How Different LLMs Behave in Unverifiable Decision Scenarios?

Zheng Jiang[1], **Wei Wang**[1,*], **Gaowei Zhang**[1], **Yang Feng**[2], **Yi Wang**[1,3,*]
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Nanjing University, Nanjing, China
[3]Shandong Key Laboratory of Advanced Computing, Inspur Computer Technology Co., Ltd, Jinan, China
Correspondence*: weiwang@bupt.edu.cn; wang@cocolabs.org

Abstract

The behaviors of Large Language Models (LLMs) as artificial social actors are largely underexplored, particularly in unverifiable scenarios where conventional benchmarking is often not applicable. Thus, examining their behaviors in such scenarios can help understand and improve LLMs' capabilities of simulating real-world social actors in many tasks such as LLM-empowered agents. We draw a typical unverifiable scenario–a simplified pull request scenario on GitHub focusing on decision-making based on Activity Overview signal–to investigate how human and LLMs behave. We introduce a method to collect, compare, and reason about human and LLMs' decisions. We reveal that there are both similarities and differences between human mind and LLMs' decisions, and proprietary LLMs generally behave more like human than open-source LLMs do. We further find that human and LLMs may rely on different information and reasoning mechanisms in decision-making. Our study thus urges more work on human and LLMs decision-making in unverifiable environments.

Keywords: Large Language Model; Decision-making; Unverifiable Environments

Introduction

In his 2025 LLM Year in Review, Andrej Karpathy pointed out that LLMs are "*at the same time a genius polymath and a confused and cognitively challenged grade schooler, seconds away from getting tricked by a jailbreak to exfiltrate your data*[1]." LLMs' jagged performance mostly results from the practice of developing intelligence through benchmarking in verifiable environments such as question answering (Kantharaj et al., 2022) or code generation (Peng et al., 2025), where an objective "ground truth" or "gold standard" exists.

However, in many real-world scenarios, it is almost impossible to construct verifiable environments to improve LLM's capability (C. Jiang et al., 2025; Toroghi et al., 2024; Zhou et al., 2023). A large proportion of human decision-making problems fall into these unverifiable scenarios, particularly in human's social interactions (Clavel, 2025; W. Zhang et al., 2025). People often need to make decisions with imperfect information and ambiguous goals, and there is almost no way to assess the consequences of their decisions directly (Jia et al., 2024; Liu et al., 2024). Therefore, verifiable benchmarks become unavailable. Given that LLMs are increasingly involved in decision-making in such scenarios, either playing the role of decision-makers or as the backbone of social agents (Piao et al., 2025; Piatti et al., 2024), it is imperative to establish a deep understanding of LLMs' behaviors relevant to human behaviors, examine these behaviors' characteristics, and inquiry their reasoning behind these behaviors. Thus, this could enable us to endow LLMs with capabilities comparable to those of humans.

In this paper, we focus on a typical unverifiable decision-making scenario, i.e., "pull request" on GitHub, a practice where LLMs are increasingly deployed as autonomous assistants or reviewers (Tao et al., 2024; Wang et al., 2024; Wölflein et al., 2025). In this scenario, source code reviewers determine which pull request shall be accepted from multiple ones correctly implementing the desired features of fixing the issues (Lin et al., 2025). Rather than merely focusing on the verifiable code, a reviewer has to make a decision with some ambiguous technical and social signals X. Zhang et al., 2022, e.g., the contributor's organizational identity or past activities, far from perfect information. Since the long-term impact of a contributor's pull request is uncertain and lacks an immediate gold-standard label (Yue et al., 2022), we cannot mathematically prove that accepting one's pull request is "*objectively better*" than accepting another's. While long-term acceptance rates might serve as a noisy, delayed ground truth, evaluating a contributor inherently necessitates subjective trade-offs (e.g., balancing project complexity with sustainable contribution patterns). Such a scenario thus provides an excellent arena to examine the limitations of LLM's intelligence in dealing with real-world decisions in unverifiable contexts.

We limit our scope to a specific signal on GitHub–Activity Overview (see Figure 1) that visualizes individual users' numerous activities of different types into a simple quadrilateral displaying on their profile page to "give viewers more context about the types of contribution" one makes (GitHub Docs, 2025). It has been proven to have significant impacts on decisions about pull-request acceptance (Z. Jiang et al., 2025; Xia et al., 2022). By separating it with other factors in a simulated pull request scenario, we can gain a deeper understanding of how LLMs behave relevant to human, through a series of experiments involving both human subjects and a diverse range of state-of-the-art LLMs. Our experiment yields 3.6k human decisions from human participants and replicates the same task with LLMs. Our analysis reveals:

1. There are both similarities and differences between human and LLMs' decisions in terms of preferences orders and consistencies. LLMs' decisions are also more likely to be

[1]https://karpathy.bearblog.dev/year-in-review-2025/

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

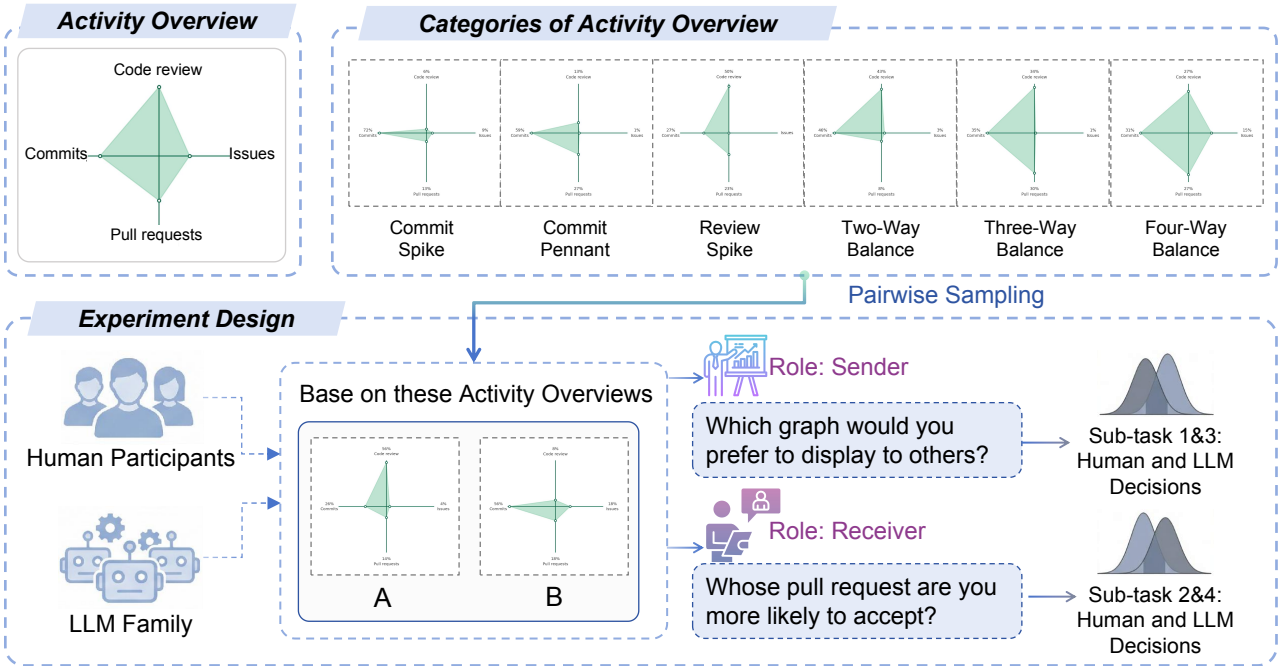

Figure 1: **Overview of the experimental framework.** The top panel illustrates an example of Activity Overview and its six main categories. The bottom panel details the pairwise comparison task, where both participants and LLMs assume distinct roles (*Sender* vs. Receiver) to make decisions.

biased by irrelevant information.

2. In general, proprietary LLMs behave more like human, while open-source LLMs are less like human.
3. Human and LLMs rely on different decision information and reasoning mechanisms. While LLMs differ in their reasoning strategies, each LLM maintains a high internal consistency in its reasoning.

Preliminaries: GitHub's Activity Overview and Its Categories

As we mentioned above, this study focuses on decision-making based on Activity Overview, a visual analytics on GitHub. It visualizes the proportions of a contributor's contributions in four types of activities: Commit, Code review, Issue, and Pull requests. Obviously, Activity Overview could exhibit different morphological patterns due to the heterogeneous nature of the distribution of contributors' activities. Z. Jiang et al. (2025) classified Activity Overviews into seven categories according to their morphological patterns, which are: Commit Spike, Commit Pennant, Review Spike, Two-Way Balance, Three-Way Balance, Four-Way Balance, and Others (Figure 1). In our study, we adopt their classification system and ask human participants and LLMs to make decisions based on a pair of Activity Overviews from two different categories. We also reuse the Activity Overviews collected by Z. Jiang et al. (2025) in our experiment tasks, which contain 1,338 valid Activity Overview snapshots from a diverse body of GitHub contributors.

Experiment Design

Research Questions

We formulate two RQs to guide our inquiry of LLMs' decision-making in unverifiable environments where the central challenge shifts away from correctness against a ground truth to the properties of LLMs' decisions, e.g., how similar they are relevant to human's, whether they exhibit coherent structure, and what leads to them. Therefore, considering *a simplified pull request decision scenario focusing on Activity Overview*, we have two research questions as follows:

RQ1. To what a degree do human and LLMs behave similarly and differently in decision-making with Activity Overview?

RQ2. What is the possible underlying reasoning of human and LLMs decisions?

Task

To answer the above RQs, we need to capture both human and LLMs' decisions. Hence, with a simplified pull-request scenario, we ask human participants and LLMs to make their decisions solely based on a contributor's Activity Overview while simply assuming other characteristics of the contributor and the contribution are equal. Given Activity Overview's nature as a social signal involving both *Senders* and *Receivers*, one may play either role in decision-making. First, contributors may be *Senders* who decide whether or not to show Activity Overview to seek the acceptance of their pull request; second, they can be *Receivers* who use it to decide whether or not to accept a contributor's pull request. Specifically, we had

two sub-tasks (1 & 2) for human participants and two similar sub-tasks (3 & 4) for LLMs (see Figure 1).

- Sub-task 1: Human Participants as Senders seeking pull request acceptance. In each trial of this sub-task, human participants are presented with two Activity Overviews from two different categories (see the lower part of Figure 1). They need to decide which one they would prefer to display to others.
- Sub-task 2: Human Participants as *Receiver* to determine a pull request's acceptance. In each trial of this sub-task, human participants are presented with two Activity Overviews from two different categories. They need to decide which pull request they are more likely to accept based on two Activity Overviews.

The only difference between sub-tasks (3 & 4) for LLMs and sub-tasks (1 & 2) is that each selected LLM simulates a human decision-maker.

Capturing Human Decisions

We employ a controlled laboratory study to capture human decisions which serve as the empirical reference for comparing with LLMs' decisions.

Participants We recruit 60 experienced developers with verified GitHub profiles. To prevent carry–over effects between roles, we use a between-subjects design: participants are randomly assigned to either the ***Sender*** **group** ($N = 30$) to perform sub-task 1 or the ***Receiver*** **group** ($N = 30$) to perform sub-task 2. Over 80% report at least one year of professional experience, and nearly all participants report high familiarity with the pull request workflow. The participants are fairly compensated at an hourly rate of about $12.

Procedure The experimental procedure is identical across both *Sender* and *Receiver* groups. Participants first give their consent and complete a structured tutorial. Then, each participant performs 60 pairwise comparison trials. In each trial of a sub-task, two Activity Overview charts are sampled from different morphological categories and presented side-by-side, with the left-right order randomized to mitigate position bias. Participants are instructed to rely on their first impression to make a decision, without access to any additional technical details or personal information. The groups differ only in their role-specific task instructions. The *Sender* group is asked to select the graph they would prefer to display to maximize their acceptance chances, while the *Receiver* group is asked to select the contributor they would be more inclined to accept. In total, we collect 3,600 human judgments (60 participants $\times$ 60 trials), yielding a diverse set of human decisions under unverifiable conditions.

Capturing LLMs' Decisions

Model Selection We select a diverse set of 20 representative LLMs as of October 2025, balancing provider diversity, accessibility, and architectural variation. Proprietary LLMs include leading vendors: OpenAI (GPT-5, GPT-4o), Anthropic (Claude-Sonnet-4, Claude-Sonnet-3.7), and Google (Gemini-2.5-Pro, Gemini-2.5-Flash). We also include top-performing open-source models from HuggingFace, spanning multiple families and parameter scales.

Prompt Strategies To ensure a fair comparison, LLM prompts are designed to strictly mirror instructions given to human participants. We first employ a role-playing system prompt to establish the persona, and then issue the same task instructions used in the human study: *"Which signal are you more inclined to present?"* for Senders and *"Whose pull request are you more inclined to accept?"* for Receivers. Models are required to make a forced choice in a standardized format, and all outputs are manually verified.

Results and Findings

RQ1: Comparing Human and LLMs' Decisions

We first compare the decisions of human and LLMs to answer the **RQ1**. We perform multiple comparisons: (1) the human participants' preference over Activity Overview categories vs. LLMs', (2) the inter-participant consistency of preference over different Activity Overview categories vs. the inter-LLM consistency, and (3) the human participants' preference of left-right order vs. LLMs'.

Preference over Activity Overview Categories To identify preference over Activity Overview categories, we transform subjective decisions in each trail's pairwise comparison into a continuous scale using the Bradley–Terry (BT) model, which yields scores and corresponding rankings over categories. The aggregated ranking is obtained by sorting categories according to BT scores estimated from all participants or all LLMs. We then compare preference patterns across human and LLMs by computing Kendall's Tau between their resulting rankings.

Human and LLMs' decisions exhibit both similarities and differences regarding their preferences over Activity Overview categories. Figure 2 shows that human participants' decisions form a clear and consistent preference pattern across both *Sender* and *Receiver* roles. They prefer more balanced activity distributions (e.g., "*Four-Way Balance*") and to overly-concentrated activity distributions (e.g., "*Commit Spike*"). In contrast, the preferences reflected by LLMs' decisions are different. In general, LLMs are more likely to put the category "*Review Spike*" over other categories. However, the order of other categories are basically very similar.

Proprietary and open-source LLMs show different preferences, and different levels of similarity to human decisions. Figure 2 gives some cues that proprietary and open-source LLMs show divergent preferences. First, the preferences of LLMs are basically very similar in both *Sender* and *Receiver* settings. For example, 5 out of 6 proprietary LLMs favor *Review Spike* most, and all put *Commit Pennant* and *Commit*

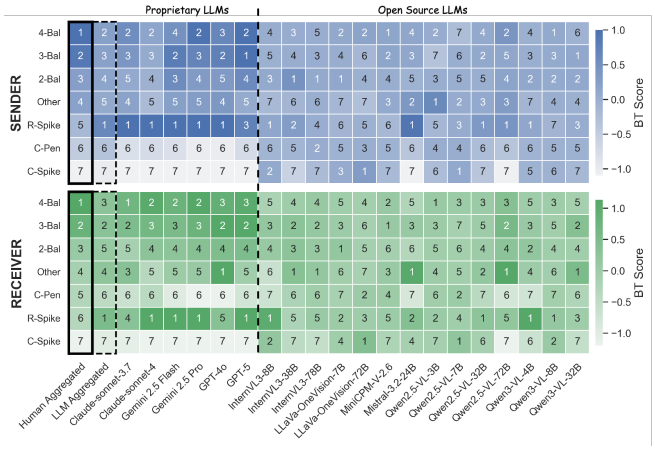

Figure 2: **Preference orders and Bradley-Terry scores across roles.** Heatmaps display preference orders for *Sender* (top) and *Receiver* (bottom) tasks. Colors represent score strength (Darker=Preferred), and numbers indicate rank (1=Best). Sorted by human preference intensity.

Spike on the bottom of the list in the *Sender* setting. However, open-source LLMs' preferences are more diverse. For example, in the *Sender* setting, only 5 out of 14 LLMs favor *Review Spike* most though it is still the category that got the most votes; meanwhile, in the *Receiver* setting, the most favorable category becomes *Other*.

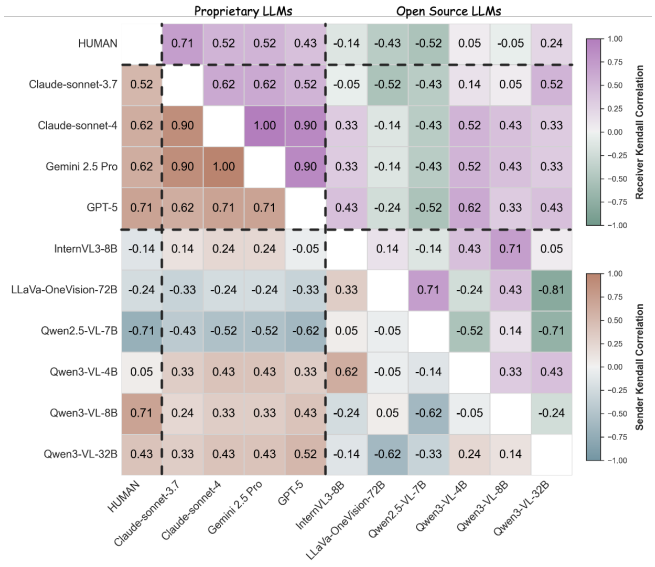

Figure 3: **Similarity of preference orders across human and LLMs.** The lower triangle reports correlations in the *Sender* setting, and the upper triangle reports correlations in the *Receiver* setting.

We further calculate the similarity among human and LLMs. Figure 3 presents the preference similarity across human and LLMs using the Kendall Correlations. Proprietary LLMs are more similar to each other, as well as strongly correlated to human's preferences. However, open-source LLMs record lower and more diverse correlations. Several open-source LLMs (e.g., LLaVa-OneVision-72B) have almost no similarity with others. It is fair to say that **proprietary LLMs behave more like human beings.** Moreover, there is no evidence that preference similarity increases with model scale, but instead varies across architectures and model families.

Inter-Human and Inter-LLM Preference Consistencies To measure inter-participant consistency of preferences, we compute Kendall's coefficient of concordance (W) over the category rankings obtained from all human participants, where values closer to 1 indicate stronger agreement. Similarly, each LLM is treated as an independent rater, and Kendall's W is computed across all model rankings to quantify inter-LLM consistency.

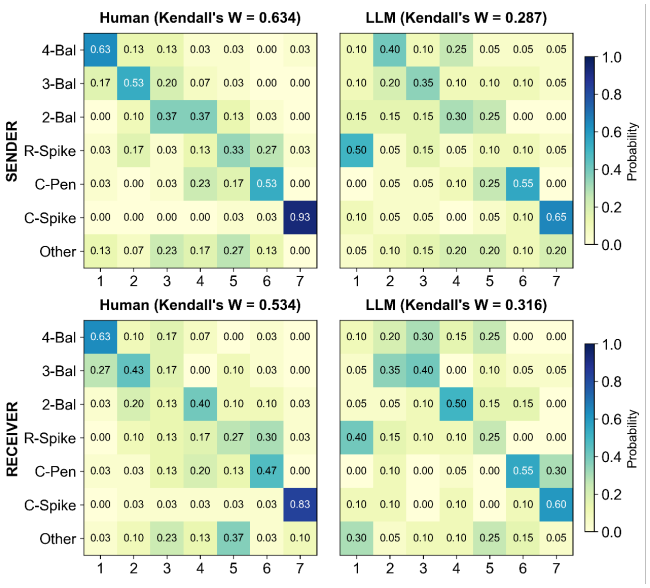

Figure 4: **Inter-rater consistency of preference rankings.** Heatmaps show the probability of each category appearing at each rank for human and LLMs in *Sender* and *Receiver*.

Human participants exhibit higher decision consistencies than LLMs do. The inter-human consistencies are relatively high in both the *Sender* (Kendall's W = 0.634) and *Receiver* settings (Kendall's W = 0.534). This agreement is particularly concentrated at the extremes of the ranking, e.g., in both roles, "*Four-way Balance*" is frequently ranked first, while "*Commit Spike*" is consistently ranked last; notably, over 83% of participants agree on ranking "*Commit Spike*" as the least preferred category. In contrast, inter-LLM consistency is remarkably lower in both settings (Kendall's W = 0.287 and 0.316). While LLMs exhibit consistency patterns that are similar to those of human, their agreement is more dispersed across ranking positions. For instance, across the seven categories, five categories do not exceed a 50% peak

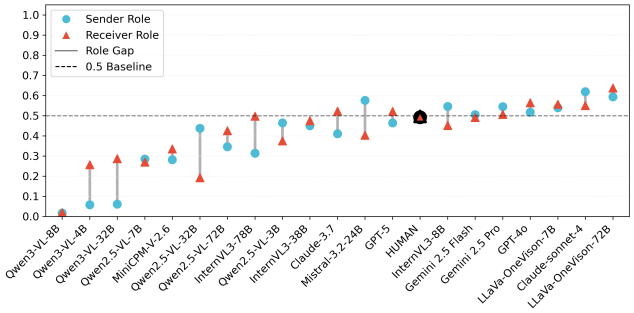

Figure 5: **Position preference across human and LLMs.** Markers show the mean probability of selecting the left option in *Sender* (blue) and *Receiver* (red) settings. Gray lines indicate role-dependent shifts. The dashed line at 0.5 denotes position-neutral decisions.

agreement in *Sender* role, and the highest agreement for any category reaches only about 65%. Similar dispersion patterns are observed in the *Receiver* condition.

Position Preference Since the left–right order in each trial has been randomized, a position neutral decision distribution should yield approximately balanced choices between the two options. We therefore measure the left-option selection rate for human and LLMs' decisions as a robustness check, to identify positional biases that are independent of stimuli.

Interestingly, LLMs exhibit position preferences absent in human decisions. Figure 5 shows that human judgments remain close to the unbiased baseline of 0.5, confirming the position neutrality. However, several LLMs exhibit apparent and consistent position preferences, either favoring the left option or shifting positional choice across roles, particularly the LLMs from the Qwen family, which consistently prefer the right side while Qwen3-VL-8B almost always chooses the right side. These deviations persist despite category randomization, indicating sensitivity to presentation order rather than task content. Notably, open-source LLMs are more likely to have position preferences than proprietary LLMs.

Answers to RQ1

The decisions of human and LLMs exhibit both similarities and differences in the simplified pull request scenario. In general, proprietary LLMs are more like human, while open-source LLMs show more heterogeneous preferences and are less similar to human. Moreover, there are many stronger inter-human consistencies than inter-LLM consistencies. Besides, LLMs have position preferences irrelevant to the Activity Overview while human do not.

Table 1: **Visual Cue Regression.** Standardized coefficients (β) for human and LLMs. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

	Activity		Geo. Aesthetics			Semantic	
Model	**Area**	**Ent**	**L/V**	**V.Sym**	**H.Sym**	**Maint**	**Sim**
Sender							
HUMAN	0.80^{***}	0.47	0.82^{***}	-0.64^{***}	-0.83^{**}	-0.40^{***}	0.09
Claude-sonnet	-0.25	1.36^{***}	2.03^{***}	-0.10	1.06^{***}	1.68^{***}	-0.06
Gemini 2.5 Pro	0.33	4.87^{***}	4.89^{***}	0.42^{*}	2.86^{***}	2.43^{***}	0.32^{*}
Qwen3-VL-32B	0.08	0.45	1.84^{***}	0.19	1.52^{***}	0.03	0.17
Receiver							
HUMAN	0.76^{***}	0.48	0.29^{*}	-0.79^{***}	-0.86^{***}	-0.39^{***}	0.01
Claude-sonnet	0.08	-0.00	0.82^{***}	-0.73^{***}	-0.72^{**}	0.02	-0.23^{***}
Gemini 2.5 Pro	1.13^{**}	4.92^{***}	4.56^{***}	0.17	2.30^{***}	3.83^{***}	0.92^{***}
Qwen3-VL-32B	0.62^{***}	-0.89^{**}	1.50^{***}	-0.82^{***}	-0.99^{***}	-0.93^{***}	0.21^{**}

RQ2: Reasoning of Human and LLMs' Decisions

To answer the **RQ2**, we further explore the possible underlying reasoning behind human and LLMs' decisions. We approach this problem from two aspects. First, we examine the correlation between Activity Overview's visual features and human and LLMs' decisions; then, we extract and analyze LLMs' chain of thought in making their decisions.

Activity Overview's Visual Features We first encode each Activity Overview into a set of interpretable features capturing different aspects of activity patterns. To improve statistical robustness and reduce multicollinearity, we filter the raw features into seven lowly correlated features in three classes: (1) **Activity Distribution**, including *Visual Area* and *Visual Entropy*; (2) **Geometric Aesthetics**, including *Long/Width Ratio*, *Vertical Symmetry*, and *Horizontal Symmetry*; and (3) **Semantic Cues**, including *Maintainer Status*, and *Cosine Similarity*. We then model each decision with a logistic regression to estimate how each feature is correlated with the decision.

Human and LLMs utilize different Activity Overviews' visual features in different ways in decision-making. Table 1 exhibits apparent differences between human and LLMs. Human tends to rely more heavily on salient and directly observable cues, such as *Area*, which can be readily perceived from the Activity Overview, while LLMs can use computed features (e.g., *Visual Entropy*) beyond human's direct observation. In addition, LLMs exhibit different tastes in geometric aesthetics from human. For example, in the *Sender* setting, their decisions are more associated with the *Long/Width Ratio*, less associated with the *Vertical Symmetry*, and in opposite directions regarding the *Horizontal Symmetry*. In addition, these features' impacts on human decisions are quite consistent across the *Sender* and *Receiver* settings but could vary significantly for LLMs' decisions. These differences suggest that **human and LLMs may have fundamentally different reasoning even when making similar decisions**.

LLMs' Chain of Thoughts To probe how models reason about their decisions, we focus on a specific comparison: *Review Spike* vs. *Commit Pennant*. We collect reasoning texts

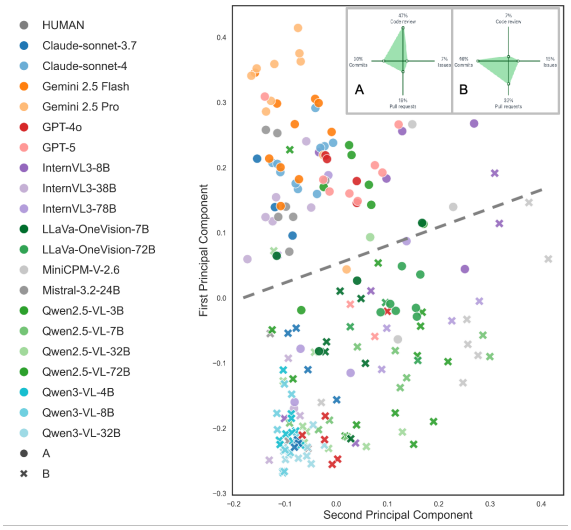

Figure 6: **Semantic embedding projection** of reasoning (*Review Spike* vs. *Commit Pennant*), with temp=1 applied to induce the variance needed for evaluating internal consistency.

using chain-of-thought (CoT) method (Wei et al., 2022) across 10 repeated runs per LLM, and project their sentence embeddings[2] to visualize similarities and differences in reasoning.

LLMs apply distinct but internally consistent reasoning. Different LLMs rely on distinct internal standards when justifying their decisions. For example, Gemini-2.5-Pro consistently treats review-heavy activity as a signal of code quality, whereas Claude-3.7-Sonnet emphasizes balance and role adaptation. In contrast, InternVL models produce more dispersed explanations, indicating less stable evaluation criteria. Meanwhile, what models think aligns closely with what they choose: explanations leading to the same decision cluster tightly in embedding space (see Figure 6), even when final choices occasionally flip. This suggests that disagreement and variability arise from competing internal preferences rather than confusion or post-hoc rationalization.

Answers to RQ2

Human and LLMs rely on different information and reasoning mechanisms in decision-making. For LLMs, their reasoning may vary, but each LLM exhibits strong internal-consistency in its reasoning.

Discussion

Human benchmarking in unverifiable environments. The past year has seen a surge of decision-making benchmarks for evaluating LLMs and revealing both their growing capabilities and their differences from human behavior. However, most existing benchmarks focus on verifiable outcomes, where performance can be measured against clearly defined ground truth and optimized objectives. In contrast, our work highlights an important yet underexplored class of problems–unverifiable decision settings–where there is no clear ground truth and the consequences of decisions cannot be directly assessed. In these settings, we find that LLM can behave quite differently from human and can follow totally different logic. Traditional correctness-based benchmarks are therefore insufficient. Therefore, human decision behaviors may become a meaningful reference in such contexts for guiding LLMs' evolution. The high level of consistency among human participants indicates that, even without verifiable outcomes, people tend to converge on shared intuitions or norms about what is most and least desirable. As a result, aggregated human decisions can serve as a stable reference for evaluating LLM decision-making in unverifiable environments, pointing toward a complementary benchmark paradigm that emphasizes behavioral patterns rather than correctness alone.

Proprietary vs. open-source LLMs in unverifiable decisions. As open-source and proprietary LLMs rapidly advance in parallel, both types of models are now able to achieve strong performances on verifiable decision benchmarks. However, our findings reveal clear differences between the two in unverifiable decision settings. Proprietary models exhibit more consistent behavior patterns and decision logic that are similar to those of human. In contrast, open-source LLMs diverge significantly from human behavior, even in this simplified task, exhibiting higher inconsistency and, in some cases, task-irrelevant biases. For example, across model sizes and versions, the Qwen family LLMs consistently produce unstable and disorganized ranking patterns though achieving huge successes on many LLM leaderboards that only consider verifiable tasks. We have no clear reason for such differences, but they are likely to stem from variation in training data, alignment strategies, and optimization objectives, which shape how models learn decision heuristics. In settings without explicit rewards or ground truth, such differences become pronounced.

Conclusion

This paper reports on our investigation of human and LLMs' decision-making in a specific unverifiable decision setting using a simplified GitHub pull-request scenario, where decisions are based on Activity Overview signals and no clear ground truth exists. Our results reveal systematic differences between human and LLMs in behavioral patterns, information and reasoning mechanisms: human decisions exhibit consistent preference structures, whereas LLMs show more variable and sometimes unstable behavior, with notable differences between proprietary and open-source LLMs. Future work can extend this approach to more unverifiable decision environments to further understand LLMs' behavior in relation to human's and identify potential improvements of LLMs' capabilities in these environments. The human and LLMs' decision data are available at: https://github.com/nicezheng/unverifiable-decision-data.

[2]Sentence embeddings from OpenAI's text-embedding-3-large API are used.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under grants 62076232 and 62172049.

References

Clavel, C. (2025, October). Understanding social interactions in the era of LLMs – the challenges of transparency. In H. L. Cardoso, R. Sousa-Silva, M. Koponen, & A. Pareja-Lora (Eds.), *Proceedings of the 2nd luhme workshop* (pp. 2–2). UP - Universidade do Porto (https://doi.org/10.21747/978-989-9193-73-4/lan2), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland; UAH - Universidad de Alcalá. https://aclanthology.org/2025.luhme-1.1/

GitHub Docs. (2025). Showing an overview of your activity on your profile [https://docs.github.com/en/account-and-profile/how-tos/setting-up-and-managing-your-github-profile/managing-contribution-settings-on-your-profile/showing-an-overview-of-your-activity-on-your-profile].

Jia, J. J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, *37*, 113360–113382.

Jiang, C., Zhang, Y., Cai, Y., Chan, C.-M., Liu, Y., Chen, M., Xue, W., & Guo, Y. (2025). Semantic voting: A self-evaluation-free approach for efficient llm self-improvement on unverifiable open-ended tasks. *arXiv preprint arXiv:2509.23067*.

Jiang, Z., Wang, W., Wang, L., Feng, Y., Zhang, M., Liu, L., Zhang, G., & Wang, Y. (2025). The paradox of being seen: Signaling contributors' unobservable activities limits their success in open collaboration. *Available at SSRN 5684204*.

Kantharaj, S., Do, X. L., Leong, R. T., Tan, J. Q., Hoque, E., & Joty, S. (2022). Opencqa: Open-ended question answering with charts. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11817–11837.

Lin, H. Y., Liu, C., Gao, H., Thongtanunam, P., & Treude, C. (2025, July). CodeReviewQA: The code review comprehension assessment for large language models. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Findings of the association for computational linguistics: Acl 2025* (pp. 9138–9166). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.476

Liu, O., Fu, D., Yogatama, D., & Neiswanger, W. (2024). Dellma: Decision making under uncertainty with large language models. *arXiv preprint arXiv:2402.02392*.

Peng, Y., Wan, J., Li, Y., & Ren, X. (2025). Coffe: A code efficiency benchmark for code generation. *Proceedings of the ACM on Software Engineering*, *2*(FSE), 242–265.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J. Y., Zhou, D., et al. (2025). Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., & Mihalcea, R. (2024). Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, *37*, 111715–111759.

Tao, W., Zhou, Y., Wang, Y., Zhang, W., Zhang, H., & Cheng, Y. (2024). Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, *37*, 51963–51993.

Toroghi, A., Guo, W., Pesaranghader, A., & Sanner, S. (2024). Verifiable, debuggable, and repairable commonsense logical reasoning via llm-based theory resolution. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6634–6652.

Wang, L., Zhou, Y., Zhuang, H., Li, Q., Cui, D., Zhao, Y., & Wang, L. (2024). Unity is strength: Collaborative llm-based agents for code reviewer recommendation. *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 2235–2239.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824–24837.

Wölflein, G., Ferber, D., Truhn, D., Arandjelovic, O., & Kather, J. N. (2025). Llm agents making agent tools. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 26092–26130.

Xia, X., Weng, Z., Wang, W., & Zhao, S. (2022). Exploring activity and contributors on github: Who, what, when, and where. *2022 29th Asia-Pacific Software Engineering Conference (APSEC)*, 11–20.

Yue, Y., Wang, Y., & Redmiles, D. (2022). Off to a good start: Dynamic contribution patterns and technical success in an oss newcomer's early career. *IEEE Transactions on Software Engineering*, *49*(2), 529–548.

Zhang, W., Liu, T., Song, M., Li, X., & Liu, T. (2025). Sotopia-: Dynamic strategy injection learning and social instruction following evaluation for social agents. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 24669–24697.

Zhang, X., Yu, Y., Gousios, G., & Rastogi, A. (2022). Pull request decisions explained: An empirical overview. *IEEE Transactions on Software Engineering*, *49*(2), 849–871.

Zhou, J. P., Staats, C. E., Li, W., Szegedy, C., Weinberger, K. Q., & Wu, Y. (2023). Don't trust: Verify–grounding llm quantitative reasoning with autoformalization. *The Twelfth International Conference on Learning Representations*.