Responsible NLP Checklist

Paper title: *AssoCiAm: A Benchmark for Evaluating Association Thinking while Circumventing Ambiguity*
Authors: *Yifan Liu, Wenkuan Zhao, Shanshan Zhong, Jinghui Qin, Mingfu Liang, Zhongzhan Huang, Wushao Wen*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
Based on the field and focus of our work, it poses lower risks for potential issues. Our work utilizes publicly available data and independently generated datasets, which minimize potential copyright concerns and reduce the likelihood of disinformation or the generation of fake profiles. Furthermore, our work is designed to be applicable to evaluation of all MLLMs in aspect that do not relate to safety issue, thereby mitigating potential risks related to security, bias, environmental impact, fairness, and privacy considerations.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
We cite Control diffusion model in section 4; DINO-v2 in section 4,7; CLIP in section 4; MLLMs in section 5; MMMU Benchmark from others in section 6; ImageNet ILSVRC12 in section 4

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We point out the artifacts are available in Section 4,5,6 but not discuss the license or terms

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We discuss that the artifacts we create is use for model evaluation in Section 1, 2, 8 and some data we use are public available and suitable for the target in Section 4, but we do not discuss the model use consistent with intended use

☒ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
The data we use do not focus on personally identifying infomation and offensive content and such contents are not usable nor acceptable for our work.

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We do not provide documentation of the artifacts, but we introduce what the benchmark and dataset we use is for.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
We discuss it in section 4,5,6

☑ **C. Did you run computational experiments?**

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We report the number of parameters in the models used in Section 6

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss the experimental setup in Section 5, 7

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We discuss the results in Section 6, 7

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
We discuss the specific models we use in experiments in Section 5, 6, 7.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We discuss what participants should do in our work in Section 4,5,7, but not include screenshots, disclaimers of any risks to participants or annotators, because our assignments are not related to their privacy or personal right.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We invite participants offline, in which we can make sure partivants are willing to finish our tasks and agree that we collect some data from them.

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
We do not explicitly discuss whether and how consent was obtained, but since the participants are invited to our assignments offline, it means that they are informed of what they should do and what will be collected and the invitation accepted indicates their agreement.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our research do not require ethics reviews.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotators in our research do not represent any specific population

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☒ E1. If you used AI assistants, did you include information about their use?
We use AI assistants for paraphrasing and polishing the our original content, which is not necessary to be disclosed, according to ACL Policy on Publication Ethics