When Researchers Say Mental Model/Theory of Mind of AI, What Are They Really Talking About?

Anonymous Author(s)
Affiliation
Address
email

Abstract

When researchers claim AI systems possess ToM or mental models, they are fundamentally discussing behavioral predictions and bias corrections rather than genuine mental states. This position paper argues that the current discourse conflates sophisticated pattern matching with authentic cognition, missing a crucial distinction between simulation and experience. While recent studies show LLMs achieving human-level performance on ToM laboratory tasks, these results are based only on behavioral mimicry. More importantly, the entire testing paradigm may be flawed in applying individual human cognitive tests to AI systems, but assessing human cognition directly in the moment of human-AI interaction. I suggest shifting focus toward mutual ToM frameworks that acknowledge the simultaneous contributions of human cognition and AI algorithms, emphasizing the interaction dynamics, instead of testing AI in isolation.

1 Introduction

Humans develop theories to explain each other's behaviors (Sellars, 1956). This ability to infer each other's mental states (Premack & Woodruff, 1978) has been called the theory of mind (ToM). However, we are in direct continuous contact with our own minds but not the mental states of others; therefore, studying human mental states is difficult. This abstract concept is not related to any specific parameters or rubric; it's the theory of an experienced mental state that we can't observe in any specific event or attribute to different, distinct events.

Kosinski (2023) argues that individual artificial neurons in LLMs function like "Chinese rooms", which follow mathematical instructions without genuine understanding. The author also suggests that complex cognitive abilities may emerge at the network level, similar to how human cognition is assumed to emerge from networks of neurons. Mechanically speaking, this comparison makes sense. However, debating whether models are merely Chinese rooms may be unproductive. The key is to study how they interact, as humans typically don't question others' cognition during normal interaction. Meanwhile, Strachan et al. (2024) claim that LLM behavior is "indistinguishable from human behavior" on ToM tasks on various dimensions. However, if you listen to an authentic interaction between two humans and an interaction between a human and an AI agent, you can tell a difference. Under the assumption that a ToM is fundamental to human social interaction, rather than a byproduct of it, this finding has motivated researchers to develop benchmarks that offer a practical and cost-effective approach to assessing ToM capabilities in LLMs.

Researchers like Gu et al. (2024) tested GPT-4 using their SimpleToM dataset. While GPT-4 achieves approximately 90% accuracy on ToM questions, performance drops to approximately 50% on behavior prediction and 15% on behavioral judgment. The test results are not ideal, and the test itself is also questionable, unless being able to answer a battery of ToM questions proves that GPT-4 has a ToM. The limitations of this kind of approaches has also been pointed out by Wang

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

et al. (2025), that current ToM tasks have three major limitations: 1) Theoretically, ToM should be multidimensional but has only been tested in one dimension; 2) Current ToM tests lack construct validity; and 3) Evaluations always use third-person static scenarios rather than spontaneous dynamic interactions. They also report that 75.5% of ToM measures focus only on beliefs rather than other considerations such as emotion, desires, and intentions. In a narrow technical sense, AI systems that track beliefs and predict behaviors could be said to have "ToM". However, this technical capability differs fundamentally from human ToM grounded in embodied experience.

2 The Fundamental Flaw in Cognitive Testing for LLMs

The trend in evaluating LLMs using cognitive tasks may show a fundamental misunderstanding about what these systems actually are and what we need to know about them. Researchers designed ToM tests, reasoning challenges, and planning problems for LLMs by assuming that humans process information like LLMs do. But as Kambhampati (2024) argues, LLMs are fundamentally "n-gram models on steroids" performing "universal approximate retrieval" rather than reasoning focused on survival that humans do.

Consider how humans may develop ToM through embodied experience. Children learn that others have different perspectives through physical interactions such as hiding objects, playing peek-a-boo, and observing emotional responses. Psychologists study how children build up their abilities using developmental scales. For example, Wellman and Liu (2004) described scales of a sequence of tasks to test children's diverse desires, diverse beliefs, knowledge access, contents false belief, explicit false belief, belief-emotion, and real-apparent emotion. While this is representative work, it doesn't mean we should test LLMs in the same way or even compare them to humans (such as the SimToM prompting framework by Wilf et al. (2024)) by assuming these researchers are correct in reducing human interaction to their identified dimensions of a ToM.

Some researchers have designed benchmarks to measure ToM abilities specifically for LLMs, such as Kanishk et al. (2023) with their social reasoning benchmark (BigToM) and Chen et al. (2024) with their ToMBench. But why should we accept that LLMs even have a belief that can be tested? More importantly, what does passing these tests tell us about how AI will function in dynamic human-AI interactions?

Many benchmarks developed for testing LLMs' ToM are derived from psychological tests such as the Sally-Anne test (Scassellati, 2001) but lack validity and reliability. As ToM is proposed as an ability to understand and respond to ongoing feedback from the social environment, the evaluation of ToM should not be static (Wang et al., 2025). The tests are measuring something about ToM, certainly, but not meaningful interaction in a valid sense. They're measuring how well statistical patterns from human cognitive behavior in the training data can be reproduced in response to prompts. In other words, it is learning patterns that we intend it to, and we are asked to accept this as having a ToM that some researchers want it to.

This becomes clearer when we examine how LLMs fail. When planning problems are presented with obfuscated names, removing the statistical patterns from training data, LLM performance plummets dramatically. They're not reasoning inductively through the structure of problems; they're deductively matching surface patterns to similar examples in their training corpus. The same limitation appears in self-critique tasks. Despite claims that LLMs can verify their own reasoning, studies show performance actually worsens with self-verification as models "hallucinate" both false positives and false negatives.

The recent phenomenon that Cuadron et al. (2025) termed the "Reasoning-Action Dilemma" further exposes this limitation. Large Reasoning Models exhibit what appears to be "overthinking", where they bias internal reasoning over environmental feedback. But this isn't genuine overthinking in the human sense, such as in anxiety-driven rumination. It's the model getting stuck in loops of text generation that statistically resemble reasoning without genuine deliberation or decision-making authority.

The test is legitimate for measuring behavioral reproduction accuracy, but not valid for inferring genuine social cognitive processes or, more importantly, for predicting that AI functions as a human does in human contexts.

3 Why Testing Misses the Point

The distinction between simulation and authentic mental processes matters, but not for reasons typically discussed. LLMs are good at simulating ToM responses because they've been trained on billions of examples of human discourse about mental states aimed at generating those responses. They can produce outputs that seem to demonstrate understanding of beliefs, desires, and intentions. However, this cannot be called "ToM" in an adaptive human sense.

This distinction becomes clearer when examining motivated reasoning. Humans don't always reason correctly, even when they possess sufficient cognitive resources. We engage in "wishful thinking", "confirmation bias", and emotionally driven reasoning. As Kunda (1990) notes, human reasoning involves "hot motivation to reach desired goals" alongside "cold motivation favoring accuracy". AI systems, following optimization objectives, will lack these intrinsic motivational states. When AI systems make errors, it's not because desire overrides logic but because they've encountered edge cases that their training and algorithm do not cover.

Proponents of AI cognition often invoke "emergence", which is a mysterious way of claiming that sufficiently complex neural networks spontaneously develop genuine understanding. They point to studies like Zhu et al. (2024) showing LLMs internally represent different agents' beliefs in their neural activations. However, finding structured representations doesn't prove consciousness exists any more than finding face-selective neurons proves the fusiform face area of the human brain experiences faces.

Debate about genuine versus simulated cognition distracts from the central issue: isolated ToM tests do not tell us what actually matters, nor how AI systems should function within human social and collaborative contexts.

4 Mutual ToM

Rather than asking whether AI "has" a ToM, we should examine how humans and AI systems can mutually develop understanding. Instead of validating AI performance on scales of ToM using isolated cognitive tests, we should test the system dynamics when AI operates within human environments. The critical question isn't whether AI can pass a false-belief task in isolation, but how human-AI interaction changes both behavior and outcomes when the human-AI system is working on a project rather than taking a test.

Wang and Goel's (2022) Mutual ToM framework offers one approach that preserves the ToM concept by focusing on three iteratively shaping elements: interpretation, feedback, and mutuality. In this framework, humans and AI agents construct representations of each other, but these representations serve different functions. Humans apply ToM to AI systems through anthropomorphism, attributing mental states that the AI doesn't possess. Meanwhile, AI systems build statistical models predicting human behavior without assuming any genuine understanding of mental states.

The key insight is that effective human-AI collaboration doesn't require AI to have a ToM but rather requires systems that support mutual adaptation and understanding. Zhang et al.'s (2024) findings support this approach. They found AI's independent ToM capability didn't significantly impact team performance. What mattered was enhancing human understanding of the agent. Surprisingly, bidirectional communication sometimes decreased performance, with participants reporting increased cognitive workload. This suggests that pretending AI has ToM might actually hinder rather than help human-AI collaboration. Considering the dynamic nature of mutual theory of mind, human-autonomy team (HAT) resilience studies employing dynamic measurements (e.g., Grimm et al., 2023) can serve as a valuable reference for future ToM studies, particularly those focusing on team-level states in human-AI collaboration.

This distinction matters for practical purposes. We should stop asking "Does this AI understand false beliefs?" and start asking "How does this AI system change human behavior in collaborative settings?" Then stop endless philosophical debates about "emergent" consciousness and ToM test scores.

5 Conclusion

When researchers test AI's ToM, they're observing sophisticated behavioral prediction, not genuine mental state attribution. But more importantly, they're asking the wrong question. The current approach of administering human cognitive tests to AI systems, whether adapted from child psychology or specially designed for LLMs, fundamentally misunderstands both what these systems are from a psychological perspective and what we need to know to use them effectively.

LLMs achieve remarkable performance on ToM tasks through statistical pattern matching across massive datasets, but this differs fundamentally from human experience, grounded in embodied experience, motivated reasoning, and genuine understanding, which led to the development of the ToM concept in the first place. However, even if AI does replicate ToM behaviors in testing conditions, these isolated assessments do not tell us about what actually matters: how human-AI systems function together as a cognitive system. Specifically, we should abandon attempts to improve AI's scores on ToM tests and instead study mutual adaptation and understanding between humans and AI systems. The mutual approach acknowledges both human tendencies to anthropomorphize and AI's statistical modeling capabilities, designing for effective collaboration within these constraints. The question isn't whether AI thinks like or simulates the human mind. It's how humans and AI can work together in the service of human survival and flourishing. After all, what matters isn't how well models score on ToM tests, but rather how good the models can serve us, humans.

References

[1] Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., Cao, Y., Hu, M., Lai, Y., Xiong, Z., & Huang, M. (2024). ToMBench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*.

[2] Cuadron, A., Li, D., Ma, W., Wang, X., Wang, Y., Zhuang, S., Liu, S., Schroeder, L. G., Xia, T., Mao, H., Thumiger, N., Desai, A., Stoica, I., Klimovic, A., Neubig, G., & Gonzalez, J. E. (2025). The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*.

[3] Gandhi, K., Fränken, J. P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 13518-13529).

[4] Grimm, D. A., Gorman, J. C., Cooke, N. J., Demir, M., & McNeese, N. J. (2023). Dynamical measurement of team resilience. Journal of Cognitive Engineering and Decision Making, 17(4), 351-382.

[5] Gu, Y., Tafjord, O., Kim, H., Moore, J., Bras, R. L., Clark, P., & Choi, Y. (2024). SimpleTom: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

[6] Kambhampati, S. (2024). Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1), 15-18.

[7] Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

[8] Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.

[9] Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515-526.

[10] Scassellati, B. M. (2001). *Foundations for a Theory of Mind for a Humanoid Robot* (Doctoral dissertation, Massachusetts Institute of Technology).

[11] Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, 1(19), 253-329.

[12] Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7), 1285-1295.

[13] Wang, Q., & Goel, A. K. (2022). Mutual theory of mind for human-AI communication. *arXiv preprint arXiv:2210.03842*.

[14] Wang, Q., Zhou, X., Sap, M., Forlizzi, J., & Shen, H. (2025). Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective. *arXiv preprint arXiv:2504.10839*.

[15] Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, 75(2), 523-541.

[16] Wilf, A., Lee, S. S., Liang, P. P., & Morency, L. P. (2023). Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. *arXiv preprint arXiv:2311.10227*.

[17] Zhang, S., Wang, X., Zhang, W., Chen, Y., Gao, L., Wang, D., Zhang, W., Wang, X., & Wen, Y. (2024). Mutual theory of mind in human-ai collaboration: An empirical study with llm-driven ai agents in a real-time shared workspace task. *arXiv preprint arXiv:2409.08811*.

[18] Zhu, W., Zhang, Z., & Wang, Y. (2024). Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496*.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   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: [NA]

Justification: It's a theoretical position paper; the whole paper was purposed to discuss the limitations of the current approach.

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 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: [NA]

Justification: No experiments.

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: [NA]

   Justification: No 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: [NA]

   Justification: No experiments.

   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: [NA]

   Justification: No experiments.

   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: [NA]

   Justification: No experiments.

   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: No harm with the paper.

   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: [Yes]

    Justification: The paper discuss both potential positive societal impacts and negative societal impacts of the work performed

    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: No data no model.

    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: [NA]

    Justification: No code no data.

    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: No new assets.

    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: Not an experimental paper.

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: No participants.

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 core method development in this research does not involve LLMs as any important, original, or non-standard components.

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.