CHAIN: Complementary Signed Graph Propagation for Uncertainty Quantification in Large Language Models

Hongxing Pan[1] **Ziwei Li**[2] **Ziyue Qiao**[3*] **Xiao Luo**[4]

[1]The Chinese University of Hong Kong, Shenzhen
[2]King Abdullah University of Science and Technology
[3]School of Computing and Information Technology,
Great Bay University, Dongguan, 523000, China
[4]University of Wisconsin–Madison

Abstract

This paper studies the problem of uncertainty quantification in large language models (LLMs), which is critical for reliable LLM deployment under the risk of hallucination. Existing approaches typically treat the response sets as unordered collections and rely on local pairwise similarity or clustering, which would overlook relational signals among responses as well as their corresponding high-order structural semantics. Towards this end, we propose a novel approach named Complementary Signed Graph Propagation (CHAIN) for uncertainty quantification in LLMs. The core idea of our proposed CHAIN is to capture complementary signals among responses using a signed graph and then extract high-order structural information using random propagation for uncertainty quantification. In particular, our proposed CHAIN builds a semantic graph where each response is considered as a node and edges come from complementary views, i.e., *enrichment relation* and *contradiction relation*. To acquire high-order structural signals, we update hidden states through both graph-based information propagation along the edges and channel fusion guided by randomized matrices. After propagation, we summarize these node states into graph-level scores for reliable uncertainty quantification. Extensive experiments on five benchmark datasets validate the effectiveness of the proposed CHAIN in comparison with competing baselines. The code is available at https://github.com/hongxing-pan/CHAIN-NeurIPS2026.

1 Introduction

As large language models (LLMs) are increasingly deployed in real-world systems where reliability matters [47, 58, 68], ensuring the trustworthiness of their outputs has become a central concern. A persistent failure mode is hallucination [20]: LLMs may produce outputs that are fluent and plausible yet factually incorrect, unsupported, or semantically inconsistent with the query [9, 20, 40]. This challenge is especially important in black-box, API-only settings, where users cannot access internal logits, hidden states, gradients, or model parameters [65]. As a result, uncertainty quantification (UQ) based on response sampling has become a practical paradigm for black-box hallucination detection [5, 67, 26, 30, 33, 44]. The basic intuition is that a confident model should generate semantically stable answers under repeated sampling, whereas an uncertain or hallucinating model is more likely to produce divergent or conflicting responses.

Existing black-box LLM Uncertainty quantification methods largely operationalize this intuition by measuring the variability of sampled responses. A large body of work compares generations

*Corresponding author: ziyuejoe@gmail.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

through semantic similarity [44, 4], entropy-based uncertainty scores [9, 46, 26, 24], semantic clustering [9], or pairwise Natural Language Inference (NLI) relations [44]. These methods provide effective tools for detecting uncertainty without requiring model internals. For example, clustering-based approaches estimate uncertainty by grouping responses into semantic equivalence classes, entropy-based approaches quantify the dispersion of answer meanings, and NLI-based approaches use entailment signals to assess whether sampled generations agree. Despite their algorithmic differences, these methods share a common assumption: uncertainty can be summarized by aggregating local, pairwise statistics over the sampled response set.

Despite their success, we identify several key limitations of existing sampling-based UQ methods as follows. (1) In much of the literature, dispersion is often taken as a proxy for disagreement, estimated through semantic similarity, entropy, clustering, and related measures [25, 9, 46, 44, 63]. However, this overlooks the directed relational structure among responses and can make the method rely on heuristic features such as lexical overlap, subsequence matching, and constituent matching[41, 35]. While some recent studies have begun to utilize directional entailment graphs to model these interactions [6], they only focus on entailment probabilities. (2) Existing methods often rely on local pairwise evidence, aggregating local similarity or NLI scores to obtain response-level uncertainty, which overlooks higher-order relationships such as support-of-support reinforcing agreement, support-of-contradiction inducing conflict, and contradiction-of-contradiction revealing indirect consistency.[7] (3) Hallucination evidence is not always reflected by holistic response-set dispersion but exposed by specific cross-response contradictions[39]. Traditional mean pooling, entropy-based summaries, or global dispersion scores can dilute these key contradictions and underestimate hallucination risk. What is missing is a mechanism for reasoning over higher-order relations among responses, so that local disagreements can be aggregated into a global inconsistency pattern rather than reduced to unordered clusters or independent pairwise scores.

To address this limitation, we propose a novel framework named Complementary Signed Graph Propagation (`CHAIN`) for black-box uncertainty quantification that models sampled responses as a signed relational structure rather than an unordered set of pairwise comparisons. `CHAIN` represents each sampled response as a node and constructs edges from two complementary semantic views: an *enrichment relation* and a *contradiction relation*. The enrichment relation captures how one response semantically supports, specifies, or extends another, while the contradiction relation captures semantic conflict among generations. On top of this complementary signed graph, `CHAIN` performs graph-based propagation to aggregate relational evidence beyond local pairwise judgments. A lightweight propagation scheme, augmented by randomized feature mixing and channel fusion, aggregates the complementary signals into a graph-level uncertainty score. In this way, `CHAIN` exposes responses that may appear locally plausible but are structurally inconsistent with the dominant semantic pattern in the sampled response set.

Our main contributions are summarized as follows:

❶ ***New Perspective:*** We reformulate black-box LLM uncertainty quantification as a structured inconsistency pattern over a signed response graph. Unlike prior methods that treat samples as unordered sets or reduce uncertainty to local pairwise agreement, our perspective captures the higher-order structural tension between semantic enrichment and contradiction.

❷ ***Novel Methodology with Theoretical Analysis:*** We propose `CHAIN`, a training-free framework that aggregates local NLI signals into higher-order evidence via complementary signed graph propagation. Crucially, we provide a theoretical analysis showing that `CHAIN` achieves strict structural separation and contradiction amplification. These properties provide a mathematical foundation for identifying hallucinations not captured by conventional pairwise-symmetric estimators.

❸ ***Empirical Validation:*** Extensive experiments across five benchmarks and four LLMs demonstrate that `CHAIN` consistently outperforms state-of-the-art black-box baselines. Ablation studies and sensitivity analyses further confirm that signed propagation and randomized feature mixing are essential inductive biases for surfacing latent hallucinations in a model-agnostic, API-only setting.

2 Related Work

Hallucination Detection in LLMs. Hallucination detection methods for LLMs can be broadly grouped into external verification, white-box probing, and black-box sampling approaches [4, 23,

20, 48]. In API-only settings, where logits and hidden states are unavailable, black-box sampling methods estimate uncertainty from multiple generated responses [65]. A representative line of work is Semantic Entropy (SE) [9] and its discrete variants [9, 57], which use Natural Language Inference (NLI) [37] to group semantically equivalent responses and compute uncertainty over the induced clusters. QTN-UQ combines semantic clustering with token-probability information, whereas CHAIN uses sampled response texts; an empirical comparison is given in Table 6. Other methods rely on lexical or semantic similarity, response graphs, or spectral statistics to measure dispersion among samples [59]. White-box methods such as EigenScore [4] and Kernel Language Entropy (KLE) [46] further exploit internal representations or kernel structures when model access is available. While effective, many of these methods reduce pairwise relations to equivalence, similarity, or dispersion, and therefore do not explicitly preserve whether two responses support or contradict each other. CHAIN targets this missing signed structure in black-box response sets.

Uncertainty Quantification. Uncertainty quantification provides a reliability signal for evaluating and calibrating generated responses [55, 33, 1, 59, 29, 8, 12, 36]. In sampling-based UQ, uncertainty is commonly estimated by aggregating multiple candidate outputs and measuring their lexical, semantic, or consistency-based variation. Recent methods move beyond hard clustering: KLE [46] models uncertainty through continuous semantic kernels, while SNNE [44] uses graph-based neighborhood entropy to capture finer response-set structure. These approaches improve over discrete equivalence classes, but their pairwise relations are still typically treated as unsigned affinities. As a result, factual opposition and harmless semantic variation may contribute similarly to the final uncertainty score. CHAIN instead keeps agreement-like and contradiction-like relations separate, enabling uncertainty to be driven by structured semantic conflict rather than only by response dispersion.

Training-free Graph Propagation. Graph neural networks provide a natural tool for modeling higher-order relations [7, 15], but supervised GNNs require task-specific labels and may suffer from over-smoothing or rank collapse [28, 49]. These limitations are undesirable in plug-and-play black-box UQ, where the detector should operate without additional training. Recent work on untrained graph neural networks and random-weight architectures suggests that useful performance can arise from graph topology, fixed filters, and random feature mixing rather than learned task-specific transformations [64, 62, 53, 43, 54, 3, 17, 52, 34, 61]. Motivated by this view, our proposed CHAIN uses NLI-derived signed edges as fixed relational filters and frozen random projections as feature mixers. This yields a training-free signed propagation mechanism that amplifies contradiction structure without requiring supervised graph optimization.

3 The Proposed CHAIN

Problem Definition. Given a prompt q, we sample a response set $\mathcal{S} = \{s_1, \ldots, s_N\}$ from a black-box LLM without accessing logits, hidden states, or model parameters. Our goal is to compute a prompt-level uncertainty score $U(q) = f(\mathcal{S})$, where larger values indicate higher hallucination risk. Since uncertainty must be inferred only from sampled responses, a reliable estimator should identify incompatible factual commitments even when the responses are lexically or semantically similar.

3.1 Framework Overview

The CHAIN pipeline models sampled responses as a signed relational object to capture high-order uncertainty signals. As shown in Figure 1, CHAIN treats sampled responses as a signed relational object rather than an unordered set. It first samples multiple responses from the black-box LLM, constructs a complementary signed graph using NLI-derived entailment and contradiction relations, propagates enrichment and contradiction signals through a training-free signed propagation module, and finally summarizes localized contradiction evidence with a polarized Top-K readout. Unlike sign-agnostic uncertainty estimators, CHAIN preserves the sign of response relations: positive enrichment edges represent semantic support or compatibility, while negative contradiction edges represent factual opposition. This separation is crucial because hallucinated response sets can appear surface-consistent while remaining internally contradictory. The following subsections define how the signed graph is constructed, how structural signals are propagated, and how contradiction evidence is summarized into $U(q)$.

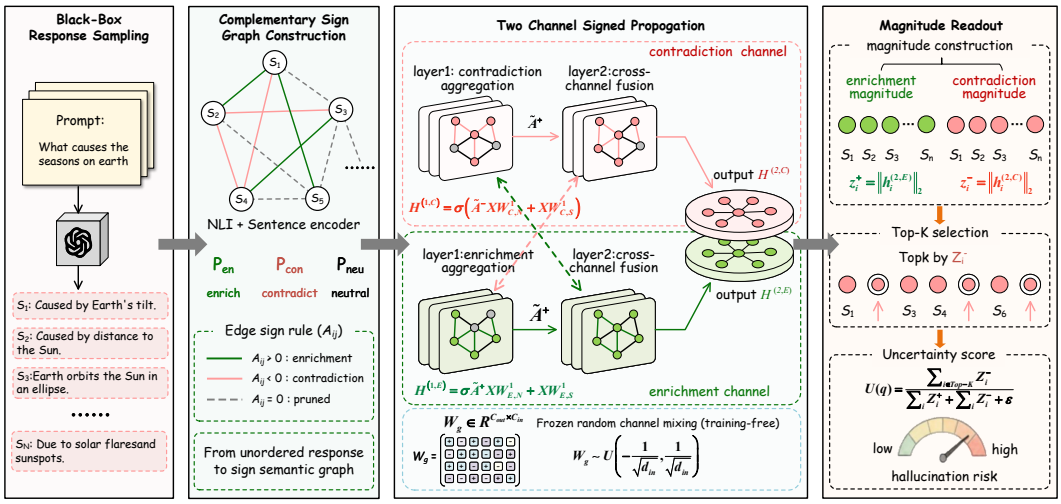

Figure 1: Overview of the CHAIN framework. Given a black-box LLM, CHAIN samples multiple responses, constructs a signed semantic graph over their enrichment and contradiction relations, amplifies structural evidence through two-layer signed propagation, and finally extracts localized contradiction signals via Top-K pooling.

3.2 Complementary Signed Graph Construction

A sampled response set contains more information than unordered semantic dispersion, but existing similarity- or clustering-based estimators often collapse different types of disagreement into the same signal. For example, two responses may differ because one provides additional details, or because they make mutually incompatible factual commitments. Treating both cases as generic low similarity loses the distinction between benign variation and hallucination-relevant contradiction. To preserve this distinction, we construct a complementary signed graph in which positive edges represent enrichment or semantic compatibility, while negative edges represent factual conflict. This graph provides the relational substrate on which later propagation can aggregate support and contradiction as separate but interacting signals.

Given the response set $\mathcal{S} = \{s_1, \ldots, s_N\}$, we encode each response s_i into a normalized sentence embedding $x_i \in \mathbb{R}^{d_{\text{in}}}$ and stack all embeddings as $X = [x_1; \ldots; x_N] \in \mathbb{R}^{N \times d_{\text{in}}}$, with $h_i^{(0)} = x_i$ as the initial node feature. To determine pairwise semantic relations, we apply a pretrained NLI model to each ordered response pair (s_i, s_j), obtaining entailment, contradiction, and neutral probabilities $p_e(i \to j)$, $p_c(i \to j)$, and $p_n(i \to j)$. Since NLI predictions can be directional, we symmetrize each relation score:

$$p_k^{\text{sym}}(i,j) = \frac{1}{2}\left(p_k(i \to j) + p_k(j \to i)\right), \qquad k \in \{e, c, n\}. \tag{1}$$

Based on these symmetric scores, we define a signed adjacency matrix $\mathcal{A} \in \mathbb{R}^{N \times N}$ by retaining only high-confidence enrichment and contradiction relations:

$$\mathcal{A}_{ij} = \begin{cases} p_e^{\text{sym}}(i,j), & p_e^{\text{sym}}(i,j) \geq \tau_{\text{ent}} \ \wedge \ p_e^{\text{sym}}(i,j) \geq \max\{p_c^{\text{sym}}(i,j), p_n^{\text{sym}}(i,j)\}, \\ -p_c^{\text{sym}}(i,j), & p_c^{\text{sym}}(i,j) \geq \tau_{\text{con}} \ \wedge \ p_c^{\text{sym}}(i,j) > p_e^{\text{sym}}(i,j), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here, τ_{ent} and τ_{con} are fixed thresholds. Positive entries in $\mathcal{A}$ correspond to enrichment edges, while negative entries correspond to contradiction edges. The strict inequality in the contradiction condition prevents the same pair from being assigned both relation types, and pairs that satisfy neither condition are pruned as ambiguous or low-confidence relations.

For propagation, we separate the signed adjacency matrix into two nonnegative components:

$$\mathbf{A}_{ij}^{+} = \max(\mathcal{A}_{ij}, 0), \qquad \mathbf{A}_{ij}^{-} = \max(-\mathcal{A}_{ij}, 0). \tag{3}$$

Thus, $\mathbf{A}^+$ stores enrichment weights and $\mathbf{A}^-$ stores contradiction weights. We denote their row-normalized forms by $\widetilde{\mathbf{A}}^+$ and $\widetilde{\mathbf{A}}^-$, with empty rows treated as zero aggregation. This decomposition preserves supportive and contradictory relations as separate graph structures, allowing the propagation module to process them through distinct but interacting channels.

3.3 Training-Free Signed Propagation

The signed graph constructed above records local pairwise relations [32], but local NLI edges alone are insufficient for reliable uncertainty estimation. First, individual NLI predictions can be noisy, as prior studies show that NLI models may rely on shallow heuristics or fail on simple controlled inference cases [41, 35, 45]. Therefore, a detector should exploit the response-set structure around each node rather than rely only on isolated pairwise judgments. Second, hallucination may appear as a higher-order pattern: a response can be supported by one group of samples while contradicting another, or a contradiction can become salient only after considering neighboring support relations. This motivates propagating enrichment and contradiction evidence through separate signed channels, following signed graph representation learning and signed message passing [7, 15]. The design converts local NLI scores into higher-order structural signals while remaining training-free, which is important in black-box settings where task-specific hallucination labels may be unavailable [65, 53, 18].

The propagation module maintains two channels: an enrichment channel H^E, which tracks support-like evidence, and a contradiction channel H^C, which tracks conflict-oriented evidence. Although the implementation uses two propagation steps, we present them as one coherent signed propagation process. The first step converts raw embeddings into relation-specific structural features, and the second fuses these features across signed neighborhoods to capture higher-order support–conflict patterns. At initialization, both channels start from the same response embeddings:

$$H^{(0,E)} = H^{(0,C)} = X.$$

The first propagation step separately aggregates enrichment and contradiction evidence from one-hop signed neighborhoods:

$$\begin{aligned} H^{(1,E)} &= \sigma\Big(\widetilde{\mathbf{A}}^+ H^{(0)} W^{(1)}_{E,\mathcal{N}} + H^{(0)} W^{(1)}_{E,\mathcal{S}}\Big), \\ H^{(1,C)} &= \sigma\Big(\widetilde{\mathbf{A}}^- H^{(0)} W^{(1)}_{C,\mathcal{N}} + H^{(0)} W^{(1)}_{C,\mathcal{S}}\Big). \end{aligned} \tag{4}$$

Here $H^{(0)} = X$, $\sigma(\cdot)$ is a nonlinear activation, and the W's are frozen feature mixing matrices, following randomized or untrained graph transformations in prior work [17, 19, 2]. The enrichment channel aggregates messages from compatible responses, while the contradiction channel aggregates messages from conflicting responses. The self terms preserve each node's own response embedding during the first structural transformation. The second propagation step performs cross-channel signed fusion:

$$\begin{aligned} H^{(2,E)} &= \sigma\Big(\widetilde{\mathbf{A}}^+ H^{(1,E)} W^{(2)}_{E,+} + \widetilde{\mathbf{A}}^- H^{(1,C)} W^{(2)}_{E,-} + H^{(1,E)} W^{(2)}_{E,0}\Big), \\ H^{(2,C)} &= \sigma\Big(\widetilde{\mathbf{A}}^+ H^{(1,C)} W^{(2)}_{C,+} + \widetilde{\mathbf{A}}^- H^{(1,E)} W^{(2)}_{C,-} + H^{(1,C)} W^{(2)}_{C,0}\Big). \end{aligned} \tag{5}$$

In this step, enrichment edges refine support-oriented evidence within the same channel, while contradiction edges transfer evidence across complementary channels. This cross-channel interaction allows a response to be interpreted through both its direct pairwise relations and the signed structural context around its neighbors.

Frozen Randomized Feature Mixing. To keep `CHAIN` training-free, all propagation weights are randomly initialized once and frozen during inference [11]. For each feature mixing matrix $W \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, we sample

$$W_{ab} \sim \mathcal{U}\left(-\frac{1}{\sqrt{d_{\text{in}}}}, \frac{1}{\sqrt{d_{\text{in}}}}\right). \tag{6}$$

This scale corresponds to the standard Kaiming-uniform initialization of linear layers [13]. The frozen projections act as lightweight channel-wise feature mixers. They do not learn dataset-specific supervision; instead, they expose structural patterns already present in the NLI-derived signed graph. This design preserves the black-box nature of the method while allowing the signed graph to produce nontrivial structural activations.

3.4 Polarized Top-K Readout

After signed propagation, we convert node-level structural evidence into a graph-level uncertainty score [34]. A global mean readout is natural, but it is poorly matched to hallucination detection because contradiction evidence is often sparse. In many sampled response sets, most generations follow the same surface template, while only a few make incompatible factual commitments. Averaging over all nodes can dilute these localized conflicts and make the response set appear more reliable than it is. We therefore use a polarized Top-K readout to emphasize the strongest contradiction-channel activations while normalizing them against graph-wide enrichment evidence.

For each response node i, we summarize the propagated representations by two channel-wise magnitudes:

$$z_i^+ = \|h_i^{(2,E)}\|_2, \qquad z_i^- = \|h_i^{(2,C)}\|_2. \tag{7}$$

Here z_i^+ measures participation in support-like structure, while z_i^- measures involvement in contradiction-oriented structure. Rather than averaging contradiction magnitudes over all nodes, `CHAIN` selects the K largest contradiction magnitudes and computes

$$U(q) = \frac{\sum_{i \in \text{Top-}K(\mathbf{z}^-)} z_i^-}{\sum_{j \in V} z_j^+ + \sum_{i \in \text{Top-}K(\mathbf{z}^-)} z_i^- + \epsilon}. \tag{8}$$

The numerator captures localized contradiction evidence, while the denominator normalizes it by the graph-wide enrichment mass and selected contradiction mass [61]. This polarized readout contrasts concentrated contradiction against graph-wide enrichment instead of treating all node magnitudes uniformly. The benefit of this design is supported by our sparse-contradiction analysis in Proposition 3.5 and empirically validated by the readout sensitivity results in Figure 2.

3.5 Theoretical Analysis

We justify `CHAIN` through a linearized surrogate of its signed propagation module. The analysis focuses on three design choices: signed response relations, contradiction-channel propagation, and Top-K readout. All formal definitions and proofs are deferred to Appendix C. Given N sampled responses, `CHAIN` constructs a signed response graph $G = (V, E^+, E^-, X)$, where E^+ denotes enrichment relations and E^- denotes contradiction relations. The encoder maintains an enrichment channel H^E and a contradiction channel H^C. For each node i, we use the contradiction-channel norm

$$z_i^- := \|H_i^C\|_2$$

as its node-level contradiction evidence. A larger z_i^- indicates that the response receives stronger structural conflict signals from the signed response graph.

Definition 3.1 (Sign-agnostic estimator)**.** *As formalized in Appendix C.4 (Definition C.6), an uncertainty estimator $\widehat{U}$ is sign-agnostic if it depends only on response embeddings or unsigned pairwise affinities, without preserving the signed decomposition (E^+, E^-).*

Theorem 3.2 (Signed separation)**.** *There exist two signed response graphs G_1 and G_2 with identical response embeddings and identical unsigned affinity structure, but different contradiction-edge sets, such that every sign-agnostic estimator $\widehat{U}$ and the linearized `CHAIN` contradiction channel satisfy*

$$\widehat{U}(G_1) = \widehat{U}(G_2), \qquad \Pr\big[z_i^-(G_1) \neq z_i^-(G_2)\big] = 1 \quad \textit{for some node } i. \tag{9}$$

Theorem 3.2 shows that contradiction is not reducible to unsigned semantic dispersion. `CHAIN` can distinguish response sets that have the same embedding geometry but differ in whether cross-response relations are neutral or contradictory. The explicit construction for this separation is provided in Theorem C.7.

Proposition 3.3 (Signed two-hop propagation)**.** *In the linearized two-layer `CHAIN` encoder, the propagated representations contain the four signed two-hop operators*

$$\mathcal{P}_2 = \left\{ \widetilde{\mathbf{A}}^+\widetilde{\mathbf{A}}^+, \widetilde{\mathbf{A}}^+\widetilde{\mathbf{A}}^-, \widetilde{\mathbf{A}}^-\widetilde{\mathbf{A}}^+, \widetilde{\mathbf{A}}^-\widetilde{\mathbf{A}}^- \right\}, \tag{10}$$

up to residual one-hop and self terms.

Thus, CHAIN captures not only direct contradiction edges, but also higher-order support–conflict patterns among response clusters. The full decomposition and derivation are given in Proposition C.5.

Theorem 3.4 (Contradiction amplification)**.** *Consider a conflicted node i with nonzero aggregate contradiction signal $\alpha_i := \|\widetilde{A}_i^- X\|_2 > 0$, and an unconflicted node j. Under the linearized first-layer contradiction channel with frozen Kaiming-uniform weights, the failure probability satisfies:*

$$\Pr(z_i^- \le z_j^-) \le \exp\left(-\frac{d_{\text{out}}}{18 d_{\text{in}}^2}\right) + 2\exp\left(-\frac{\alpha_i^2 d_{\text{out}}}{1152 d_{\text{in}}^2}\right) + 2\exp\left(-\frac{\alpha_i^4 d_{\text{out}}}{1152 d_{\text{in}}^2}\right). \tag{11}$$

As established in Theorem 3.4, for a fixed $\alpha_i > 0$, the failure probability decays exponentially with respect to the output dimension d_{out}. This result justifies using $\|H_i^C\|_2$ as robust contradiction evidence, ensuring that nodes receiving structural conflict signals become distinguishable in the negative channel with high probability as the random feature width increases. The complete proof is provided in Appendix C.5.

Proposition 3.5 (Top-K readout under sparse contradiction)**.** *Suppose exactly K^* nodes have high contradiction magnitude z_h, while the remaining nodes have low magnitude z_l, with $z_h > z_l > 0$. Among normalized nonnegative linear readouts, assigning all weight to high-contradiction nodes maximizes the hallucination-to-safe contrast. In particular, any Top-K readout with $K \le K^*$ and the mean readout satisfy the following:*

$$\rho_{\text{Top-}K} = \frac{z_h}{z_l}, \qquad \rho_{\text{mean}} = 1 + \frac{K^*}{N}\left(\frac{z_h}{z_l} - 1\right) < \frac{z_h}{z_l}, \quad K^* < N. \tag{12}$$

Proposition 3.5 explains why CHAIN reads out the largest contradiction magnitudes: when contradiction evidence is sparse, mean aggregation dilutes the signal. The formal proof is given in Proposition C.9. Together, these results show that CHAIN preserves signed contradiction information, propagates higher-order support–conflict structure, amplifies negative-channel evidence, and uses Top-K readout to avoid dilution. The analysis is limited to linearized surrogates and does not fully characterize the nonlinear two-layer encoder used in experiments.

The contradiction-channel ℓ_2 norm is invariant to orthogonal changes of feature coordinates and scales linearly with the representation. The amplification result concerns node-level ordering under its stated assumptions; it does not by itself guarantee ordering of the final normalized graph score. A graph-level separation additionally requires a sufficient margin between the conflicting and unconflicted nodes, and the enrichment-channel normalization must also be taken into account. The independently initialized, frozen random mixers transform the self, enrichment, and contradiction messages into different feature coordinates without training additional parameters. Kaiming-uniform initialization is used to control the scale of these transformations; the ablations assess their empirical contribution.

4 Experiment

4.1 Experimental Setup

Baselines. We compare CHAIN against representative black-box uncertainty estimators, including Predictive Entropy (PE) [38], LexSim [10], and Discrete Semantic Entropy (DSE) [25, 9], as well as graph-based and kernel-based estimators such as EigenScore [4], SNNE [44], and KLE [46].

Datasets. Our evaluation spans five question-answering benchmarks covering different reasoning and retrieval settings. We use SQuAD [50] and CoQA [51] for reading comprehension, NQ-Open [27] and TriviaQA [22] for factual retrieval, and HotpotQA [66] for multi-hop question answering.

Evaluation Target. Given a designated response and alternative generations, CHAIN uses the signed response graph as contextual evidence for the designated response's reliability. SentSim and ROUGE-L are reference-based proxies for that response, not factuality labels for the entire graph. A correct response may coexist with conflicting alternatives, while a consistent set may still be incorrect. We report fixed-direction AUROC without post-hoc score flipping. **Implementation Details.** We evaluate CHAIN on four target generators: OPT-6.7B [69], Qwen3-8B-Instruct[2], Mistral-7B-Instruct [21], and Phi-3.5-mini [42]. We sample $N{=}10$ responses per prompt at temperature

[2] https://huggingface.co/Qwen/Qwen3-8B

Table 1: AUROC on open and closed source LLMs. Open-source results are averaged over four target LLMs. GPT, Gemini, and Opus denote `gpt-4o-mini`, `gemini-2.5-flash-lite`, and `opus-4.8`, respectively. AUC_s and AUC_r denote AUROC under the SentSim- and ROUGE-L-based labels, respectively. Blue cells indicate the best result in each column.

LLMs	Method	CoQA		SQuAD		NQ-Open		TriviaQA		HotpotQA	
		AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r
Open-source	PE	0.660	0.623	0.586	0.583	0.708	0.696	0.768	0.767	0.609	0.578
	EigenScore	0.700	0.678	0.708	0.686	0.668	0.671	0.709	0.692	0.645	0.622
	LexSim	0.734	0.767	0.680	0.682	0.732	0.749	0.772	0.778	0.710	0.707
	DSE	0.782	0.759	0.727	0.723	0.755	0.749	0.767	0.785	0.757	0.723
	SNNE	0.806	0.779	0.827	0.821	0.787	0.743	0.772	0.782	0.802	0.767
	KLE	0.753	0.746	0.719	0.718	0.750	0.739	0.810	0.786	0.721	0.716
	`CHAIN`	**0.830**	**0.814**	**0.836**	**0.829**	**0.828**	**0.773**	**0.817**	**0.792**	**0.848**	**0.799**
LLMs	**Method**	**GPT–NQ**		**GPT–HotpotQA**		**Gemini–Trivia**		**Gemini–SQuAD**		**Opus–CoQA**	
		AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r
Closed-source	LexSim	0.727	0.713	0.546	0.728	0.710	0.637	0.785	0.677	0.779	0.608
	DSE	0.762	**0.734**	0.719	0.690	0.796	0.828	0.787	0.709	0.708	0.825
	SNNE	0.762	0.685	0.741	0.669	**0.811**	0.711	0.754	0.695	**0.862**	0.807
	`CHAIN`	**0.767**	0.711	**0.746**	**0.739**	0.789	**0.831**	**0.853**	**0.735**	0.860	**0.862**

T=1, with top-p=1.0 and maximum generation length 64. For each prompt, `CHAIN` builds a signed response graph using cross-encoder DeBERTa-v3-base NLI [14] to compute pairwise entailment and contradiction probabilities, with thresholds τ_{ent}=0.35 and τ_{con}=0.80. Node features are initialized with sentence embeddings, and the final score is computed using a polarized Top-K readout with K=3. All stochastic components are averaged over three random seeds.

Metrics. Following prior work[25, 9, 5, 4, 46, 44, 30], we evaluate all uncertainty estimators by their ability to distinguish correct from hallucinated responses using AUROC. We consider two hallucination labels based on SentSim [56] and ROUGE-L [31], and report the corresponding scores as AUC_s and AUC_r, respectively. Additional implementation details, per-model results, and full ablations are provided in Appendix A.

4.2 Main Results

Performance on Open-Source LLMs. Table 1 reports mean fixed-direction AUROC over four target LLMs. The results show that (1) `CHAIN` is the top method in every averaged dataset–metric column (five datasets $\times$ two label oracles), indicating that its performance is not tied to a single dataset or evaluation oracle; (2) the gains of `CHAIN` are larger on open-domain and multi-hop benchmarks, suggesting that signed graph modeling is particularly useful when sampled responses contain subtle factual conflicts or unstable evidence chains; (3) `CHAIN` benefits more under the SentSim-based label than the ROUGE-L-based label, which is consistent with its use of signed semantic relations among responses; and (4) `CHAIN` remains effective in settings where strong baselines already achieve high AUROC. On SQuAD and TriviaQA, `CHAIN` obtains the best scores under both labels despite strong performance from competitors such as SNNE and KLE.

Performance on Closed-Source LLMs. The lower part of Table 1 reports the results on closed-source LLMs. The results show that (1) `CHAIN` achieves the best performance in most evaluated settings and remains competitive in the others, indicating that its effectiveness extends beyond open-source models; (2) `CHAIN` performs consistently across different model families and benchmarks, suggesting that its use of complementary response relations is not restricted to a particular generator.

4.3 Ablation Studies

To evaluate each component in `CHAIN`, we introduce six variants: (1) `CHAIN` w/o Mixer, replacing frozen randomized feature mixing with an identity mapping; (2) `CHAIN` w/o E^+, removing enrichment edges; (3) `CHAIN` w/o E^-, removing contradiction edges; (4) `CHAIN` w/o Prop., removing signed propagation layers; (5) `CHAIN` w/o embedding L2 normalization, disabling ℓ_2 row normalization of sentence embeddings; and (6) `CHAIN` w/o cross-channel fusion, disabling cross-channel neighbor mixing in the second signed convolution layer. We compare these variants with the full model across

Table 2: Ablation studies of our proposed CHAIN. AUROC is macro-averaged over five benchmarks. Our full model achieves the best performance on both metrics.

Metric	CHAIN	CHAIN w/o Mixer	CHAIN w/o E^+	CHAIN w/o E^-	CHAIN w/o Prop.	CHAIN w/o L2	CHAIN w/o Fusion
SentSim	**0.832**	0.618	0.822	0.817	0.758	0.638	0.643
ROUGE-L	**0.801**	0.580	0.795	0.799	0.737	0.590	0.591

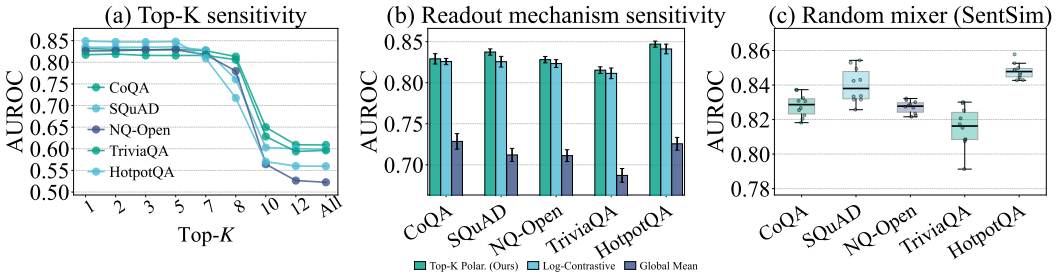

Figure 2: (a) Impact of K on detection performance; (b) Comparison of different readout mechanisms; (c) Performance stability across multiple random mixer initializations.

all benchmarks in Table 2. The results show that (i) removing randomized feature mixing or signed propagation causes the largest drops, highlighting the importance of structural projection and higher-order message passing; (ii) the full model outperforms single-relation variants on SentSim, supporting complementary signed relations; (iii) ROUGE-L is more tolerant to removing contradiction edges, but the full model remains the most stable across oracles; and (iv) removing embedding normalization or cross-channel fusion further reduces AUROC, indicating that both input geometry and cross-channel message passing are important for conflict readouts.

4.4 Furthur Analysis

Sensitivity Analysis. We analyze the influence of K by varying it from $K = 1$ to All ($N = 15$). As shown in Figure 2(a), AUROC remains high and stable when $K \leq 8$, suggesting that hallucination evidence is concentrated in a localized subset of responses. However, performance drops sharply as K approaches N, confirming that aggregating over the full response set dilutes structural conflict signals. Figure 2(b) further shows that the polarized Top-K readout consistently outperforms both log-contrastive and global mean baselines across all five benchmarks. We also evaluate alternative edge thresholds: performance remains close to the default in most tested settings, although a larger decrease occurs for one configuration (Table 4). These results empirically support the use of localized contradiction readout rather than global response-level aggregation. This stability is consistent with the polarized NLI probability distribution in Figure 4.

Robustness Analysis. To test whether CHAIN depends on a particular frozen randomized mixer, we run 10 trials with different random seeds. As shown in Figure 2(c) and (d), AUROC exhibits some randomseed dependent variation, especially on certain dataset and label combinations, but remains stable across the five benchmarks. For most datasets, the inter-seed spread is moderate relative to the gap between CHAIN and weaker ablations, indicating that random feature mixing is not merely a lucky initialization effect. These results indicate that the improvement of our CHAIN mainly come from signed propagation and the induced support and conflict evident, while the frozen random initialization mixer provides a lightweight, easy and robust feature transformation.

4.5 Qualitative Study

Figure 3 shows a representative HotpotQA example where the gold answer is "2003", but Qwen3-8B produces superficial consistent responses with conflicting final commitments, including 2003, 2002, 2002 to 2003, and 2007. Because these responses share a repetitive reasoning template and overlapping draft and year evidence, lexical-overlap and sign-agnostic uncertainty methods fail to flag the case as highly inconsistent. In contrast, CHAIN preserves signed NLI relations among responses: the pairwise relation matrix separates support, contradiction, and neutral links, while the thresholded signed graph exposes localized contradiction edges between incompatible answers. The contradiction

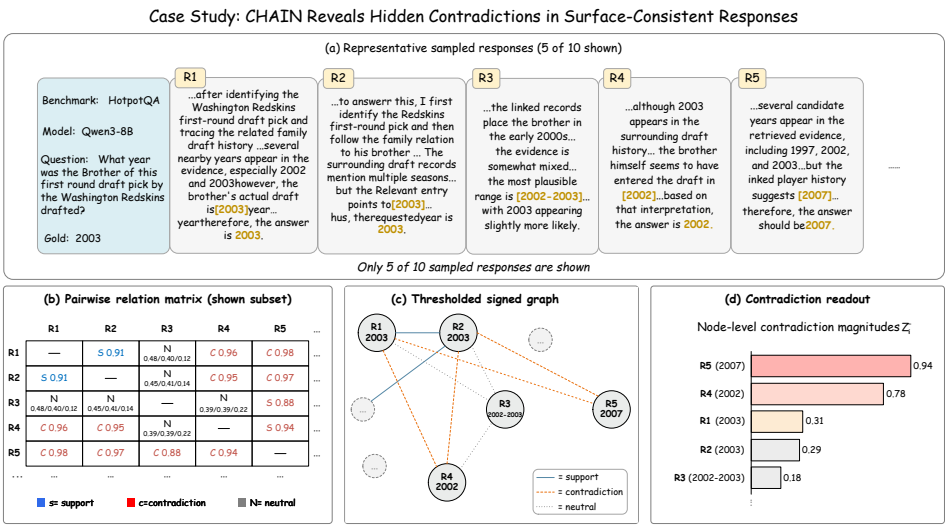

Figure 3: Qualitative case study of hidden contradictions. An example where surface-consistent sampled responses make incompatible answer commitments. CHAIN reveals these hidden contradictions.

readout assigns the largest node-level magnitudes to the most inconsistent responses, especially the 2007 and 2002 answers, demonstrating that CHAIN can detect hallucination-relevant disagreement that is hidden by superficial consistency. It also suggests that the key signal for black-box hallucination detection is not merely whether responses are different, but whether their differences form signed conflict patterns around the final factual commitment.

5 Conclusion

This paper studies black-box hallucination detection from sampled responses and shows that hallucination risk is not always captured by semantic dispersion: responses may appear similar while making incompatible factual commitments. To address this, CHAIN represents responses as a signed graph with enrichment and contradiction edges, propagates signed structural evidence, and reads out localized contradiction signals through a polarized Top-K score. Across five QA benchmarks and four target generators, CHAIN achieves the strongest average AUROC in all averaged dataset–metric settings. Ablations further show that the gains come from the combination of signed topology, feature mixing, graph propagation, and localized Top-K readout; removing feature mixing or propagation causes clear degradation, while global mean pooling dilutes sparse contradiction evidence. This suggests that hallucination may arise as a relational inconsistency among sampled responses, and that preserving signed response relations provides a useful inductive bias for black-box detection.

6 Broader Impacts

Potential Positive Impacts: CHAIN enhances the safety of LLM deployments in risk-sensitive domains without requiring internal model access. By effectively identifying structural inconsistencies, it helps mitigate AI-generated misinformation and fosters public trust in generative systems. **Potential Negative Impacts:** First, our reliance on multi-sampling increases the computational overhead and environmental footprint during inference. Second, it poses a risk of *automation bias*, where users might over-rely on the UQ scores and diminish their own critical oversight.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their constructive comments and suggestions. The work of Ziyue Qiao was supported in part by the National Natural Science Foundation of China (Grant Nos. 62676075 and 62406056) and in part by the Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2025A1515140238 and 2024A1515140114).

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[2] Thu Bui, Carola-Bibiane Schönlieb, Bruno Ribeiro, Beatrice Bevilacqua, and Moshe Eliasof. On the effectiveness of random weights in graph neural networks. *ArXiv*, abs/2502.00190, 2025.

[3] Weipeng Cao, Xizhao Wang, Zhong Ming, and Jinzhu Gao. A review on neural networks with random weights. *Neurocomputing*, 275:278–287, 2018.

[4] 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. *ArXiv*, abs/2402.03744, 2024.

[5] Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, 2024.

[6] Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. LLM uncertainty quantification through directional entailment graph and claim level response augmentation. *CoRR*, abs/2407.00994, 2024.

[7] Tyler Derr, Yao Ma, and Jiliang Tang. Signed graph convolutional networks. *2018 IEEE International Conference on Data Mining (ICDM)*, pages 929–934, 2018.

[8] Boyang Fan, Ziyue Qiao, Wei Ju, and Xiao Luo. Context-anchored asymmetric conformal prediction for uncertainty-aware reasoning over knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 25165–25185, 2026.

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

[10] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:539–555, 2020.

[11] Claudio Gallicchio and Simone Scardapane. Deep randomized neural networks. In *Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*, pages 43–68. Springer, 2020.

[12] Jiawei Gu, Ziyue Qiao, and Xiao Luo. Neubm: Mitigating model bias in graph neural networks through neutral input calibration. *arXiv preprint arXiv:2505.15180*, 2025.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034. IEEE Computer Society, 2015.

[14] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

[15] Yixuan He, Gesine Reinert, Songchao Wang, and Mihai Cucuringu. SSSNET: semi-supervised signed network clustering. In Arindam Banerjee, Zhi-Hua Zhou, Evangelos E. Papalexakis, and Matteo Riondato, editors, *Proceedings of the 2022 SIAM International Conference on Data Mining, SDM 2022, Alexandria, VA, USA, April 28-30, 2022*, pages 244–252. SIAM, 2022.

[16] José M Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014.

[17] Changqin Huang, Ming Li, Feilong Cao, Hamido Fujita, Zhao Li, and Xindong Wu. Are graph convolutional networks with random weights feasible? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2751–2768, 2022.

[18] Tianjin Huang, Tianlong Chen, Meng Fang, Vlado Menkovski, Jiaxu Zhao, Lu Yin, Yulong Pei, Decebal Constantin Mocanu, Zhangyang Wang, Mykola Pechenizkiy, and Shiwei Liu. You can have better graph neural networks by not training weights at all: Finding untrained gnns tickets. In *LOG IN*, 2022.

[19] Tianjin Huang, Tianlong Chen, Meng Fang, Vlado Menkovski, Jiaxu Zhao, Lu Yin, Yulong Pei, Decebal Constantin Mocanu, Zhangyang Wang, Mykola Pechenizkiy, and Shiwei Liu. You can have better graph neural networks by not training weights at all: Finding untrained gnns tickets. In Bastian Rieck and Razvan Pascanu, editors, *Learning on Graphs Conference, LoG 2022, 9-12 December 2022, Virtual Event*, Proceedings of Machine Learning Research, page 8. PMLR, 2022.

[20] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023.

[21] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

[22] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics, 2017.

[23] Hazel Kim, Tom A Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting llm hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32298–32310, 2025.

[24] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *ArXiv*, abs/2406.15927, 2024.

[25] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[26] Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, Ken Fukuda, and Masahiro Hamasaki. Evidential semantic entropy for llm uncertainty quantification. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7107–7122, 2026.

[27] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019.

[28] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3538–3545. AAAI Press, 2018.

[29] Zhaoye Li, Huan Chen, Huibin Tan, Long Lan, Yize Sui, and Jing Ren. Uncertainty quantification for black-box llms via star graphs connectivity: Exploring alternatives for semantic density. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 273–289. Springer, 2025.

[30] Zhaoye Li, Siyuan Shen, Wenjing Yang, Ruochun Jin, Huan Chen, Ligong Cao, and Jing Ren. Enhancing uncertainty quantification in large language models through semantic graph density. In *Conference on Uncertainty in Artificial Intelligence*, 2025.

[31] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[32] Haoxin Liu, Ziwei Zhang, Peng Cui, Yafeng Zhang, Qiang Cui, Jiashuo Liu, and Wenwu Zhu. Signed graph neural network with latent groups. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1066–1075, 2021.

[33] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen-Yu Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025.

[34] Yixin Liu, Kaize Ding, Qinghua Lu, Fuyi Li, Leo Yu Zhang, and Shirui Pan. Towards self-interpretable graph-level anomaly detection, 2023.

[35] Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. Simple but challenging: Natural language inference models fail on simple sentences. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Findings of ACL, pages 3449–3462. Association for Computational Linguistics, 2022.

[36] Junyu Luo, Jinsheng Huang, Yang Xu, Lutong Zou, Xiao Luo, Bohan Wu, Yifan Wang, Wei Ju, and Ming Zhang. Conformalized hierarchical calibration for uncertainty-aware adaptive hashing. In *International Conference on Learning Representations*, volume 2026, pages 105223–105246, 2026.

[37] Bill MacCartney. *Natural language inference*. Stanford University, 2009.

[38] Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[39] Potsawee Manakul, Adian Liusie, and Mark 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*, pages 9004–9017, 2023.

[40] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics, 2020.

[41] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics, 2019.

[42] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219, 2024.

[43] Meenal V. Narkhede, Prashant P. Bartakke, and Mukul S. Sutaone. A review on weight initialization strategies for neural networks. *Artif. Intell. Rev.*, 55(1):291–322, 2022.

[44] Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

[45] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics, 2020.

[46] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *ArXiv*, abs/2405.20003, 2024.

[47] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[48] Hongxing Pan, Yingying Guo, Wenqing Kuang, and Jiashi Lu. Mind the unseen mass: Unmasking llm hallucinations via soft-hybrid alphabet estimation, 2026.

[49] Feifei Qian, Lu Bai, Lixin Cui, Ming Li, Hangyuan Du, Yue Wang, and Edwin R. Hancock. Exploring the over-smoothing problem of graph neural networks for graph classification: An entropy-based viewpoint. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 3235–3244. ijcai.org, 2025.

[50] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016.

[51] Siva Reddy, Danqi Chen, and Christopher D. Manning. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguistics*, 7:249–266, 2019.

[52] Emanuele Rossi, Fabrizio Frasca, Ben Chamberlain, Davide Eynard, Michael M. Bronstein, and Federico Monti. SIGN: scalable inception graph neural networks. *CoRR*, abs/2004.11198, 2020.

[53] Ryoma Sato. Training-free graph neural networks and the power of labels as features. *Trans. Mach. Learn. Res.*, 2024, 2024.

[54] Simone Scardapane and Dianhui Wang. Randomness in neural networks: an overview. *WIREs Data Mining Knowl. Discov.*, 7(2), 2017.

[55] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3):63:1–63:38, 2026.

[56] Yurun Song, Junchen Zhao, and Lucia Specia. Sentsim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3143–3156, 2021.

[57] Qiyao Sun, Xingming Li, Xixiang He, Ao Cheng, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Efficient hallucination detection: Adaptive bayesian estimation of semantic entropy with guided semantic exploration. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 33117–33125. AAAI Press, 2026.

[58] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025.

[59] Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025.

[60] Pragatheeswaran Vipulanandan, Kamal Premaratne, and Dilip Sarkar. Semantic uncertainty quantification of hallucinations in llms: A quantum tensor network based method. *arXiv preprint arXiv:2601.20026*, 2026.

[61] Yili Wang, Yixin Liu, Xu Shen, Chenyu Li, Rui Miao, Kaize Ding, Ying Wang, Shirui Pan, and Xin Wang. Unifying unsupervised graph-level anomaly detection and out-of-distribution detection: A benchmark. In *International Conference on Learning Representations*, volume 2025, pages 91903–91929, 2025.

[62] Felix Wu, Tianyi Zhang, Amauri H. Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. *CoRR*, abs/1902.07153, 2019.

[63] Zhiping Xiao, Yusheng Zhao, Qixin Zhang, Jiaye Xie, Wanjia Zhao, Weizhi Zhang, Xiao Luo, Philip Yu, and Ming Zhang. Sample lottery: Unsupervised discovery of critical instances for llm reasoning. In *International Conference on Learning Representations*, volume 2026, pages 8919–8949, 2026.

[64] Peng Xu, Lin Zhang, Xuanzhou Liu, Jiaqi Sun, Yue Zhao, Haiqin Yang, and Bei Yu. Do not train it: A linear neural architecture search of graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 38826–38847. PMLR, 2023.

[65] Yihao Xue, Kristjan H. Greenewald, Youssef Mroueh, and Baharan Mirzasoleiman. Verify when uncertain: Beyond self-consistency in black box hallucination detection. *ArXiv*, abs/2502.15845, 2025.

[66] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018.

[67] Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by flexible evidential deep learning. *arXiv preprint arXiv:2510.18322*, 2025.

[68] Zhiyin Yu, Bo Zhang, Chong Chen, Zhonghai Wu, and Xiao Luo. Mask: Multi-agent collaboration for optimizing adversarial attacks on multi-agent systems. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

[69] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022.

A Experiment Details

Experimental protocol. We evaluate hallucination detection in a primary black-box setting: for each prompt q, the target LLM is queried multiple times under identical decoding settings, and an uncertainty score is computed from the sampled responses. Our proposed `CHAIN` operates solely on response texts, without access to model logits or parameters. We also include methods requiring additional model information as performance references.

Unless stated otherwise, we sample $N{=}10$ responses per prompt at temperature $T{=}1$. The benchmarks are SQuAD [50], CoQA [51], NQ-Open [27], TriviaQA [22], and HotpotQA [66]. The

open-source target generators are OPT-6.7B, Qwen3-8B-Instruct[3], Mistral-7B-Instruct, and Phi-3.5-mini [69, 21, 42]. The same CHAIN pipeline is applied to every model–dataset pair.

For the closed-source experiments, we sample 400 questions per evaluated dataset and generate ten responses per question using `gpt-4o-mini`, `gemini-2.5-flash-lite`, or `opus-4.8`. All compared methods use the same sampled responses within each model–dataset setting; CHAIN accesses only the response texts.

Labels and metrics. We evaluate whether a prompt-level score derived from the sampled response graph predicts the reliability of a designated response. SentSim and ROUGE-L provide reference-based reliability labels for that response, rather than factuality labels for the entire sampled set. A correct designated response may coexist with conflicting alternatives, while a consistent response set may still be incorrect.

For each prompt, we retain up to ten deduplicated sampled completions to construct the response graph. The SentSim [56] and ROUGE-L [31] labels are applied to a single designated completion: we use `most_likely_generation` when it is non-empty, and otherwise the first entry in `generations`. Under SentSim, the designated completion is labeled faithful if its maximum ℓ_2-normalized cosine similarity with any reference answer is at least $\tau_s{=}0.9$. Under ROUGE-L, it is labeled faithful if its maximum ROUGE-L F1 against any reference is at least $\tau_r{=}0.5$. The corresponding hallucination indicator equals one when this threshold check fails.

We report fixed-direction AUROC under these two labels as AUC_s and AUC_r, respectively. Score directions follow each method's uncertainty convention and are fixed before evaluation, without post-hoc direction flipping. Table 3 provides the per-model results.

Baselines. We compare CHAIN with uncertainty estimators using different information sources, including text-based methods and methods requiring additional model information. **Predictive Entropy (PE)** [16] aggregates token-level predictive entropy over sampled continuations. **EigenScore** [4] measures spectral dispersion of a response similarity matrix. **Lexical Similarity (LexSim)** uses mean pairwise lexical overlap across sampled responses as an uncertainty signal. **DSE** (Discrete Semantic Entropy) uses NLI to define semantic equivalence between responses, partitions them into equivalence classes, and computes entropy over class proportions. **SNNE** (Semantic Nearest-Neighbor Entropy) [44] estimates semantic dispersion through ROUGE-L-based neighborhood structure. **KLE** (Kernel Language Entropy) [46] constructs an NLI-based response graph, applies a heat-kernel transformation, and computes the von Neumann entropy of the resulting operator.

Configuration. The open-source sampling and evaluation experiments are run on a workstation with eight NVIDIA GeForce RTX 4090 GPUs under CUDA 12.x and a matching PyTorch build. Responses are collected from the Hugging Face `validation` splits using ten i.i.d. continuations per prompt, temperature $T{=}1$, top-$p{=}1$, and `max_new_tokens`$=64$. Input contexts are truncated to 6000 characters. We use automatic mixed precision where applicable and store sampled continuations without generating a separate greedy anchor.

The default CHAIN configuration uses a cross-encoder DeBERTa-v3-base NLI verifier [14] with batch size 24. Response embeddings are obtained with `all-MiniLM-L6-v2`, using mean pooling, maximum input length 256, batch size 48, and ℓ_2-normalized 384-dimensional vectors. The edge thresholds are $\tau_{\mathrm{ent}}{=}0.35$ and $\tau_{\mathrm{con}}{=}0.80$. The readout uses Top-$K$ negative-channel aggregation with $K{=}3$ in the plain ratio form. Results involving stochastic components are averaged over three random seeds. Additional implementation details and evaluation configurations are provided in the accompanying code.

Per-model results. Table 3 reports fixed-direction AUROC for each open-source target model. CHAIN achieves strong performance across the four generators and five datasets, obtaining the best result in most model–dataset–label settings. Some baselines remain stronger in individual cases, indicating that relative performance varies with the generator and reliability label. These results complement the averaged main-table comparison by showing the performance of each method before aggregation.

[3] https://huggingface.co/Qwen/Qwen3-8B

Table 3: Fixed-direction AUROC across four open-source LLMs. AUC_s and AUC_r use the SentSim- and ROUGE-L-based labels, respectively. Blue cells with bold values indicate the best result in each column within a model block.

	CoQA		SQuAD		NQ-Open		TriviaQA		HotpotQA	
Method	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r
OPT-6.7B										
PE	0.708	0.660	0.662	0.673	0.673	0.668	0.742	0.708	0.623	0.609
EigenScore	0.670	0.624	0.724	0.672	0.605	0.668	0.649	0.625	0.646	0.640
LexSim	0.742	0.694	0.704	0.716	0.741	0.754	0.778	0.749	0.728	0.684
DSE	0.769	0.718	0.793	0.768	0.788	**0.781**	0.764	0.744	0.778	0.709
SNNE	0.782	0.703	0.842	0.777	0.788	0.759	0.780	0.759	0.781	0.724
KLE	0.815	0.741	0.700	0.730	0.733	0.780	0.785	0.752	0.752	0.704
`CHAIN` (Ours)	**0.828**	**0.815**	**0.855**	**0.819**	**0.801**	0.776	**0.786**	**0.773**	**0.791**	**0.751**
Qwen3-8B-Instruct										
PE	0.640	0.533	0.539	0.546	0.723	0.737	0.764	0.777	0.593	0.544
EigenScore	0.740	0.733	0.650	0.656	0.658	0.679	0.724	0.726	0.705	0.647
LexSim	0.635	0.782	0.547	0.581	0.748	**0.755**	0.738	0.751	0.745	0.738
DSE	0.756	0.742	0.620	0.634	0.781	0.747	0.756	0.796	0.752	0.755
SNNE	0.727	0.782	0.768	**0.853**	0.861	0.736	0.751	0.776	0.818	0.809
KLE	0.727	0.790	0.734	0.728	0.784	0.723	0.860	0.789	0.706	0.713
`CHAIN` (Ours)	**0.770**	**0.792**	**0.769**	0.839	**0.912**	0.740	**0.865**	**0.803**	**0.874**	**0.826**
Mistral-7B-Instruct										
PE	0.603	0.666	0.565	0.521	0.703	0.671	0.780	0.802	0.558	0.529
EigenScore	0.682	0.687	0.753	0.729	0.718	0.688	0.750	0.709	0.616	0.583
LexSim	0.733	0.775	0.772	0.733	0.711	0.721	0.770	**0.815**	0.641	0.661
DSE	0.780	0.783	0.817	0.786	0.700	0.699	0.764	0.813	0.754	0.724
SNNE	0.850	0.807	0.903	0.874	0.719	0.703	0.766	0.795	**0.865**	**0.841**
KLE	0.708	0.716	0.736	0.704	0.737	0.710	0.802	0.805	0.707	0.744
`CHAIN` (Ours)	**0.854**	**0.812**	**0.930**	**0.875**	**0.815**	**0.801**	**0.803**	0.804	0.859	0.837
Phi-3.5-mini										
PE	0.688	0.697	0.576	0.591	0.732	0.709	0.785	0.783	0.661	0.629
EigenScore	0.709	0.666	0.705	0.685	0.689	0.649	0.714	0.708	0.611	0.617
LexSim	0.828	0.819	0.699	0.697	0.730	0.766	0.801	0.797	0.728	0.746
DSE	0.824	0.793	0.679	0.702	0.750	0.771	0.786	0.788	0.746	0.706
SNNE	0.867	0.826	**0.795**	0.780	0.781	0.774	0.789	0.796	0.744	0.695
KLE	0.762	0.736	0.704	0.708	0.746	0.741	0.793	**0.799**	0.717	0.704
`CHAIN` (Ours)	**0.868**	**0.836**	0.789	**0.781**	**0.782**	**0.775**	**0.815**	0.790	**0.866**	**0.782**

B Additional Experiments

B.1 Robustness

Edge thresholds. We vary τ_{ent} and τ_{con} to examine how relation selection affects downstream detection. Table 4 reports changes in macro-averaged SentSim AUROC relative to the default configuration $(0.35, 0.80)$. Most tested combinations remain close to the default at the displayed precision. However, performance decreases for $(0.45, 0.70)$, showing that threshold robustness does not hold uniformly across the entire tested range. The default configuration lies within a region of relatively stable performance rather than an isolated favorable setting.

Figure 4 provides a complementary view through NLI probabilities pooled over response pairs across models and datasets. Contradiction probabilities have substantial mass near zero and one, so moderate threshold changes can preserve many high-confidence contradiction edges. Entailment probabilities are more diffuse, and changing their threshold can alter the supporting graph structure. Together with the downstream sweep, these observations suggest that retaining strong contradictions can help stabilize the readout, while simultaneous changes to both edge types can still reduce performance.

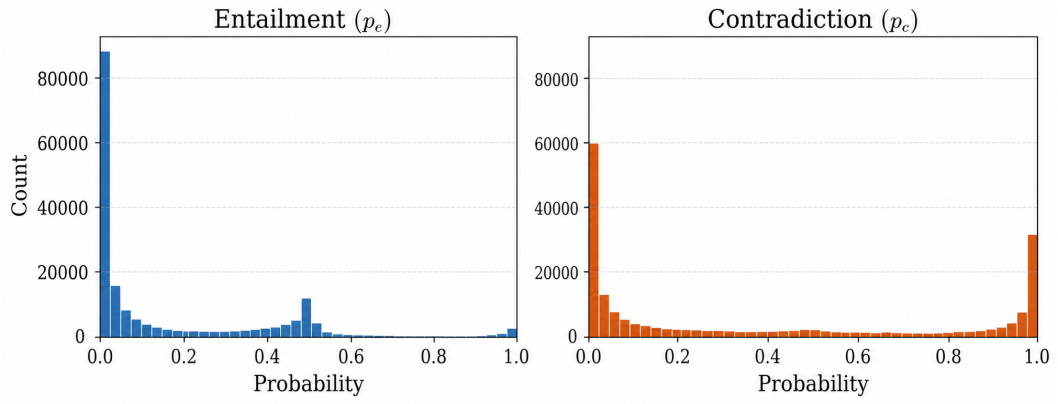

Figure 4: NLI probability distributions pooled across response pairs. Left: entailment probabilities p_e. Right: contradiction probabilities p_c, with substantial mass near zero and one. These distributions help interpret the threshold sensitivity results in Table 4.

Table 4: Change in macro-averaged SentSim AUROC relative to the default edge thresholds $(0.35, 0.80)$. Values are rounded to two decimal places; 0.00 denotes a change that rounds to zero.

$\tau_{\text{ent}} \backslash \tau_{\text{con}}$	0.70	0.75	0.80	0.85	0.90
0.25	0.00	0.00	0.00	0.00	0.00
0.30	0.00	0.00	0.00	0.00	+0.01
0.35	0.00	0.00	0.00	0.00	0.00
0.40	0.00	0.00	0.00	0.00	0.00
0.45	−0.06	−0.01	0.00	0.00	0.00

B.2 Computational Efficiency

We compare the cost of uncertainty estimation using the same response-sampling protocol as the main experiments. Table 5 reports per-prompt wall-clock measurements on 200 CoQA prompts with ten retained responses, measured on a single NVIDIA H100 80GB GPU. The timing configuration and implementation details are provided in the code.

`CHAIN` requires 90 directed NLI comparisons for ten responses and has a measured runtime between DSE and KLE in this comparison. Methods without NLI comparisons are faster, illustrating the additional cost of estimating semantic relations. Within `CHAIN`, signed propagation and readout take approximately 0.9 ms per prompt after relation construction. Thus, the propagation stage itself is lightweight, although the complete pipeline incurs additional cost for obtaining and processing response relations. Pairwise relation estimation requires $N(N-1)$ directed comparisons. Consequently, increasing the number of sampled responses raises this cost quadratically, even when propagation and readout remain inexpensive. The measurements characterize the practical cost at the ten-response setting used in the main experiments.

Table 5: Wall-clock comparison on 200 CoQA prompts with ten responses per prompt, measured on one H100 80GB GPU. NLI pairs counts the comparisons performed by each implementation.

Method	NLI pairs	Time per prompt
EigenScore	0	5.4 ms
LexSim	0	35.0 ms
SNNE	0	41.5 ms
DSE	90	65.1 ms
`CHAIN`	90	83.2 ms
KLE	200	153.4 ms

B.3 Comparison with QTN-UQ

We additionally compare CHAIN with QTN-UQ [60], which incorporates token-probability information into semantic clustering. We use Mistral-7B-Instruct, for which token probabilities are available, and retain the generation settings, reliability labels, and fixed-direction AUROC protocol of the main comparison. QTN-UQ uses its required token-probability inputs, whereas CHAIN continues to use only sampled response texts. Additional implementation details are provided in the code.

Table 6 shows that neither method uniformly dominates across datasets and reliability labels. CHAIN performs better under both labels on CoQA and SQuAD and under the SentSim-based label on TriviaQA and HotpotQA. QTN-UQ performs better under both labels on NQ-Open and under the ROUGE-L-based label on TriviaQA and HotpotQA. The comparison therefore reveals variation across evaluation settings rather than a universal advantage for either method.

These results complement the main comparisons by showing that text-based signed graph modeling remains competitive with an estimator that additionally uses token probabilities. The settings favoring QTN-UQ also indicate that CHAIN does not subsume probability-based uncertainty estimation. Its practical advantage is applicability when only sampled response texts are available, including the closed-source setting evaluated in the main paper.

Table 6: Fixed-direction AUROC comparison with QTN-UQ using Mistral-7B-Instruct. AUC_s and AUC_r use the SentSim- and ROUGE-L-based labels, respectively. Blue cells indicate the best result in each column.

	CoQA		SQuAD		NQ-Open		TriviaQA		HotpotQA	
Method	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r	AUC_s	AUC_r
QTN-UQ	0.817	0.749	0.840	0.785	**0.831**	**0.931**	0.781	**0.913**	0.814	**0.934**
CHAIN	**0.854**	**0.812**	**0.930**	**0.875**	0.815	0.801	**0.803**	0.804	**0.859**	0.837

C Theoretical Analysis of CHAIN

This appendix provides the formal derivations supporting Section 3.5. Throughout, all frozen random hidden matrices in CHAIN are assumed to be drawn from PyTorch's default Kaiming-uniform scheme, i.e. each entry is i.i.d. from

$$\mathcal{U}\big(-1/\sqrt{d_{\mathrm{in}}},\, 1/\sqrt{d_{\mathrm{in}}}\big),$$

where d_{in} is the input dimension of that layer. Entries then have mean 0, variance $1/(3d_{\mathrm{in}})$, and support contained in $[-1/\sqrt{d_{\mathrm{in}}},\, 1/\sqrt{d_{\mathrm{in}}}]$.

We formalize six claims: (i) graph propagation admits a linearized fixed-operator interpretation; (ii) frozen random hidden mappings can remain meaningful at the representation level; (iii) the proposed two-channel signed encoder mixes all four classes of signed two-hop paths; (iv) CHAIN is strictly more expressive than sign-agnostic estimators on an explicit construction; (v) signed propagation amplifies contradiction signals with exponentially decaying failure probability; (vi) in an idealized sparse-conflict model, the Top-K readout is optimal. These are design-justifying results on simplified, mostly linearized surrogates, not a full characterization of the nonlinear two-layer CHAIN encoder.

C.1 Linearized Graph Propagation

Proposition C.1. *Consider a K-layer graph convolution model with fixed propagation matrix S, input feature matrix X, and layer weights $\Theta^{(1)}, \ldots, \Theta^{(K)}$. If all intermediate nonlinearities are removed, then the network reduces to*

$$\hat{Y} = \mathrm{softmax}\big(S^K X\, \Theta\big), \qquad \Theta := \Theta^{(1)}\Theta^{(2)} \cdots \Theta^{(K)}. \tag{13}$$

Equivalently, the encoder decomposes into a fixed feature-propagation step $\bar{X} = S^K X$ followed by a single linear map.

Proof. Removing the nonlinearity σ from each layer gives $H^{(k)} = SH^{(k-1)}\Theta^{(k)}$ with $H^{(0)} = X$. By induction,

$$H^{(K)} = S^K X \Theta^{(1)} \cdots \Theta^{(K)}.$$

Applying the final softmax yields Eq. (13). □

Proposition C.1 motivates the structural view adopted by CHAIN: after the signed topology is fixed, a substantial portion of the inductive bias can come from signed propagation itself rather than from heavily optimized hidden graph filters. We do not claim that CHAIN inherits this reduction exactly, since CHAIN retains nonlinearities between layers; the proposition only establishes that a standard linearized graph model has this form.

C.2 Frozen Random Hidden Mappings

We next formalize the random-hidden-mapping argument used in the main text. CHAIN's propagation layers use rectangular weight matrices, so we state the results for general rectangular factors using the Moore–Penrose pseudo-inverse.

Theorem C.2. *Consider a linearized L-layer graph propagation model*

$$\mathbf{H}^{(L)} = \mathbf{A}^L \mathbf{X} \prod_{l=1}^{L} \mathbf{W}_l, \tag{14}$$

where $\mathbf{W}_l \in \mathbb{R}^{d_{l-1} \times d_l}$ may be rectangular. Let $\mathbf{W}_l(0)$ denote frozen random hidden matrices and let $\mathbf{W}_l^*, \mathbf{W}_o^*$ denote the hidden and output weights of a hypothetical fully trained model. Define

$$\mathbf{P} := \prod_{l=1}^{L} \mathbf{W}_l(0) \in \mathbb{R}^{d_0 \times d_L}, \qquad \mathbf{Q} := \prod_{l=1}^{L} \mathbf{W}_l^*.$$

If

$$\operatorname{col}\big(\mathbf{A}^L \mathbf{X} \mathbf{Q} \mathbf{W}_o^*\big) \subseteq \operatorname{col}\big(\mathbf{A}^L \mathbf{X} \mathbf{P}\big),$$

then there exists an output matrix $\widetilde{\mathbf{W}}_o$ such that

$$\mathbf{A}^L \mathbf{X} \mathbf{P}\, \widetilde{\mathbf{W}}_o = \mathbf{A}^L \mathbf{X} \mathbf{Q}\, \mathbf{W}_o^*. \tag{15}$$

A sufficient condition for this containment is that $\mathbf{P}$ has full row rank, i.e. $\operatorname{rank}(\mathbf{P}) = d_0$. Since

$$\operatorname{rank}(\mathbf{P}) \le \min_{0 \le l \le L} d_l,$$

a concrete dimension-level sufficient condition is

$$d_l \ge d_0 \qquad \text{for all } l = 0, 1, \ldots, L. \tag{16}$$

Under Kaiming-uniform initialization, Proposition C.3 shows that (16) implies $\operatorname{rank}(\mathbf{P}) = d_0$ almost surely.

Proof. Set

$$B := \mathbf{A}^L \mathbf{X} \in \mathbb{R}^{N \times d_0}, \qquad M := B\mathbf{P} \in \mathbb{R}^{N \times d_L}, \qquad T := B\mathbf{Q}\mathbf{W}_o^* \in \mathbb{R}^{N \times c}.$$

We seek $\widetilde{\mathbf{W}}_o \in \mathbb{R}^{d_L \times c}$ satisfying

$$M \widetilde{\mathbf{W}}_o = T.$$

This linear system is solvable if and only if $\operatorname{col}(T) \subseteq \operatorname{col}(M)$. Under the theorem's column-space containment assumption, one valid exact solution is

$$\widetilde{\mathbf{W}}_o := M^+ T,$$

where M^+ is the Moore–Penrose pseudo-inverse of M. Since $T \in \operatorname{col}(M)$, we have $MM^+T = T$.

We now verify the sufficient full-row-rank condition. Suppose $\operatorname{rank}(\mathbf{P}) = d_0$. Then $\mathbf{P}$ admits a right-inverse

$$\mathbf{P}^\dagger = \mathbf{P}^\top (\mathbf{P}\mathbf{P}^\top)^{-1} \in \mathbb{R}^{d_L \times d_0}, \qquad \mathbf{P}\mathbf{P}^\dagger = \mathbf{I}_{d_0}.$$

Set

$$\widetilde{\mathbf{W}}_o := \mathbf{P}^\dagger \mathbf{Q} \mathbf{W}_o^*.$$

Then

$$M \widetilde{\mathbf{W}}_o = B\mathbf{P}\mathbf{P}^\dagger \mathbf{Q}\mathbf{W}_o^* = B\mathbf{Q}\mathbf{W}_o^* = T.$$

Thus full row rank of $\mathbf{P}$ implies the desired representation equivalence. Moreover, in this case $\operatorname{col}(B\mathbf{P}) = \operatorname{col}(B)$, and $T = B\mathbf{Q}\mathbf{W}_o^*$ is necessarily contained in $\operatorname{col}(B)$. □

Theorem C.2 shows that frozen random hidden transformations are not automatically meaningless: under the stated column-space condition, their effect on propagated representations can be exactly compensated by an appropriate output map. We next state the probabilistic guarantee for the rank condition.

Proposition C.3. *Let $\mathbf{W}_l(0) \in \mathbb{R}^{d_{l-1} \times d_l}$ be hidden-layer weight matrices whose entries are sampled independently from an absolutely continuous probability distribution, such as Kaiming-uniform initialization. Then:*

1. *Each $\mathbf{W}_l(0)$ has $\operatorname{rank}(\mathbf{W}_l(0)) = \min(d_{l-1}, d_l)$ almost surely.*
2. *$\operatorname{rank}(\mathbf{P}) = \min(d_0, d_1, \ldots, d_L)$ almost surely. In particular, if $d_l \geq d_0$ for all $l = 0, 1, \ldots, L$, then $\mathbf{P}$ has full row rank almost surely.*

Proof. **Claim 1.** For $\mathbf{M} \in \mathbb{R}^{m \times n}$ with independent absolutely continuous entries, every $\min(m, n) \times \min(m, n)$ minor is a polynomial in the entries. At least one such minor is not the zero polynomial, and the zero set of a non-trivial polynomial has Lebesgue measure zero. Therefore $\operatorname{rank}(\mathbf{M}) = \min(m, n)$ almost surely.

Claim 2. Let

$$\mathbf{P}_l := \prod_{k=1}^{l} \mathbf{W}_k(0) \in \mathbb{R}^{d_0 \times d_l}.$$

We prove by induction that

$$\operatorname{rank}(\mathbf{P}_l) = \min(d_0, d_1, \ldots, d_l)$$

almost surely.

The base case $l = 1$ follows from Claim 1. For the induction step, suppose

$$\operatorname{rank}(\mathbf{P}_{l-1}) = \min(d_0, \ldots, d_{l-1}) =: r$$

almost surely. Condition on a realization of $\mathbf{P}_{l-1}$ with rank r. There exist invertible matrices R and S such that

$$R\mathbf{P}_{l-1}S = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}.$$

Multiplication by invertible matrices does not change rank, so

$$\operatorname{rank}(\mathbf{P}_{l-1}\mathbf{W}_l(0)) = \operatorname{rank}\left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} S^{-1}\mathbf{W}_l(0) \right).$$

Since $\mathbf{W}_l(0)$ has independent absolutely continuous entries and S^{-1} is fixed under the conditioning, the first r rows of $S^{-1}\mathbf{W}_l(0)$ form an $r \times d_l$ random matrix whose columns are independent and have absolutely continuous distributions on $\mathbb{R}^r$. Such a matrix has rank $\min(r, d_l)$ almost surely by the same polynomial-minor argument as in Claim 1. Therefore

$$\operatorname{rank}(\mathbf{P}_l) = \min(r, d_l) = \min(d_0, \ldots, d_l)$$

almost surely. Iterating to $l = L$ gives

$$\operatorname{rank}(\mathbf{P}) = \min(d_0, \ldots, d_L)$$

almost surely. If $d_l \geq d_0$ for all $l = 0, 1, \ldots, L$, this minimum equals d_0, so $\mathbf{P}$ has full row rank almost surely. □

Remark C.4. *Kaiming-uniform initialization is absolutely continuous, so Proposition C.3 applies to `CHAIN`. Discrete alternatives, such as Achlioptas sparse $\pm\sqrt{3}/0$ projections, do not satisfy the absolute-continuity assumption and require a separate rank analysis.*

These results do not directly characterize `CHAIN`'s norm-based Top-K readout. Their role is narrower: they justify that, at the representation level, frozen random hidden transformations drawn from a continuous distribution can act as feature mixers without destroying the structural information produced by signed propagation.

C.3 Signed Two-Hop Path Decomposition

Proposition C.5. *Let $\mathbf{H}^{(1,E)}$ and $\mathbf{H}^{(1,C)}$ denote the first-layer enrichment and contradiction representations. Let the second layer be*

$$\mathbf{H}^{(2,E)} = \sigma\Big(\widetilde{\mathbf{A}}^{+}\mathbf{H}^{(1,E)}\mathbf{W}^{(2)}_{E,\mathcal{N}^+} + \widetilde{\mathbf{A}}^{-}\mathbf{H}^{(1,C)}\mathbf{W}^{(2)}_{E,\mathcal{N}^-} + \mathbf{H}^{(1,E)}\mathbf{W}^{(2)}_{E,\mathcal{S}}\Big),$$
$$\mathbf{H}^{(2,C)} = \sigma\Big(\widetilde{\mathbf{A}}^{+}\mathbf{H}^{(1,C)}\mathbf{W}^{(2)}_{C,\mathcal{N}^+} + \widetilde{\mathbf{A}}^{-}\mathbf{H}^{(1,E)}\mathbf{W}^{(2)}_{C,\mathcal{N}^-} + \mathbf{H}^{(1,C)}\mathbf{W}^{(2)}_{C,\mathcal{S}}\Big). \tag{17}$$

Ignoring the nonlinearities, $\mathbf{H}^{(2,E)}$ and $\mathbf{H}^{(2,C)}$ are linear combinations of the four signed two-hop operators

$$\widetilde{\mathbf{A}}^{+}\widetilde{\mathbf{A}}^{+}, \qquad \widetilde{\mathbf{A}}^{+}\widetilde{\mathbf{A}}^{-}, \qquad \widetilde{\mathbf{A}}^{-}\widetilde{\mathbf{A}}^{+}, \qquad \widetilde{\mathbf{A}}^{-}\widetilde{\mathbf{A}}^{-},$$

together with residual one-hop and self terms.

Proof. Linearize the first layer:

$$\mathbf{H}^{(1,E)} = \widetilde{\mathbf{A}}^{+}\mathbf{X}\mathbf{W}^{(1)}_{E,\mathcal{N}} + \mathbf{X}\mathbf{W}^{(1)}_{E,\mathcal{S}},$$
$$\mathbf{H}^{(1,C)} = \widetilde{\mathbf{A}}^{-}\mathbf{X}\mathbf{W}^{(1)}_{C,\mathcal{N}} + \mathbf{X}\mathbf{W}^{(1)}_{C,\mathcal{S}}. \tag{18}$$

Substituting into the enrichment branch of Eq. (17) yields terms involving

$$\widetilde{\mathbf{A}}^{+}\widetilde{\mathbf{A}}^{+}\mathbf{X} \qquad \text{and} \qquad \widetilde{\mathbf{A}}^{-}\widetilde{\mathbf{A}}^{-}\mathbf{X},$$

plus residual one-hop and self terms. Substituting into the contradiction branch analogously yields terms involving

$$\widetilde{\mathbf{A}}^{+}\widetilde{\mathbf{A}}^{-}\mathbf{X} \qquad \text{and} \qquad \widetilde{\mathbf{A}}^{-}\widetilde{\mathbf{A}}^{+}\mathbf{X}.$$

Hence all four signed two-hop classes appear across the two branches. □

C.4 Structural Separation from Sign-Agnostic Estimators

Definition C.6 (Sign-agnostic estimator)**.** *An uncertainty estimator $\hat{U}$ is sign-agnostic if it depends on the response set only through (a) the node embedding matrix $X \in \mathbb{R}^{N\times d}$ and/or (b) an unsigned symmetric pairwise-relation matrix $M \in \mathbb{R}^{N\times N}$ that does not separate entailment and contradiction. Examples include cosine similarity, Gaussian kernels, k-NN graphs, and bidirectional-entailment equivalence collapsed to a single non-negative score.*

This class covers the sign-agnostic baselines used in our experiments: EigenScore, KLE, SNNE, LexSim, PE, and DSE. DSE uses NLI only to determine a binary equivalence relation and therefore does not exploit the separate contradiction component.

Theorem C.7 (Separation)**.** *There exist two response graphs $G_1 = (X, A^+, A_1^-)$ and $G_2 = (X, A^+, A_2^-)$ with the same node embeddings X and the same enrichment edges A^+ such that:*

1. *Every sign-agnostic estimator in Definition C.6 assigns $\hat{U}(G_1) = \hat{U}(G_2)$.*
2. *The linearized `CHAIN` model produces different contradiction-channel magnitudes on G_1 and G_2 for almost every draw of the frozen random hidden matrices; consequently, a readout using these magnitudes can separate the two graphs.*

Proof. Let $N = 4$ with distinct unit-norm embeddings arranged as two clusters: x_1, x_2 are close and x_3, x_4 are close, while all cross-cluster similarities are equal and lower. Such a configuration exists by taking two nearly orthogonal unit vectors e_1, e_2 and placing x_1, x_2 near e_1 and x_3, x_4 near e_2 with symmetric perturbations.

Define common enrichment edges $(1, 2)$ and $(3, 4)$ with positive weight. For G_1, add negative edges on all four cross-cluster pairs

$$\{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}.$$

For G_2, set $A_2^- = \mathbf{0}$ and interpret the cross-cluster difference as neutrality rather than contradiction.

Claim 1. By construction, X is identical for G_1 and G_2. Any estimator that uses only embeddings or unsigned affinities that do not separate entailment from contradiction receives the same information in both cases. Therefore every estimator falling under Definition C.6 produces identical output on G_1 and G_2.

Claim 2. In the linearized `CHAIN`, the first-layer contradiction channel is

$$h_i^{(1,C)} = \widetilde{A}_i^- X W_{C,\mathcal{N}}^{(1)} + x_i W_{C,\mathcal{S}}^{(1)}.$$

On G_2, $\widetilde{A}^- = \mathbf{0}$, so

$$h_1^{(1,C)}\big|_{G_2} = x_1 W_{C,\mathcal{S}}^{(1)}.$$

On G_1, using uniform negative edge weights from node 1 to nodes 3 and 4,

$$\widetilde{A}_1^- X = (x_3 + x_4)/2,$$

and hence

$$h_1^{(1,C)}\big|_{G_1} = x_1 W_{C,\mathcal{S}}^{(1)} + v, \qquad v := \tfrac{1}{2}(x_3 + x_4)^\top W_{C,\mathcal{N}}^{(1)}.$$

Because $(x_3 + x_4)/2 \neq 0$ and $W_{C,\mathcal{N}}^{(1)}$ has an absolutely continuous initialization, $v \neq 0$ almost surely.

It remains to rule out the possibility that the two vectors differ but have exactly the same norm. Condition on any realization with $v \neq 0$ and let

$$a := x_1 W_{C,\mathcal{S}}^{(1)}.$$

The event

$$\|a + v\|_2 = \|a\|_2$$

is equivalent to

$$2\langle a, v\rangle + \|v\|_2^2 = 0.$$

For fixed nonzero v, this is a non-trivial affine hyperplane condition on a. Since $x_1 \neq 0$ and $W_{C,\mathcal{S}}^{(1)}$ has absolutely continuous entries independent of $W_{C,\mathcal{N}}^{(1)}$, the probability that a falls exactly on this hyperplane is zero. Therefore

$$\|h_1^{(1,C)}\|_2\big|_{G_1} \neq \|h_1^{(1,C)}\|_2\big|_{G_2}$$

almost surely.

Thus `CHAIN`'s signed contradiction representation separates G_1 and G_2 almost surely, while any sign-agnostic estimator assigns them the same value by Claim 1. A readout that depends on contradiction-channel magnitudes, including the Top-K readout used by `CHAIN`, can therefore distinguish the two constructions. □

The semantic interpretation is clear: G_1 represents two paraphrased clusters making opposing factual claims, while G_2 represents two internally consistent but topically unrelated groups. Any estimator operating on embeddings or unsigned similarities alone collapses the two; `CHAIN` separates them via its signed decomposition.

C.5 Contradiction Amplification under Signed Propagation

Theorem C.8 (Contradiction Amplification)**.** *Consider the linearized single-layer `CHAIN` model. Assume $\|x_i\|_2 = 1$ for all nodes i. Let*

$$W_{C,\mathcal{N}}^{(1)}, W_{C,\mathcal{S}}^{(1)} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$$

be two independent frozen random matrices with i.i.d. entries from

$$\mathcal{U}(-1/\sqrt{d_{\text{in}}},\, 1/\sqrt{d_{\text{in}}}).$$

Let

$$\mathcal{C} = \{i : |\mathcal{N}_i^-| \geq 1\}, \qquad \mathcal{F} = V \setminus \mathcal{C},$$

and define

$$z_i^- := \|h_i^{(1,C)}\|_2, \qquad h_i^{(1,C)} = \widetilde{A}_i^- X W_{C,\mathcal{N}}^{(1)} + x_i W_{C,\mathcal{S}}^{(1)}.$$

Fix $i \in \mathcal{C}$ and $j \in \mathcal{F}$, and set

$$\alpha_i := \|\widetilde{A}_i^- X\|_2.$$

Assume $\alpha_i > 0$, i.e., the weighted aggregate of contradiction-neighbor embeddings at node i is nonzero. Then

$$\Pr\big(z_i^- \le z_j^-\big) \le \exp\!\left(-\frac{d_{\text{out}}}{18 d_{\text{in}}^2}\right) + 2\exp\!\left(-\frac{\alpha_i^2 d_{\text{out}}}{1152 d_{\text{in}}^2}\right) + 2\exp\!\left(-\frac{\alpha_i^4 d_{\text{out}}}{1152 d_{\text{in}}^2}\right). \tag{19}$$

In particular, for fixed $\alpha_i > 0$, $\Pr(z_i^- \le z_j^-) \to 0$ exponentially fast as $d_{\text{out}} \to \infty$.

Proof. Write

$$W_{\mathcal{S}} = W_{C,\mathcal{S}}^{(1)}, \qquad W_{\mathcal{N}} = W_{C,\mathcal{N}}^{(1)},$$

and denote by $w_{\mathcal{S},k}$ and $w_{\mathcal{N},k}$ their k-th columns. The columns of each matrix are independent, and the two matrices are mutually independent.

For $j \in \mathcal{F}$, $\widetilde{A}_j^- = 0$, so

$$h_j^{(1,C)} = x_j W_{\mathcal{S}}, \qquad (z_j^-)^2 = \|x_j W_{\mathcal{S}}\|_2^2.$$

For $i \in \mathcal{C}$, set

$$u_i := x_i W_{\mathcal{S}}, \qquad v_i := (\widetilde{A}_i^- X) W_{\mathcal{N}}.$$

Then

$$(z_i^-)^2 = \|u_i\|_2^2 + \|v_i\|_2^2 + 2\langle u_i, v_i\rangle.$$

Define $u_j := x_j W_{\mathcal{S}}$. The event $\{z_i^- > z_j^-\}$ is equivalent to

$$\|v_i\|_2^2 + 2\langle u_i, v_i\rangle + \underbrace{(\|u_i\|_2^2 - \|u_j\|_2^2)}_{=:\Delta_{ij}} > 0. \tag{20}$$

Uniform entry bounds. Every entry of $W_{\mathcal{S}}$ and $W_{\mathcal{N}}$ lies in $[-1/\sqrt{d_{\text{in}}}, 1/\sqrt{d_{\text{in}}}]$. Thus, for any column w and any $a \in \mathbb{R}^{d_{\text{in}}}$,

$$|a^\top w| \le \|a\|_1 \|w\|_\infty \le \sqrt{d_{\text{in}}}\|a\|_2 \cdot \frac{1}{\sqrt{d_{\text{in}}}} = \|a\|_2.$$

Applying this to $a = x_i$ and $a = \widetilde{A}_i^- X$ gives

$$|(u_i)_k| \le 1, \qquad |(v_i)_k| \le \alpha_i.$$

Therefore $(v_i)_k^2 \in [0, \alpha_i^2]$ and $(u_i)_k (v_i)_k \in [-\alpha_i, \alpha_i]$.

Term 1: concentration of $\|v_i\|_2^2$. Let $y_i := \widetilde{A}_i^- X$, so $\|y_i\|_2 = \alpha_i$. For each k,

$$(v_i)_k = y_i^\top w_{\mathcal{N},k}$$

has mean zero and variance $\alpha_i^2/(3d_{\text{in}})$. The variables $(v_i)_k$ are independent across k. Thus

$$\mu_v := \mathbb{E}\|v_i\|_2^2 = \frac{\alpha_i^2 d_{\text{out}}}{3 d_{\text{in}}}.$$

Since $(v_i)_k^2 \in [0, \alpha_i^2]$, Hoeffding's inequality gives

$$\Pr\!\left(\|v_i\|_2^2 < \frac{\mu_v}{2}\right) \le \exp\!\left(-\frac{2(\mu_v/2)^2}{d_{\text{out}}\alpha_i^4}\right) = \exp\!\left(-\frac{d_{\text{out}}}{18 d_{\text{in}}^2}\right). \tag{21}$$

Term 2: concentration of $\langle u_i, v_i\rangle$. Since $W_{\mathcal{S}}$ and $W_{\mathcal{N}}$ are independent, $(u_i)_k$ and $(v_i)_k$ are independent for every k. The products $(u_i)_k (v_i)_k$ are independent across k, have mean zero, and lie in $[-\alpha_i, \alpha_i]$. By Hoeffding's inequality,

$$\Pr\big(|\langle u_i, v_i\rangle| > t_1\big) \le 2\exp\!\left(-\frac{t_1^2}{2\alpha_i^2 d_{\text{out}}}\right). \tag{22}$$

Term 3: concentration of Δ_{ij}. Write

$$\Delta_{ij} = \sum_{k=1}^{d_{\text{out}}} \delta_k, \qquad \delta_k = (x_i^\top w_{\mathcal{S},k})^2 - (x_j^\top w_{\mathcal{S},k})^2.$$

The δ_k are independent across k because the columns of $W_{\mathcal{S}}$ are independent. Since $\|x_i\|_2 = \|x_j\|_2 = 1$,

$$\mathbb{E}\delta_k = \frac{\|x_i\|_2^2 - \|x_j\|_2^2}{3d_{\text{in}}} = 0.$$

The uniform bound above gives

$$|x_i^\top w_{\mathcal{S},k}| \le 1, \qquad |x_j^\top w_{\mathcal{S},k}| \le 1.$$

Thus both squared terms lie in $[0, 1]$, so

$$\delta_k \in [-1, 1].$$

By Hoeffding's inequality,

$$\Pr\big(|\Delta_{ij}| > t_2\big) \le 2\exp\left(-\frac{2t_2^2}{4d_{\text{out}}}\right) = 2\exp\left(-\frac{t_2^2}{2d_{\text{out}}}\right). \tag{23}$$

Combining. Set

$$B := \mu_v/2 = \frac{\alpha_i^2 d_{\text{out}}}{6d_{\text{in}}}, \qquad t_1 = t_2 = \frac{B}{4} = \frac{\alpha_i^2 d_{\text{out}}}{24 d_{\text{in}}}.$$

On the intersection of the events

$$E_1 : \|v_i\|_2^2 \ge B, \qquad E_2 : |\langle u_i, v_i\rangle| \le t_1, \qquad E_3 : |\Delta_{ij}| \le t_2,$$

we have

$$\|v_i\|_2^2 + 2\langle u_i, v_i\rangle + \Delta_{ij} \ge B - 2t_1 - t_2 = B - \frac{B}{2} - \frac{B}{4} = \frac{B}{4} > 0.$$

Thus (20) holds. Plugging the chosen t_1, t_2 into (21)–(23) gives

$$\Pr(\bar{E}_1) \le \exp\left(-\frac{d_{\text{out}}}{18 d_{\text{in}}^2}\right), \tag{24}$$

$$\Pr(\bar{E}_2) \le 2\exp\left(-\frac{\alpha_i^2 d_{\text{out}}}{1152 d_{\text{in}}^2}\right), \tag{25}$$

$$\Pr(\bar{E}_3) \le 2\exp\left(-\frac{\alpha_i^4 d_{\text{out}}}{1152 d_{\text{in}}^2}\right). \tag{26}$$

A union bound yields Eq. (19). Each term decays exponentially in d_{out} for fixed $\alpha_i > 0$. □

Theorem C.8 states a conditional result on the linearized single-layer model: whenever the weighted contradiction-neighbor aggregate at node i is nonzero, its negative-channel magnitude stochastically dominates that of an unconflicted node j, with failure probability decaying exponentially in d_{out}. This provides design-level justification for reading out through a Top-K selection on z^-, but does not characterize the full nonlinear two-layer `CHAIN`.

C.6 Optimality of Top-K Readout in a Sparse-Conflict Model

Proposition C.9 (Top-K optimality, idealized two-level model)**.** *Let $\mathbf{z}^- = (z_1^-, \dots, z_N^-)$, and assume an idealized two-level contrast: exactly K^* nodes satisfy $z_i^- = z_h$ for $i \in \mathcal{C}$, and all remaining nodes satisfy $z_j^- = z_l$ for $j \in \mathcal{F}$, with $z_h > z_l > 0$. Consider weighted linear readouts*

$$R_{\mathbf{w}}(\mathbf{z}^-) = \sum_{i=1}^N w_i z_i^-, \qquad w_i \ge 0, \qquad \sum_{i=1}^N w_i = 1.$$

Define the contrast ratio

$$\rho(\mathbf{w}) := \frac{R_{\mathbf{w}}(\mathbf{z}^-_{\text{hallu}})}{R_{\mathbf{w}}(\mathbf{z}^-_{\text{safe}})},$$

where $\mathbf{z}^-_{\text{hallu}}$ has K^* entries equal to z_h and the rest equal to z_l, while $\mathbf{z}^-_{\text{safe}}$ is the all-z_l vector. Then:

1. *For every valid* $\mathbf{w}$, $\rho(\mathbf{w}) \le z_h/z_l$, *with equality if and only if* $\mathbf{w}$ *is supported on* $\mathcal{C}$.
2. *The Top-*K *readout with* $K \le K^*$ *attains* $\rho = z_h/z_l$.
3. *The mean readout satisfies*

$$\rho_{\text{mean}} = 1 + \frac{K^*}{N}\left(\frac{z_h}{z_l} - 1\right) < \frac{z_h}{z_l}$$

whenever $K^* < N$.

Proof. For any valid $\mathbf{w}$,

$$R_{\mathbf{w}}(\mathbf{z}^-_{\text{safe}}) = z_l.$$

Moreover,

$$R_{\mathbf{w}}(\mathbf{z}^-_{\text{hallu}}) = z_h \sum_{i\in\mathcal{C}} w_i + z_l \sum_{j\in\mathcal{F}} w_j = z_l + (z_h - z_l)\sum_{i\in\mathcal{C}} w_i.$$

Thus

$$\rho(\mathbf{w}) = 1 + \frac{z_h - z_l}{z_l}\sum_{i\in\mathcal{C}} w_i.$$

The maximum over the simplex is attained when $\sum_{i\in\mathcal{C}} w_i = 1$, i.e., when all weight is supported on $\mathcal{C}$. This gives Claim 1. If $K \le K^*$, the Top-K readout places its weight on entries equal to z_h, hence all selected nodes lie in $\mathcal{C}$ and Claim 2 follows. For the mean readout, $\sum_{i\in\mathcal{C}} w_i = K^*/N$, which gives Claim 3. □

Remark C.10. *Proposition C.9 is an optimality statement for the idealized two-level contrast model only. Together with Theorem C.8, it explains why a localized Top-*K *readout is natural when contradiction evidence is concentrated on a small subset of nodes. For the actual nonlinear* `CHAIN` *with a continuum of* z^- *values, the proposition should be read as a design rationale rather than a global optimality guarantee.*

Proposition C.9 is consistent with the empirical Top-K sweep in Figures **??**–**??**: small K preserves concentrated contradiction signal, while large K approaches full-coverage averaging and can dilute it by mixing in low-magnitude consensus nodes.

C.7 Summary

The results in this appendix provide complementary design-level justifications for `CHAIN`. Proposition C.1 casts graph propagation as a fixed structural operator in the linearized setting. Theorem C.2 and Proposition C.3 show that, under Kaiming-uniform initialization and the non-bottleneck width condition $d_l \ge d_0$ for all $l = 0, 1, \ldots, L$, frozen random hidden mappings can be compensated at the representation level. Proposition C.5 confirms that the two-layer signed encoder covers all four classes of signed two-hop paths. Theorem C.7 exhibits an explicit pair of response graphs on which sign-agnostic estimators cannot distinguish two qualitatively different scenarios while `CHAIN`'s signed contradiction representation almost surely can. Theorem C.8 gives an exponentially decaying failure probability for contradiction amplification in the linearized single-layer model, under the condition that the weighted contradiction-neighbor aggregate is nonzero. Proposition C.9 shows that, in an idealized sparse-conflict model, the Top-K readout is optimal in contrast ratio.

These results justify `CHAIN`'s design choices—signed propagation, frozen random hidden maps, and Top-K readout—on simplified surrogates; they do not claim a full characterization of the nonlinear two-layer encoder, which is studied empirically in the main experiments.

D Licenses of Existing Assets

In this work, we utilize several publicly available datasets and pre-trained language models. We deeply respect the intellectual property of the original creators. All assets were used strictly for academic research purposes, in full compliance with their respective licenses. Below is a summary of the licenses for the primary assets used in our experiments.

Table 7: Summary of assets and their respective licenses used in this work.

Asset Name	Asset Type	License
Datasets		
SQuAD [50]	QA Dataset	CC BY-SA 4.0
CoQA [51]	QA Dataset	CC BY-SA 4.0 & Various Domain Licenses
NQ-Open [27]	QA Dataset	Apache License 2.0 & CC BY-SA 3.0
TriviaQA [22]	QA Dataset	Apache License 2.0
HotpotQA [66]	QA Dataset	CC BY-SA 4.0
Models		
OPT-6.7B [69]	LLM	OPT License (Non-commercial research)
Qwen3-8B-Instruct[4]	LLM	Apache License 2.0
Mistral-7B-Instruct [21]	LLM	Apache License 2.0
Phi-3.5-mini [42]	LLM	MIT License
DeBERTa (NLI Base)	NLI Model	MIT License

Limitations. Although CHAIN improves black-box hallucination detection by modeling signed relations among sampled responses, it still has several limitations. First, the method relies on multiple sampled responses and pairwise NLI inference, which introduces an $O(N^2)$ computational cost as the number of samples increases. This cost is manageable for the sample size used in our experiments, but may become expensive for large-scale deployment or very long generations. Second, CHAIN depends on the quality of the external NLI model: noisy entailment or contradiction predictions can affect the constructed signed graph, especially for ambiguous, multi-hop, or domain-specific answers. Third, like other sampling-based methods, CHAIN may fail when all sampled responses consistently make the same incorrect factual commitment, since little cross-response contradiction is observable in that case. Finally, our theoretical analysis is based on a linearized surrogate of the signed propagation module and does not fully characterize the nonlinear encoder used in experiments. Future work could explore more efficient graph construction, stronger domain-adaptive NLI models, and hybrid detectors that combine signed response inconsistency with external evidence retrieval.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction explicitly state the scope of our work, namely black-box LLM uncertainty quantification, and our main contributions, including the proposed `CHAIN` framework, its theoretical properties, and empirical effectiveness, which are supported by the subsequent method, theory, and experimental sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss limitations of our work, including the need for multiple sampled responses, dependency on the quality of the underlying NLI model, the focus on short- and medium-form QA benchmarks, and failure cases where all sampled responses are consistently wrong.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions.
   - The authors should reflect on the scope of the claims made.
   - The authors should reflect on the factors that influence the performance of the approach.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete and correct proof?

   Answer: [Yes]

   Justification: The main paper states the key theoretical results, including signed separation, signed two-hop propagation, contradiction amplification, and Top-K readout behavior. Appendix C provides the full assumptions, formal definitions, and complete proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper, regardless of whether the code and data are provided or not?

   Answer: [Yes]

   Justification: We disclose the experimental settings needed to reproduce the main results, including the evaluated LLMs, decoding parameters, NLI model, graph-construction thresholds, readout hyperparameters, benchmark splits, and evaluation metrics in Appendix A. We also provide code and reproduction scripts at https://github.com/hongxing-pan/CHAIN-NeurIPS2026.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Reproducibility can be accomplished by releasing code and data, providing detailed instructions, hosted models, model checkpoints, or other appropriate means.
   - In the case of closed-source models, it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide a public code repository at https://github.com/hongxing-pan/CHAIN-NeurIPS2026. The repository contains the implementation of CHAIN, scripts for graph construction and evaluation, and instructions for reproducing the reported results. The experiments use publicly available benchmark datasets and publicly accessible model checkpoints or APIs, with data access and preparation described in the repository and Appendix A.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines https://neurips.cc/public/guides/CodeSubmissionPolicy for more details.
   - While code and data release is encouraged, papers cannot be rejected simply for not including code unless this is central to the contribution.
   - The instructions should contain the exact command and environment needed to reproduce the results.
   - The authors should provide instructions on data access and preparation.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines, or state which experiments are omitted and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions if applicable.
   - Providing as much information as possible in supplemental material is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details necessary to understand the results?

Answer: [Yes]

Justification: The main paper and Appendix A specify the benchmark splits, response sampling protocol, evaluated LLMs, NLI thresholds, Top-K readout setting, random seeds, and evaluation metrics. Additional reproduction commands and implementation details are provided in the code repository at https://github.com/hongxing-pan/CHAIN-NeurIPS2026.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report results over multiple benchmark–model settings and include seed-based variability for stochastic components in the ablation and sensitivity analyses. The corresponding figures report error bars over independent random seeds, and the appendix states how the reported variability is computed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests.
- The factors of variability that the error bars are capturing should be clearly stated.
- The method for calculating the error bars should be explained.
- The assumptions made should be given.
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

Answer: [Yes]

Justification: We report our hardware setup and key implementation details, including GPU type, batch sizes, and approximate runtime, in Appendix A. Peak GPU memory per GPU is below 20 GB under the reported configuration, and each dataset$\times$LLM run takes roughly 1 hour end-to-end.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our research aims to improve the reliability of LLM outputs by detecting hallucination risk from sampled responses. We use public benchmark datasets and publicly accessible models, and do not release new high-risk models or sensitive datasets.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We include a Broader Impact statement in Appendix 6, discussing how hallucination detection may improve reliability while also noting that uncertainty scores should not be treated as complete factuality guarantees.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Authors should consider possible harms from intended use, incorrect use, and misuse.
- If there are negative societal impacts, authors may discuss possible mitigation strategies.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?

Answer: [N/A]

Justification: Our work focuses on an evaluation and uncertainty quantification framework. We do not release new high-risk foundation models, scraped datasets, or generative systems with increased misuse potential.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse should be released with necessary safeguards.
- Datasets that have been scraped from the Internet could pose safety risks.
- Authors are encouraged to take safeguards into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets used in the paper properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All pre-trained LLMs, NLI models, sentence encoders, and QA datasets are properly cited. The code repository documents the external assets used by the experiments and points to their original sources and applicable licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license should be included for each asset.
- For scraped data from a particular source, the copyright and terms of service should be provided.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset should be provided.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the implementation of `CHAIN` as a new code asset through an code repository at https://github.com/hongxing-pan/CHAIN-NeurIPS2026. The repository includes documentation and reproduction instructions. We do not introduce a new dataset or a new foundation model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets if applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?

    Answer: [N/A]

    Justification: Our research uses existing benchmark datasets and automated evaluation with NLI and similarity metrics. It does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board approvals or an equivalent approval/review was obtained?

    Answer: [N/A]

    Justification: No human subjects were involved in this research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval or equivalent may be required for any human subjects research.
    - For initial submissions, do not include any information that would break anonymity.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are the target generators evaluated in this work, and their usage is described in the experimental setup, including the evaluated model families and decoding configuration. We do not use LLMs as undisclosed tools for generating experimental labels, modifying results, or developing the core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to the NeurIPS LLM policy for what should or should not be described.