SEMNORM: SEMANTIC NORMALIZATION FOR REASONING TRACES EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As large reasoning models generate increasingly lengthy and seemingly coherent natural language outputs, the difficulty of evaluating them at the trace-level has increased. In tandem, the verification of lengthy traces no longer provide sufficiently granular information about specific step-by-step errors or premise validity. To address these limitations in natural language reasoning traces, SemNorm (SemNorm) introduces two components: **1. Atomic extraction**, to process reasoning traces into atomic claims-level statements; and **2. Directed graph construction**, which builds a knowledge graph from the pairwise entailment between atomic statements. This framework is founded on two observations. First, lexical-based methods are inadequate for establishing information content in diverse natural language generations. Second, embedding-based retrieval methods establish undirected semantic similarity, which leaves causal and implicative statements indistinguishable. To verify the information content in the extracted graph, SemNorm is applied as a scoring metric for reasoning quality and validity, and benchmarked against embedding-based methods as a semantic equivalence measure in perturbation-based experiments. Thus, this work discusses methods for estimating trace-level metrics, and introduces a directed reasoning graph representation for computing graph-derived scores which are comparably monotonic and on par with existing NLI-based scorers and embedding-based scorers for reasoning trace lexical, quality and correctness indicators and as a retriever model for RAG.

1 INTRODUCTION

The emergent reasoning capabilities of large language models (LLMs) have expanded their use in knowledge-intensive natural language generation (NLG) tasks, including medical systemic review (Tang et al., 2025) and research paper generation (Yamada et al., 2025). Yet high-risk or specialized industries such as healthcare still require human evaluation and approval, constraining the throughput of LLM-based applications and limiting practical deployment. For example, LLM usage to summarize patient information from a large amount of clinical notes would still require manual human verification of the source documents for each output statement, given the propensity of LLMs to hallucinate. While existing metrics have been introduced to assist in quantifying hallucination (Kuhn et al., 2023) or evaluate source attribution (Tang et al., 2025), they rely on mechanisms that fail to scale efficiently to longer evaluations. Thus, there exists a evaluation bottleneck as tasks scale from short question answering to longer reasoning traces and document-length outputs, with an increasingly burdensome search for human evaluators to identify logical inconsistencies or factual errors.

To address this limitation, frameworks have been developed to provide verification and scoring for traces, whether by part or in whole (Lee & Hockenmaier, 2025), but these are typically verified in specific domains, typically mathematical reasoning (Mukherjee et al., 2025), or code generation (Jain et al., 2025), where there are well-defined success indicators to verify against.

In this paper, we investigate reasoning-trace evaluation as a problem of establishing semantic equivalence, or a lack thereof. Natural language inference (NLI) provides a useful primitive because entailment is explicitly directional: it evaluates whether one statement follows from another rather than only whether the two statements are similar. We opt to use smaller NLI models in our framework for broader usability; these remain highly competitive against state-of-the-art LLMs Wang et al. (2024b). By leveraging NLI models to construct a directed reasoning graph, which we term

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

as *semantic normalization* (SemNorm), we can evaluate the directed entailment relations between generated statements, and establishing the effectiveness of composite graph scores built from the SemNorm graph.

2 METHODOLOGY

2.1 PRELIMINARIES: SEMANTIC EQUIVALENCE AND LOGICAL ENTAILMENT

Each reasoning trace $R_g = \{s_0, s_1, ..., s_i\}$ composed of i sentences can be represented as an unordered set of sentences $\mathbb{R}^s = \{s_0, s_1, ..., s_j\}, j \leq i$. By passing each medical reasoning query into an LLM, we can obtain generated reasoning traces R_{Gen} and compare them with existing or verified reference reasoning trace R_{GT}. Next, consider that each sentence s_i would need to be de-duplicated into the unordered set. One way this can be done is via atomic decomposition and pairwise equivalence classification. First, an atomic decomposition function f can be applied to convert each sentence into its constituent, indivisible factual statements $f(s_i) = \{a_0, a_1, ..., a_k\}$. Then, equivalence can be established between the atomic-level sets for each reasoning trace $\mathbb{R}^a_{GT} \equiv \mathbb{R}^a_{Gen}$ by matching each $a_k^{Gen} \overset{g}{\equiv} a_k^{GT}$ given a classifier g. With this definition, we can thus define the bounds of our equivalence measurements as such: if $\mathbb{R}^a_{GT} \equiv \mathbb{R}^a_{Gen}$, an equivalence score of 1.0 is assigned, and if $\mathbb{R}^a_{GT} \cap \mathbb{R}^a_{Gen} = \emptyset$ based on having no equivalent constituent items $a_k^{Gen} \overset{g}{\equiv} a_j^{GT}$, a score of 0.0 can be assigned. By this definition, the classifier g should return a score between 0 and 1 for each pair of inputs.

However, this lower bound of 0 is usually difficult to achieve in natural language classification, and potentially undesirable, as useful semantic information may still be present in non-equivalent statements. Consider the two atomic statements a_k^{Gen} = The quick brown fox jumps over the lazy dog and a_k^{GT} = The fast brown dog jumps over the fox. On a cursory inspection, these two statements are semantically distinct and $a_k^{Gen} \not\equiv a_j^{GT}$. Based on the semantic equivalence definition, we should assign a score of 0.0; However, it is clear that there is an entity-level overlap, with the same participants of a fox and a dog in both statements. Thus, it might be preferable for a "good" classifier to return a non-zero equivalence score given the overlap in entities.

When applied to sets of factual atomic statements, the definition for semantic equivalence can also be extended to investigate unsupported generated claims, which exist when $\mathbb{R}^a_{Gen} \cap (\mathbb{R}^a_{GT})^c \neq \emptyset$, and similarly establish a lack of generated support relative to the reference trace when $\mathbb{R}^a_{Gen} \subsetneq \mathbb{R}^a_{GT}$. Thus, we can assign penalization terms based on these set differences for irrelevant generated facts and missing generated support. However, semantic equivalence is not sufficient for establishing whether two reasoning traces are identical; logical entailment also needs to be addressed. Semantic equivalence has been established above as an undirected classification scoring function $g(x, y) \rightarrow [0, 1]$, but factual statements can have unidirectional semantic relations. For example when investigating an acyclic cause c and effect e, the entailment classifier E should return $E(a \rightarrow e) = 1.0$ and $E(e \rightarrow a) = 0.0$. Thus, the entailment classifier provides a direction-conditioned score on pairs of statements.

With these definitions we focus on two key questions: How well do semantic equivalence measures and entailment scorers apply to reasoning trace scoring? And does applying a suitable atomic decomposition function f improve scorer performance, or is a direct evaluation from the raw text sufficient?

2.2 SEMANTIC NORMALIZATION (SEMNORM)

2.2.1 NLI-BASED MODELS

NLI-based models are fine-tuned to return the classification for a given pair of text inputs A and B with the following output classes: Entailment (A entails B), Contradiction (A contradicts B), or Neutral (A has no relation with/is neutral with respect to B). These models are traditionally trained on large-scale datasets such as MNLI (Williams et al., 2018). SemNorm utilizes an existing NLI model, Deberta-large-mnli (He et al., 2021), which is specifically fine-tuned for the NLI task, and applies softmax to the output logits to obtain class probabilities P_e for entailment, P_c for contradiction, and P_n for neutral, which are used to construct textual entailment scores.

2.2.2 ATOMIC EXTRACTION

Given that the classification output of NLI models, the entailment score is highly dependent on the necessary evidence being provided. Thus, we also pre-process the input text into atomic relational statements via an LLM inference step. These statements are meant to be non-decomposable; once a complex reasoning chain is broken down into its atomic constituents, a semantically equivalent chain should also break down into the same atomic statements. A similar LLM pipeline is used to extract atomic (subject, relation, object) triplets for constructing a triplet-level (subject and object nodes, relational edges) semantic and episodic memory graphs from text in prior work (Anokhin et al., 2025). In SemNorm, the pipeline is extended to also output a natural language statement of the triplet which are used as the nodes, and the prompt is updated to include causal statements in addition to relational statements.

2.2.3 NLI-BASED EQUIVALENCE AND SEMNORM CONSTRUCTION

Graph construction is initialized by assigning each atomic statement to a separate node, $\mathbb{N}_0 = \{a_0\}$. For each pair of atomic statements a_0 and a_1, we perform NLI in both directions to estimate $P_e(a_0 \rightarrow a_1)$ and $P_e(a_1 \rightarrow a_0)$. If both entailment probabilities exceed a predetermined threshold t, we treat the statements as equivalent and merge them into a single node, $\mathbb{N} = a_0, a_1$. In contrast, if entailment holds in only one direction, we retain the statements in separate nodes and add the corresponding directed entailment edge, $\mathbb{N}_0 \rightarrow \mathbb{N}_1$ or $\mathbb{N}_1 \rightarrow \mathbb{N}_0$.

For nodes with more than one equivalent statement, an embedding model m is applied to each statement to obtain their corresponding embedding vector $m(a_0) = \boldsymbol{v}_0$, and the closest vector to the centroid of their embedding vectors is assigned as the reference statement for that node. This selection criterion assumes that proximity to the embedding centroid reflects semantic representativeness within the equivalence set; validating this assumption is left to future work.

Thus, by comparing each statement in $\mathbb{R}^a_{GT} = \{a_0, a_1, ..., a_k\}$ and $\mathbb{R}^a_{Gen} = \{a_0, a_1, ..., a_l\}$ in a pairwise manner, and retaining each provenance source $p = GT, Gen$ as a node attribute, we can construct an directed graph with source attributions, merging equivalent statements and establishing logical entailment edges. Furthermore, in addition to P_e, the NLI model also provides P_c and P_n.

In the experiments below, the above 3 probabilities for each pairwise NLI comparison is used for edge construction. Specifically, we define the directed pairwise Semantic Normalization score (SemScore) as $S_{\text{SN}}(a_i, a_j) = P_e(a_i \rightarrow a_j) - P_c(a_i \rightarrow a_j)$ and aggregate these scores using a bidirectional Signed-Max metric that selects the largest-magnitude SN relation for each statement while preserving its sign. This is applied in Experiment 1 and Experiment 1.1. In contrast, SemNorm based on bipartite graph matching (SemNormBP) is reported in Experiment 1.1 and Experiment 2, and operates directly on the NLI probabilities, combining maximal bidirectional entailment support between $\mathbb{R}_{GT} \leftrightarrow \mathbb{R}_{Gen}$ with an all-pairs contradiction penalty. Lastly, the full SemNormKG method in Experiment 3 follows the above thresholded edge construction process to construct a canonical directed graph within $\mathbb{R}^a_{GT}$ nodes, followed by bi-directional merger of each generated node into that base graph using the same threshold. This allows for transitive node merger, where $a^{GT}_{k1} \leftrightarrow a^{Gen}_{l}$ and $a^{GT}_{k2} \leftrightarrow a^{Gen}_{l}$ would merge all 3 items $a^{GT}_{k1} \leftrightarrow a^{Gen}_{l} \leftrightarrow a^{Gen}_{l}$ into one node, even if $a^{GT}_{k1} \leftrightarrow a^{GT}_{k2}$ was not established in the initial construction. Then, edges are materialized for the remainder a^{GT} items only if they successfully exceed the threshold for any of the current merged nodes in the graph. For Experiment 3, the graph is exported back into a SemNorm reasoning trace as context for multi-turn updates, via continuity-aware Kahn topological traversal (Kahn, 1962) on the largest weakly connected component. Full definitions and scoring procedures are described in Algorithm 1-4 in Appendix A.1.

2.3 EVALUATION DETAILS

Overall, we investigate the performance of NLI-based models and derivative scores versus existing embedding-based models and evaluation metrics. The following embedding models are evaluated based on their the performance on STS task on the MTEB leaderboard (Muennighoff et al., 2023): `GIST-small-Embedding-v0` (denoted as GIST) (Solatorio, 2024) and `KaLM-embedding-multilingual-mini-instruct-v2.5` (denoted as KaLM) (Zhao et al., 2025). We restrict our experiments to these models are they are the most comparable in

parameter size to `DeBERTa-Large-MNLI` (He et al., 2021), which we selected as the NLI model, following prior works that established its applicability for semantic equivalence determination (Kuhn et al., 2023; Qiu & Miikkulainen, 2024). The inference LLMs evaluated were `Qwen2.5-32B` (Team, 2024), `Qwen3.5-9B` (Qwen Team, 2026), and `Gemma4-31B` (Team et al., 2026), with serving and sampling configurations reported in Appendix A.

Wherever applicable, the base atomic statements for each reasoning trace are provided as the unit of measurement, along with sentence-level measurements as a secondary ablation source.

Experiment 1: Sentence-level semantic equivalence By investigating sentence-level pairwise entailment, contradiction, and graph construction over 8 expanded sentence types per input sentence, how well do the pairwise similarity methods perform in this controlled setting? For this experiment, K-Soap (Li et al., 2025a) and ACI-Bench (Yim et al., 2023) are used as source datasets for clinical notes, from which sentences are extracted as individual samples for additional perturbation to expand to 8 positive and negative generated sentence types. `Qwen2.5-32B` is used as the synthetic generator for all samples.

Exp1.1: Note-level monotonicity analysis By inserting each input sentence's contradiction sentence back into the clinical note sequentially, we can create increasingly perturbed versions of a given clinical note. This allows us to investigate how well the above scoring methods track the monotonic perturbation progression compared with the original note as perturbation increases. In addition to the embedding-based methods, we also report BertScore (Zhang et al., 2020) and SemNormBP.

Experiment 2: Trace-level evaluations After investigating contradiction perturbation in a controlled setting we investigate trace-level evaluation for LLm-generated explanations against reference explanations. Experiments are conducted on the MedExQA dataset (Kim et al., 2024), which has questions with 2 reference explanations and a final multiple-choice question (MCQ) option as the ground truth label. A chain-of-thought (CoT) prompt is used to evaluate explanation generation, with answer elicitation after generation has completed. Thus, this takes two turns in the standard chat template; however, results are reported as Turn 1 (Explanation Generation, Final Answer). The following metrics are used as baseline performance metrics: Final Answer correctness after explanation generation, ROUGE-L (Lin, 2004), and LLM-as-a-judge (Li et al., 2024; Gu et al., 2025), implemented separately for quality judge against reference explanations and correctness judge against reference explanations + question + 4-options + final answer. Each continuous metric score is binarized by assigning positive class to scores $i >= \mathrm{Median}(I)$ for each base indicator I, and the area under the Receiver Operating Characteristic curve (AUROC) is computed for each method against the binarized indicators. The follow continuous scoring methods are evaluated: word-level lexical F1, ROUGE-L, BertScore, GIST/KaLM with the same bipartite maximum support matching as SemNormBP, Factscore (Min et al., 2023), and SemNormBP.

Experiment 3: Multi-turn information updates From the Turn 1 results obtained above, we extend to a Turn 2 information update. By retrieving context from the reference explanations + and generated explanation, specific evidence which might have been missed in the generation, or exists but would be beneficial to repeat can be introduced. The following methods are evaluated: Baseline consistency prompt adapted from (Li et al., 2025b), Retrieval-Augmented Generation (RAG) with `all-mpnet-base-v2` (Song et al., 2020; HuggingFace, 2025) as the retriever, and SemNormKG in the modes below:

- **Elimination (E)**: All generated node statements that do not have any entailment edges with nodes containing GT statements are removed. Then, the remaining generated node statements are presented to the LLM in Turn 3 as the "user-cleaned" generation. This allows us to simulate an expert user removing irrelevant statements and providing an updated reasoning trace.
- **Retrieval (R)**: For the same SemNorm graph, we retrieve up to K ($K = 2$) GT nodes that do not have any entailment edges with nodes containing generated statements. These uncovered GT atomics are ranked in ascending order by maximum bidirectional entailment support, breaking ties by their source index, and retrieve the first K ($K = 2$), which prioritizes missing information with the weakest entailment support. This mode simulates user-provided external information that the generated explanation missed.

• **SemNorm (S)**: The entire composed SemNorm paragraph is returned as the knowledge update. This simulates a full rewrite of the generated explanation and composition with the ground truth references to provide all relevant information.
• **Adversarial-Retrieval (A)**: The same K statements retrieved are provided as (R), but with a "contradictory" prompt stating that the user is trying to provide misinformation. This mode simulates additional useful external information provided by the user, role-playing as an unhelpful user.
• **SemNormKG as RAG (SemNormRAG)**: The same method as (R), but with the analogous K-threshold tuning from RAG applied.

3 Experiment 1: Sentence-level semantic equivalence

Table 1: **NLI-based models perform better than embedding-based models on 3-chain graph recovery.** Evaluations on KSOAP-Expanded and ACI-Bench-Expanded datasets across pairwise contradiction classification, pairwise entailment classification, and frozen 3-chain graph recovery. Contradiction classification uses two positive and two negative curated pairs per three-sentence sample; entailment classification uses seven original-to-candidate pairs per eight-sentence sample. Three-chain recovery is evaluated over all 56 ordered candidate pairs from eight non-contradiction sentences. Candidate F1 measures recovery of the four canonical directed edges; full-chain recovery is the fraction of samples with all four edges recovered; spurious-edge rate is the fraction of admitted entailment edges that are not canonical edges. For GIST and KaLM, edge acceptance is evaluated against undirected edges; NLI-based methods compare against directed edges.

KSOAP-Expanded				
Model	$\text{AUROC}_{\text{ent}}$	$\text{AUROC}_{\text{contra}}$	Candidate F1	Spurious-edge rate (↓)
GIST	0.68 (0.184)	0.52 (0.014)	0.19 (0.007)	0.85 (0.019)
KaLM	0.69 (0.194)	0.49 (0.015)	0.19 (0.003)	0.86 (0.012)
SemNorm	0.85 (0.003)	0.64 (0.013)	0.38 (0.004)	**0.61 (0.021)**
SemNorm-mednli	0.84 (0.002)	0.68 (0.018)	0.25 (0.005)	0.82 (0.003)
SemNorm-xlarge	**0.86 (0.004)**	**0.69 (0.015)**	**0.40 (0.002)**	**0.61 (0.015)**
ACI-Bench-Expanded				
Model	$\text{AUROC}_{\text{ent}}$	$\text{AUROC}_{\text{contra}}$	Candidate F1	Spurious-edge rate
GIST	0.84 (0.001)	0.56 (0.001)	0.22 (0.002)	0.82 (0.005)
KaLM	0.85 (0.001)	0.53 (0.001)	0.20 (0.001)	0.84 (0.010)
SemNorm	**0.86 (0.001)**	0.74 (0.006)	**0.38 (0.001)**	**0.61 (0.006)**
SemNorm-mednli	0.85 (0.001)	0.70 (0.005)	0.17 (0.000)	0.90 (0.000)
SemNorm-xlarge	**0.86 (0.001)**	**0.75 (0.006)**	**0.38 (0.001)**	**0.61 (0.018)**

3.1 Discussion: Experiment 1

For each pairwise experiment, we see that semantic similarity based on cosine similarity is unable to reliably classify binary labels for entailment and contradiction separately. In the 3-chain recovery, we also see the undirected embedding methods admitting a higher rate of spurious edges, which is reflective of their inability to discriminate between sentence-level entailment of cause to original statement and statement to downstream effect in the presence of distractor sentences in the all-to-all comparison between the set of 8 sentences per original clinical note statement. Additionally, the pairwise and 3-chain experiment revealed model-specific differences for NLI models, with `DeBERTa-v2-Xlarge-mnli` (xlarge) (He et al., 2021) being the best over both datasets, while `PubMedBERT-MNLI-MedNLI` (mednli) (Deka et al., 2023) had markedly higher spurious-edge rate than the other two models.

For the remainder of the experiments, we use the default DeBERTa-large-mnli model for the NLI-based methods (reported in the table as SemNorm).

3.2 EXP1.1: NOTE-LEVEL MONOTONICITY ANALYSIS

Since the original sentences for the above experiment were extracted from clinical notes, each generated contradiction statement can thus be deterministically assigned to a specific clinical note, at the specified sentence index. Thus, we repurposed the (original, contradiction) sentence pairs by reconstructing the clinical notes with increasing number of contradiction sentences inserted. For a clinical note N with k sentences, we can define it as such: $N = (s_0, s_1, ...s_k)$. If the generated contradictions only exist for index 0 and $k-3$, then the set of contradictions can then be represented as $C : (c_o \rightarrow s_0, c_{k-3} \rightarrow s_{k-3})$. For each sentence with a paired contradiction, we can thus replace each s_i with c_i to create a perturbed note $N_{\{c_i\}}$. Thus, by inserting the contradictory sentence recursively, we can monotonically score the semantic equivalence $SE(a, b)$ of the notes as $SE(N, N) = 1.0, SE(N, N_{\{c_0\}}) < SE(N, N), SE(N, N_{\{c_0, c_{k-3}\}}) < SE(N, N_{\{c_0\}})$. This thus allows us to investigate the monotonicity of baseline scorers against each note-level perturbation, with the number of pairwise reference SE scores per note equal to number of contradictions for that node + 1.

A LLM-as-a-judge filtering stage was put in place to verify contradictions were valid on the note-level, because the generator only focused on returning a statement that contradicts at one specific sentence-index based on another sentence index, and is thus not indicative of note-level contradiction. We report the filtering process and the contradictions per note after filtering for KSoap-Expanded and ACIBench-Expanded in Appendix B. For the accepted contradictions we report the note-level metrics below based on the direct raw scores of each method.

3.3 RESULTS AND DISCUSSION: NOTE-LEVEL STAGED PERTURBATION ANALYSIS

As seen from the results, the embedding-based methods track contradictions better on the sentence level compared to NLI-based methods. There is also a clear improvement in monotonic scoring for sentence-level evaluations of semantic equivalence between the source note and the perturbed notes compared to trace-level evaluation, as reported on KaLM. We report the remainder trace-level results and chunk-level results for the other models in Appendix B, but note that all other models exceeded their maximum context window of 512 tokens and the results are thus confounded by this limit on input lengths. This exposes a fundamental flaw of legacy BERT-based models, which is their smaller context windows, while also providing preliminary evidence that the vector space for embedding-based models is well-defined for this monotonic change in semantic content as measured from a given reference.

Table 2: Note-level perturbation monotonicity metrics for KSoap-Expanded and ACIBench-Expanded. Each row reports pairwise comparison scores between the original note against each perturbed note as the number of contradictions increases. Rate is adjacent monotonicity: the proportion of neighboring perturbation steps where score decreases or stays unchanged (n/N). ρ_{sim} is the per-note Spearman rank correlation between model scores across the perturbation ladder and the expected similarity targets $(1 - k/K)$ is averaged across individual note trajectories. All-pairs concordance instead checks every a $<$ b pair of adjacent indexes in a trajectory for score[a] $>$ score[b].

Method (KSoap-E/ACIBench-E)	Adjacent Monotonic Rate	ρ_{sim}	Concordance
GIST-Sentence	*98.70* / **99.07**	*99.94* / **99.90**	*99.87* / **99.89**
KaLM-Sentence	96.52 / 97.31	99.74 / *99.75*	99.64 / 99.66
BERTScore-Sentence	**99.13** / 97.73	**99.97** / 99.47	**99.91** / 99.71
SemNormBP-Sentence	98.26 / *99.03*	99.77 / 99.69	99.69 / *99.87*
GIST-Trace	53.91 / 63.64	79.40 / 86.90	79.65 / 81.89
KaLM-Trace	82.61 / 84.97	92.86 / 96.35	96.28 / 96.19
BERTScore-Trace	64.35 / 66.96	92.52 / 93.82	86.37 / 84.41
SemNormBP-Trace	39.57 / 36.95	78.96 / 76.99	53.74 / 54.56

4 EXPERIMENT 2: TRACE-LEVEL EVALUATIONS

Given that trace-level monotonic degradation behavior was well-modeled by aggregated sentence-level scores for clinical notes, are we able to extend this finding to comparing claims and evidences between reasoning trace sources? This focuses on two questions:

- Is measuring semantic equivalence, or lack thereof, indicative of other aspects of performance for reasoning, such as quality and correctness?
- For the above are aggregated sentence-level measures indicative of trace-level reference labels? Or should trace-level semantic measures be used?

Overall, a pair-wise DeLong test (DeLong et al., 1988) with the directed hypothesis that method 1 (M1) performs better than method 2 (M2) is presented in Figure 1. The test confirms that SemNormBP-Full is sometimes significantly better than other methods on quality (Qwen2.5-32B as judge), with the row-level green cells indicating significantly better than other methods and column-level green cells indicating that other methods were better than that column's method. However, the Full mode was only significantly better than the Trace mode on 4 of 24 experiments. Thus, while the Full mode was strictly superior to the Trace mode (Trace was significantly better on 0 of the 24 experiments), the additional complexity of computing sentence and atomic nodes is not well-supported. The full plots for all indicators and methods are provided in Appendix C.

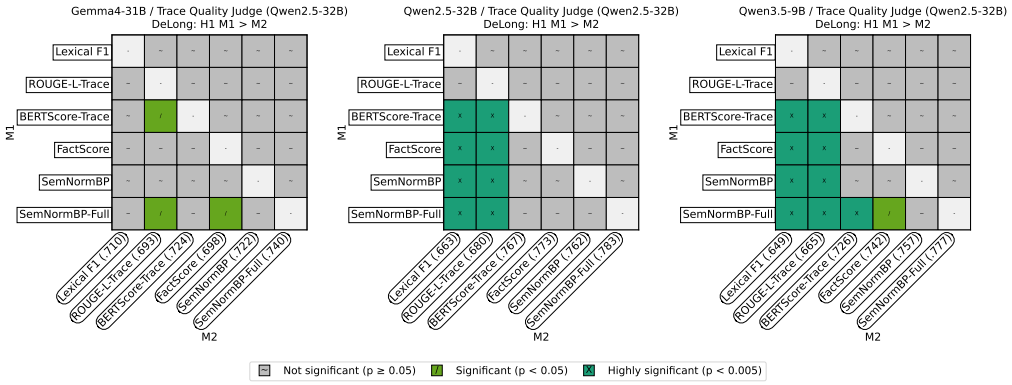

Figure 1: Pairwise DeLong Significance for selected method on Quality Judge ($Q_{2.5}$).

The results presented in Table 3 also support the claim that semantic measures generally outperform lexical measures when compared against binarized reference indicators. However, inference on Gemma4-31B provides the opposite trend, with lexical measures having a better AUROC on 4 of 8 indicators. At the same time, the maximum AUROC for each of the 8 indicator is also lower for Gemma4-31B, which could mean that generated explanations from Gemma4-31B were less lexically and semantically similar to the reference generations. Together with the final answer accuracy of 84.3% reported in Table 4, which was the highest among the 3 inference models evaluated, this provides preliminary but inconclusive evidence that Gemma4-31B might have lower CoT faithfulness (Lanham et al., 2023) in this setting.

5 EXPERIMENT 3: MULTI-TURN INFORMATION UPDATES

5.1 METHODOLOGY

By composing all the available information across reference explanations and the generated explanation for each sample, we can filter for only important and relevant information to include. This is done via standard retrieval-augmented generation with top-K chunks, which retrieves the K highest scoring chunks of context to insert as the "explanation".

The retrieval methods are implemented with a hyperparameter search over $K = [1, 10]$ on the validation set based on highest answering accuracy and a secondary tie-breaker of lowest K, with the following ranking setup for each method:

Table 3: **Aggregating sentence to trace-level NLI generally improves alignment with LLM-As-A-Judge quality and correctness indicators:** Trace-level metrics are reported in the table for all methods, except Full-level, which indicates scoring was aggregated across sentence, atomic, and trace-level nodes. Ans_{acc} is the final answer accuracy measured against ground truth MCQ labels, $ROUGE_{\text{bin}}$ is the binarized ROUGE-L score, C and Q are correctness and quality indicators respectively, and G, Q2.5, and Q3.5 are Gemma4-31B, Qwen-2.5-32B, and Qwen-3.5-9B respectively which are used as judge models for C and Q. All metrics are binarized.

Gemma4-31B								
Method (AUROC)	Ans_{acc}	$ROUGE_{\text{bin}}$	C_{G}	Q_{G}	$C_{\text{Q3.5}}$	$Q_{\text{Q3.5}}$	$C_{\text{Q2.5}}$	$Q_{\text{Q2.5}}$
Lexical F1	0.60	**0.89**	0.61	*0.69*	0.61	*0.69*	0.61	0.71
ROUGE-L	**0.62**	-	**0.62**	0.67	**0.62**	0.67	0.62	0.69
BERTScore	0.57	*0.80*	0.62	**0.69**	0.62	**0.69**	0.62	*0.72*
GIST	0.57	0.74	0.57	0.67	0.57	0.67	0.58	0.71
KaLM	0.57	0.76	0.58	0.67	0.58	0.67	0.59	0.70
FactScore	*0.60*	0.72	0.62	0.67	0.62	0.67	**0.64**	0.70
SemNormBP	0.57	0.65	*0.62*	0.67	*0.62*	0.67	*0.63*	0.72
SemNormBP-Full	0.58	0.67	0.62	0.67	0.62	0.67	0.63	**0.74**
Qwen2.5-32B								
Method	Ans_{acc}	$ROUGE_{\text{bin}}$	C_{G}	Q_{G}	$C_{\text{Q3.5}}$	$Q_{\text{Q3.5}}$	$C_{\text{Q2.5}}$	$Q_{\text{Q2.5}}$
Lexical F1	0.60	**0.90**	0.63	0.72	0.64	0.68	**0.67**	0.65
ROUGE-L	0.61	-	0.65	0.70	0.64	0.69	*0.66*	0.67
BERTScore	0.61	*0.79*	0.72	0.73	0.68	0.73	0.64	0.73
GIST	0.65	0.71	0.75	0.72	0.67	0.72	0.63	0.70
KaLM	0.64	0.72	0.73	0.72	0.65	0.73	0.62	0.71
FactScore	**0.71**	0.63	**0.80**	0.70	*0.69*	0.74	0.64	0.74
SemNormBP	0.64	0.58	0.73	*0.74*	0.67	*0.75*	0.62	*0.76*
SemNormBP-Full	*0.67*	0.61	*0.78*	**0.74**	**0.71**	**0.76**	0.64	**0.78**
Qwen3.5-9B								
Method	Ans_{acc}	$ROUGE_{\text{bin}}$	C_{G}	Q_{G}	$C_{\text{Q3.5}}$	$Q_{\text{Q3.5}}$	$C_{\text{Q2.5}}$	$Q_{\text{Q2.5}}$
Lexical F1	0.55	**0.88**	0.59	0.69	0.59	0.71	0.62	0.66
ROUGE-L	0.56	-	0.62	0.69	0.59	0.69	0.62	0.68
BERTScore	0.55	*0.79*	0.68	0.76	0.64	**0.76**	0.63	0.77
GIST	0.54	0.72	0.64	0.72	0.62	0.70	0.61	0.73
KaLM	0.55	0.73	0.62	0.73	0.61	0.70	0.60	0.72
FactScore	**0.63**	0.68	**0.73**	**0.77**	**0.69**	0.75	*0.65*	*0.77*
SemNormBP	0.61	0.65	0.70	*0.76*	*0.67*	*0.75*	0.64	0.76
SemNormBP-Full	*0.61*	0.65	*0.71*	**0.77**	0.66	0.74	**0.65**	**0.78**

• RAG: Each chunk i is compared against the query Q by computing the cosine similarity of their embeddings after passing through an embedding model. Higher cosine similarity has higher ranking.
• SemNormRAG: Each chunk i is compared against each other chunk j in a pairwise manner using an NLI model. For each eligible chunk, the retrieval score is the maximum bidirectional entailment score with its corresponding chunks, $r(i) = \max_j \frac{P_e(i \to j) + P_e(j \to i)}{2}$. Chunks are ranked in ascending order of (r(i)), such that chunks with lower entailment scores are retrieved first.

5.2 Discussion

The baselines for CoT with the two reference explanations followed by answer elicitation has the highest answering accuracy, with Reference 1 being preferred to Reference 2 in all 3 evaluation models tested. On the other end, the raw generated CoT from each of the models had the lowest answering accuracy. In this direct explanation followed by answering setting, which is termed as Turn 1 (T_1), we attempt to recover performance from the generated explanation towards the reference explanation with various methods by retrieving specific statements from reference explanations and generated explanations. Retrieving with SemNorm-based scoring is on par with standard embedding-based RAG. In Turn 2, we continue the conversation by appending additional context after the generated explanation and answer. In this setting, RAG performed slightly better. Thus, the SemNorm-based

Table 4: Multi-turn accuracy across reference, SemNorm, and RAG configurations. Turn 1 refers to CoT generation and answer elicitation, while Turn 2 is the follow-up information update with direct answer elicitation

Model	T_1^{Gen}	T_1^{Ref*}	$T_1^{SemNormKG[S]}$	$T_1^{SemNormRAG-GT*}$	$T_1^{RAG-GT*}$
Gemma4-31B	84.3	**89.3 [r1]**	85.9	*87.8 [s]*	87.1 [s]
Qwen2.5-32B	79.5	**89.3 [r1]**	84.9	*87.8 [s]*	*87.8 [s]*
Qwen3.5-9B	78.8	**87.7 [r1]**	83.1	85.4 [s]	*87.2 [s]*
$T_1^{Gen} \rightarrow$ **		T_2^{Base}	$T_2^{SemNorm*}$	$T_2^{SemNormRAG-GT*}$	$T_2^{RAG-GT*}$
Gemma4-31B		83.8	86.5 [R]	**87.9 [s]**	*87.8 [s]*
Qwen2.5-32B		77.7	85.0 [R/A/S]	*86.9 [s]*	**87.1 [s]**
Qwen3.5-9B		79.1	84.4 [R]	*86.4 [s]*	**86.9 [s]**

* Starred columns report the best-performing configuration among the corresponding submodes.
For T_1^{Ref*}, [r1] and [r2] denote reference prompts 1 and 2, respectively.
For $T_1^{SemNorm*}[S]$, [S] indicates the full SemNorm-constructed text block. For all {-RAG} methods, [a] denotes atomic statements and [s] sentences as chunks.
For $T_2^{SemNorm*}$, [E], [R], [A], and [S] denote Elimination, Retrieval@K=2, Adversarial-Retrieval, and full SemNorm, respectively.
Multiple subtags, e.g. [A/S] or [A/S/E/R], indicate tied best-performing configurations. $**T_1^{Gen}$ is used as the initial generated response in Turn 1 to which all reported Turn 2 updates are appended to.

scoring is inferior to RAG based on the overall answering accuracy. Interestingly, the baseline method that provided no actual information update had mixed results compared to the Turn 1 accuracy, which provides some evidence to support multi-turn consistency issues as studied in previous works. Overall, the atomic extraction appears to support metric scoring, but was not selected in any of the presented modes; for RAG over all reference + generated statements, atomics had the higher performance in 1 of 12 experiments, but the All mode was inferior to GT-only mode on 11 of 12 paired comparisons between each base method. Full results are presented in Appendix D.

6 Related work

Process reward models (PRMs) To train state of the art large reasoning models that can solve complex problems, a standard component of reward-based reinforcement learning is a separate process reward model, which is typically trained to assign step-level scores to each step of a reasoning trace. This can then be used to validate generated trajectories from reasoning models and optimize them towards obtaining the final answer while also generating the highest scoring reasoning trajectory. While existing works focus on verifying step-by-step mathematical reasoning (Wang et al., 2024a; Mukherjee et al., 2025), with more recent work proposing multi-domain PRMs (Zeng et al., 2025), the development of a generally applicable step-level scorer for the medical reasoning remains outstanding. SemNorm provides an initial approach to composing trace-level scores from step-level evaluations, but its applicability compared to PRMs remains an open area for further investigation, especially with respect to step-ordering and human verification of the recovered reasoning chain.

Factual grounding and verification Fact-based verification metrics target generative issues such as ungrounded attributions (Tang et al., 2025) and hallucinations (Lee & Hockenmaier, 2025). Previous works utilizing NLI-based models have largely focused on measuring factuality in specific tasks such as summarization (Laban et al., 2022) and short-form QA (Honovich et al., 2021), with datasets covering sentence-level verification. In our work, we utilize sentence-level scoring to build an automatic evaluator for trace-level correctness, providing empirical evidence that sentence-level entailment can be utilized to estimate trace-level performance metrics.

7 Conclusion

SemNorm closes the gap in reasoning trace evaluations by establishing the effectiveness of NLI-based graph construction for scoring generated reasoning traces against references. In particular, it outperforms embedding-based models, and provides non-inferior retrieval scoring when applied as an RAG method, setting the stage for further works on directed reasoning evaluation.

8 Reproducibility Statement

The code repository to reproduce the experiments as well as the augmented dataset are scheduled for public release alongside the camera-ready version. Prompt formats, sampling configurations, and experimental details have been included as far as possible, and any missing information required for reproducibility will be made available in the supplemental code release.

9 AI Disclosure

In this work, we used generative AI tools to generate synthetic datasets by augmenting existing datasets and to re-implement method from existing works. We have not used generative AI tools to support the design and refinement of research methodology and experiments, develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in writing mathematical proofs, propose or refine hypotheses, translate content, clean or reformat datasets, conduct qualitative or thematic data analysis, or interpret experimental results.

Additionally, we used generative AI tools to assist with software code, including refactoring of code and the positioning of content for figures, and review portions of the manuscript for stylistic content. All AI-generated synthetic data used in the experiments was reviewed via the the filtering and validation procedures (LLM-As-A-Judge) described in the paper. AI-assisted code was reviewed and tested by the authors, and all AI-assisted manuscript content was reviewed and judiciously added as necessary. We take responsibility for the final content of this work, including all text, claims, and other artifacts produced with the aid of generative AI.

References

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. Arigraph: learning knowledge graph world models with episodic memory for llm agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/2. URL https://doi.org/10.24963/ijcai.2025/2.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. Multiple evidence combination for fact-checking of health-related information. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pp. 237–247, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.bionlp-1.20.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2531595.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL https://aclanthology.org/2021.emnlp-main.619/.

HuggingFace. all-mpnet-v2-hf, 2025. URL https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Arnav Kumar Jain, Gonzalo Gonzalez-Pumariega, Wayne Chen, Alexander M Rush, Wenting Zhao, and Sanjiban Choudhury. Multi-turn code generation through single-step rewards. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aJeLhLcsh0.

Arthur B. Kahn. Topological sorting of large networks. *Communications of the ACM*, 5:558 – 562, 1962. URL https://api.semanticscholar.org/CorpusID:16728233.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. MedExQA: Medical question answering benchmark with multiple explanations. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 167–181, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.14. URL https://aclanthology.org/2024.bionlp-1.14/.

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*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1789–1814, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.94. URL https://aclanthology.org/2025.findings-emnlp.94/.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL https://arxiv.org/abs/2412.05579.

Yizhan Li, Sifan Wu, Christopher Smith, Thomas Lo, and Bang Liu. Improving clinical note generation from complex doctor-patient conversation, 2025a. URL https://arxiv.org/abs/2408.14568.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6679–6700, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.347. URL https://aclanthology.org/2025.findings-acl.347/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023. URL https://arxiv.org/abs/2210.07316.

Sagnik Mukherjee, Abhinav Chinta, Takyoung Kim, Tarun Anoop Sharma, and Dilek Hakkani Tur. Premise-augmented reasoning chains improve error identification in math reasoning with LLMs. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4tYckHNVXV.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LOH6qzI7T6.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Aivin V. Solatorio. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning, 2024. URL https://arxiv.org/abs/2402.16829.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf.

Xuemei Tang, Xufeng Duan, and Zhenguang Cai. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1602–1617, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.83. URL https://aclanthology.org/2025.emnlp-main.83/.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja

Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Yuxia Wang, Minghan Wang, and Preslav Nakov. Rethinking STS and NLI in large language models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 965–982, St. Julian's, Malta, March 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.65. URL https://aclanthology.org/2024.findings-eacl.65/.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Nature Scientific Data*, 2023.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, and Kangwook Lee. VersaPRM: Multi-domain process reward model via synthetic reasoning data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=l19DmXbwPK.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model, 2025. URL https://arxiv.org/abs/2506.20923.

A Appendix: Reproducibility details

All experiments were conducted on 8*Nvidia A100-SXM4-80GB on one A100 server for the Gemma4-31B experiments. The vLLM library (Kwon et al., 2023) is used to serve each model on separate endpoints and the experiments were run in parallel on each endpoint and then consolidated before post-processing. The exact snapshots for Qwen-3.5-9B and Gemma-4-31B are reported in the supplementary code. The sampling parameters used for all LLM inference steps (synthetic data generation, explanation generation, atomic extraction, answer elicitation) follow the recommended configurations on their respective model cards, and are consolidated in Table 5 below.

Table 5: Sampling parameters for LLM inference. RP and PP are Repetition Penalty and Presence Penalty respectively.

Model	**Temp**	**Top-p**	**Top-k**	**Min-p**	**RP**	**PP**	**Thinking**
Gemma4-31B	1.0	0.95	64	–	–	–	No
Qwen3.5-9B	1.0	0.95	20	0.0	1.0	1.5	No
Qwen2.5-32B	0.7	0.8	20	–	1.05	–	No

A.1 Semantic Normalization scoring constructors

Algorithm 1 Shared All-to-All NLI Cache

Require: Reference statements $\mathbb{R}_{\mathrm{GT}} = \{a_1^{\mathrm{GT}}, \ldots, a_n^{\mathrm{GT}}\}$, generated statements $\mathbb{R}_{\mathrm{Gen}} = \{a_1^{\mathrm{Gen}}, \ldots, a_m^{\mathrm{Gen}}\}$, NLI model $\mathcal{M}$
Ensure: Entailment matrix $\boldsymbol{E} \in [0,1]^{(n+m)\times(n+m)}$ and contradiction matrix $\boldsymbol{C} \in [0,1]^{(n+m)\times(n+m)}$

$\mathbb{R} \leftarrow \mathbb{R}_{\mathrm{GT}} \cup \mathbb{R}_{\mathrm{Gen}}$
for $i = 1, \ldots, |\mathbb{R}|$ **do**
  for $j = 1, \ldots, |\mathbb{R}|$ **do**
    $E_{i,j} \leftarrow P_{\mathcal{M}}(E \mid a_i, a_j)$
    $C_{i,j} \leftarrow P_{\mathcal{M}}(C \mid a_i, a_j)$
  end for
end for
▷ $E_{i,j}$ and $C_{i,j}$ correspond to the directed relation $a_i \to a_j$.
▷ All downstream methods reuse $\boldsymbol{E}$ and $\boldsymbol{C}$ without additional NLI calls.
return $(\boldsymbol{E}, \boldsymbol{C})$

Algorithm 2 Experiment 1: Bidirectional Semantic Normalization Score

Require: $\mathbb{R}_{\mathrm{GT}}$, $\mathbb{R}_{\mathrm{Gen}}$, shared NLI matrices $(\boldsymbol{E}, \boldsymbol{C})$ from Algorithm 1
Ensure: Pairwise SemScore matrix $\boldsymbol{S} \in [-1,1]^{m\times n}$

for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, n$ **do**
    $s_{i,j}^{\rightarrow} \leftarrow P_{\mathcal{M}}(E \mid a_i^{\mathrm{Gen}}, a_j^{\mathrm{GT}}) - P_{\mathcal{M}}(C \mid a_i^{\mathrm{Gen}}, a_j^{\mathrm{GT}})$
    $s_{i,j}^{\leftarrow} \leftarrow P_{\mathcal{M}}(E \mid a_j^{\mathrm{GT}}, a_i^{\mathrm{Gen}}) - P_{\mathcal{M}}(C \mid a_j^{\mathrm{GT}}, a_i^{\mathrm{Gen}})$
    $S_{i,j} \leftarrow \frac{1}{2}\left(s_{i,j}^{\rightarrow} + s_{i,j}^{\leftarrow}\right)$
  end for
end for
▷ All probabilities are retrieved from $(\boldsymbol{E}, \boldsymbol{C})$; notation above makes the NLI direction explicit.
return $\boldsymbol{S}$

Algorithm 3 Experiment 2: Full Directed NLI Graph Score

Require: Generated statements $\mathbb{R}_{\mathrm{Gen}}$, reference sets $\mathbb{R}_{\mathrm{GT}}^{(1)}$ and $\mathbb{R}_{\mathrm{GT}}^{(2)}$, shared NLI matrices $(\boldsymbol{E}, \boldsymbol{C})$
Ensure: Pairwise, entailment, contradiction, and full directed-NLI scores

for $q \in \{1, 2\}$ **do**
  $\mathbb{R}_{\mathrm{GT}} \leftarrow \mathbb{R}_{\mathrm{GT}}^{(q)}$
  if $|\mathbb{R}_{\mathrm{Gen}}| = 0$ **or** $|\mathbb{R}_{\mathrm{GT}}| = 0$ **then**
    $s_{\mathrm{pair}}^{(q)} \leftarrow 0$, $s_{\mathrm{ent}}^{(q)} \leftarrow 0$, $c^{(q)} \leftarrow 0$, $s_{\mathrm{full}}^{(q)} \leftarrow 0$
    continue
  end if
  Construct $\boldsymbol{E}^{\mathrm{G}\to\mathrm{R}}, \boldsymbol{E}^{\mathrm{R}\to\mathrm{G}}, \boldsymbol{C}^{\mathrm{G}\to\mathrm{R}}, \boldsymbol{C}^{\mathrm{R}\to\mathrm{G}}$ as the corresponding submatrices of $(\boldsymbol{E}, \boldsymbol{C})$
  $s_{\mathrm{pair}}^{(q)} \leftarrow \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{E_{i,j}^{\mathrm{G}\to\mathrm{R}} + E_{j,i}^{\mathrm{R}\to\mathrm{G}} - C_{i,j}^{\mathrm{G}\to\mathrm{R}} - C_{j,i}^{\mathrm{R}\to\mathrm{G}}}{2}$
  $s_{\mathrm{ent}}^{(q)} \leftarrow \frac{1}{4}\Big(\frac{1}{m}\sum_{i=1}^{m}\max_j E_{i,j}^{\mathrm{G}\to\mathrm{R}} + \frac{1}{m}\sum_{i=1}^{m}\max_j E_{j,i}^{\mathrm{R}\to\mathrm{G}}$
    $+\frac{1}{n}\sum_{j=1}^{n}\max_i E_{i,j}^{\mathrm{G}\to\mathrm{R}} + \frac{1}{n}\sum_{j=1}^{n}\max_i E_{j,i}^{\mathrm{R}\to\mathrm{G}}\Big)$
  $c^{(q)} \leftarrow \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\max\left(C_{i,j}^{\mathrm{G}\to\mathrm{R}}, C_{j,i}^{\mathrm{R}\to\mathrm{G}}\right)$
  $s_{\mathrm{full}}^{(q)} \leftarrow s_{\mathrm{ent}}^{(q)} - c^{(q)}$
end for
$s_{\mathrm{pair}} \leftarrow \frac{1}{2}\left(s_{\mathrm{pair}}^{(1)} + s_{\mathrm{pair}}^{(2)}\right)$
$s_{\mathrm{ent}} \leftarrow \frac{1}{2}\left(s_{\mathrm{ent}}^{(1)} + s_{\mathrm{ent}}^{(2)}\right)$
$s_{\mathrm{con}} \leftarrow -\frac{1}{2}\left(c^{(1)} + c^{(2)}\right)$
$s_{\mathrm{full}} \leftarrow \frac{1}{2}\left(s_{\mathrm{full}}^{(1)} + s_{\mathrm{full}}^{(2)}\right)$
return $(s_{\mathrm{pair}}, s_{\mathrm{ent}}, s_{\mathrm{con}}, s_{\mathrm{full}})$

Algorithm 4 Experiment 3: SemNormKG Construction

Require: Ground-truth atomics $\mathbb{R}_{\mathrm{GT}} = \{a_1^{\mathrm{GT}}, \ldots, a_n^{\mathrm{GT}}\}$, generated atomics $\mathbb{R}_{\mathrm{Gen}} = \{a_1^{\mathrm{Gen}}, \ldots, a_m^{\mathrm{Gen}}\}$, shared NLI matrices $(\boldsymbol{E}, \boldsymbol{C})$, entailment threshold τ
Ensure: Canonical semantic graph $\mathcal{G} = (\mathbb{V}, \mathcal{E})$
$\mathbb{R} \leftarrow \mathbb{R}_{\mathrm{GT}} \,\|\, \mathbb{R}_{\mathrm{Gen}}$
Initialize a disjoint-set structure over $\mathbb{R}$

Construct canonical components
for each pair (a_i, a_j), $i < j$, such that $\{a_i, a_j\} \nsubseteq \mathbb{R}_{\mathrm{Gen}}$ **do**
  if $E_{i,j} \geq \tau$ **and** $E_{j,i} \geq \tau$ **then**
    UNION(i, j)
  end if
end for
$\mathbb{U} \leftarrow$ components induced by the disjoint-set structure
Generated–generated pairs are excluded; union closure permits transitivity.

Select canonical representatives
for each component $\mathbb{U}_\ell \in \mathbb{U}$ **do**
  for each $a_i \in \mathbb{U}_\ell$ **do**
    if $|\mathbb{U}_\ell| = 1$ **then**
      $d_i \leftarrow 1$
    else
      $d_i \leftarrow \frac{1}{|\mathbb{U}_\ell| - 1} \sum_{\substack{a_j \in \mathbb{U}_\ell \\ j \neq i}} \frac{E_{i,j} + E_{j,i}}{2}$
    end if
  end for
  $i^\star \leftarrow \arg\max_{i : a_i \in \mathbb{U}_\ell} d_i$
Break ties by GT provenance, source index, then text.
  Create canonical node $v_\ell \in \mathbb{V}$ represented by $a_{i^\star}$
end for

Materialize directed entailment edges
for each ordered pair $(a_i, a_j) \in \mathbb{R} \times \mathbb{R}$ such that $\{a_i, a_j\} \nsubseteq \mathbb{R}_{\mathrm{Gen}}$ **do**
  if $v(a_i) \neq v(a_j)$ **and** $E_{i,j} \geq \tau$ **then**
    Add $(v(a_i), v(a_j))$ to $\mathcal{E}$ with weight $E_{i,j}$
  end if
end for
Contradiction scores in $\boldsymbol{C}$ remain pair evidence and do not induce graph connectivity.
$\mathcal{G} \leftarrow (\mathbb{V}, \mathcal{E})$
return $\mathcal{G}$

B Appendix: Experiment 1 extended results

B.1 Additional Experiment 1 results for pairwise and 3-chain sentence-level experiments

Table 6: KSOAP-Expanded three-sentence samples: pairwise contradiction classification on two positive and two negative curated pairs per sample

Model	AUROC	Accuracy	Precision	Recall	F1
GIST	0.68 (0.184)	0.59 (0.170)	0.56 (0.138)	0.90 (0.112)	0.67 (0.071)
KaLM	0.69 (0.194)	0.60 (0.174)	0.56 (0.138)	0.91 (0.095)	0.68 (0.079)
SemNorm	0.85 (0.003)	**0.75 (0.004)**	**0.67 (0.007)**	0.84 (0.019)	0.75 (0.005)
SemNorm-mednli	0.84 (0.002)	0.73 (0.017)	0.65 (0.030)	0.83 (0.047)	0.73 (0.004)
SemNorm-xlarge	**0.86 (0.004)**	0.75 (0.006)	0.64 (0.007)	**0.92 (0.011)**	**0.76 (0.004)**

Table 7: KSOAP-Expanded 8-sentence samples: pairwise entailment classification on seven original-to-candidate pairs per sample

Model	AUROC	Accuracy	Precision	Recall	F1
GIST	0.52 (0.014)	0.51 (0.001)	0.51 (0.001)	**0.98 (0.026)**	0.67 (0.006)
KaLM	0.49 (0.015)	0.51 (0.004)	0.50 (0.002)	0.97 (0.039)	0.66 (0.011)
SemNorm	0.64 (0.013)	0.53 (0.019)	0.52 (0.012)	0.96 (0.049)	0.67 (0.006)
SemNorm-mednli	0.68 (0.018)	0.56 (0.016)	0.54 (0.017)	0.92 (0.090)	0.68 (0.016)
SemNorm-xlarge	**0.69 (0.015)**	**0.59 (0.034)**	**0.56 (0.028)**	0.89 (0.075)	**0.68 (0.005)**

Table 8: KSOAP-Expanded 8-sentence samples: frozen 3-chain graph recovery over all 56 ordered candidate pairs from eight non-contradiction sentences. Candidate F1 measures recovery of the four canonical directed edges; full-chain recovery is the fraction of samples with all four edges recovered; spurious-edge rate is the fraction of admitted entailment edges that are not canonical edges. For GIST and KaLM, edge acceptance is evaluated against undirected edges; NLI-based methods compare against directed edges.

Model	Candidate F1	Full-chain recovery	Spurious-edge rate
GIST	0.19 (0.007)	0.02 (0.016)	0.85 (0.019)
KaLM	0.19 (0.003)	0.01 (0.006)	**0.86 (0.012)**
SemNorm	0.38 (0.004)	0.01 (0.006)	0.61 (0.021)
SemNorm-mednli	0.25 (0.005)	**0.11 (0.016)**	0.82 (0.003)
SemNorm-xlarge	**0.40 (0.002)**	0.02 (0.009)	0.61 (0.015)

B.2 Experiment 1.1 reproducibility

Table 9: ACI-Bench-Expanded three-sentence samples: pairwise contradiction classification on two positive and two negative curated pairs per sample

Model	AUROC	Accuracy	Precision	Recall	F1
GIST	0.84 (0.001)	0.73 (0.010)	0.65 (0.019)	0.82 (0.028)	0.72 (0.001)
KaLM	0.85 (0.001)	0.72 (0.003)	0.62 (0.005)	0.89 (0.011)	0.73 (0.001)
SemNorm	0.86 (0.001)	**0.77 (0.003)**	**0.68 (0.005)**	0.88 (0.005)	0.77 (0.001)
SemNorm-mednli	0.85 (0.001)	0.76 (0.004)	0.68 (0.008)	0.83 (0.013)	0.75 (0.001)
SemNorm-xlarge	**0.86 (0.001)**	0.76 (0.001)	0.65 (0.002)	**0.94 (0.006)**	**0.77 (0.001)**

Table 10: ACI-Bench-Expanded 8-sentence samples: pairwise entailment classification on seven original-to-candidate pairs per sample

Model	AUROC	Accuracy	Precision	Recall	F1
GIST	0.56 (0.001)	0.52 (0.005)	0.51 (0.003)	0.98 (0.016)	0.67 (0.001)
KaLM	0.53 (0.001)	0.51 (0.002)	0.50 (0.001)	**0.99 (0.010)**	0.67 (0.002)
SemNorm	0.74 (0.006)	0.64 (0.026)	0.60 (0.023)	0.87 (0.047)	0.71 (0.005)
SemNorm-mednli	0.70 (0.005)	0.59 (0.015)	0.56 (0.017)	0.91 (0.055)	0.69 (0.007)
SemNorm-xlarge	**0.75 (0.006)**	**0.67 (0.009)**	**0.63 (0.013)**	0.83 (0.029)	**0.71 (0.004)**

Table 11: ACI-Bench-Expanded 8-sentence samples: frozen 3-chain graph recovery over all 56 ordered candidate pairs from eight non-contradiction sentences. Candidate F1 measures recovery of the four canonical directed edges; full-chain recovery is the fraction of samples with all four edges recovered; spurious-edge rate is the fraction of admitted entailment edges that are not canonical edges. For GIST and KaLM, edge acceptance is evaluated against undirected edges; NLI-based methods compare against directed edges.

Model	Candidate F1	Full-chain recovery	Spurious-edge rate
GIST	0.22 (0.002)	0.01 (0.002)	0.82 (0.005)
KaLM	0.20 (0.001)	0.01 (0.006)	0.84 (0.010)
SemNorm	0.38 (0.001)	0.01 (0.002)	0.61 (0.006)
SemNorm-mednli	0.17 (0.000)	**0.67 (0.016)**	**0.90 (0.000)**
SemNorm-xlarge	**0.38 (0.001)**	0.02 (0.007)	0.61 (0.018)

Table 12: Metadata for generated contradictions and accepted contradictions on note-level reconstruction

Dataset	KSoap-Expanded	ACI-Bench-Expanded
Quality-gate candidates (all notes)	556	4,608
Candidate notes	20	207
Candidates in evaluated cohort	556	4,602
Evaluated notes	20	203*
Candidates / note (mean ± SD)	27.80 ± 16.94	22.67 ± 13.14
Accepted after judge	230	2,376
Accepted / note (mean ± SD)	11.50 ± 9.14	11.70 ± 6.82
Min accepted / note	1	1
Max accepted / note	44	35
Ladder members	250	2,579
Score rows	2,500	25,790
Primitive pairs	78,060	528,434

* 4 notes had 0 contradictions after filtering; these were excluded from analysis

C Appendix: Experiment 2 extended results

The aggregate statistics for the pariwise Delong comparisons between methods are provided below in Table 13.

Table 13: Directional counts aggregated across 24 plots (8 reference indicators × 3 inference models). Each cell is stacked as $p < 0.05$ / $p < 0.005$; rows are M1 and columns are M2, with the test conducted on $H1 : M1 > M2$.

M1 / M2	Lexical F1	ROUGE-L	BERTScore	GIST	KaLM	FactScore	SemNormBP	SemNormBP-Full
Lexical F1	-	2 / 0	*4 / 3*	4 / 3	4 / 3	*3 / 3*	*4 / 3*	**3 / 3**
ROUGE-L	5 / 3	-	*4 / 3*	7 / 3	7 / 3	*3 / 3*	3 / 3	**3 / 3**
BERTScore	*9 / 7*	*11 / 7*	-	*10 / 4*	9 / 6	**4 / 3**	3 / 3	**3 / 3**
GIST	7 / 3	5 / 2	1 / 0	-	1 / 0	2 / 1	3 / 3	**3 / 3**
KaLM	5 / 4	5 / 2	0 / 0	2 / 0	-	*3 / 2*	3 / 3	**3 / 3**
FactScore	**11 / 9**	*11 / 9*	**5 / 4**	**13 / 7**	**13 / 6**	-	**5 / 4**	*2 / 1*
SemNormBP	*9 / 7*	9 / 7	1 / 1	9 / 3	10 / 2	1 / 0	-	0 / 0
SemNormBP-Full	**11 / 8**	**14 / 8**	*4 / 3*	**13 / 4**	*11 / 7*	*3 / 0*	*4 / 1*	-

D APPENDIX: EXPERIMENT 3 EXTENDED RESULTS

Table 14: Turn 1 accuracy across reference, SemNorm, and RAG configurations.

Model	T_1^{Gen}	T_1^{Ref*}	$T_1^{SemNormKG}$	$T_1^{SemNormRAG-GT*}$	$T_1^{SemNormRAG*}$	$T_1^{RAG-GT*}$	T_1^{RAG*}
Gemma4-31B	84.3	**89.3 [r1]**	85.9	*87.8 [s]*	85.9 [s]	87.1 [s]	86.0 [s]
Qwen-35-9B	78.8	**87.7 [r1]**	83.1	85.4 [s]	83.1 [s]	*87.2 [s]*	83.0 [a]
Qwen2-32B	79.5	**89.3 [r1]**	84.9	*87.8 [s]*	82.7 [a]	*87.8 [s]*	87.0 [s]

* Starred columns report the best-performing configuration among the corresponding submodes.
For T_1^{Ref*}, [r1] and [r2] denote reference prompts 1 and 2, respectively.
For T_1^{RAG*} and $T_1^{SemNormRAG*}$, [a] denotes atomic statements and [s] sentences as chunks.
Multiple subtags, e.g. [r1/r2] or [a/s], indicate tied best-performing configurations.

Table 15: Turn 2 accuracy relative to the Turn 1 generation starting point.

Model	T_1^{Gen}	T_2^{Base}	$T_2^{SemNorm*}$	$T_2^{SemNormRAG-GT*}$	$T_2^{SemNormRAG*}$	$T_2^{RAG-GT*}$	T_2^{RAG*}
Gemma4-31B	84.3	83.8	86.5 [R]	**87.9 [s]**	86.2 [s]	*87.8 [s]*	87.5 [s]
Qwen-35-9B	78.8	79.1	84.4 [R]	*86.4 [s]*	84.0 [s]	**86.9 [s]**	*86.4 [s]*
Qwen2-32B	79.5	77.7	85.0 [R/A/S]	86.9 [s]	82.1 [s]	*87.1 [s]*	**87.3 [s]**

T_1^{Gen} is used as the initial generated response in Turn 1 to which all reported Turn 2 updates are appended to.
* Starred columns report the best-performing configuration among the corresponding submodes.
For T_2^{RAG*} and $T_2^{SemNormRAG*}$, [a] denotes atomic statements and [s] sentences as chunks.
For $T_2^{SemNorm*}$, [E], [R], [A], and [S] denote Elimination, Retrieval, Adversarial-Retrieval, and full SemNorm, respectively.
Multiple subtags, e.g. [A/S] or [A/S/E/R], indicate tied best-performing configurations.