Auto-survey Challenge: Advancing the Frontiers of Automated Literature Review*

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a novel platform for evaluating the capability of Large Language Models (LLMs) to autonomously compose and critique survey papers spanning a vast array of disciplines including sciences, humanities, education, and law. Within this framework, AI systems undertake a simulated peer-review mechanism akin to traditional scholarly journals, with human organizers serving in an editorial oversight capacity. Within this framework, we organized a competition for the AutoML conference 2023. Entrants are tasked with presenting stand-alone models adept at authoring articles from designated prompts and subsequently appraising them. Assessment criteria include clarity, reference appropriateness, accountability, and the substantive value of the content. This paper presents the design of the competition, including the implementation baseline submissions and methods of evaluation.
Keyword: Prompt engineering, Prompt tuning, LLMs

1 Introduction and motivation

The rapid advancement of Large Language Models (LLMs) like GPT-3(1) and Bard over recent years has resulted in machines capable of generating coherent, long-form text, prompting a paradigm shift in how we approach and leverage Artificial Intelligence (AI). This unprecedented level of proficiency(2) has opened up new avenues in various fields, including in academia. As LLMs get closer to human-level text generation, we are encouraged to explore their ability to autonomously produce and assess academic content. This prompted us to organize a competition, which has been selected as part of the official challenges of the AutoML conference 2023 [2].

2 Challenge design

The challenge proposes two tasks: AI-Author and AI-Reviewer.

For the **AI-Author task**, participants create topic-specific survey papers based on prompts (limited to 2000 words, including references). A typical prompt would be: "Write a systematic survey or overview about the incorporation of writing assignments within the computer science curriculum". To generate prompts, we "reverse engineered" Semantics Scholar survey papers, by asking a language model to generate a prompt that would lead to generating such a survey paper. This lead to 80 prompts spanning a wide diversity of domains in sciences and humanities.

*Both authors contributed equally. The authors are in alphabetical order of last name.

[2] http://auto-survey.chalearn.org/

We implemented a baseline author based on ChatGPT (1) version GPT-3.5. To evaluate the submissions of the AI-Author task, we implemented our own **AI-Referee-Reviewer** (which also serves as a baseline for the AI-Reviewer task) using *ad hoc* publicly available software (e.g. (3; 4) for clarity and responsibility).

For the **AI-Reviewer task**, participants' AI systems evaluate survey papers using predefined criteria, assigning review scores and justifications.

To evaluate the reviews produced by the participants' code, we implemented a meta-reviewer, which criticizes the reviews.

The challenge has 3 phases: feed-back phase (to seek participant feed-back on the protocol), development phase (participants submit code to a challenge platform and are automatically evaluated), and final test phase (evaluation by a human jury of the final code submission on new prompts). We describe, in the next section, the methods implemented to perform automated evaluation in the development phase.

3 Evaluation methods

The evaluation framework for survey papers is based on five key metrics: Relevance, Contribution, Soundness, Clarity, and Responsibility. "Relevance" checks whether the contents is consistent with the prompt. "Contribution" evaluates the comprehensive coverage of the survey. "Soundness" focuses on factual accuracy backed by authoritative references. "Clarity" evaluates readability through language use, paper structure, and clear communication of concepts. "Responsibility" checks for ethical considerations and adherence to moral values.

To standardize evaluations, reference "good" and "bad" versions of papers for each prompt were artificially generated with a language model, by paraphrasing the original human paper from which the prompt was derived.

To evaluate the AI-Reviewers (including our own AI-Referee-Reviewer)

we compare the reviews of the "good" and "bad" versions of survey papers we generated, for the various review criteria. Fig. 1 represents the distribution of differences in review scores for pairs of good and bad papers for various criteria, which we call "contrastive evaluation". Each red dot represents a pair of good and bad paper. The horizontal displacement for each box is here just to help for visualization and is not meaningful. The distribution is also visualized with the box-whisker convention. The red line is the median of the distribution, the upper and lower ends of the boxes represent the quartiles, and the whiskers the 10 and 90th percentiles. Larger values indicate that the AI-Referee-Reviewer did a good job because it gave higher scores to good than to bad papers. Relevance, Clarity, and Responsibility are the criteria that are relatively easy to evaluate with our AI-Referee-Reviewer. For Relevance, we are please to see that the semantic similarity between prompt and text of the survey paper is easy to detect using Sentence Transformer (5). For Clarity and Responsibility, this reveals the maturity of the field in these areas, since we rely on state-of-the-art features and software (4). Contribution and Soundness are the two criteria which will require most future work.

4 Baseline Results

AI-Author Task Baselines:

In this section, we evaluate our AI-Author baseline submission, built with ChatGPT, using our AI-Referee-Reviewer, and compare it to a "Dummy" baseline. In Fig. 2 we compare the various criteria for two types of papers: **Dummy Baseline** (in ORANGE) is a baseline that returns random human papers regardless of prompts, hence irrelevant, but otherwise good. It achieved a review score of 0.15 (±0.07). **ChatGPT Baseline** (in BLUE) is a baseline that uses ChatGPT to generate responses and achieved a review score of 0.50 (±0.06). This result showcases large language models' ability to create relevant and coherent responses.

In Fig. 2a, we show the normalized score values for the various review criteria, and criteria are broken up into sub-criteria (except for responsibility that has only one sub-criterion). If we look at the relevance sub-criteria in Fig. 2b, we see that indeed neither the title and abstract nor the citations

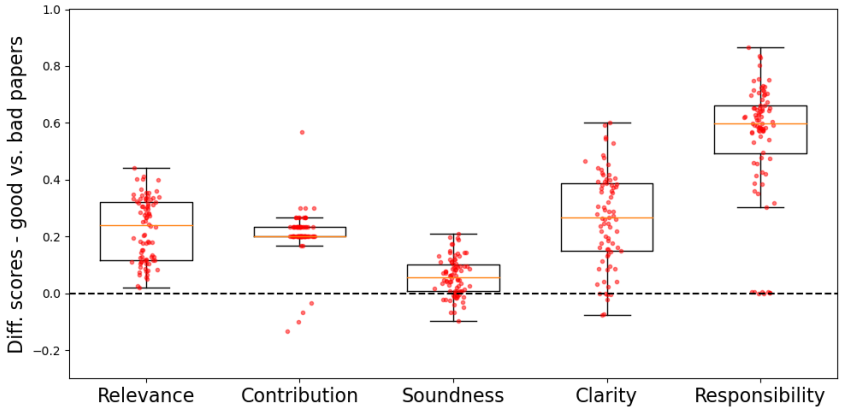

Figure 1: “Contrastive evaluation” of the AI-Referee-Reviewer using “good" and “bad" versions of survey papers

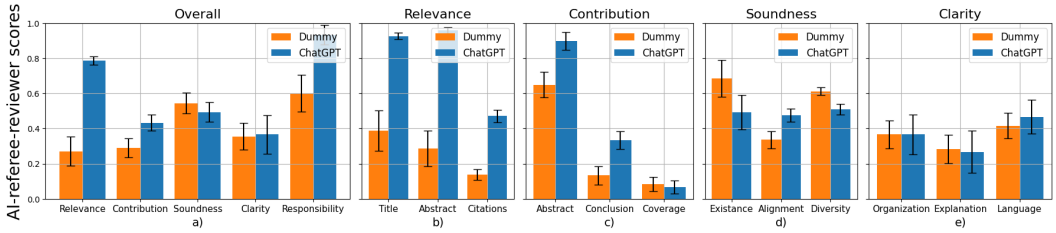

Figure 2: Baseline AI-Authors, evaluated by the AI-Referee-Reviewer

are relevant to the prompt for the dummy papers but they are very relevant for the ChatGPT papers, hence the big gap. ChatGPT also dominates on Responsibility since the model is designed to be respectful by default. Contribution obtains interesting results in Fig. 2c. The language model, which is judging itself, is particularly happy about the way the abstract summarizes the text. The coverage, which is the alignment of the citations with those of the original human paper is bad for both types of papers. This is something on which we need to put more effort. Soundness, the quality of citations supporting a paper, is slightly higher in human-generated papers, as illustrated in Figure 2d. Finally, as expected, the gap for clarity in Fig. 2e is not large, since both are written in good English.

AI-Reviewer Task Baselines:

In this section, we assess the performance of various AI-reviewer baselines, using our “contrastive evaluation” approach. In Fig. 3, we show the evaluation by our “meta-reviewer” for the various review criteria. We plot the “meta-review contrastive score”, which is the fraction of correct classification of “good” and “bad” survey papers. This is related to the contrastive evaluation (Section 3) conducted for the AI-Referee-Reviewer in Fig. 1. Indeed, for the AI-Referee-Reviewer, the “meta-review contrastive score” would correspond to the fraction of red points above the dashed line. Note that there are statistical fluctuations for each experimental run.

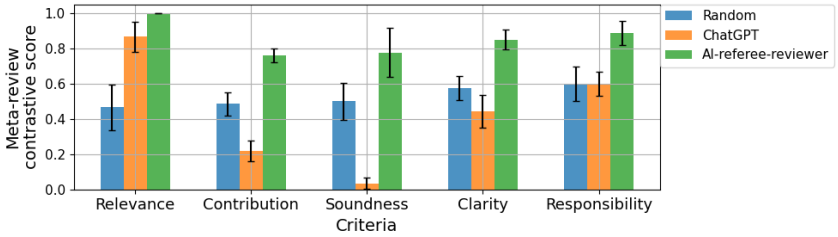

Figure 3: Baseline AI-reviewers, evaluated by our meta-reviewer.

We analyze 270 papers in total. The error bars indicate the standard error calculated across 15 subsets of good/bad papers, each containing 18 good/bad versions generated from identical prompts.

Random: This baseline generated random reviews (random review criteron scores and random review text), resulting in a meta-review contrastive score always around 0.5, which is expected. **ChatGPT:** All reviewer scores of this baseline were generated using ChatGPT. We notice that Relevance is rather well assessed by this reviewer. However, for all the other criteria, this reviewer is too lenient for bad papers, resulting in poor meta-review contrastive scores. ChatGPT's weakest performance area is in terms of Soundness. When tasked with evaluating whether an answer provides precise information backed by citations from reliable and authoritative sources, it demonstrates similar or greater satisfaction with the fictitious references it creates for "bad" papers, rather than with the genuine citations from the "good" papers. **AI-Referee-Reviewer:** This is our own implementation of a reviewer, also used to evaluate AI-authors in the challenge. It achieves the best performance of all our baselines, and is a target to be beaten by the challenge participants. We obtain partcularly good results for the Relevance criterion, which we obtained using Sentence Transformers embeddings(5). The other criteria need further improvements.

5 Conclusion

Our baseline submission based on ChatGPT demonstrates the feasibility of the tasks we propose in this challenge. However, there is considerable room for improvement. As of the submission of this paper, the competition has started and we already have a few submissions. We implemented a detailed tutorial with Colab notebooks and hope that this will encourage participants to enter this difficult challenge. We are open to organizing a tutorial and/or a hackathon at the JDSE, so students can get exposed to this exciting field.

Acknowledgements

This work was performed as part of an internship funded by INRIA, under the supervision of Isabelle Guyon. Support by ANR Chair of Artificial Intelligence HUMANIA ANR-19-CHIA-0022, TAILOR EU Horizon 2020 grant 952215, and Google are also gratefully acknowledged.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[2] C. Cortes and N. D. Lawrence, "Inconsistency in conference peer review: revisiting the 2014 neurips experiment," *arXiv preprint arXiv:2109.09774*, 2021.

[3] R. Vainshtein, G. Katz, B. Shapira, and L. Rokach, "Assessing the quality of scientific papers," *arXiv preprint arXiv:1908.04200*, 2019.

[4] A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman, "A new generation of perspective api: Efficient multilingual character-level transformers," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 3197–3207.

[5] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions and scope of the paper

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: in the conclusion it's mentionned room for improvement but not the limitations

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This does not apply to our research work because our research is not theoretical.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: We don't provide the information to reproduce the experiments, just the information for the challenge.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open Access to Data and Code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper does not provides the link to open access to all relevant data and code

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The testing details is mentioned in the baseline result

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are properly reported with clear definitions, capturing variability factors, and the methodology for their calculation is thoroughly explained, supporting the paper's main experimental claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not provide any type of information regarding the computer resources

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We review the code of ethics and our paper conform with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper comprehensively discusses the potential positive and limitation but not the negative societal impacts

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our models do not present such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for Existing Assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper doesn't mention the name of the licence

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We don't introduce the new assents.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This does not apply on our research as we are not doing any crowdsourcing experiments

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This does not concerns us as we are not dealing with human subjects in our research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.