

Explanation Of Revisions for Resubmission of article “*None of the Others*: a General Technique to Distinguish Reasoning from Memorization in Multiple-Choice LLM Evaluation Benchmarks”

We have carefully revised the manuscript, especially addressing the following points:

Regarding methodology:

- We have integrated the model proposed in Wicked¹ to improve dataset filtering and ensure that each question has only one correct answer, further reducing potential ambiguities.
- The corresponding tables and figures have been updated accordingly, although the results remain largely unchanged.

Regarding writing and presentation:

- We have softened the tone of our proposal, clarifying that the aim is not to introduce a definitive method for evaluating reasoning, but rather to complement existing work with an automatic and domain-independent technique that helps disentangle memorization effects, even if it cannot fully isolate them.
- These nuances, along with other concerns raised by the reviewers, are now more clearly acknowledged in the Limitations section.
- We have restructured the Related Work section to make it more concise and directly focused on the context and contributions of our work, as suggested by the reviewers.
- We have updated several references that were previously cited as preprints, replacing them with their corresponding published versions.

¹ Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. WiCkeD: A Simple Method to Make Multiple Choice Benchmarks More Challenging. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1183–1192, Vienna, Austria. Association for Computational Linguistics.