Reversing The Lens: Using Explainable AI To Understand Human Expertise

Anonymous Author(s)
Affiliation
Address
email

Abstract

Both humans and machine learning models learn from experience, particularly in safety- and reliability-critical domains. While psychology seeks to understand human cognition, the field of Explainable AI (XAI) develops methods to interpret machine learning models. This study bridges these domains by applying computational tools from XAI to analyze human learning. We modeled human behavior during a complex real-world task – tuning a particle accelerator – by constructing graphs of operator subtasks. Applying techniques such as community detection and hierarchical clustering to archival operator data, we reveal how operators decompose the problem into simpler components and how these problem-solving structures evolve with expertise. Our findings illuminate how humans develop efficient strategies in the absence of globally optimal solutions, and demonstrate the utility of XAI-based methods for quantitatively studying human cognition.

1 Introduction

As the horizon broadens for the application of machine learning (ML) and large language models (LLMs) in physical sciences, the importance of understanding AI reasoning in complex environments is at its highest. Real-world, physical environments are inherently complex and pose formidable challenges for both human and artificial agents. To ensure safe and reliable deployment of AI systems in such environments, we must first understand and then improve their reasoning capabilities. However, current benchmarks of reasoning often fail to reflect the complexity of real-world tasks and detailed explanations remain limited to relatively simple tasks [1–3]. Moreover, AI reasoning remains fragmented across domains (e.g., mathematical, spatial, or general reasoning), which hinders an integrated understanding.

We believe a general set of methods to model and investigate reasoning in complex environments can help greatly. Despite immense progress in AI reasoning, humans consistently outperform AIs in complex environments [e.g., 4–8]. This superiority stems from their ability to reason efficiently in solving problems of enormous complexity with limited computational resources. General representations of complex reasoning can help learn from human solutions in improving AI reasoning, as well as improving human-AI alignment.

A rich line of research on human reasoning indicates that complex reasoning must be investigated and explained through processes or methods that are feasible to implement, given the environmental complexity [9–16]. Graph theory provides a general framework for achieving such process-level explanations of reasoning for both human and AI agents. Here, we implement a set of graph-based methods, commonly used in ML and XAI, to model how humans solve a complex real-world task at various levels of experience. The experimental task we use is tuning a particle accelerator, which requires complex reasoning in a large and uncertain search space. We represent the whole task as weighted graphs of its parameters for three experience groups. Thereafter, we examine (1) the

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

processes as subsets of task parameters using community detection algorithms and (1) the organization of the task parameters through hierarchical clustering. We find that the operators divide the task parameters into three subsets regardless of their experience level. However, we also find fine-grained changes in the structure underneath the similarity of partitions of the task.

2 Modeling and Explaining Behavior in Complex Environments

ML has been applied to a wide range of physical sciences, such as statistical physics, particle and quantum physics, quantum computing, and chemistry [17]. As an example relevant to our experimental paradigm, in particle accelerator operations, ML methods enable simulations of control systems, anomaly detection, uncertainty quantification, system design, and active control [18–22].

Importantly, we need to explain AI reasoning not just based on the inputs and the outputs, but also the formation of higher-level concepts that are necessary for efficient problem-solving [2, 23, 24]. The problems in physical sciences are generally complex and uncertain, which eliminates the possibility of finding optimal solutions using the traditional views of rationality [10, 25]. For such problems, humans use *bounded rationality*; that is, they approximate good enough solutions using heuristics that frequently outperform the state-of-the-art optimization algorithms in complex environments [4, 6, 26]. Crucially, the ML models are not immune to the complexity; thus, in improving their performance in complex environments, we need to teach them to reason efficiently as humans do.

We believe graph models of complex behavior provide a promising path to general explanations. Graph-theoretic models serve as the foundation for cognitive network science, which has been exceptionally successful in explaining complex problem solving and reasoning of humans [27, 28]. Graph-based methods also serve as a bedrock for improving and explaining the performance of neural networks [2, 29, 30]. In this work, we use graphs to model human performance in the complex task of tuning a particle accelerator and demonstrate the efficacy of graph-theoretic measures in capturing how humans navigate and master the task.

3 Methods Used

3.1 Experimental Task: Tuning Particle Accelerators

The accelerator we use is the Linac Coherent Light Source, a Free Electron LASER (FEL) at SLAC National Accelerator Laboratory. The goal of FEL tuning is to maximize the pulse intensity of the resultant X-ray beams, using a set of 27 tuning parameters. The search space of parameter values is enormous, making the task extremely complex. For illustration, there are $27! \approx 1.09 \times 10^{28}$ possible sequences to adjust the parameters in and $\approx 5.45 \times 10^{20}$ ways to partition the set into subsets.

3.2 Dataset and Participants

To examine how the operators deal with this complexity, we use a large archive of about 350000 texts logged by them on operations between 2009 and 2022. We obtained Institutional Review Board (IRB) approval for using this dataset for our study, and detailed measures were adopted to anonymize the data. To extract information related to FEL tuning and identify the parameters used, we parsed the logs with a host of natural language processing methods. For details of the steps and links to the data and code, please see [31]. The resulting data were divided into three groups based on experience level: (1) Novices (<1 year of experience), (2) Intermediates (1-4 years), and Experts (>4 years).

3.3 Graph Construction and Analysis

For each group, the graphs were constructed using the 27 parameters as nodes and the co-occurrences as edge weights between parameters. Thereafter, we examined the presence of groups (using community detection) and the organization of the parameters (using hierarchical clustering).

3.3.1 Community Detection

Communities are local structures in graphs, consisting of a subset of nodes that have high edge density within the subset and low density elsewhere. As community detection is an NP-hard problem, finding optimal partitions is intractable beyond small graphs, and heuristic-based approaches are used for large graphs [32]. We used two popular algorithms for community detection: (1) the Louvain algorithm, and (2) spectral clustering. The strength of partitions is measured by modularity, which compares the actual density of edges within communities to the density expected at random [33]. Modularity ranges between $[-1, 1]$. Values close to 0 indicate partitions no better than random, and values of 1 represent perfectly separated partitions. Values between 0.3-0.7 are considered to indicate *strong* partitions [33, 34].

3.3.2 Hierarchical Clustering

Communities represent sets of nodes that cluster together, but do not reveal the structure of the nodes; for this purpose, hierarchical clustering is a widely used method. We use agglomerative hierarchical clustering based on linkage methods [35, 36]; that is, we begin with individual nodes at the lowest level and cluster nodes based on pairwise distances as we progress to increasingly higher levels, until the cluster encompasses all nodes and converges to the entire graph.

4 Results

4.1 Consistent Communities in Graphs across levels of experience

Figure 1 displays the graphs for three groups of operators, where the nodes are colored according to the communities detected by the Louvain algorithm and verified using spectral clustering. For all three groups, modularity values of the partitions are well above 0.30, suggesting strong partitions. Importantly, the three groups demonstrate remarkable similarities in categorizing the subtasks into communities. We find exactly three communities in the networks for all groups. These communities are also largely similar, with only one or two subtasks being classified differently across groups (e.g., Parameter 0 for the experts and Parameters 3 & 4 for the intermediates). Upon consulting with domain experts, we learned that the community denoted in green consists of parameters related to beam transport and steering, the purple community corresponds to parameters that affect beam energy and compression, and the pink set consists of all other parameters.

The strong partitions indicate that humans divide the complex task into parts of manageable complexity. The similarities of communities demonstrated by the groups are quite striking, considering the extremely large number of possible partitions ($\approx 5.45 \times 10^{20}$). These similarities strongly suggest that operators at all stages of expertise can effectively recognize and categorize parameters into similar groups. Therefore, any differences in tuning performance with expertise are unlikely to stem from improvements in categorizing different parameters into communities.

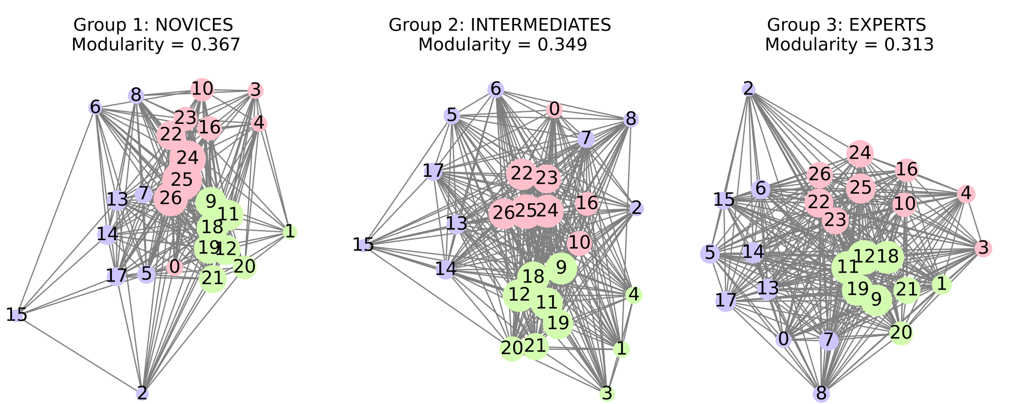

Figure 1: Graphs of FEL tuning for three groups of operators. The node sizes represent the PageRank values, and the distances between nodes represent the edge weights. Communities (denoted by colors) were identified using the Louvain algorithm and verified using spectral clustering.

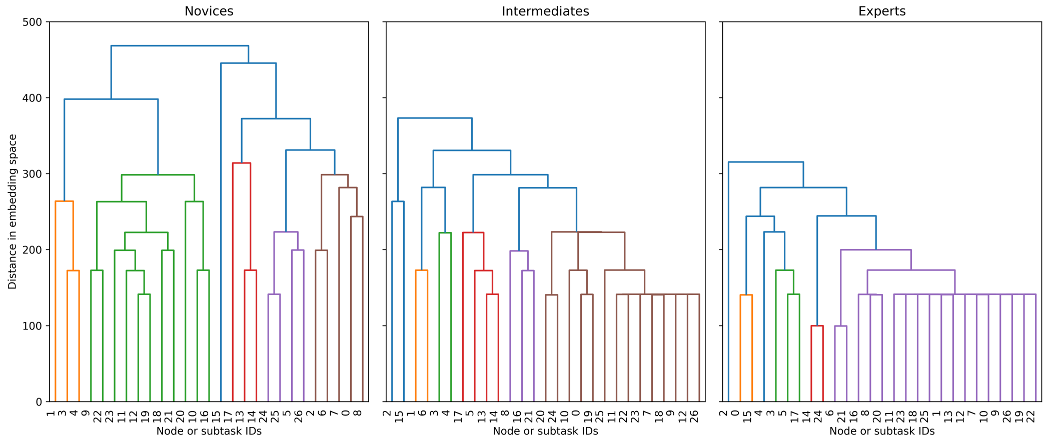

Figure 2: Hierarchical clustering of FEL tuning graphs for three experience groups. The height shows the distance between elements in the embedding space. The elements are clustered together based on this distance (denoted by the same color).

4.2 Evolving Hierarchical Structures within Communities with Experience

To examine how subtasks are organized into communities, we performed hierarchical clustering based on linkage methods that aim to group elements based on their distances in graph embedding space. The results are shown as dendrograms in Figure 2. The height of the dendrograms represents the distance at which two nodes are considered to belong to the same cluster. The horizontal connections mark the distance at which the elements connect or become part of the same group.

As we see, the dendrograms decrease in height with increased expertise, indicating that with experience, the subtasks became closer in distance and the graphs became denser. For novice operators, the subtasks are grouped together at much higher distances than for the other groups. Moreover, the distances varied considerably more for the novices than for others. Finally, the differences in structure appear to be larger between the novices and the intermediates than between the intermediates and the experts, reflecting a steep learning curve for novice operators.

5 Conclusions

Our two sets of results indicate that, underneath the similarities in the communities, the frequency and sequencing of subtasks may change considerably with expertise. Surprisingly, the communities remained the same at all experience levels, despite the large scope of differences. The modularity values also indicate *strong* partitions that are unlikely to be found at random. These results strongly support a divide-and-conquer strategy often observed in human problem solving. As optimizing parts of complex systems does not guarantee global optimality, this strategy is a boundedly rational approach, one that enables us to solve problems of enormous complexity using limited computational resources. To improve AI reasoning in complex environments, we need to train models to be efficient in resource use, for which human performance provides a roadmap.

While our study needs to be replicated for a larger set of tasks to generalize the findings, it highlights the need to examine and explain AI reasoning with models that accurately reflect the complexity of the task at hand. As there are numerous paths of reasoning, we need to specify the actual processes used by the AI agents, as we do for human agents. Otherwise, we may expect abilities or processes beyond the agents for the given problem. Notably, in cognitive models, human process or strategy selection is often modeled and explained as rational meta-reasoning among alternatives based on some form of reinforcement learning [8, 37–40], but at the cost of modeling a part of the process as a *blackbox* [14]. Therefore, general methods to probe intelligent behavior and reasoning in complex environments may lead to an integrated understanding and accurate benchmarks, helping to maximize the effectiveness of human-AI teams in complex, uncertain environments of the real world.

References

[1] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[2] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[3] Lance Ying, Katherine M Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L Griffiths, and Joshua B Tenenbaum. On benchmarking human-like intelligence in machines. *arXiv preprint arXiv:2502.20502*, 2025.

[4] Gerd Gigerenzer. Why heuristics work. *Perspectives on Psychological Science*, 3(1):20–29, 2008.

[5] Peter Bossaerts and Carsten Murawski. Computational complexity and human decision-making. *Trends in Cognitive Sciences*, 21(12):917–929, 2017.

[6] Gerd Gigerenzer. What is bounded rationality? In *Routledge Handbook of Bounded Rationality*, pages 55–69. Routledge, 2020.

[7] Roussel Rahman. Dynamics of individual learning (Publication No. 29261428) [Doctoral dissertation, Rensselaer Polytechnic Institute], 2022.

[8] Catherine Sibert and Roussel Rahman. The need for speed? exploring the contribution of motor speed to expertise in a complex, dynamic task. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[9] Allen Newell, John Calman Shaw, and Herbert A Simon. Elements of a theory of human problem solving. *Psychological Review*, 65(3):151–166, 1958.

[10] Herbert A Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6):467–482, 1962.

[11] Herbert A Simon and Allen Newell. Human problem solving: The state of the theory in 1970. *American Psychologist*, 26(2):145–159, 1971.

[12] Yuichiro Anzai and Herbert A Simon. The theory of learning by doing. *Psychological Review*, 86(2):124–140, 1979.

[13] Herbert A Simon. What we know about learning. *Journal of Engineering Education*, 87(4):343–348, 1998.

[14] Gerd Gigerenzer. How to explain behavior? *Topics in Cognitive Science*, 12(4):1363–1381, 2020.

[15] Roussel Rahman and Wayne D Gray. Spotlight on dynamics of individual learning. *Topics in Cognitive Science*, 12(3):975–991, 2020.

[16] Roussel Rahman and Wayne D Gray. Towards precise measures of individual performance in complex tasks. In Terrence C Stewart, editor, *Proceedings of the 19th international conference on cognitive modeling*, pages 227–233. Applied Cognitive Science Lab, Penn State., 2021.

[17] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

[18] Auralee Edelen, Christopher Mayes, Daniel Bowring, Daniel Ratner, Andreas Adelmann, Rasmus Ischebeck, Jochem Snuverink, Ilya Agapov, Raimund Kammering, Jonathan Edelen, et al. Opportunities in machine learning for particle accelerators. *arXiv preprint arXiv:1811.03172*, 2018.

[19] Auralee Edelen, Nicole Neveu, C Mayes, C Emma, and D Ratner. Machine learning models for optimization and control of x-ray free electron lasers. In *NeurIPS Machine Learning for the Physical Sciences Workshop*, 2019.

[20] Aashwin Ananda Mishra, Auralee Edelen Linscott, and Adi Hanuka. Bayesian neural networks for uncertainty estimation in particle accelerator applications. In *Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.*, pages 1–6, 2020.

[21] Lipi Gupta, Auralee Edelen, Nicole Neveu, Aashwin Mishra, Christopher Mayes, and Young-Kee Kim. Improving surrogate model accuracy for the lcls-ii injector frontend using convolutional neural networks and transfer learning. *Machine Learning: Science and Technology*, 2(4):045025, 2021.

[22] Joseph Duris, Dylan Kennedy, Adi Hanuka, Jane Shtalenkova, Auralee Edelen, P Baxevanis, Adam Egger, T Cope, M McIntire, S Ermon, et al. Bayesian optimization of a free-electron laser. *Physical review letters*, 124(12):124801, 2020.

[23] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[24] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[25] Herbert A Simon. From substantive to procedural rationality. In *25 Years of Economic Theory: Retrospect and prospect*, pages 65–86. Springer, 1976.

[26] Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118, 1955.

[27] Cynthia SQ Siew, Dirk U Wulff, Nicole M Beckage, Yoed N Kenett, et al. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019, 2019.

[28] Yoed N Kenett, Nicole M Beckage, Cynthia SQ Siew, and Dirk U Wulff. Cognitive network science: A new frontier. *Complexity*, 2020:1–4, 2020.

[29] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159, 2024.

[30] Muhan Zhang. Graph neural networks: link prediction. *Graph Neural Networks: Foundations, Frontiers, and Applications*, pages 195–223, 2022.

[31] Anonymous Anonymous. Anonymous title. *Annyomous*, 1000.

[32] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[33] Mark EJ Newman. Analysis of weighted networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 70(5):056131, 2004.

[34] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[35] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[36] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[37] Falk Lieder and Thomas L Griffiths. Strategy selection as rational metareasoning. *Psychological Review*, 124(6):762–794, 2017.

[38] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43(1):e1: 1–60, 2020.

[39] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635, 2003.

[40] Ron Sun. Introduction to computational cognitive modeling. *Cambridge handbook of computational psychology*, pages 3–19, 2008.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions: applying XAI methods to model human expertise in tuning particle accelerators, revealing both the stable and the evolving problem-solving structures. These claims are supported by graph-based analyses and are reflected in the results and conclusions. Please see Sections 2, 4, and 5.

   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: The paper discusses limitations in terms of dataset scope (archival operator data), assumptions about task decomposition, and generalizability to other domains. These are mentioned in the Conclusions section and also throughout the paper.

   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: NA

   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: The paper describes the dataset (archival operator logs), graph construction methods, and clustering techniques in detail. These are sufficient to reproduce the main findings. See Section 3.

   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 dataset and all associated code are already shared in open-source repositories, but we do not include the links to maintain anonymity, as required for the double-blind process. If accepted, the links to the dataset and the codes would be included in the final paper.

   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: Please see Section 3 for a description that helps understand the results of the paper. Due to the page limit, we were unable to include all details. However, they are included in another article, which we anonymized for the double-blind process.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: The statistical significance of the graph partitions is provided in the form of modularity values, which distinguish significant or strong partitions from randomly expected ones. Specifically, modularity values equal to or more than 0.3 are considered proofs of strong partitions. Moreover, these results were verified using two different community detection algorithms. Please see Sections 3 and 4.

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 data processing and the model development were done on a moderately powered, consumer-level laptop. As we expect the results to be reproducible using almost any modern machine (or even Google Colab with a free tier), we did not discuss the computational hardware here.

   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: This study adheres to the NeurIPS Code of Ethics. It utilizes anonymized versions of archival data with institutional review board approval, ensuring that it avoids harm, respects privacy, and promotes transparency and reproducibility.

   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: We believe our study has only positive societal impacts, which is reflected throughout the paper. Our study focuses on developing appropriate explanations of intelligent behavior, a crucial area of research in light of the emergence of the LLMs. As LLMs and other AI models are frequently compared against human performance, our work lays the foundation for fair and accurate comparisons that can limit spurious claims of superhuman performance by AIs.

    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: These concerns are well addressed in the IRB approval for our study on human subjects, a reference to which will be included in the final version. However, we did not discuss these points in the main paper due to the 4-page limit.

    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 are the original creators of the dataset, which is mentioned in Section 3. As mentioned earlier, we refrain from including the links to the dataset to retain anonymity.

    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: The assets (e.g., data and code) are well documented and available in open-source repos. However, references to these assets are anonymized for the review process.

    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: NA

    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: [Yes]

    Justification: The study was reviewed and approved by the Institutional Review Board (IRB), which is mentioned in Section 3. It was determined that no foreseeable harm was likely to result for participants. Upon acceptance, full details of the IRB protocol will be provided in the Appendix.

    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: NA

    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.