A AtomWorld Setup Details

A.1 Supported action prompts

Table 2: Examples of actions and the corresponding action prompts for point-based tasks.

Action name	Action prompt
`move`	Move the point at index {index} by displacement {displacement}.
`move_towards`	Move the point at index {from_index} towards the point at index {to_index} by {distance}.
`insert_between`	Insert a new point between points at indices {index1} and {index2}, {distance} units away from point {index1}.
`rotate_around`	Rotate all points by {angle_deg} degrees around the axis {axis}, with the point at index {center_index} as the center of rotation. The rotation follows the right-hand rule.

Table 3: Examples of actions and the corresponding action prompts for AtomWorld.

Action name	Action prompt
`add`	Add one {symbol} atom at the Cartesian coordinate {position} to the cif file.
`move`	Move the atom at index {index} by {d_pos} angstrom in the cif file.
`move_towards`	Move the atom at index {index1} towards the atom at index {index2} by {distance} angstrom in the cif file.
`insert_between`	Insert a {symbol} atom in the line between atoms at indices {index1} and {index2}, and the inserted atom must be {distance:.2f} angstrom from atom at {index1} in the cif file.
`rotate_around`	Rotate all surrounding atoms within {radius} angstrom of the center atom at index {index} by {angle} degree around the axis {axis} in the cif file. The rotation should following the right-hand rule.

In addition to the actions listed above, we have also implemented several others, including `remove`, `swap`, `delete_around`, `move_selected`, etc. These actions are not presented here, as they have not yet undergone systematic evaluation.

A.2 Evaluation metrics

Two primary metrics are used to evaluate the correctness of the LLM-generated structures: the error rate and the mean maximum distance (`max_dist`).

The error rate is defined as the number of test cases exhibiting any of the following errors divided by the total number of test cases. These errors are categorized into three hierarchical levels:

1. **Wrong output format.** The LLM's response must enclose the generated structure within a predefined tag so that it can be correctly extracted from the textual output. Failure to do so constitutes an output format error.
2. **Wrong structure format.** Even if the structure is successfully extracted, its file format may still be invalid or incompatible with downstream processing tools. Such cases are counted as structure format errors.
3. **Mismatch of structures.** For structurally valid outputs, we compare them with the target structures using `StructureMatcher` with a site tolerance of 0.5. Any generated structure whose site matching exceeds this tolerance is considered a mismatch.

The second primary metric - mean `max_dist` - is computed only for structurally valid outputs that pass the tolerance check. For each matched pair of structures, we calculate the maximum pairwise atomic displacement after optimal alignment, and then average this value across all test cases. The `max_dist` metric is used because it is generally more significant than the RMSD value in our cases. This is because only a few or even a single atom is "moved" while others remain unchanged, making the maximum displacement a more representative indicator of the structural difference.

A.3 Full Prompt Templates

Listing 1: A prompt example for a specific task of AtomWorld

```
You are a CIF operation assistant. You will be given an input CIF
content and an action prompt. Your task is to apply the action
described in the action prompt to the initial CIF content. The
coordinates in the action are in Cartesian format. Return the modified
 CIF content in cif format within <cif> and </cif> tags.

Please ensure the output is a valid CIF file, with correct formula,
and atom positions.

Input CIF content:
{The specific CIF file is inserted here}

Action prompt: Insert Lu between atoms at indices 6 and 5 that is 4.03
 angstrom from atom 6.
```

A.4 Illustrative example of the framework

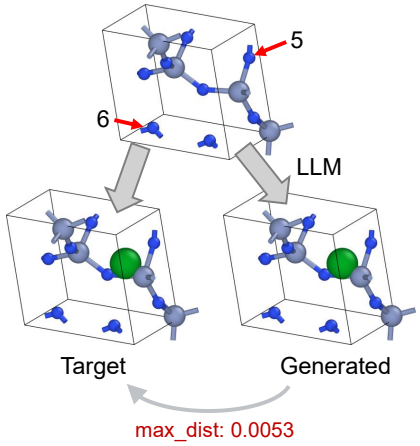

Figure 3: The workflow of a specific `insert_between` task.

To provide a concrete understanding of our proposed AtomWorld Bench, we present an illustrative example of its workflow. This case study focuses on a specific task: inserting a Lu atom between the fifth and the sixth atoms in the specific CIF structure. The prompt used here is listed in Appendix A.3. The workflow randomly selects the atom indices and determines the position of the atom to be inserted based on the selected atoms. Based on the initialized action, the framework gives out a target structure. The LLM will also generate a structure after processing the prompt, as shown in Figure 3. In this example, the two structures are nearly identical, with a `max_dist` of 0.0053 Å, indicating high accuracy.

B Full Evaluation Results

B.1 Tables for error rates and mean `max_dist`

Table 4: Model performances of `add` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	10.0	0.0140
Qwen3 32b	22.0	0.0352
ChatGPT o3	20.0	**0.0015**
ChatGPT o4-mini	**3.20**	0.0060
DeepSeek V3-0324	34.0	0.1221
Llama3 70b	49.6	0.1315

Table 5: Model performances of `move` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	**19.2**	0.0293
Qwen3 32b	30.8	0.0605
ChatGPT o3	23.2	**0.0102**
ChatGPT o4-mini	48.0	0.0734
DeepSeek V3-0324	32.4	0.1274
Llama3 70b	48.8	0.1315

Table 6: Model performances of `move_towards` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	**22.0**	0.0513
Qwen3 32b	34.0	**0.0201**
ChatGPT o3	37.6	0.0425
ChatGPT o4-mini	35.2	0.1063
DeepSeek V3-0324	51.2	0.2467
Llama3 70b	70.4	0.1613

Table 7: Model performances of `insert_between` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	**32.0**	**0.0444**
Qwen3 32b	37.2	0.1082
ChatGPT o3	40.4	0.0778
ChatGPT o4-mini	52.8	0.1501
DeepSeek V3-0324	54.4	0.2004
Llama3 70b	70.8	0.1921

Table 8: Model performances of `rotate_around` action

Model	Error rate (%)	mean `max_dist` (Å)
Gemini 2.5 Pro	88.0	0.0790
Qwen3 32b	96.0	0.0900
ChatGPT o3	**87.2**	0.0933
ChatGPT o4-mini	88.4	0.1832
DeepSeek V3-0324	93.2	0.3607
Llama3 70b	96.0	0.3557

B.2 The `max_dist` violin plots

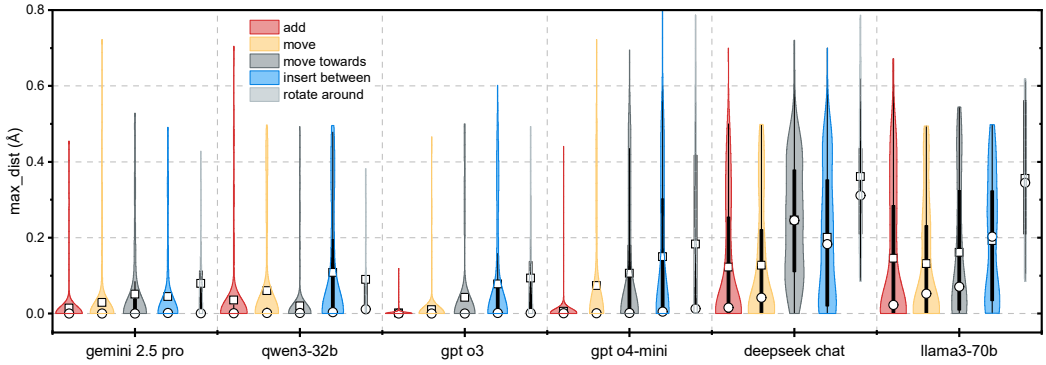

Figure 4: The violin plots of `max_dist` of evaluation results. The hollow squares indicate the mean values, and the hollow circles indicate the medians.

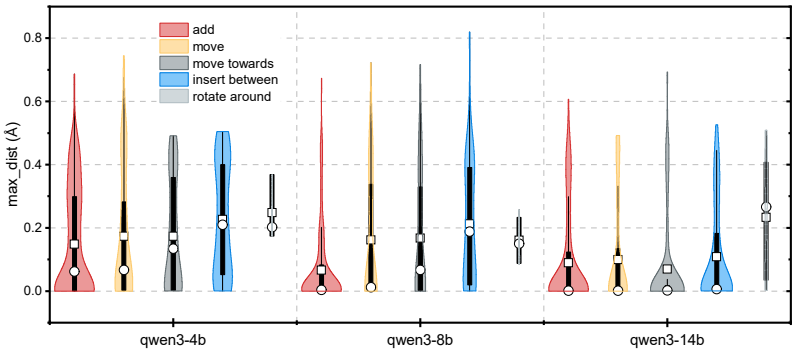

Figure 5: The violin plots of `max_dist` of evaluation results.

B.3 Chemical Competence Score

The Chemical Competence Score (CCS) is designed to assess a model's latent chemical knowledge by evaluating its precision in distinguishing chemically accurate from inaccurate descriptions of crystal structures. Following the methodology of Bran *et al.*[22], the dataset was constructed by sampling 600 unique crystal structures from the Materials Project, with corresponding descriptions generated using Robocrystallographer[23]. An inaccurate dataset was then created by replacing one sentence in each original description with a sentence describing a different crystal. Because the CCS is computed from the token log-likelihoods at the model's final layer, access to these probabilities is required; consequently, the score was calculated only for the open-source models benchmarked in this study. The resulting scores are reported in Table 9 and visualised in Fig. 6.

Table 9: CCS score of open-source models

Model	CCS
Qwen3 4B	0.768
Qwen3 8B	0.829
Qwen3 14B	1.061
Qwen3 32B	**1.141**
Llama3 70b	0.987

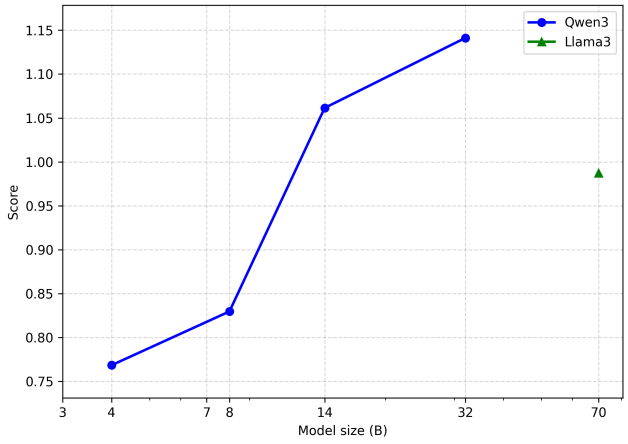

Figure 6: Line plot illustrating the relationship between CCS and model size for open-source models

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 claim that we have provided a new benchmark and playground for testing LLMs' spatial understanding of materials.

   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 have mentioned the limitations that will be further solved in the discussion and conclusion parts.

   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: The current paper does not include theoretical results.

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: We provide the full benchmark code along with a detailed README. The experiments can be reproduced directly by running the provided scripts, without requiring manual parameter tuning.

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: We will release the benchmark code and data.

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 benchmark details are included in the methodology, as well as the readme file inside the code.

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: Due to skewed and task-dependent distributions, mean $\pm$ SD may be misleading. Instead, raw data and distribution figures are provided in the appendix.

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 experiments are mainly based on API calls to external LLM providers, but we do not provide detailed statistics on runtime or compute usage.

   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: The research conforms fully with the NeurIPS Code of Ethics, and no deviations are present.

   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: [NA]

    Justification: The paper mainly focuses on the usage of LLMs in materials science, which does not directly cause the societal impact.

    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: The paper poses no 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: [Yes]

    Justification: We rely on the Materials Project database, pymatgen, etc., which are properly cited and used under their respective licenses.

    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: We introduce a new benchmark implementation, including code, dataset, and prompt templates. These will be released in a GitHub repository.

    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: The paper 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.
    - 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: The paper 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 rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The benchmark is designed to test LLMs, and we have also used LLMs for revising the manuscript and some of the coding. But we did not develop the core method with LLMs.

    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.