Responses to Reviewer 1

> The distinction between Background and Related Work sections is not entirely clear; merging them could improve readability.

- Change: have now merged these two sections together into a "Background and Related Work" section

> The paper devotes a considerable amount of space to discussing prior work, which somewhat overshadows the explanation of the proposed methodology. A stronger balance in favor of methodological clarity would help.

- Change: trimmed down excessive coverage of prior work, including completely cutting out the "Probabilistic Techniques for Hallucination Detection" subsection, as this was not referenced in any of the subsequent text.

> A central component of the approach, the Weisfeiler-Lehman Graph Kernel, is not explained with sufficient concreteness. The appendix description feels too abstract, and it is not entirely clear how the extracted triples are represented and compared within this framework. Providing a more intuitive walkthrough or a worked example would make the contribution much more accessible.

- Change: added a worked example in a subsection under the appendix's description. Referenced this in the main text at the end of the "Graph Kernel-based Comparison of Knowledge Graphs" section

Responses to Reviewer 2

> The Related Work requires a deeper focus on the interpretability limitations of existing models and a more robust justification for the preference of the proposed approach

- Change: added sentences stating interpretability limitations for relevant prior work (LLM-based methods, open-domain & closed-domain KG-based methods)

> Regarding Algorithm 1, clarification is needed for the condition "attributes differ". I assumed that "attributes differ" refers to a discrepancy in the relation label (the r in an (h,r,t) triple) even if the head and tail entities are identical, however in the discussion the authors refer to a conflict at entity-level (Paris and Rome).

- Change: updated algorithm and added a short clarification above it to clarify that our method will construct an edit operation for a difference between any element of the triples.

> The paper clearly states an LLM generates the natural language explanation, yet it doesn't specify the LLM used for this task or provide an example of the specific prompt used for explanation generation.

- Change: updated the text to explicitly mention that we use gpt-4o-mini in our experiments, and added a subsection for the relevant prompt in the appendix

> For the triple notation, e.g., (h,r,t), using italics to distinguish variable names from prose is recommended for clarity.

- Change: updated to write these variable names in italics.

> The necessary use of empirically chosen graph kernel similarity thresholds that vary significantly by task introduces complexity. This sensitivity to different domains and tasks means real-world deployment would likely require re-optimizing the threshold before use.

- Response: we agree that this is a limitation of our method, and we have now covered this in the Limitations and Future Research section. In particular, we put forward a potential solution of setting aside some labelled data from the target domain for use in optimising this threshold.

> The framework incurs a significant practical computational burden. I think that programmatically constructing ground truth KGs and, for open-domain tasks, retrieving relevant facts from Wikidata on the fly via SPARQL queries can be time-consuming, which could be a limitation for real-time applications. Is this limitation worth discussing?

- Change: added a paragraph to address this in the Limitations and Future Research section: “Third, the requirements of having to construct ground truth KGs can induce a significant computational burden, which may make our method less effective in time-constrained applications and domains. This is especially prevalent in the open-domain, where many relevant entities must first be retrieved from Wikidata via SPARQL queries.”

> This is more of a suggestion: the structured nature of the contrastive explanation (based on graph edit operations) is ideally suited to generating follow-up prompts (e.g. to the same LLM that generated the hallucinated sentence) to actively correct the hallucination. This could be a valuable direction for future exploration.

- Change: Thank you for this suggestion. We agree that this could be a valuable direction for future exploration. Although this would not be considered a ground-truth source of knowledge, probing the LLM in a structured way (based upon the graph edit operations) would likely provide a complementary explanation and potential correction for the hallucination. We have outlined this as an interesting direction at the end of the Limitations and Future Research section: “Future research could explore alternative explanation and correction strategies, such as through generation of follow-up queries to the LLM that generated the hallucinated sentence, which would further enhance the method's transparency and practical utility, as well as have the potential to correct hallucinated facts.”

> The technique used for filtering the ground-truth KG by maximizing the cosine similarity of SBERT embeddings (arg max) with the claim KG triples. From my understanding, this reliance on argmax is a concern because it could select irrelevant context triples as long as they are the closest available in the embedding space, potentially retaining triples that are only somewhat relevant (similar domains) but not directly tied to the entities being evaluated. A strict similarity threshold alongside argmax could make this filtering process more robust.

- Response: we appreciate this point and agree that argmax alone can, in principle, select the “least bad” triple when none are very close. While across the benchmarks used in our work, we did not observe the case where a triple would not be properly matched, we do have some downstream controls to prevent this from having a major impact on the end result if it were to occur. For instance, the WL subtree kernel

rewards matched *subtrees* much more heavily than isolated triples. Weakly-matched entity pairs chosen by argmax rarely form matching neighbourhood patterns and therefore contribute very little to the normalised kernel score. Additionally, in the explanation generation stage of the methodology, we already apply cosine-based thresholds for entity and relation agreement when identifying contradictory pairs, which gates spurious alignments. Importantly, one advantage of leaving the filter without a hard cutoff is that it helped us prioritise recall, especially in open-domain settings where missing a relevant supporting triple risks false negatives.

> The observation that explanation quality monotonically declines as hallucination severity decreases is important. It suggests the current method struggles with subtle inconsistencies, as it is too reliant on finding concrete conflicting triples, which are sparser in nuanced hallucinations.

- Response: We agree, and view this as a known limitation of a contradiction-driven explanation method. We invite future work to improve upon this, perhaps by relying on the follow-up query method that you suggested.

> I think that the performance of the approach is heavily reliant on the quality of the actual ground truth KG, which, in turn, is heavily conditioned by several possible failure points. These include the performance of the SentenceBERT embeddings, the SpaCy Entity Linker, and the empirically chosen Similarity Thresholds. Performance comparison with baselines indicates that the model's value may ultimately lie more in the benefit of structured explanations than in detection performance alone.
> Testing the open-domain hallucination detection only on the WikiBio dataset is somewhat limiting, given its modest size. Furthermore, while competitive, the detection performance might appear suboptimal compared to certain baselines.

- Response: We agree that our method is dependent on construction of the ground truth KG (although this choice represents a trade-off that we deem to be worthwhile, compared to sources which serve as a proxy to the ground-truth), and wish to clarify that we also see our contribution primarily in the production of a human-understandable explanation based on comparison with ground-truth data, without sacrificing significant classification performance when compared to current state-of-the-art methods. In response to this, we have updated our title from "KEA Explain: A Neurosymbolic Framework for **Detecting and Explaining** LLM Hallucinations" to now emphasise the explanation component: "KEA Explain: A Neurosymbolic Framework for **Explaining** LLM Hallucinations". In addition, we have also made small edits throughout the paper to increase the clarity of this focus to the reader. For example, in the Discussion section, we change "We also tested our method's ability to add interpretability through the use of graph-based knowledge representation" to "Although we have shown performance close to state-of-the-art detection methods, our key contribution is represented in our method's ability to add interpretability through the use of graph-based knowledge representation."