Do LLMs Understand Context? A Knowledge Graph-Based Evaluation Framework

Subavarshana Arumugam[1*] **Mamta Nallaretnam**[1*] **Kithuni Wickramasinghe**[1*]
Chamath Gunapala[1*] **Pragatheeswaran Vipulanandan**[2] **Kamal Premaratne**[2]
Uthayasanker Thayasivam[2]

[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]Department of Electrical and Computer Engineering, University of Miami, USA
{subavarshanaa.21, nallaretnam.21, kithuni.21, chamathg.21}@cse.mrt.ac.lk

Abstract

While large language models (LLMs) have achieved remarkable linguistic capabilities, a profound question lingers at their core: do these models truly comprehend context or simply excel at pattern matching on an unprecedented scale? Contextual understanding in LLMs refers to the ability to correctly extract relevant information from a given context, integrate it into a coherent internal representation, and reason over it to produce factually consistent and contextually grounded responses. However, traditional methods such as BiLingual Evaluation Understudy (BLEU) and perplexity simply measure surface-level performance. This reveals a critical gap in question answering (QA), where responses must be contextually grounded rather than simply being memorized associations. To fill this void, we propose a novel knowledge graph (KG) based evaluation framework for LLM contextual understanding in QA. Central to this is Semantic Structural Similarity for KGs (S3KG), a hybrid similarity measure combining structural and semantic signals into a single score. In addition, a diagnostic analysis framework is developed to identify and categorize reasoning errors at the triplet level, enabling fine-grained analysis of model failures. Together, across nine benchmarks, S3KG achieves F1 gains of up to $+7.6$ points over the strongest baseline and AUROC up to 0.973.

1 Introduction

LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) (Reimers and Gurevych, 2019), the Generative Pre-trained Transformer (GPT) models (Radford et al., 2018), and their advanced variants, have fundamentally transformed natural language processing (NLP). These models exhibit extraordinary proficiency in generating coherent, human-like text, answering complex questions, and executing a broad spectrum of tasks (Brown et al., 2020). However, a profound question lingers at the core of these models: do they genuinely understand the content they process or do they merely produce plausible outputs without true contextual comprehension (Zhu et al., 2024). Hallucination detection literature has largely bypassed this dimension, focusing instead on output-level signals such as semantic entropy and token sequence probabilities (Vipulanandan et al., 2026).

Traditional evaluation metrics—such as perplexity, BLEU (Papineni et al., 2002), and other token-level matching methods on standardized benchmarks—primarily measure syntactic- or surface-level performance and fail to capture the depth of semantic comprehension or contextual understanding (Reiter, 2018). This limitation raises concerns about the reliability of LLMs in high-stakes scenarios such as QA systems in medicine and healthcare, defense, and legal analysis.

2 Related Work

Evaluating how LLMs understand and utilize contextual information remains a key challenge in QA systems, as fluent and plausible responses do not necessarily reflect faithful use of the provided context (Wallat et al., 2024). Zhu et al. (2024) evaluate LLMs across four core tasks—coreference resolution, discourse relation classification, dialogue state tracking, and query rewriting—showing that while they capture general contextual patterns, they fail to reveal which parts of the context are misunderstood, and their analysis does not extend to the QA domain. Complementing this, Yan et al. (2024) probe LLM reasoning by manipulating in-context examples, including logical modifications such as swapping "AND"/"OR", finding that models do not consistently obey formal reasoning rules nor exhibit clearly identifiable error patterns. Together, these works establish that current LLMs demonstrate strong surface-level comprehension yet still

[*]The first authors contributed equally to this work.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 479–492
November 6-10, 2026 ©2026 Association for Computational Linguistics

struggle with fine-grained contextual interpretation and logical reasoning. Evaluating these limitations in long-form LLM answers is particularly challenging, making the use of KGs which involve accurate extraction of relational triplets from text a promising evaluation approach.

KG construction has progressed from fixed-schema supervised pipelines to joint extraction architectures enabled by pretrained language models (Shang et al., 2022). Embedding models such as TransE (Bordes et al., 2013), which represents relations as vector translations, and RotatE (Sun et al., 2019), which models relations as complex-space rotations to capture symmetry and composition, makes them well-suited for link prediction on static KGs but ill-suited for cross-graph similarity. The Weisfeiler-Lehman (WL) kernel (Shervashidze et al., 2011) compares graphs by iteratively aggregating neighbourhood labels into histograms, while the Wasserstein WL (WWL) kernel (Togninalli et al., 2019) replaces histogram comparison with Wasserstein distance for better handling of continuous attributes. Both treat node labels as opaque symbols, so semantically equivalent but syntactically distinct labels receive zero credit—a key limitation. Haskins and Adams (2025) partially address this through SBERT-based (Reimers and Gurevych, 2019) semantic clustering with few-shot instruction tuning, yet the alignment remains lossy and falls short of a principled similarity measure for heterogeneous, independently constructed KGs.

3 Our Contributions

Our work makes three main contributions.

- **Semantic Structural Similarity for KGs (S3KG)** is a hybrid semantic-structural similarity metric that converts LLM responses and reference answers into KG triplets and produces a single interpretable evaluation score.
- **Contextual Understanding Score (CUS)** is a model-level aggregate of two complementary dimensions, factual accuracy (GoldSim) and contextual faithfulness (CtxSim), enabling cross-model comparison across benchmarks.
- **Triplet Analyzing Unit (TAU)** is a diagnostic component that identifies and classifies reasoning failures at the triplet level for fine-grained behavioral analysis of model outputs.

4 Methodology

Our framework evaluates LLM contextual understanding using a KG-based pipeline (see Figure 1). Given a question and its supporting context, the LLM generates a response, from which KGs are constructed alongside those derived from the gold (or reference or ground truth) answer and context. These KGs are then compared to measure similarity. Low scoring pairs are further analyzed using a triplet analyzing unit to identify reasoning errors.

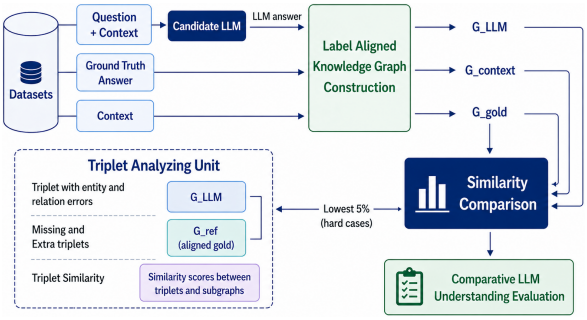


Figure 1: Methodology pipeline for LLM comparison and evaluation.

4.1 LLM Answer Collection

Each model is prompted with a question paired with its associated supporting context, and the generated response is recorded alongside the gold answer to form the inputs for downstream evaluation. To establish a reproducible baseline, responses are first generated deterministically at temperature zero, yielding outputs that closely adhere to the provided context; subsequent runs are at progressively higher temperature settings allowing for us to examine how increasing generation diversity affects the model's ability to retain and utilize contextual information. These collected responses—together with the gold answers and supporting contexts—feed directly into the KG construction stage.

4.2 KG Construction

For each QA instance, knowledge graphs are constructed from three sources: the gold answer, the model-generated response, and the supporting context. To ensure that the resulting KGs are comparable, we adopt the single few-shot prompting strategy with instruction tuning used in Sansford et al. (2024) and Haskins and Adams (2025), applying a shared extraction prompt uniformly across all three sources. This ensures that entities and relations are extracted under the same schema, encouraging a

consistent entity and relation label space across all 3 KGs. Full details of the extraction prompt appear in Appendix B.

Following extraction, an additional NLP normalization step is applied uniformly to every entity and relation label across all 3 KGs. This includes lowercasing, lemmatization, and whitespace normalization, ensuring that any residual syntactic variation introduced during extraction is not retained in the final KG representations. Together, the consistency-aware prompting and post-extraction normalization ensure that the 3 KGs are structurally compatible and ready for meaningful comparison using S3KG.

4.3 S3KG: Semantic Structural Similarity for KGs

We denote a KG as $\mathcal{G} = \mathcal{G}(\mathcal{T})$, where $\mathcal{T} = \{(h_i, r_i, t_i),\ i \in \mathcal{I}\}$ is a collection of triplets enumerated via a finite index set $\mathcal{I}$. In the i-th triplet (h_i, r_i, t_i), h_i, r_i, and t_i are the head entity, relation, and tail entity, respectively.

S3KG computes the similarity between two KGs $\mathcal{G}_1(\mathcal{T}_1)$ and $\mathcal{G}_2(\mathcal{T}_2)$. Here, for $k = 1, 2$, $\mathcal{T}_k = \{(h_{k,i}, r_{k,i}, t_{k,i}),\ i \in \mathcal{I}_k\}$. Rather than comparing KGs by a single criterion, S3KG operates at two levels: (1) a *structural score* S_{WL} computed node- and edge-wise using the WL kernel over soft-label aligned KGs, and (2) a *semantic score* $\mathrm{S}_{\mathrm{SBERT}}$ computed triplet-wise from mean-pooled SBERT embeddings (Reimers and Gurevych, 2019). These are blended via a mixing coefficient α to get the final combined similarity score (see (2)).

Triplet-Level Matching. Triplets are serialised as natural language (NL) strings. SBERT embeddings are computed for all triplets in both sets $\mathcal{T}_1$ and $\mathcal{T}_2$ using `paraphrase-MPNet-base-v2`. For each triplet in $\mathcal{T}_1$, the most semantically similar triplet in $\mathcal{T}_2$ is selected by cosine similarity, producing a filtered set $\widehat{\mathcal{T}}_2 \subseteq \mathcal{T}_2$ that anchors the comparison to semantically relevant content. This unidirectional matching strategy is adopted deliberately. By treating $\mathcal{T}_1$ as the reference set, the method primarily evaluates how well the content of $\mathcal{T}_1$ is covered by $\mathcal{T}_2$, thereby emphasising recall while not penalising additional triplets present in $\mathcal{T}_2$. In contrast, a symmetric bidirectional formulation—computed by averaging matches from both $\widehat{\mathcal{T}}_2$ and $\widehat{\mathcal{T}}_1$, similar to the F1 formulation of BERTScore (Zhang et al., 2020)—would account for both recall and precision by also penalising unmatched surplus triplets. Investigating the empirical differences between the unidirectional and bidirectional variants is left for future work.

Soft Label Alignment. The Standard WL kernel compares graphs by matching node labels exactly: two labels contribute to the similarity score only if they are syntactically identical strings. This means that semantically equivalent but syntactically different entity or relation labels (e.g., *"found"* and *"discovered"*) are treated as entirely distinct, and no credit is awarded for any semantic equivalence. S3KG resolves this through a soft label alignment step applied before kernel computation. Each node in $\mathcal{G}$ carries an entity label, and each edge carries a relation label. These are aligned independently to prevent cross-type collisions: for each entity label ℓ in $\mathcal{G}_1$, if its maximum cosine similarity, computed via SBERT embeddings $\phi(\cdot)$ to any entity label in $\mathcal{G}_2$ exceeds a threshold $\tau = 0.65$, it is mapped to the canonical identifier of the best-matching label in $\mathcal{G}_2$ (e.g., `node_0`, `node_1`); otherwise it is left unchanged. The same procedure is applied independently to relation labels (e.g., `rel_0`, `rel_1`). After alignment, syntactically different but semantically equivalent labels share the same canonical identifier, allowing the WL kernel to recognise them as matching.

WL Kernel Structural Similarity. Following soft label alignment, the two KGs are compared using the WL graph kernel (Shervashidze et al., 2011). The WL kernel operates iteratively: at iteration 0, each node is characterised by its initial (soft aligned) entity label. At each subsequent iteration k, every node aggregates its current label with the multiset of its neighbours' labels and the connecting relation labels, producing a new refined label that encodes the node's k-hop neighbourhood structure. We use $K = 5$ iterations, so each node's final label summarises structural patterns up to 5 hops away. The kernel score is the normalised inner product between the resulting label-count histograms of the two graphs, yielding a structural similarity score $\mathrm{S}_{\mathrm{WL}} \in [0, 1]$ regardless of KG size.

SBERT Mean-Pool Semantic Similarity. Each triplet $(h, r, t) \in \mathcal{T}$ is encoded by SBERT into an embedding $\phi(h, r, t)$; the graph-level representation is the mean $\bar{\mathbf{e}}_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{T}} \phi(h, r, t)$. Semantic similarity is the cosine between the mean-pooled representations of $\mathcal{T}_1$ and $\widehat{\mathcal{T}}_2$ (the semantically filtered reference triplets from Step 1), clipped to

$[0, 1]$:

$$\mathrm{S}_{\mathrm{SBERT}}(\mathcal{T}_1, \widehat{\mathcal{T}}_2) = \max\left(0, \frac{\bar{\mathbf{e}}_{\mathcal{T}_1} \cdot \bar{\mathbf{e}}_{\widehat{\mathcal{T}}_2}}{\|\bar{\mathbf{e}}_{\mathcal{T}_1}\| \cdot \|\bar{\mathbf{e}}_{\widehat{\mathcal{T}}_2}\|}\right). \quad (1)$$

This captures sentence-level meaning that discrete WL label refinement cannot.

Combined Score. The structural and semantic scores are combined via mixing coefficient α as

$$\mathrm{S}_{\mathrm{S3KG}} = (1-\alpha)\,\mathrm{S}_{\mathrm{WL}} + \alpha\,\mathrm{S}_{\mathrm{SBERT}}.\ \alpha \in [0, 1]. \quad (2)$$

We use $\alpha = 0.5$ so that the score equally weights structural fidelity from WL neighbourhood aggregation over aligned labels and semantic coherence from SBERT mean-pooling, thus encoding both local relational patterns and global meaning within a single score.

4.4 Contextual Understanding Score (CUS)

With the 3 KGs—KG_{LLM} associated with the model-generated response, KG_{gold} associated with the gold answer, and KG_{ctx} associated with the supporting context—in hand, for each QA instance q, we apply S3KG to get (1) $\mathrm{GoldSim}(q) = \mathrm{S}_{\mathrm{S3KG}}(KG^{(q)}_{\mathrm{LLM}}, KG^{(q)}_{\mathrm{gold}})$ which measures factual accuracy by comparing the LLM response against the gold answer; and (2) $\mathrm{CtxSim}(q) = \mathrm{S}_{\mathrm{S3KG}}(KG^{(q)}_{\mathrm{LLM}}, KG^{(q)}_{\mathrm{ctx}})$ which measures contextual faithfulness by comparing the LLM response against the supporting context. Since neither dimension alone reflects true understanding, we employ the harmonic mean to generate a *Contextual Understanding Score (CUS)* as

$$\mathrm{CUS}(q) = \frac{2 \cdot \mathrm{GoldSim}(q) \cdot \mathrm{CtxSim}(q)}{\mathrm{GoldSim}(q) + \mathrm{CtxSim}(q)}. \quad (3)$$

This harmonic mean penalises imbalanced profiles, ranking a model having one strong and one weak score below one having a pair of moderate scores. The dataset-level CUS is the mean of $\mathrm{CUS}(q)$ over all N samples.

4.5 Triplet Analysis Unit (TAU)

For the 5% of lowest-scoring QA pairs, we apply a triplet analysis unit (TAU) to identify where and how the LLM generated KG diverges from the gold KG. Each triplet is converted into an NL sentence and encoded using a sentence transformer model. Cosine similarity is computed between gold and LLM triplet embeddings; aligned triplets are identified by thresholding the cosine similarity between triplet sentence embeddings.

After removing aligned triplets, residual pairs are categorized into interpretable error classes based on component-wise cosine similarities for head, relation, and tail: (1) *Relation mismatch*: entities match, but the relation differs. (2) *Entity mismatch*: the relation aligns, but the entity pair is inconsistent. (3) *Extra triplets*: hallucinated or additional triplets generated by the LLM. (4) *Missing triplets*: relevant triplets that were not extracted by the LLM. Full TAU evaluation details appear in Appendix C.

5 Experiments

5.1 Datasets

Two context-rich QA datasets containing long-form answers are used for evaluation purposes. (1) **PubMedQA** (Jin et al., 2019) contains 273,518 biomedical QA pairs drawn from research articles, with answers typically exceeding 100 words, providing a testbed for domain-specific detailed response evaluation. (2) **MesaQA** (Wang et al., 2025) comprises approximately 6,100 QA pairs from consumer healthcare documents, featuring abstractive answers averaging 70 words that require multi-span evidence integration. Together, these two datasets provide evaluation coverage across academic biomedical reasoning and practical healthcare knowledge synthesis.

5.2 LLM Answer Collection

We evaluate 4 instruction-tuned open-source language models with 7-billion parameters: `Llama-2-7b-chat-hf`, `Gemma-7b-it`, `Mistral-7B-Instruct-v0.2`, and `Falcon-7B-Instruct`. These specific models are selected because they operate at a similar scale, thus allowing for a fair comparison without the influence of model size. Each model is prompted with a question and its associated supporting context, and the model-generated response is recorded alongside the gold answer for evaluation. To establish a baseline, responses are first generated deterministically at 0 temperature, yielding outputs that closely adhere to the provided context. Additional responses are then generated at higher temperature settings (we use 0.3, 0.7, and 1.0) to examine how increasing generation diversity influences the model's ability to retain and utilize contextual information. Full temperature analysis appears in Appendix A.

5.3 Benchmarking Dataset Collection

To assess generalisation across diverse text types, we use 10 datasets spanning three structural categories. Each dataset is cast as a binary classification task: given a text pair (s_1, s_2), predict whether the pair is semantically equivalent (label $= 1$) or not (label $= 0$). All datasets are balanced at N positive and N negative pairs (we use $N = 400$). Performance is measured by maximum F1 score (obtained via threshold sweep) and AUROC.

Short-Text with Human-Annotated. We use 3 datasets containing sentence pairs averaging 10–22 words with crowd-sourced or expert equivalence labels: **MRPC** (Dolan and Brockett, 2005), consisting of news sentence pairs with paraphrase labels; **PAWS-Wiki** (Zhang et al., 2019), adversarially constructed paraphrase pairs from Wikipedia where lexical overlap is deliberately an unreliable signal; and **STS12** (Agirre et al., 2012), sentence similarity pairs drawn from multiple NLP tasks.

KG-Perturbed Paragraphs. We use 6 datasets containing paragraphs averaging 69–126 words, constructed by perturbing entity relationships in KG-derived paragraph representations. The 5 evaluated datasets—**SK-Codex 400**, **SK-Combined**, **SK-FindKG**, **SK-GloBI**, and **SK-Oregano**—differ in their underlying KG source, covering general encyclopaedic (Codex (Safavi and Koutra, 2022)), financial/economic (FindKG (Li and Sanna Passino, 2024)), biological interaction (GloBI (Poelen et al., 2014)), and food ontology (Oregano (Boudin et al., 2023)) domains.

NLP-Perturbed Paragraphs. NLP-Perturbed Paragraphs. The **Wikipedia Entity-Swap** dataset (399 pairs) replaces named entities in Wikipedia passages using four NLP-based perturbations—node replacement, node deletion, edge deletion, and edge replacement via WordNet antonyms (Miller, 1992)—with no KG involvement at any stage (Rico et al., 2016; Wei and Zou, 2019). It serves as an anti-circularity probe: were our gains an artifact of circular evaluation, performance here should collapse, but it does not.

5.4 Methods Evaluated

S3KG is evaluated with a mixing coefficient α (see (2)) which controls the blend between KG structural signal and sentence-transformer signal: $\alpha = 0.0$ recovers a pure KG structural embedding ; $\alpha = 1.0$ recovers a pure dense sentence-transformer representation. Pure KG structural embeddings capture relational and ontological structure between concepts but lack linguistic flexibility and contextual expressiveness, whereas sentence transformers excel at contextual and semantic similarity yet remain blind to the underlying KG topology. S3KG bridges this gap by interpolating between both signals, enabling richer matching that is sensitive to both conceptual structure and NL meaning. For each dataset, we report the best-performing variant selected by maximum F1 across the sweep $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. Results for a more complete per-dataset α sweep appear in Appendix E. We compare against 7 standard baselines: ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), MiniLM (Wang et al., 2020), and sentence-T5-base (Ni et al., 2022). Table 1 provides a concrete worked example demonstrating how S3KG detects relational reversals that all 7 baselines fail to distinguish.

All experiments[1] and were conducted on a workstation equipped with 64 GB RAM and an NVIDIA A6000 GPU.

6 Results

We evaluate S3KG against seven baselines on 9 benchmarks spanning short-text paraphrase detection, KG-perturbed paragraphs, and an anti-circularity entity-swap control. Table 2 summarises F1 and AUROC across every dataset $\times$ method cell; the per-dataset best α together with the headline scores appear in Table 3. Full per-dataset performance tables (short-text, KG-perturbed paragraph, Wikipedia entity-swap) appear in Appendix D; the complete α sweep appears in Appendix E.

Text Richness Drives KG Performance. S3KG performance scales with text length and relational density. On short texts (MRPC, STS12; 10–22 words average), the structural KG signal is sparse because only a few well-formed triplets can be extracted, and S3KG is competitive but trails sentence-T5-base by 6–7 F1 points. On paragraph-level datasets (69–126 words), S3KG reaches top-1 performance on 4 of 5 KG-perturbed benchmarks—SK-Codex 400 (F1 $= 0.932$, $+5.7$ over MiniLM), SK-Combined (0.834, $+6.4$ over sentence-T5-base), SK-GloBI (0.892, $+7.6$ over BERTScore), and SK-Oregano (0.812, $+1.2$ over sentence-T5-base)—confirming that sufficient rela-

[1]are available at: https://github.com/aaivu/knowledge-xtraction

Table 1: Similarity scores on two PAWS-Wiki pairs (threshold $= 0.5$;✓ = correct, × = incorrect).The positive pair differs only in word order; the negative pair swaps the subject and object of the winning relation. All seven baselines assign near-identical high scores to both pairs, failing on the negative example. For the **positive pair**, S3KG achieves the optimal score of 1.00, whereas surface-form methods such as BLEU (0.58) and ROUGE-L (0.85) underperform by penalising inconsequential word-order variation.For the **negative pair**, S3KG scores 0.48—the only sub-threshold result—correctly predicting *Not Similar*. The score is not zero because the sentences share substantial content; only the relational direction differs. The KG component isolates this reversal via the directed triplet, reducing the score from the near-1.0 surface baseline to just below the decision threshold, while $\alpha = 0.5$ balances surface and structural similarity.

	Positive Example	**Negative Example**
Text s_1	*His father returned as a finished violinist of the Russian School to Bombay.*	*Renzo Furlan won 6–3, 6–4 against Thomas Johansson in the finals.*
Text s_2	*His father returned to Bombay as a finished violinist of the Russian school.*	*Thomas Johansson won 6–3, 6–4 against Renzo Furlan in the finals.*
Triplet (s_1)	(father, returned_to, Bombay)	(Renzo Furlan, won_against, Thomas Johansson)
Triplet (s_2)	(father, returned_to, Bombay)	(Thomas Johansson, won_against, Renzo Furlan)
True Label	**1** (Similar)	**0** (Not Similar)
Similarity scores (✓ = correct prediction; × = incorrect prediction)		
S3KG (Ours)	**1.00** ✓	**0.48** ✓
ROUGE-1	1.00 ✓	1.00 ×
ROUGE-2	0.83 ✓	0.75 ×
ROUGE-L	0.85 ✓	0.69 ×
BLEU	0.58 ✓	0.70 ×
BERTScore	0.98 ✓	0.98 ×
MiniLM	1.00 ✓	0.94 ×
sentence-T5-base	1.00 ✓	0.99 ×

tional content is required for the structural channel to pay off. On the adversarial PAWS-Wiki paraphrase benchmark, S3KG still leads (F1 $= 0.766$) while ROUGE-1 collapses to near-random (AUC $= 0.490$), reflecting the well-known failure of n-gram overlap on surface-form-preserving rephrasings.

KG Extraction Quality is a Bottleneck. On SK-FindKG, where financial and economic vocabulary degrades triplet extraction quality, sentence-T5-base leads (F1 $= 0.848$) and S3KG drops to 0.767. The best α for this dataset is 0.0 (pure KG), but the absolute score remains capped by noisy triplets, evidence that S3KG's gains depend on the underlying graph being faithfully recoverable from text.

Anti-Circularity Validation. On the Wikipedia Entity-Swap control, which is constructed independently of any KG used during S3KG development, S3KG attains the best meaningful score (F1 $= 0.872$, AUC $= 0.890$), exceeding MiniLM (F1 $= 0.821$) and sentence-T5-base (F1 $= 0.762$). ROUGE-1 is excluded from the meaningful comparison because entity-swapped pairs share nearly all surrounding tokens, making unigram overlap trivially near-perfect, a dataset artifact rather than a real signal. The result rules out circularity as an explanation for S3KG's KG-perturbed gains.

Optimal α is Dataset-Dependent. Lower α values favour datasets where perturbations are primarily structural (SK-FindKG: $\alpha = 0.0$; STS12, Wiki Swap: $\alpha = 0.1$). Higher values are preferred when KG and dense signals are complementary (SK-Codex 400 and SK-Combined: $\alpha = 0.5$; SK-GloBI: $\alpha = 0.6$). The full sweep appears in Appendix E.

Baseline Behaviour. Token-overlap baselines (ROUGE, BLEU) are strong on KG-perturbed datasets where perturbations alter surface form, but unreliable on adversarial datasets (PAWS-Wiki: ROUGE-1 AUC $= 0.490$; Wiki Swap: ROUGE-L AUC $= 0.311$). BERTScore is more stable but consistently underperforms S3KG on KG-perturbed data. MiniLM and sentence-T5-base are

Table 2: F1 Score and ROC-AUC of S3KG and seven baselines across all nine benchmark datasets. The best-performing method per dataset is **bolded**. S3KG is shown using the best α variant per dataset, selected by maximum F1 over α swept over $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$ (full sweep in Appendix E). S3KG attains the highest score on 6 of 9 datasets and the highest *meaningful* score on the Wikipedia Entity-Swap anti-circularity control, with gains of up to $+7.6$ F1 over the strongest baseline on KG-rich paragraph datasets.

F1 Score

Dataset	S3KG (Ours)	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	MiniLM	sent-T5-base
MRPC	0.692	0.745	0.720	0.729	0.687	0.758	0.723	**0.765**
PAWS-Wiki	**0.766**	0.678	0.715	0.735	0.716	0.691	0.687	0.674
Semantic-KG Combined	**0.834**	0.732	0.707	0.722	0.715	0.757	0.770	0.774
Wiki Swap	0.872	**1.000**	0.860	0.729	0.868	0.747	0.821	0.767
Semantic-KG Codex 400	**0.932**	0.835	0.822	0.792	0.806	0.823	0.875	0.872
Semantic-KG FindKG	0.767	0.745	0.717	0.719	0.711	0.739	0.802	**0.848**
Semantic-KG GloBI	**0.892**	0.784	0.776	0.763	0.775	0.816	0.780	0.730
Semantic-KG Oregano	**0.812**	0.745	0.752	0.792	0.745	0.743	0.773	0.800
STS12	0.786	0.725	0.681	0.703	0.671	0.682	0.833	**0.856**

AUROC

Dataset	S3KG (Ours)	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	MiniLM	sent-T5-base
MRPC	0.673	0.784	0.721	0.760	0.677	**0.816**	0.748	**0.816**
PAWS-Wiki	0.795	0.490	0.721	**0.807**	0.747	0.702	0.638	0.668
Semantic-KG Combined	**0.829**	0.728	0.711	0.717	0.708	0.792	0.789	0.828
Wiki Swap	**0.890**	1.000	0.772	0.311	0.745	0.645	0.811	0.790
Semantic-KG Codex 400	**0.973**	0.917	0.894	0.855	0.884	0.916	0.943	0.945
Semantic-KG FindKG	0.796	0.745	0.706	0.721	0.710	0.761	0.844	**0.902**
Semantic-KG GloBI	**0.935**	0.833	0.833	0.800	0.819	0.871	0.817	0.761
Semantic-KG Oregano	**0.892**	0.782	0.791	0.835	0.794	0.798	0.814	0.871
STS12	0.834	0.754	0.656	0.710	0.644	0.636	0.894	**0.928**

Table 3: Best S3KG variant per dataset, selected by maximum F1 via grid search over $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$. KG-rich paragraph datasets gain up to $+7.6$ F1 over the strongest baseline; sparse or noisy KGs yield reduced margins.

Dataset	Best α	F1	AUROC
MRPC	0.3	0.692	0.673
PAWS-Wiki	0.5	0.766	0.795
STS12	0.1	0.786	0.834
SK-Codex 400	0.5	0.932	0.973
SK-Combined	0.5	0.834	0.829
SK-FindKG	0.0	0.767	0.796
SK-GloBI	0.6	0.892	0.935
SK-Oregano	0.4	0.812	0.892
Wiki Swap	0.1	0.872	0.890

the strongest baselines overall but require full fine-tuned transformer inference, whereas S3KG's KG component is comparatively lightweight at inference time.

6.1 LLM Comparison on QA Datasets

Having validated S3KG as a reliable KG similarity measure, we apply it as an evaluation instrument to address the following question: *given a question and its supporting context, to what extent does an LLM capture the relational knowledge of the reference answer, and how faithfully does its response reflect the provided context?* Traditional metrics such as BLEU or exact match are insufficient for this purpose, as they assess syntactic-level token overlap rather than the relational knowledge structure of a response.

Using the CUS evaluation pipeline in Section 4.4, Table 4 reports mean GoldSim, CtxSim, and CUS for four 7B-parameter models on PubMedQA and MesaQA ($N = 400$, $\alpha = 0.5$).

MesaQA Results. Gemma-7B, Llama-2-7B, and Mistral-7B achieve similar CUS scores (0.675–0.678), while Falcon-7B scores notably lower (0.606). Gemma-7B leads on GoldSim (0.689), while Mistral-7B leads on CtxSim (0.736) and achieves the best overall CUS (0.678) by balancing both dimensions. CtxSim consistently exceeds GoldSim across all models, indicating that models draw effectively from context but add content not present in the reference answer.

PubMedQA Results. Performance drops substantially, with CUS ranging from 0.480 (Falcon-7B) to 0.592 (Mistral-7B), reflecting the difficulty

Table 4: LLM Evaluation Results (mean over N=400 samples,, α=0.5). GoldSim measures factual alignment with the reference answer; CtxSim measures faithfulness to the supporting context; CUS (Equation 3) is their harmonic mean, penalising imbalanced profiles. Mistral-7B achieves the best CUS on both datasets, reflecting consistently balanced factual and contextual understanding, while Falcon-7B underperforms across all metrics and both domains.

Dataset	Model	GoldSim	CtxSim	CUS
MesaQA	Gemma-7B	**0.6889**	0.7031	0.6774
	Llama-2-7B	0.6570	0.7236	0.6752
	Mistral-7B	0.6491	**0.7356**	**0.6780**
	Falcon-7B	0.6036	0.6458	0.6063
PubMedQA	Gemma-7B	**0.5235**	0.6401	0.5587
	Llama-2-7B	0.5220	0.6567	0.5651
	Mistral-7B	0.5138	**0.7331**	**0.5923**
	Falcon-7B	0.4541	0.5560	0.4800

of matching precise biomedical reference answers. Mistral-7B again leads in CUS, supported by the highest CtxSim (0.733). Gemma-7B scores highest on GoldSim (0.524) but lowest on CtxSim (0.640), indicating closer alignment with reference content but weaker use of biomedical context. Falcon-7B is weakest across all metrics on both datasets.

Cross-Dataset Observations. CtxSim exceeds GoldSim in all 8 model–dataset combinations, consistent with instruction-tuned models elaborating on context rather than producing concise reference-style responses. The $\sim$10-point CUS gap between MesaQA and PubMedQA across all models points to domain complexity as the primary factor, with biomedical vocabulary and reasoning posing challenges irrespective of model architecture.

Triplet Analysis of the Least Similar KGs. Across the 5% lowest-similarity cases in both datasets, the four models show clear differences, and we focus on this tail subset because these challenging instances make model failures more diagnostic and reveal systematic weaknesses that can be masked by strong average scores. Mistral is consistently the most reliable on the hard examples, preserving more reference facts and producing more aligned triplets than the other models. Llama generally falls in the middle, while Gemma shows the weakest performance, especially on the general health QA set where it often fails to recover many gold-aligned facts. A second consistent pattern is that, in these difficult cases, the generated KGs tend to align more closely with the contextual KG than with the reference KG. This suggests that many failures are not simply random errors, but cases where the model either drifts toward a different interpretation, omits key reference facts, or introduces unsupported additions. Finally, the PubMed setting is noticeably harder for all models: aligned-triplet recovery drops sharply, indicating that technical biomedical terminology, abbreviations, and entity variability dominate the failure modes in the hardest cases.

7 Discussion

The proposed KG-based framework provides a structured approach to evaluate LLM understanding, moving beyond syntactic-level metrics to assess relational and structural fidelity. S3KG's type-separated, one-to-one label alignment addresses fundamental limitations of clustering-based approaches while maintaining the WL kernel's ability to capture multi-hop neighborhood similarity.

The consistent performance gap between PubMedQA and MesaQA scores across all 3 models reveals a measurable difference in LLM capability for domain-specific versus general health knowledge. The two-dimensional S3KG evaluation further distinguishes factual accuracy (gold similarity) from contextual faithfulness (context similarity), exposing model-specific trade-offs that aggregate metrics cannot capture. The TAU additionally enables pinpointing specific reasoning failures, whether due to incorrect entity substitution, relation errors, or broader inconsistencies.

8 Conclusion

We presented a novel KG-based evaluation framework for assessing LLM contextual understanding in QA. By constructing canonicalized KGs from LLM outputs, gold answers, and context, and comparing them using S3KG, we move beyond syntactic-level accuracy toward verifiable, graph-theoretic comprehension measurement. S3KG achieves best-per-dataset F1 of 0.766–0.932 and AUROC up to 0.973, consistently outperforming lexical and neural baselines on KG-rich datasets while remaining competitive on short-text settings. The TAU provides interpretable, fine-grained diagnostics of reasoning failures. Evaluations on PubMedQA and MesaQA demonstrate consistent model-specific strengths and weaknesses, establishing a reproducible pipeline that can be extended to other datasets and tasks to support trustworthy AI development. In this sense, whether LLMs un-

derstand context becomes empirically testable by measuring how well their generated responses preserve the relational knowledge expressed in the reference answer and supporting context.

Limitations

Key limitations of this work include: (1) sensitivity of similarity scores to KG extraction quality, as noisy or incomplete triplet extraction directly degrades S3KG performance; (2) computational cost of SBERT inference at scale, which may be prohibitive for very large evaluation sets without GPU acceleration; and (3) evaluation is currently restricted to open-source 7B-parameter models; extending to larger proprietary models such as GPT-4 remains future work. Additionally, the current alignment scheme does not handle directional semantic equivalence (e.g., *daughter_of* vs. *mother_of*), which would require attention-based mechanisms.

Acknowledgements. The work of Kamal Prematratne (KP) was supported by the French/US joint project LUCAS between the Agence Nationale de la Recherche (ANR) (grant ANR-25-CE23-2189) and the U.S. National Science Foundation (NSF) (grant numbers 2530255 and 2530256).

References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, pages 2787–2795. Curran Associates, Inc.

Marina Boudin, Gayo Diallo, Martin Drancé, and Fleur Mougin. 2023. The OREGANO knowledge graph for computational drug repurposing. *Scientific Data*, 10:871. Food ontology and natural compound knowledge graph for drug repurposing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

W. B. Dolan and C. Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Reilly Haskins and Benjamin Adams. 2025. Kea explain: Explanations of hallucinations using graph kernel analysis. arXiv preprint arXiv:2507.03847.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Xiaohui Victor Li and Francesco Sanna Passino. 2024. FinDKG: Dynamic knowledge graphs with large language models for detecting global trends in financial markets. In *Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF 2024)*, pages 573–581. Financial knowledge graph extracted from news articles using LLMs.

C.-Y. Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Jianmo Ni, Noah Constant Gustavo Hernandez Abrego, Keith Hall Ji Ma, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

J. H. Poelen, J. D. Simons, and C. J. Mungall. 2014. GloBI: Global biotic interactions. [Online]. Available: `https://www.globalbioticinteractions.org`. Accessed: Jan. 15, 2025.

Alec Radford, Kiran Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Sennrich Rico, Haddow Barry, and Birch Alexandra. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

T. Safavi and D. Koutra. 2022. Codex: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, page 8328–8350.

Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. Grapheval: A knowledge-graph based llm hallucination evaluation framework. arXiv preprint arXiv:2407.10793.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. OneRel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11285–11293.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. 2011. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. 2019. Wasserstein Weisfeiler–Lehman graph kernels. In *Advances in Neural Information Processing Systems*, volume 32, pages 6439–6449. Curran Associates, Inc.

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

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2024. Correctness is not faithfulness in rag attributions. *arXiv preprint arXiv:2412.18004*.

Jui-I Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2025. MESAQA: A dataset for multi-span contextual and evidence-grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10891–10901, Abu Dhabi, UAE. Association for Computational Linguistics.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Junbing Yan, Chengyu Wang, Jun Huang, and Wei Zhang. 2024. Do large language models understand logic or just mimick context? arXiv preprint arXiv:2402.12091.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Y. Zhang, J. Baldridge, and L. He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018.

A Effect of Temperature on Contextual Understanding

Tables 5 and 6 report mean GoldSim and ContextSim respectively for each model across temperature settings $T \in \{0.0, 0.3, 0.7, 1.0\}$ on both datasets. Most models show little sensitivity to temperature, with score variations within ± 0.01–0.02 across all settings. The exception is Falcon-7B on MesaQA, where GoldSim drops substantially from 0.6036 at $T = 0.0$ to 0.4662 at $T = 1.0$, indicating that higher sampling randomness significantly degrades factual alignment for this model. Scores tend to peak mildly at $T = 0.3$ for most models before declining at $T = 1.0$, suggesting that a small degree of randomness can marginally improve contextual grounding without sacrificing factual accuracy. PubMedQA scores are notably more stable across temperatures than MesaQA, likely due to the constrained nature of biomedical answers.

B KG Construction Prompt

KGs are extracted using a structured chat-style prompt inspired by Sansford et al. (2024) and Hask-

Table 5: Mean GoldSim per model across temperatures. CUS remains stable across temperature settings, with $T = 0.0$ serving as a reliable default for controlled evaluation.

Dataset	Model	$T = 0.0$	$T = 0.3$	$T = 0.7$	$T = 1.0$
MesaQA	Llama-2-7B	0.6570	**0.6588**	0.6580	0.6406
	Gemma-7B	0.6889	**0.7060**	0.7015	0.6931
	Mistral-7B	0.6491	**0.6567**	0.6523	0.6444
	Falcon-7B	0.6036	**0.6085**	0.5558	0.4662
PubMedQA	Llama-2-7B	**0.5220**	0.5185	0.5143	0.5087
	Gemma-7B	**0.5235**	0.5149	0.5153	0.5204
	Mistral-7B	**0.5138**	0.5121	0.5085	0.5023
	Falcon-7B	**0.4541**	0.4330	0.4184	0.3852

Table 6: Mean ContextSim per model across temperatures. CtxSim is more sensitive to temperature than GoldSim, yet remains stable for most models, with $T = 0.3$ yielding peak contextual faithfulness across both datasets before declining at higher temperatures.

Dataset	Model	$T = 0.0$	$T = 0.3$	$T = 0.7$	$T = 1.0$
MesaQA	Llama-2-7B	0.7236	0.7212	**0.7248**	0.7026
	Gemma-7B	0.7031	**0.7217**	0.7045	0.7013
	Mistral-7B	0.7356	**0.7585**	0.7287	0.7223
	Falcon-7B	0.6458	**0.6571**	0.6122	0.5195
PubMedQA	Llama-2-7B	0.6567	0.6563	**0.6594**	0.6468
	Gemma-7B	0.6401	**0.6540**	0.6434	0.6319
	Mistral-7B	0.7331	**0.7352**	0.7314	0.7003
	Falcon-7B	**0.5560**	0.5165	0.5120	0.4571

ins and Adams (2025). The prompt instructs the model to perform four sequential steps across all three input texts:

1. **Entity detection:** Extract all named entities, concepts, attributes, quantities, dates, locations, and roles comprehensively.
2. **Coreference resolution:** Replace all pronouns with their referent entity names, using consistent labels across all three texts.
3. **Relation extraction:** Identify semantic relationships as simple, concise phrases, decomposing compound sentences into one triplet per fact.
4. **Knowledge graph refinement:** Where the same entity or relation appears across multiple graphs, use the same label consistently without merging distinct facts.

The model is instructed to return a JSON object with exactly 3 keys (`knowledge_graph1`, `knowledge_graph2`, and `knowledge_graph3`), each containing a list of `[subject, relation, object]` triples. Few-shot examples are included in the system prompt to ground the expected output format and label consistency behaviour. The user turn specifies the three input sources explicitly:

- `TEXT1` — reference answer.
- `TEXT2` — model-generated response.
- `TEXT3` — supporting context.

The complete prompt, including few-shot examples, is available at https://github.com/aaivu/knowledge-xtraction.

C TAU Diagnosing Context and Answer Deviations

C.1 Dataset and Annotation

We evaluate the TAU on a manually annotated subset of the **MessaQA** dataset, which contains general health-related questions with long form answers. For each QA instance, two KGs are constructed: (i) a gold KG extracted from the reference answer, and (ii) an LLM generated (Llama and Gamma was used) KG extracted from the model output. The task of the triplet analysis unit is to identify semantically aligned triplet pairs between these two graphs. To obtain reliable evaluation labels, we created a gold set of aligned triplet pairs (GT $\leftrightarrow$ LLM). Alignment was independently annotated by three medical students following a fixed guideline defining semantic equivalence at the triplet level. Disagreements were resolved through adjudication.

C.2 Annotation Quality

Inter annotator agreement was measured over candidate aligned triplet pairs. The results indicate strong consistency, with percent agreement of 0.9433, pairwise F1 scores of 0.9708, 0.9825, and 0.9882, and pairwise Jaccard scores of 0.9433, 0.9657, and 0.9766, confirming the reliability of the annotations.

C.3 Evaluation Protocol

Performance is evaluated against the annotated alignments using precision, recall, and F1 score, reported using both micro averaged metrics (aggregated over all triplets) and macro averaged metrics (computed per QA instance and averaged), capturing both overall performance and consistency across samples.

Table 7: Comparison of TAU alignment performance between our method and the KEA baseline, evaluated on manually annotated MesaQA and PubMed subsets.

Metric	KEA Baseline	Ours
Micro Precision	**0.853**	0.847
Micro Recall	0.738	**0.949**
Macro Precision	**0.953**	0.803
Macro Recall	0.641	**0.946**
Micro F1	0.836	**0.895**
Macro F1	0.622	**0.782**

C.4 Results

As shown in Table 7, the KEA baseline achieves higher precision due to its conservative component wise matching, but exhibits low recall, missing many valid alignments. In contrast, our sentence-level alignment approach significantly improves recall (+34.8%), resulting in higher Micro and Macro F1 scores. This improvement arises from robustness to lexical variation in relations. For example, semantically equivalent relations such as *"treats"* and *"used for"* are often not aligned by KEA, whereas our method captures such equivalence through sentence level semantic similarity. The lower precision reflects the expected trade-off when moving from strict lexical matching to semantic matching, while the overall F1 improvement indicates a better balance between sensitivity and specificity. Additionally, our method enables residual error analysis by categorizing mismatches into relation wrong and entity wrong types.

D Per-Dataset Performance Tables

This appendix decomposes the headline results of Table 2 into the 3 benchmark families, with F1 and AUROC shown in adjacent columns so that threshold-dependent and ranking behaviour can be inspected side by side. All datasets share the same protocol: $N = 400$ balanced pairs, per-method threshold sweep for F1, and AUROC on the raw scores. For S3KG, the best α identified in Appendix E is used per dataset, and the best score per column is **bolded**. The Wikipedia Entity-Swap table (Table 9) additionally reports precision and recall to expose ROUGE-1's near-perfect score as a token-overlap artifact rather than a meaningful signal.

ROUGE-1 achieves a near-perfect score on this dataset due to a known artifact: entity-swapped pairs differ only in the swapped entity tokens while sharing nearly identical surrounding surface-form tokens, making unigram overlap trivially high. ROUGE-1 is therefore excluded from the meaningful comparison.

Table 8: Performance on Short-Text Datasets. S3KG leads on PAWS-Wiki (F1 0.766, AUROC 0.795) but trails sentence-T5-base on MRPC and STS12, where dense semantic representations have a natural advantage over structural signals.

	MRPC		PAWS-Wiki		STS12	
Method	**F1**	**AUROC**	**F1**	**AUROC**	**F1**	**AUROC**
S3KG (Ours)	0.692	0.673	**0.766**	**0.795**	0.786	0.834
ROUGE-1	0.745	0.784	0.678	0.490	0.725	0.754
ROUGE-2	0.720	0.721	0.715	0.721	0.681	0.656
ROUGE-L	0.729	0.760	0.735	0.807	0.703	0.710
BLEU	0.687	0.677	0.716	0.747	0.671	0.644
BERTScore	0.758	0.816	0.691	0.702	0.682	0.636
MiniLM	0.723	0.748	0.687	0.638	0.833	0.894
sentence-T5-base	**0.766**	**0.816**	0.674	0.668	**0.853**	**0.928**

Table 9: Results on Wikipedia Entity-Swap ($N = 400$). S3KG attains the highest F1 and AUROC across all retained baselines; ROUGE-1 is excluded as its near-perfect score is a token-overlap artifact rather than a meaningful signal.

Method	F1	AUROC
S3KG (Ours)	**0.872**	**0.890**
ROUGE-2	0.860	0.772
ROUGE-L	0.729	0.311
BLEU	0.868	0.745
BERTScore	0.747	0.645
MiniLM	0.821	0.811
sentence-T5-base	0.762	0.806

E Hyper-parameter Selection

The mixing coefficient α in (2) is the sole hyper-parameter of S3KG, controlling the trade-off between the structural WL kernel signal ($\alpha = 0.0$) and the SBERT semantic signal ($\alpha = 1.0$), with intermediate values blending both. We tune α per dataset via grid search over $\{0.0, 0.1, \ldots, 1.0\}$, selecting the value maximising binary-classification F1; AUROC is reported alongside to confirm the result is not an artefact of threshold sensitivity.

Table 11 consolidates the full sweep across all nine benchmark datasets. The best α per dataset (selected by maximum F1) is **bolded** together with its F1 / AUROC entries. Three patterns emerge: (i) datasets dominated by perturbations (SK-FindKG, Wiki Swap, STS12) favour

Table 10: Performance on KG-Perturbed Paragraph Datasets. C400: SK-Codex 400; Comb.: SK-Combined; Find: SK-FindKG; Oreg.: SK-Oregano. S3KG leads on four of five datasets; sentence-T5-base outperforms on SK-FindKG, the only dataset where pure dense embeddings dominate ($\alpha = 0.0$ is optimal).

	C400		**Comb.**		**Find**		**GloBI**		**Oreg.**	
Method	**F1**	**AUROC**	**F1**	**AUROC**	**F1**	**AUROC**	**F1**	**AUROC**	**F1**	**AUROC**
S3KG (Ours)	**0.932**	**0.973**	**0.834**	**0.829**	0.767	0.796	**0.892**	**0.935**	**0.812**	**0.892**
ROUGE-1	0.835	0.917	0.732	0.728	0.745	0.745	0.784	0.833	0.745	0.782
ROUGE-2	0.822	0.894	0.707	0.711	0.717	0.706	0.776	0.833	0.752	0.791
ROUGE-L	0.792	0.855	0.722	0.717	0.719	0.721	0.763	0.800	0.792	0.835
BLEU	0.806	0.884	0.715	0.708	0.711	0.710	0.775	0.819	0.745	0.794
BERTScore	0.823	0.916	0.757	0.792	0.739	0.761	0.816	0.871	0.743	0.798
MiniLM	0.875	0.943	0.770	0.789	0.802	0.844	0.780	0.817	0.773	0.814
sentence-T5-base	0.876	0.944	0.770	0.827	**0.848**	**0.902**	0.728	0.760	0.797	0.871

low $\alpha \in \{0.0, 0.1\}$, where the WL signal carries most of the discriminative power; (ii) datasets where both relational and lexical paraphrase signals are simultaneously informative (SK-Codex 400, SK-Combined, PAWS-Wiki) peak in the balanced range $\alpha \in [0.4, 0.6]$; (iii) the AUROC surface is consistently flatter than the F1 surface, indicating that α primarily reshapes the score distribution near the decision boundary.

Table 11: Full S3KG α sweep across short-text and combined datasets (F1 / AUROC). The best α per dataset (by maximum F1) is **bolded**. Balanced datasets such as PAWS-Wiki and SK-Combined peak at $\alpha \in [0.4, 0.6]$, while STS12 favours the KG-heavy end ($\alpha = 0.1$).

	MRPC		**PAWS-Wiki**		**STS12**		**SK-Codex 400**		**SK-Combined**	
α	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC
0.0	0.676	0.654	0.694	0.739	0.783	0.828	0.871	0.969	0.776	0.801
0.1	0.681	0.664	0.745	0.781	**0.786**	**0.834**	0.922	0.973	0.791	0.825
0.2	0.683	0.669	0.760	0.790	0.780	0.834	0.927	0.973	0.819	0.829
0.3	**0.692**	**0.673**	0.764	0.793	0.780	0.831	0.927	0.973	0.828	0.830
0.4	0.680	0.674	0.764	0.795	0.780	0.827	0.929	0.973	0.833	0.829
0.5	0.680	0.675	**0.766**	**0.795**	0.775	0.824	**0.932**	**0.973**	**0.834**	**0.829**
0.6	0.681	0.674	0.764	0.795	0.768	0.820	0.932	0.973	0.832	0.828
0.7	0.688	0.674	0.764	0.795	0.764	0.815	0.932	0.973	0.833	0.828
0.8	0.684	0.672	0.764	0.795	0.762	0.810	0.932	0.973	0.832	0.828
0.9	0.683	0.671	0.764	0.795	0.756	0.805	0.932	0.973	0.828	0.827
1.0	0.681	0.671	0.764	0.731	0.756	0.790	0.932	0.932	0.828	0.799

Table 12: Full S3KG α sweep across KG-perturbed and entity-swap datasets (F1 / AUROC). The best α per dataset (by maximum F1) is **bolded**. Structure-dominated datasets (SK-FindKG, Wiki Swap) favour low $\alpha \in \{0.0, 0.1\}$, confirming the WL kernel's advantage on relational perturbations.

	SK-FindKG		**SK-GloBI**		**SK-Oregano**		**Wiki Swap**	
α	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC
0.0	**0.767**	**0.796**	0.811	0.898	0.777	0.884	0.821	0.892
0.1	0.760	0.798	0.883	0.934	0.803	0.892	**0.872**	**0.890**
0.2	0.762	0.793	0.888	0.937	0.811	0.892	0.869	0.890
0.3	0.748	0.788	0.889	0.936	0.812	0.892	0.868	0.890
0.4	0.744	0.783	0.891	0.936	**0.812**	**0.892**	0.865	0.890
0.5	0.741	0.779	0.890	0.935	0.810	0.892	0.868	0.890
0.6	0.736	0.776	**0.892**	**0.935**	0.812	0.892	0.865	0.889
0.7	0.734	0.772	0.892	0.935	0.812	0.891	0.865	0.889
0.8	0.731	0.769	0.892	0.935	0.810	0.891	0.865	0.889
0.9	0.729	0.765	0.892	0.934	0.810	0.891	0.865	0.889
1.0	0.728	0.756	0.892	0.917	0.810	0.770	0.865	0.852