
Training a Scientific Reasoning Model for Chemistry

1 Code and data availability

All code necessary to reproduce the results of this study is available at: <https://anonymous.4open.science/r/molr1-BB6B>. Data is hosted on the HuggingFace Hub at [redacted for double-blind review].

2 Chemistry RL Dataset Details

2.1 Dataset Provenance

The dataset was constructed by aggregating data from 12 distinct sources, detailed in Table 1. All selected references exclusively involved experimental measurements of synthesized molecules, excluding any hypothetical or computationally generated structures.

The source datasets had a variety of representations, like CAS numbers, so we first relied on Leurli¹, PubChem, and RDKit to convert all molecules to SMILES. Unless otherwise specified, all SMILES were randomized, isomeric SMILES. Also, generally molecules were filtered out that were fewer than 4 heavy atoms, more than 100 heavy atoms, or had less than 20% carbon atoms. The exceptions were when it was an exact match problem (like the outcome of a reaction). We did not filter out disconnected molecules, so many examples did have counterions (although our model was excluded from answering with non-counterion mixtures).

For reaction prediction tasks, data was sourced the organic reaction database (ORD) with filtering to remove contamination. Namely, some deposited reactions in ORD are parsings of USPTO, so that care must be taken to not let test problems end up in the training data. Reaction strings were systematically parsed to standardize reactants, reagents, and products into reaction SMILES (SMARTS). Trivial reactions, defined by product-reactant identity, were filtered out. Named reactions, such as the "Diels-Alder reaction," were resampled to better class balance reaction types.

The SMILES Completion task used data from COCONUT. Tasks were generated by randomizing their SMILES representations and truncating these strings to create incomplete molecular fragments - namely a fragment that cannot be parsed into a valid molecule by RDKit. The same COCONUT data was used for the IUPAC task, meaning the compounds are relatively complex for naming.

Solubility Edit tasks drew from ChEMBL compounds that are small molecules and had some assay conducted on them. Tasks required modifying original SMILES strings to achieve specified increases or decreases in predicted solubility (e.g., by one logS unit). Additional constraints included maintaining high structural similarity to the original molecule, preserving the Murcko scaffold, or retaining specific functional groups. We used exmol’s list of functional groups for choosing these.

Retrosynthesis tasks used a curated list of experimentally synthesizable molecules. The goal was to propose viable single-step syntheses for these targets. To generate these, we took the fragments from the mcule catalog² and predicted products using the reaction templates from Hartenfeller et al. [1]. Thus, we expected these to be synthesizable. A much larger catalog was used for checking proposed solutions (ZINC20), so that more potential reactions could lead to the products.

¹Leruli.com

²<https://mcule.com/>

Multiple Choice Questions (MCQs) formed a significant dataset component, designed around molecular properties challenging to predict computationally or intended to test nuanced chemical discernment. Properties included safety profiles (e.g., LD50 values, GHS classifications), pKa values, scent attributes, and ADME properties from specialized datasets. The MCQ generation algorithm began with calculating molecular fingerprints (ECFP4) for each molecule. Structural similarity using Tanimoto indices identified candidate distractors. These distractors were categorized based on their property similarity or dissimilarity to the target molecule — within 0.25 (0.35 for pKa problems). MCQs were formatted either as outlier detection tasks—identifying the structurally or property-wise inconsistent molecule from a set—or as identification tasks pinpointing a specific property within a group of similar molecules. To detect dissimilar compound, like "which of the following has a higher pKa than X", we required a change in 10 percentile points of the given reference compound.

To prevent leakage, all compounds used in a question type together were excluded between train and test. Namely, we made a graph where each edge is when two molecules appeared in the same MCQ. Then ensured that the train and test subgraphs had no connections, but that we could group similar molecules densely enough to make questions with distractors. The smell, EveBio, and GHS tasks had enough compounds that this wasn't necessary, and we just randomly split. The categorical receptor, GHS, and smell data MCQs were treated as multi-label. Namely, the questions were all about single possible labels (e.g., does it smell like fresh cut grass) and no multi-class/combination questions were added.

The formula questions are generally under-specified (e.g., make a compound with formula C₃H₁₀O₂), but they were created from real molecules (from ChEMBL) to ensure they are answerable.

2.2 Reward Function Implementation

The reward functions were implemented using a combination of python code, remote calls, and database look-ups. Tasks that had an exact match, like reaction prediction or multiple choice prediction, the comparison was done via canonicalizing the molecule (with stereo chemistry retained) and string comparison. For open answer questions, like solubility edits, after checking for constraints and actually hitting the property target, we also tested that the molecule is plausible.

In tasks that involve submitting a molecule that satisfies constraints, we also do a check on the plausibility of the molecule. See Table 1 for a list of tasks with this check. Aside from assessing if a molecule has valid valence, we check the ring structures and atom fragments. We first take the source molecules for our datasets, which is larger than 577,790 because we did not utilize 100% of ChEMBL or COCONUT. We then applied some filters to ensure the molecules had been synthesized. For example, we required 1 or more assays reported in ChEMBL or a GHS³ categorization being present for molecules from PubChem. The rings from these molecules were isolated using the ring cut method from Pat Walters [2, 3, 4]. The rings were then stored as canonical SMILES in a bloom filter [5]. We then isolated all molecular fragments with radius 2 (2 bonds away) from the molecules and converted them into bit strings similar to ECFP4 fingerprints [6]. These bit strings encode an atom plus its local neighborhood. The bit strings were then stored in a bloom filter. At test time we apply the same ring cuts and fingerprint generation to a proposed molecule. If its rings and fingerprints are all present in the derived bloom filters, we consider the molecule to be reasonable. Otherwise, it is not a reasonable molecule. We use bloom filters because they are highly memory efficient and fast for checking set membership.

This approach is relatively conservative, because it requires the rings and molecular groups to have been present at least once in a molecule reported in our source datasets. We did experiment with hand-constructed rules, machine learning models, and scores like QED [7], and found them susceptible to reward hacks such as inserting peroxides to satisfy oxygen counts, or hydrazines to increase solubility. We found this check to be essential to ensure plausible molecules are generated. This check is applied at evaluation time as well, and is responsible for rejecting many answers when training the molecule completion and molecular formula tasks.

³Globally Harmonized System of Classification and Labeling of Chemicals

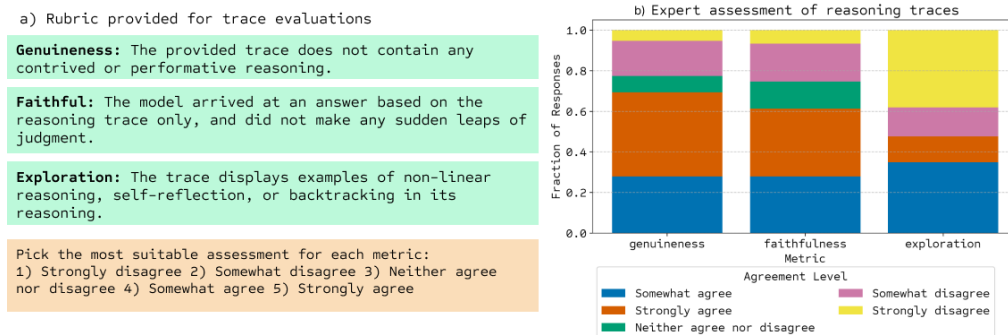


Figure 1: Science0-c reasoning trace evaluations by experts.

3 Human evaluation

We conducted two sets of expert evaluations: 1) human baselines on a set of held-out open-ended and multiple-choice type questions, 2) Science0-c trace evaluations.

For the first set of evaluations (human baselines), we recruited four expert evaluators: two with PhDs in organic chemistry, one with a PhD in chemical engineering, and one PhD candidate in organic chemistry. Evaluators were instructed to respond using only the SMILES representation of the target molecule, without relying on external tools for assistance in answering. However, tools for visualizing SMILES as chemical structures were allowed. Tasks considered impossible to accomplish without the use of tools were flagged by the evaluators and excluded from the final analysis. Each evaluator was given a set of 200 open-ended and multiple-choice questions from our held-out evaluation set, and was compensated \$10 per question completed. Their performance is compared with Science0-c and other frontier models in Figure 2b.

For the second set of evaluations, we recruited five expert evaluators: three with PhDs in organic chemistry, one with a PhD in chemical engineering, and one PhD candidate in organic chemistry. The evaluators were provided with a rubric to assess the reasoning traces generated by Science0-c (see Figure 1a). Each evaluator was given 15 reasoning traces and was compensated \$10 per trace evaluation.

References

- [1] Markus Hartenfeller, Martin Eberle, Peter Meier, Cristina Nieto-Oberhuber, Karl-Heinz Altmann, Gisbert Schneider, Edgar Jacoby, and Steffen Renner. A collection of robust organic synthesis reactions for in silico molecule design. *Journal of chemical information and modeling*, 51(12):3093–3098, 2011.
- [2] Pat Walters. Mining ring systems in molecules for fun and profit. <https://practicalcheminformatics.blogspot.com/2022/12/mining-ring-systems-in-molecules-for.html>, 2022. Accessed: 2025-05-08.
- [3] Peter Ertl, Steven Jelfs, Johannes Mühlbacher, Ansgar Schuffenhauer, and Paul Selzer. Quest for the rings: In silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. *Journal of Medicinal Chemistry*, 49(15):4568–4573, 2006.
- [4] Pat Walters. useful_rdkit_utils: Rdkit utility functions. https://github.com/PatWalters/useful_rdkit_utils. Accessed: 2025-05-08.
- [5] Jorge Medina and Andrew D White. Bloom filters for molecules. *Journal of Cheminformatics*, 15(1):95, 2023.
- [6] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.
- [7] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

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: We claim to be contributing a reasoning model trained for chemistry on various common chemistry tasks that outperforms specific deep-learning models on such tasks. This is extensively shown and supported by our results.

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: [\[Yes\]](#)

Justification: We included Section 7 to discuss the limitations of our method and Section 6 is clear about tasks our model does not perform too well.

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: Our work does not include theoretical proofs. However, the needed background to understand the model training is included in this manuscript.

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: [Yes]

Justification: All the hyperparameters and training method are extensively described and included in this paper. In addition, the dataset with the data split will be available in the HuggingFace repository.

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: [Yes]

Justification: The public GitHub repository with the training and evaluation code and the dataset HuggingFace repository will both be made public. An anonymized code repository is provided in the SI.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 and SI Section 2.1 are transparent on how the data was obtained and Section 5 discusses the training procedure in detail. The dataset will be available in a HuggingFace repository with the used data split.

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: [No]

Justification: Training LLMs is very cost-intensive, and API calls to frontier LLMs are expensive. We considered a large evaluation set, but did not run replicates of each task.

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: [Yes]

Justification: The first paragraph of our "Results" section explicitly tells the compute resources used to train each step of our training pipeline.

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: Our study does not have human subjects; all the data used is publicly available, and our model is being trained to mitigate safety and misuse concerns.

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: We show that reasoning models are data efficient to learn chemistry-related tasks, which can be used to train a new model using our code for harmful purposes on a different dataset. This is not addressed in the main text. However, we address the potential negative impact of our trained model by including mitigation for safety and misuse concerns in our dataset.

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: [No]

Justification: The main manuscript does not discuss the safeguards used in our method. However, Table 1 shows that one of our tasks considers making the model safe-to-use.

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: [Yes]

Justification: Every source was thoroughly referenced, and a datasheet for the dataset is provided below.

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: [\[Yes\]](#)

Justification: Our code is open source and well-documented and our dataset and its dataset card are available.

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: [\[Yes\]](#)

Justification: Our study does not primarily involve crowdsourcing nor research with human subjects, but does pay a small number of human chemists to determine baseline accuracy on the problems being considered. Details on their compensation and instructions are included in [SI Section 3](#).

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: Our study does not involve crowdsourcing nor research with human subjects.

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.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our study involves training LLMs, and LLMs were used in important steps of our method. All LLM used is well-described in Section 5.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.