001]	yes	1.00
TrQA Qw7B	NONTOPO→NONTOPO⊕SINK	0.881	0.882	+0.001	[-0.001, +0.003]	yes	1.00
TrQA Qw7B	HIDDEN→HIDDEN⊕0D	0.853	0.854	+0.001	[-0.001, +0.002]	yes	1.00
TrQA Qw7B	LOGPROB→LOGPROB⊕0D	0.847	0.846	-0.000	[-0.003, +0.002]	yes	1.00
TrQA Qw7B	SINK→SINK⊕0D	0.713	0.732	+0.019	[+0.009, +0.029]	no	0.60

Table 21: **The coning defect as a detector** (§5.2). AUC of the per-layer defects (δ_0, $\min_s \delta_s$) and of per-head δ_0 (δ_h); then ΔAUC from adding them to per-head sink mass (early fusion, $|$), and from stacking them onto SINK, per-head sink mass or the non-topological bank (late fusion, $\oplus$). Marks as in Table 22.

Setting	AUC δ	AUC δ_h	δ_h $\|$Sink$_h$	$\delta_h \oplus$Sink$_h$	$\delta \oplus$Sink	$\delta_h \oplus$NT	$\delta \oplus$NT
TQA Qw3B	0.541	0.820	$+0.015^{*}$	$+0.018^{*}$	$-0.001^{\equiv}$	$+0.002^{*\equiv}$	$0.000^{\equiv}$
TQA Qw1.5B	0.568	0.763	$+0.006^{\equiv}$	$+0.010^{*\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA Phi-3	0.523	0.840	$+0.005^{\equiv}$	$+0.009^{*\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$	$0.000^{*\equiv}$
TQA TinyLl.	0.498	0.736	$+0.002^{\equiv}$	$+0.006^{*\equiv}$	$+0.001^{\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$
TQA SmolLM	0.499	0.691	$+0.004$	$+0.020^{*}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA Mis7B	0.496	0.846	$-0.001^{\equiv}$	$+0.004^{*\equiv}$	$-0.001^{*\equiv}$	$0.000^{\equiv}$	$-0.001^{\equiv}$
TQA Mis7B (chat)	0.497	0.851	$-0.001^{\equiv}$	$+0.004^{*\equiv}$	$-0.001^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA Qw7B	0.584	0.751	$+0.001^{\equiv}$	$+0.019^{*}$	$+0.002^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
HE Qw3B	0.654	0.957	$+0.015^{*}$	$+0.018^{*}$	$+0.002^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
HE Qw1.5B	0.641	0.898	$+0.018^{*}$	$+0.025^{*}$	$+0.002^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TrQA Qw3B	0.598	0.781	-0.004	$+0.013^{*}$	$+0.006^{\equiv}$	$+0.004^{*\equiv}$	$0.000^{\equiv}$
TrQA Qw1.5B	0.614	0.766	$+0.019^{*}$	$+0.027^{*}$	$+0.003^{\equiv}$	$+0.004^{\equiv}$	$0.000^{\equiv}$
TrQA Phi-3	0.505	0.789	-0.012	$+0.011^{*}$	$0.000^{\equiv}$	$+0.002^{\equiv}$	$0.000^{*\equiv}$
TrQA TinyLl.	0.644	0.784	$+0.004$	$+0.010^{*}$	$+0.002^{\equiv}$	$+0.008^{*\equiv}$	$-0.001^{\equiv}$
TrQA Mis7B	0.561	0.769	-0.008	$+0.010^{*}$	$-0.001^{\equiv}$	$+0.009^{*\equiv}$	$0.000^{*\equiv}$
TrQA Qw7B	0.671	0.782	$+0.015^{*}$	$+0.029^{*}$	$+0.017^{*}$	$+0.005^{*\equiv}$	$+0.001^{\equiv}$

Table 22: **Key paired comparisons** (ΔAUC). "X$-$Y" is AUC(X)$-$AUC(Y); "X$|$Y" is AUC(Y+X)$-$AUC(Y) under early fusion; "X$\oplus$Y" is the same under late fusion (stacked out-of-fold logits). $P_{0,h}$: per-head MST total. Defl.: sink-deflated 0D+1D PH; NT: all non-topological features. $^{\equiv}$: equivalent within ± 0.015 (90% interval inside the margin); *: 95% interval excludes 0; $^{*\equiv}$: both, a real effect smaller than the margin. Intervals and TOST power for every cell are in Appendix M.

Setting	Sink$-$0D	0D$\|$Sink	$P_{0,h}$ $\|$Sink$_h$	Defl.$\|$Sink	1D$|$0D	0D$\|$NT	0D$\oplus$NT	Defl.$\|$NT
TQA Qw3B	$+0.001^{\equiv}$	$+0.003^{\equiv}$	$+0.009^{*\equiv}$	$+0.029^{*}$	$+0.033^{*}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$
TQA Qw1.5B	$-0.003^{\equiv}$	$+0.012^{*}$	$+0.011^{*\equiv}$	$+0.034^{*}$	$-0.008^{*\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA Phi-3	$+0.005^{\equiv}$	$+0.006^{\equiv}$	$0.000^{\equiv}$	$+0.033^{*}$	$-0.003^{\equiv}$	$0.000^{\equiv}$	$0.000^{*\equiv}$	$+0.001^{\equiv}$
TQA TinyLl.	-0.010	$+0.020^{*}$	$+0.003^{\equiv}$	$+0.045^{*}$	$+0.007^{\equiv}$	$0.000^{\equiv}$	$0.000^{*\equiv}$	$-0.001^{\equiv}$
TQA SmolLM	$+0.028^{*}$	$+0.030^{*}$	$+0.106^{*}$	$+0.041^{*}$	$+0.005^{\equiv}$	$0.000^{\equiv}$	$0.000^{*\equiv}$	$-0.001^{\equiv}$
TQA Mis7B	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$	$+0.055^{*}$	$0.000^{\equiv}$	$0.000^{*\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA Mis7B (chat)	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.001^{\equiv}$	$+0.067^{*}$	$0.000^{\equiv}$	$0.000^{*\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TQA Qw7B	$+0.070^{*}$	$+0.027^{*}$	$+0.060^{*}$	$+0.038^{*}$	$+0.021^{*}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
HE Qw3B	$+0.018^{*}$	$+0.028^{*}$	$+0.004^{*\equiv}$	$+0.058^{*}$	$+0.015^{*}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
HE Qw1.5B	$+0.031^{*}$	$+0.012^{*}$	$+0.008^{*\equiv}$	$+0.052^{*}$	$+0.026^{*}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TrQA Qw3B	-0.015^{*}	$+0.021^{*}$	$+0.004^{*\equiv}$	$+0.002$	$-0.003^{\equiv}$	$+0.001^{\equiv}$	$+0.002^{\equiv}$	$0.000^{\equiv}$
TrQA Qw1.5B	-0.009	$+0.010$	$+0.004^{*\equiv}$	$+0.012$	$-0.002^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$-0.001^{\equiv}$
TrQA Phi-3	$-0.004^{\equiv}$	$+0.006^{\equiv}$	$+0.002^{\equiv}$	$+0.029^{*}$	$-0.005^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$+0.002^{*\equiv}$
TrQA TinyLl.	$+0.001^{\equiv}$	$+0.004^{\equiv}$	$-0.001^{\equiv}$	$+0.014$	$-0.002^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$
TrQA Mis7B	$-0.003^{\equiv}$	$+0.002^{\equiv}$	$+0.002^{\equiv}$	$+0.043^{*}$	-0.007	$+0.001^{\equiv}$	$+0.003^{\equiv}$	$0.000^{\equiv}$
TrQA Qw7B	$+0.046^{*}$	$+0.022^{*}$	$+0.031^{*}$	$+0.020^{*}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$	$0.000^{\equiv}$