Investigating Human and LLMs' Decisions in Unverifiable Environments: A Case Study with GitHub Activity Overview

Zheng Jiang[1], Wei Wang[1,*], Gaowei Zhang[1], Yang Feng[2], Yi Wang[1,*],
[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] Nanjing University, Nanjing, 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 has little to help improve their abilities. 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 social 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 systematic method to collect, compare, and reason about human and LLMs' decisions. Our results reveal that there are both similarities and differences between human 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 future work on human and LLMs decision-making in unverifiable environments.

1 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 Unverifiable data*[1]." LLMs' jagged performance mostly results from the practice of developing intelligence through benchmarking in verifiable environments such as question answering (Xu et al., 2025) 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 (Toroghi et al., 2024; Jiang et al., 2025a; Zhou et al., 2023). A large proportion of human decision-making problems fall into these unverifiable scenarios, particularly in human's social interactions (Zhang et al., 2025; Clavel, 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 (Piatti et al., 2024; Piao et al., 2025), 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, we might be able to equip LLMs with the ability similar to human.

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; Wölflein et al., 2025; Wang et al., 2024). 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., 2025a). Rather than merely focusing on the verifiable code, a reviewer has to make a decision with some ambiguous technical and social signals (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

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

Findings of the Association for Computational Linguistics: ACL 2026, pages 27591–27618
July 2-7, 2026 ©2026 Association for Computational Linguistics

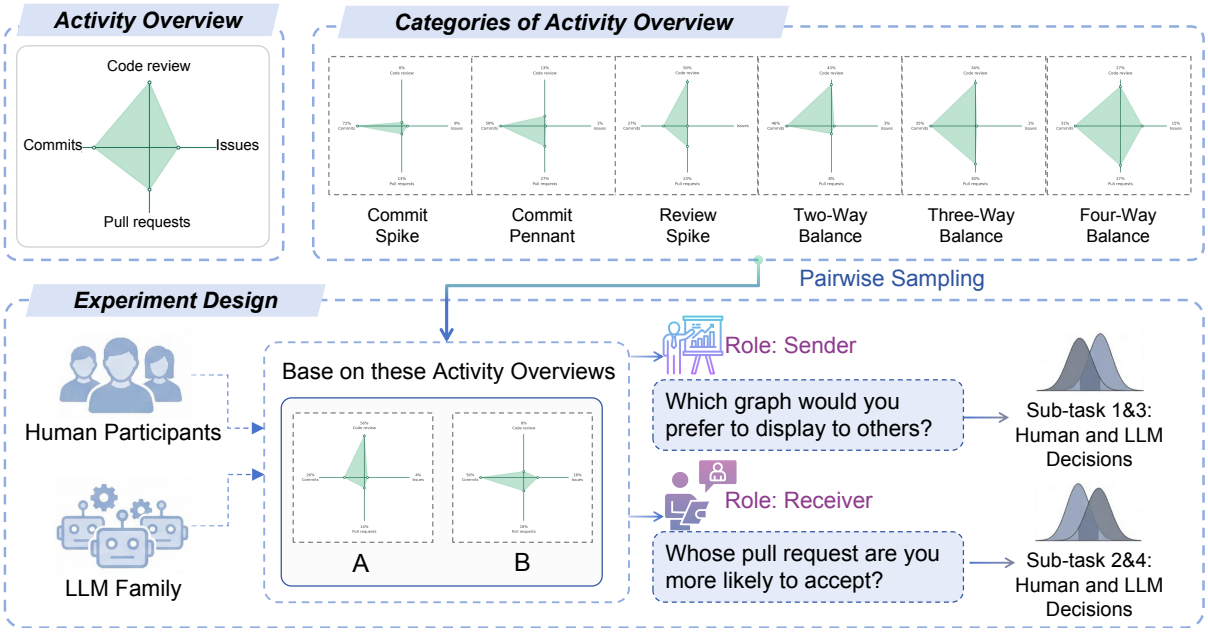


Figure 1: **Overview of the experimental framework.** The top panel illustrates an example of Activity Overview and its six main categories. The bottom panel details the pairwise comparison task, where both participants and LLMs assume distinct roles (*Sender* vs. Receiver) to make decisions.

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 (Xia et al., 2022; Jiang et al., 2025b). 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 of human and LLMs' decisions 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 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.

2 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. Jiang et al. (2025b) 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). A more detailed explanation of these categories is available in the Appendix A. 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 Jiang et al. (2025b) in our experiment tasks, which contain 1,338 valid Activity Overview snapshots from a diverse body of GITHUB contributors.

3 Experiment Design

3.1 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. Understanding these properties is critical for assessing whether LLM outputs can support informed decisions rather than arbitrary decisions. Therefore, considering *a simplified pull request decision scenario focusing on Activity Overview*, we have two research questions as follows:

RQ1. To what 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?

3.2 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 Activity Overview 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 the two sub-tasks (3 & 4) for LLMs that each selected LLM simulates a human decision-maker.

3.3 Capturing Human Decisions

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

3.3.1 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. Their details are in Appendix B. The participants are fairly compensated at an hourly rate of about $12.

Ethics Considerations. The study has been reviewed and approved as minimal risk by the Academic Ethics Committee of the authors' institution.

3.3.2 Procedure

The experimental procedure is identical across both groups. Participants first give their consent and complete a structured tutorial. Then, each participant performs 60 pairwise comparison trials. As we mentioned in Section 4.1, in each trial, 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 (see Appendix B). The *Sender* group is asked to select

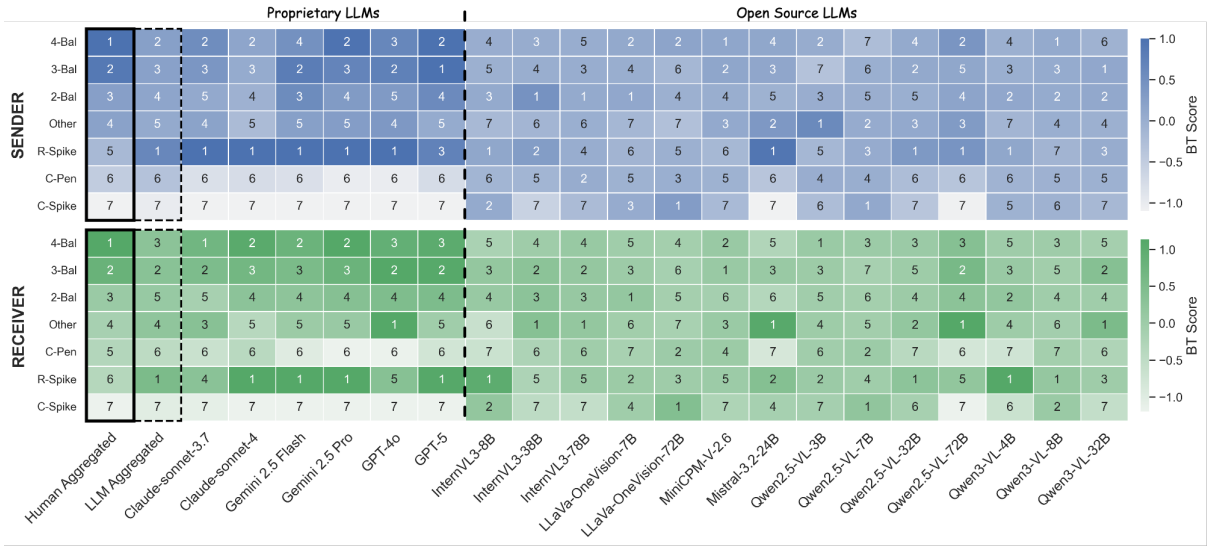


Figure 2: **Preference orders and Bradley-Terry scores across roles.** Heatmaps display preference orders for *Sender* (top) and *Receiver* (bottom) tasks. Colors represent score strength (Darker=Preferred), and numbers indicate rank (1=Best). Categories are sorted by human preference intensity.

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.

3.4 Capturing LLMs' Decisions

3.4.1 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. To disentangle the impact of visual perception from semantic reasoning, we further incorporate 9 pure-text LLMs (e.g., GPT-3.5-Turbo, DeepSeek-V3) as the text-only control group. More details are listed in Appendix C.

3.4.2 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. Full prompt templates for both roles are provided in Appendix D.

4 Results and Findings

4.1 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'.

4.1.1 Preference over Activity Overview Categories

To identify preference over Activity Overview categories, we transform subjective decisions in each trial'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. Complete metric definitions see Appendix G.

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 is basically very similar. Detailed scores for all models see Appendix Figure 12.

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 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 got most votes; meanwhile, in the *Receiver* setting, the most favorable category becomes *Other*.

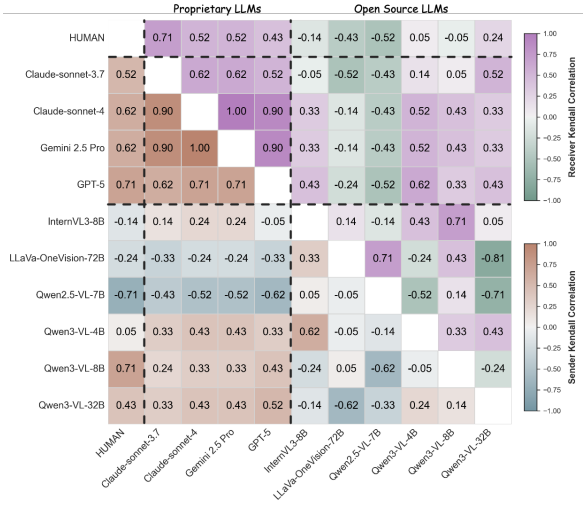


Figure 3: **Similarity of preference orders across human and LLMs.** The lower triangle reports correlations in the *Sender* setting, and the upper triangle reports correlations in the *Receiver* setting.

We further calculate the similarity among human and LLMs. Figure 3 (more in Appendix Figure 13), 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.

4.1.2 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 (see Appendix G), 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.

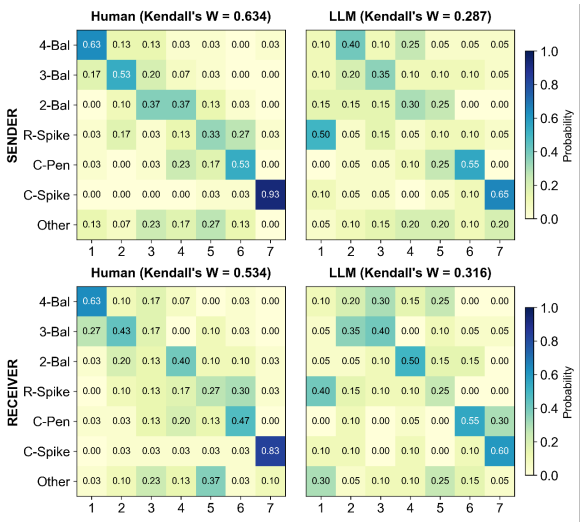


Figure 4: **Inter-rater consistency of preference rankings.** Heatmaps show the probability of each category appearing at each rank for human and LLMs in the *Sender* and *Receiver* settings.

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

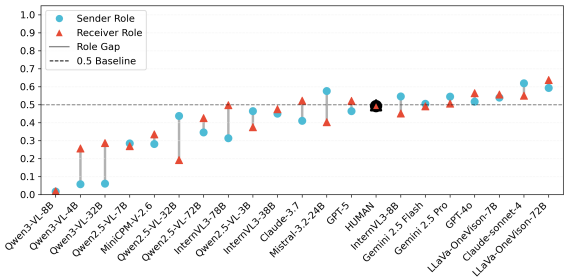


Figure 5: **Position preference across human and LLMs.** Markers show the mean probability of selecting the left option in *Sender* (blue) and *Receiver* (red) settings. Gray lines indicate role-dependent shifts. The dashed line at 0.5 denotes position-neutral decisions.

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

4.1.3 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 decision 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. It is easy to notice that 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 less similar to human. Moreover, there are stronger inter-human consistencies than inter-LLM consistencies. Besides, LLMs have position preferences irrelevant to the Activity Overview while human do not.

4.2 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.

4.2.1 Activity Overview's Visual Features

We first encode each Activity Overview into a set of interpretable features that capture different aspects of the activity pattern. 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*. Detailed definitions and statistics are provided in Appendix E. 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 indicates that there are 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

Table 1: **Visual Cue Regression (Selected Results).** Standardized coefficients (β) for a representative set of human and LLMs. Shaded rows denote section groupings. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Activity Distribution		Geometric Aesthetics			Semantic Cues	
Human&LLMs	**Area**	**Entropy**	**L/W Ratio**	**V.Sym**	**H.Sym**	**Maintainer**	**Similarity**
Sender							
HUMAN	0.80^{***}	0.47	0.82^{***}	-0.64^{***}	-0.83^{**}	-0.40^{***}	0.09
Claude-sonnet-3.7	-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-3.7	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^{**}

Sender setting, their decisions are more associated with by *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**. Appendix E provides all LLMs' regression results, which offer additional support to this argument and further indicate that open-source LLMs are less consistent than proprietary LLMs.

4.2.2 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 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 shifts its reasoning with role framing (see Appendix F). 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

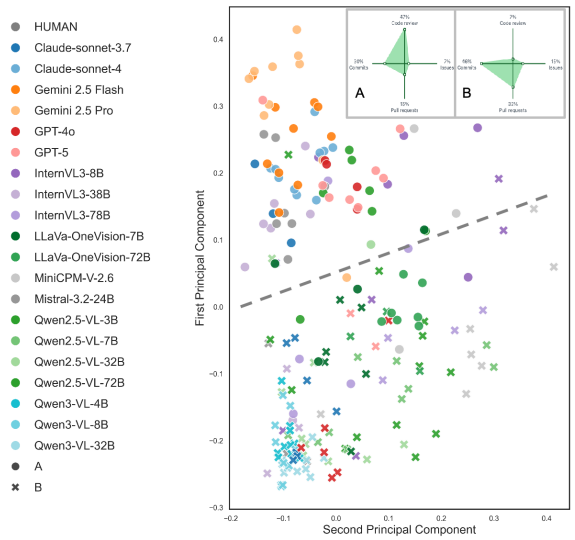


Figure 6: **Semantic embedding projection** of reasoning for *Review Spike* vs. *Commit Pennant*.

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.

4.3 Impact of Modality: Text-Only Control Experiment

To rigorously isolate logical reasoning from visual perception, we conduct a text-only control experiment. We losslessly convert the 4-dimensional visual data of the Activity Overview into a text-only format. We then evaluate several representative MLLMs on this text-only data, alongside pure text LLMs, to determine if the observed discrepancies stem from the models' reasoning capabilities or their visual encoders.

Removing visual cues affects MLLMs differently, but fundamental reasoning differences

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

remain . As shown in Table 8, models such as GPT-4o and the Qwen3-VL series no longer exhibit a significant reliance on the visually salient "Area" feature when processing text. Conversely, Claude-3.7-Sonnet and Mistral-3.2-24B remain largely unaffected by the modality shift. Notably, regardless of the input modality, all MLLMs consistently rely on computed features like "Visual Entropy"–a feature that human participants largely ignore.

Pure text LLMs exhibit an alignment gap with human independent of visual processing capabilities. The results of pure text LLMs in Table 9 further indicate that the divergence in decision-making persists at the semantic level. Similar to MLLMs, pure text LLMs heavily rely on computed metrics (e.g., Entropy) beyond human's direct observation, rather than adopting the heuristic cues prioritized by human participants. This reinforces our core finding that human and LLMs may have fundamentally different reasoning even when making similar decisions.

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.

5 Related Work

A large body of recent benchmark work evaluates LLM intelligence in decision-making systems, including multiple-choice question answering (Manakul et al., 2023), game-theoretic tasks (Mao et al., 2025), and risk decision problems (Jia et al., 2024). For example, LLMs are commonly evaluated in strategic or interactive games such as text-based games (Topsakal et al., 2024) or negotiation games (Abdelnabi et al., 2024), where LLM performance is quantified by maximized reward or reaching a predefined goal. These benchmarks reveal the gap between human and LLMs and allow controlled model comparisons. With such clearly predefined objectives, a range of mitigation strategies for improving LLM performance through techniques such as prompting strategies (Lin et al., 2025b), reinforcement learning with verifiable rewards (Wen et al., 2025), and LLM alignment (Tennant et al., 2024). While LLMs perform well in these verifiable systems, this evaluation paradigm does not extend well to unverifiable decision settings, where outcomes are uncertain and no gold standard exists. In many real-world situations, such as aesthetic evaluation (Di Dio et al., 2025), moral judgment (Garcia et al., 2024), or impression management on social media platforms (Yang and Zhang, 2022), decision quality cannot be objectively verified, and multiple choices may be equally reasonable. In these contexts, key factors such as preference (Li et al., 2024), internal consistency (Lee et al., 2025), and sensitivity to decision cues (Eigner and Händler, 2024) play a central role, rather than correctness with respect to a single answer. Despite the importance of such social scenarios, we still have limited evidence on how LLMs' decision behaviors compare to human behaviors in unverifiable social contexts. Our study addresses this underexplored regime by examining human and LLMs' decision-making within a controlled and well-defined scenario, providing insight on their potential for developing more general and socially grounded intelligence.

6 Discussion

Human benchmarking in unverifiable environments. The past year has seen a surge of decision-making benchmarks for LLMs, evaluating them from multiple perspectives 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 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 do and can follow totally different logic. Traditional correctness-based benchmarks are therefore insufficient. Therefore, human decision behaviors may become a meaningful benchmark in such contexts for identifying the potential directions to improve LLMs and guiding their evolution. The high consistency observed 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 decision-making in unverifiable environ-

ments, 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 more significantly from human behavior, even in this simple 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 especially visible.

7 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 relevant to human's and identify potential improvements of LLMs' capabilities in these environments. The human and LLMs' decision data supporting the results in this paper are available at: https://github.com/nicezheng/unverifiable-decision-data.

Limitations

This study has several limitations. First, our experiments focus on a simplified, well-controlled decision scenario centered on GITHUB Activity Overview presentation, which may not capture the complexity of an unverifiable decision environment in real-world. But even such a simplified environment reveals many differences between human and LLMs in decision-making, thus urging future work in further comparing human and LLMs in unverifiable environments. Second, although we compare a diverse set of proprietary and open-source LLMs, model behavior may evolve rapidly with new training data, alignment strategies, or inference techniques, and our results should be interpreted as a snapshot rather than a definitive characterization. Third, our analysis may also be restricted by our expertise. We may not be able to reveal more interesting or profound insights from our data. In addition, due to the space constraints, not all analyses and results can be reported within the page limit. We thus release all human and LLMs' decision data, so readers may play with it to develop more findings and insights.

Acknowledgments

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

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599.

Chloé Clavel. 2025. Understanding social interactions in the era of LLMs – the challenges of transparency. In *Proceedings of the 2nd LUHME Workshop*, pages 2–2, Bologna, Italy. 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 and UAH - Universidad de Alcalá.

Cinzia Di Dio, Martina Ardizzi, Sara Valentina Schieppati, Davide Massaro, Gabriella Gilli, Vittorio Gallese, and Antonella Marchetti. 2025. Art made by artificial intelligence: The effect of authorship on aesthetic judgments. *Psychology of Aesthetics, Creativity, and the Arts*, 19(5):1164.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Basile Garcia, Crystal Qian, and Stefano Palminteri. 2024. The moral turing test: Evaluating human-llm alignment in moral decision-making. *arXiv preprint arXiv:2410.07304*.

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.

Jingru Jessica Jia, Zehua Yuan, Junhao Pan, Paul McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, 37:113360–113382.

Chunyang Jiang, Yonggang Zhang, Yiyang Cai, Chi-Min Chan, Yulong Liu, Mingming Chen, Wei Xue, and Yike Guo. 2025a. Semantic voting: A self-evaluation-free approach for efficient llm self-improvement on unverifiable open-ended tasks. *arXiv preprint arXiv:2509.23067*.

Zheng Jiang, Wei Wang, Liu Wang, Yang Feng, Min Zhang, Libo Liu, Gaowei Zhang, and Yi Wang. 2025b. The paradox of being seen: Signaling contributors' unobservable activities limits their success in open collaboration. *Available at SSRN 5684204*.

Noah Lee, Jiwoo Hong, and James Thorne. 2025. Evaluating the consistency of llm evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10650–10659.

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024. Dissecting human and llm preferences. *arXiv preprint arXiv:2402.11296*.

Hong Yi Lin, Chunhua Liu, Haoyu Gao, Patanamon Thongtanunam, and Christoph Treude. 2025a. CodeReviewQA: The code review comprehension assessment for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9138–9166, Vienna, Austria. Association for Computational Linguistics.

Wenye Lin, Jonathan Roberts, Yunhan Yang, Samuel Albanie, Zongqing Lu, and Kai Han. 2025b. Gamebot: Transparent assessment of llm reasoning in games. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7656–7682.

Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. 2024. Dellma: Decision making under uncertainty with large language models. *arXiv preprint arXiv:2402.02392*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Mqag: Multiple-choice question answering and generation for assessing information consistency in summarization. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. Alympics: Llm agents meet game theory. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2845–2866.

Yun Peng, Jun Wan, Yichen Li, and Xiaoxue Ren. 2025. Coffe: A code efficiency benchmark for code generation. *Proceedings of the ACM on Software Engineering*, 2(FSE):242–265.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2024. Moral alignment for llm agents. *arXiv preprint arXiv:2410.01639*.

Oguzhan Topsakal, Colby Jacob Edell, and Jackson Bailey Harper. 2024. Evaluating large language models with grid-based game competitions: an extensible llm benchmark and leaderboard. *arXiv preprint arXiv:2407.07796*.

Armin Toroghi, Willis Guo, Ali Pesaranghader, and Scott Sanner. 2024. Verifiable, debuggable, and repairable commonsense logical reasoning via llm-based theory resolution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6634–6652.

Luqiao Wang, Yangtao Zhou, Huiying Zhuang, Qingshan Li, Di Cui, Yutong Zhao, and Lu Wang. 2024. Unity is strength: Collaborative llm-based agents for code reviewer recommendation. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 2235–2239.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, and 1 others. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelovic, and Jakob Nikolas Kather. 2025. Llm agents making agent tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26092–26130.

Xiaoya Xia, Zhenjie Weng, Wei Wang, and Shengyu Zhao. 2022. Exploring activity and contributors on github: Who, what, when, and where. In *2022 29th Asia-Pacific Software Engineering Conference (APSEC)*, pages 11–20. IEEE.

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Haitao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2025. Let llms take on the latest challenges! a chinese dynamic question answering benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10435–10448.

Hongjun Yang and Shengtai Zhang. 2022. Social media affordances and fatigue: The role of privacy concerns, impression management concerns, and self-esteem. *Technology in Society*, 71:102142.

Yang Yue, Yi Wang, and David Redmiles. 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.

Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025. Sotopia-: Dynamic strategy injection learning and social instruction following evaluation for social agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24669–24697.

Xunhui Zhang, Yue Yu, Georgios Gousios, and Ayushi Rastogi. 2022. Pull request decisions explained: An empirical overview. *IEEE Transactions on Software Engineering*, 49(2):849–871.

Jin Peng Zhou, Charles E Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. 2023. Don't trust: Verify–grounding llm quantitative reasoning with autoformalization. In *The Twelfth International Conference on Learning Representations*.

Appendices

A Dataset Construction

A.1 Data Source

We construct our dataset from real-world developer activity on GitHub, building on developers used in prior large-scale studies (Jiang et al., 2025b). These developers are sampled from the 200 most-starred public repositories, which span a wide range of software domains and include contributors from diverse geographic regions. For each developer, we collect an Activity Overview snapshot from their GitHub profile, which summarizes their relative engagement across four activity types: commits, pull requests, code reviews, and issues. Profiles with missing activity dimensions or insufficient contribution history are filtered out. After filtering, the dataset contains 1,338 valid Activity Overview snapshots.

A.2 Activity Pattern Categorization

We follow an established categorization scheme and employ a card-sorting procedure to group Activity Overview snapshots into six major activity patterns along with an additional *Other* category for less frequent or irregular shapes. This categorization is based on relative activity structure rather than absolute counts. Formal category definitions, distributions, and representative examples are provided in Table 2.

B Human Participants

B.1 Statistics of Human Participants

A total of 60 human participants took part in the study, evenly split between the *Sender* and *Receiver* roles ($N = 30$ each). As shown in Figure 7, participants exhibit a diverse demographic and technical background. The majority have substantial development experience, with most reporting three or more years of experience, and a large proportion actively using GitHub, having submitted pull requests or performed code reviews. In addition, while familiarity with the Activity Overview feature varies, participants overall possess sufficient domain knowledge to meaningfully engage with the task.

B.2 Task Instructions

Figure 8 shows the onboarding tutorial provided to human participants before the experiment, which introduces GitHub's Activity Overview visualization and explains how to interpret its four activity dimensions. After completing the tutorial, participants proceeded to the pairwise decision task. Figure 9 illustrates the human interface used in the experiment, showing an example trial from the *Sender* role, where participants compare two Activity Overview graphs and make a choice based on their first impression.

C Model Selection

This study evaluates a total of 20 MLLMs and 9 pure-text LLMs released or actively used around October 2025. The proprietary models are drawn from three major providers (OpenAI, Anthropic, and Google), and include multiple generations and variants within each family. For open-source models, we select high-performing multimodal LLMs from HuggingFace, covering a range of parameter scales and model families (e.g., Qwen, InternVL, LLaVA, MiniCPM, Mistral). All models are evaluated under a standardized prompting and default decoding setup. A complete list of models, along with their providers, model sizes, and inference pipeline, is provided in Table 3.

D Prompt Template

Figure 10 and Figure 11 present the full prompt templates used for LLM evaluation in the *Sender* and *Receiver* roles, respectively. To ensure a fair comparison with human participants, the prompts closely mirror the instructions used in the human study, including role specification, task description, and forced-format requirements. These prompts serve as the sole input for LLM decision-making in the pairwise task.

E Activity Overview Visual Features

Table 4 lists the complete set of features along with their descriptions and computational definitions, while Table 5 reports their basic statistics across all sampled Activity Overview instances. These features serve as the inputs for the regression analyses presented in the main text. Table 6 and Table 7 present the complete regression results for the *Sender* and *Receiver* settings

F Decision Explanations

Figure 14 and Figure 15 present illustrative examples of explanations produced by several representative LLMs in the *Sender* and *Receiver* settings,

respectively. In the *Sender* setting, GPT-4o and Gemini-2.5-Pro consistently justify their choices by emphasizing high code review activity as a signal of quality, while Claude-sonnet-3.7 focuses more on balanced contribution patterns. In contrast, Qwen models often emphasizing commits or pull requests. In the *Receiver* setting, Claude-sonnet-3.7 alternates between emphasizing review experience and balanced contribution profiles, and GPT-4o focuses more on pull request as an indicator of practical experience. These clustered explanations indicate that each model applies a consistent but model-specific internal criterion.

To exclude the confounding effect of varying visual understanding capabilities of LLMs, we have conducted a control experiment by prompting models to describe the content of the "Activity Overview" graphs. The results show that almost all models (both proprietary and open-source) achieved over 95% accuracy in describing the 4-dimensional metrics. Given that the baseline visual understanding of these graphs is saturated across the models, we argue that the observed differences stem primarily from internal decision-making logic rather than visual capabilities.

G Evaluation Metrics

G.1 Bradley–Terry model

The Bradley–Terry (BT) model transforms pairwise comparisons into a latent preference scale. Given two categories i and j with scores s_i and s_j, the probability that i is preferred over j is

$$P(i \succ j) = \frac{\exp(s_i)}{\exp(s_i) + \exp(s_j)}.$$

We estimate s_i by maximizing the likelihood over observed pairwise judgments. The resulting scores induce a total ranking over categories, enabling comparison of preference structures across humans and models.

G.2 Kendall's Tau

Kendall's τ measures the ordinal agreement between two rankings. Given two ranked preferences x and y, it is defined as

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j),$$

where n is the number of AO categories. $\tau = 1$ indicates identical rankings, $\tau = 0$ indicates no correlation, and $\tau = -1$ indicates reversed orderings. We use Kendall's τ to quantify similarity between preference rankings across agents and across repeated evaluations.

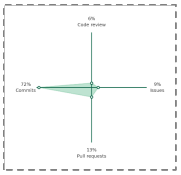
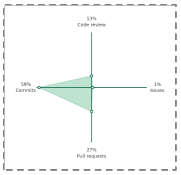
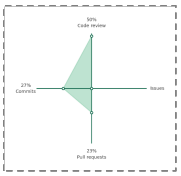
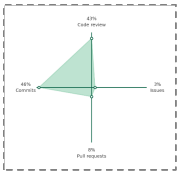
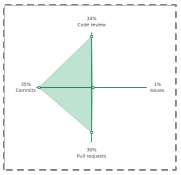
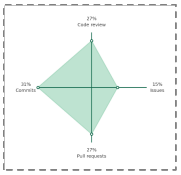
G.3 Kendall's coefficient of concordance (W)

Kendall's coefficient of concordance (W) measures how consistently decision-makers rank the categories. In our setting, each rater corresponds to a human participant or an LLM model (m), and each item corresponds to one of the seven categories (n). Let R_j denote the sum of ranks assigned to category j across all raters, and $\bar{R}$ the mean of these rank sums. Kendall's W is defined as:

$$W = \frac{12 \sum_{j=1}^{n} (R_j - \bar{R})^2}{m^2(n^3 - n)}.$$

Larger values of W indicate stronger consensus among raters.

Table 2: Morphological categories of Activity Overview. The six major Activity Overview morphological categories, detailed descriptions, frequencies (N and percentage), and a representative example.

Morphological Categories	**Descriptions**	*N* (%)	**Examples**
Commit Spike	*Commits* account for a vast majority in a contributor's activities.	289 (21.6%)	
Commit Pennant	*Commits* accounts for a majority in a contributor's activities, while there are certain amounts of *Code review* and *Pull requests*.	515 (38.5%)	
Review Spike	*Code review* accounts for an overwhelming majority in a contributor's activities.	139 (10.4%)	
Two-Way Balance	Two out of four types of activities account for substantial and similar proportions in a contributor's activities, while the other two only account for minimal proportions.	140 (10.5%)	
Three-Way Balance	Three out of four types of activities account for substantial and similar proportions in a contributor's activities, while the other only accounts for a minimal proportion.	146 (10.9%)	
Four-Way Balance	All four activities (*Commits, Code review, Issues, Pull requests*) take substantial shares with none dominating.	85 (6.4%)	

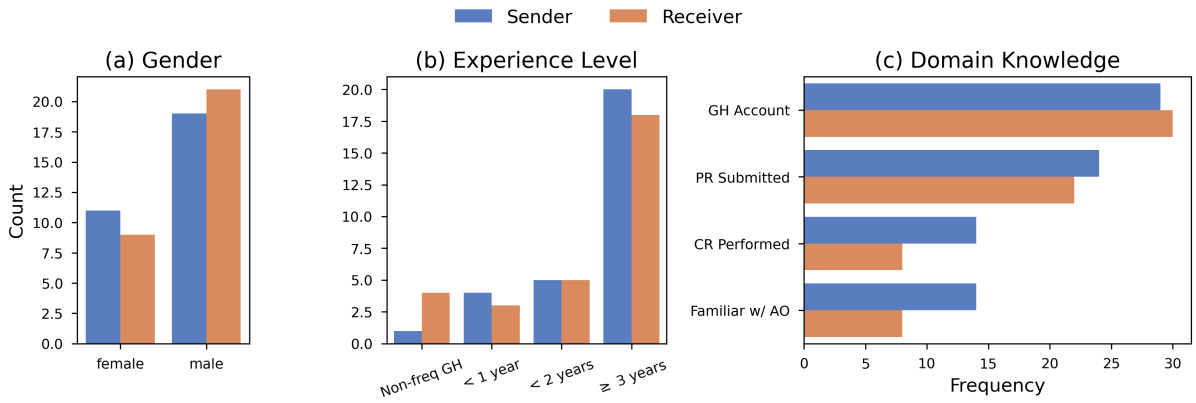
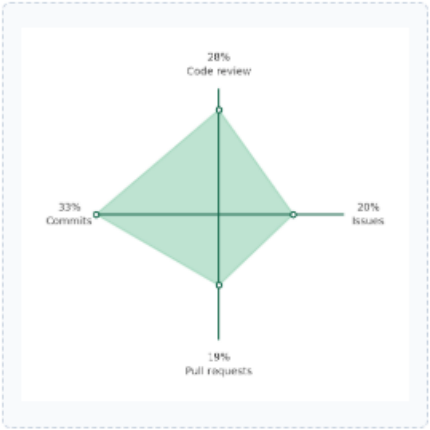


Figure 7: Demographic and background statistics of human participants.. (a) Distribution of gender across Sender ($N = 30$) and Receiver ($N = 30$) groups. (b) Development experience levels, where "*Non-freq GH*" indicates infrequent GitHub usage; "*< 1 year*", "*< 2 years*", and "$\geq$ *3 years*" denote the years of development experience. (c) Self-reported domain knowledge, including GitHub account ownership, experience submitting pull requests (PR Submitted), performing code reviews (CR Performed), and familiarity with the Activity Overview feature.

About Activity Overview

Activity Overview (AO) is a visualization feature introduced by GitHub in August 2018. AO provides a highly summarized view of a user's four main activities in the software development process: Code Review, Pull Request, Commits, and Issues. Through this graphical representation, developers can intuitively understand the relationships and workflows between these activities.

Activity Overview Graphic



Code Review

Team members review code submitted by others, checking for errors, ensuring quality standards are met, and providing suggestions for improvement.

Pull Request

Developers submit modified code to the team for discussion and review before merging.

Commits

When writing code, developers make a "commit" each time they make changes to record the content and reason for the code changes.

Issues

Issues are used to track various problems in a project, such as bugs, feature requests, or task assignments, helping the team track and manage work.

Example Interpretation

As shown in the figure above, the Activity Overview uses a diamond chart to display the relationships between the four activities:

- Pull Request and Code Review indicate the developer's activity level in modifying and reviewing code.
- Commits reflect the frequency of code changes and the developer's productivity.
- Issues show the developer's involvement in raising and resolving problems.

Figure 8: Onboarding tutorial shown to human participants prior to the experiment. The screenshot illustrates the explanation of GitHub's Activity Overview and guidance on how to interpret its visual representation.

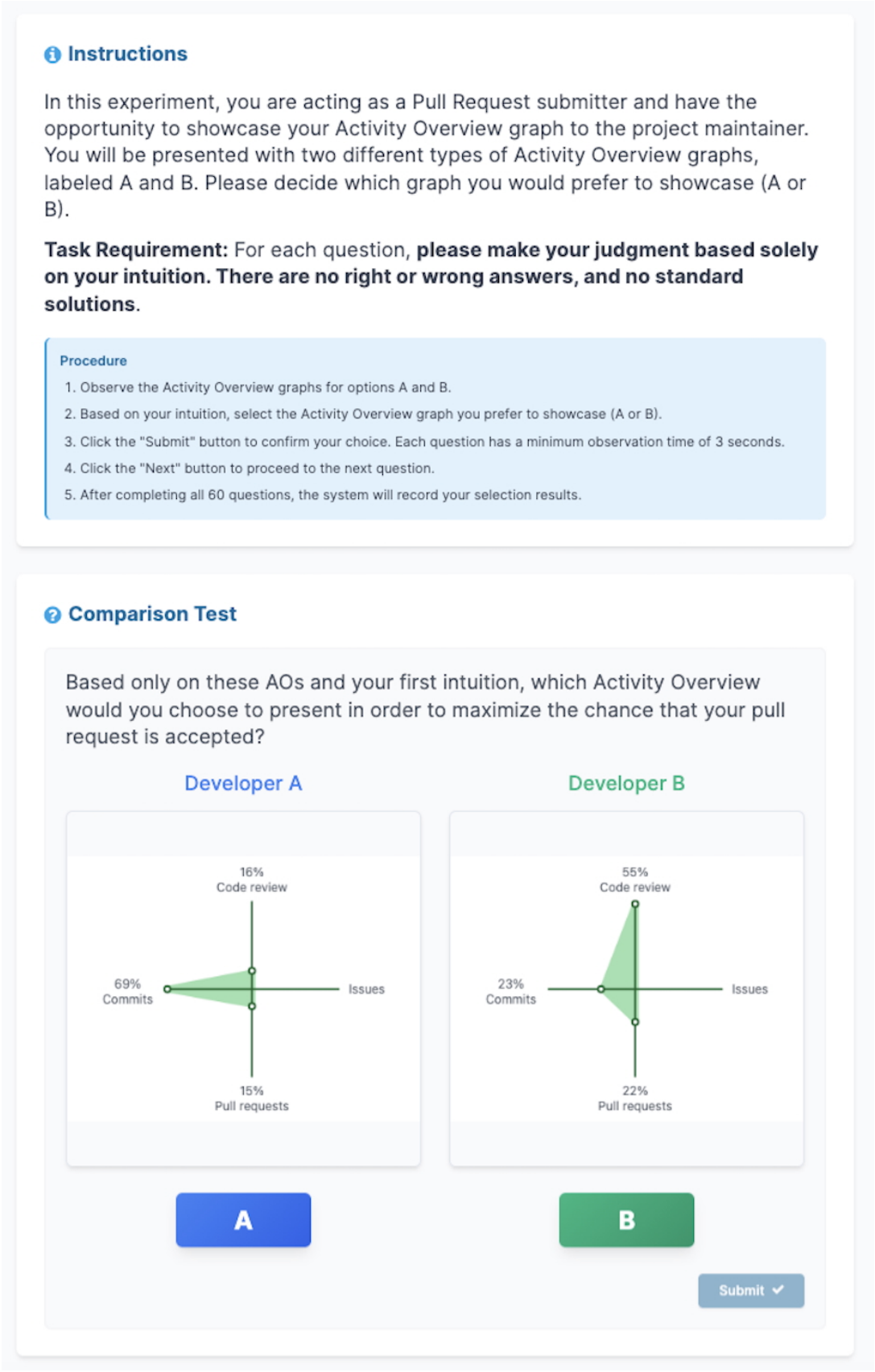


Figure 9: Human interface for the pairwise decision task. The screenshot shows an example trial from the *Sender* role, where participants compare two Activity Overview graphs and select their preferred option.

Table 3: Configuration details of Proprietary and Open-source LLMs

Label	Provider	Model	Release	Version	Inference Pipeline
Proprietary LLMs					
MLLM	OpenAI	GPT-5	2025-08	`gpt-5-2025-08-07`	API
		GPT-4o	2024-11	`gpt-4o-2024-11-20`	API
	Google	Gemini 2.5 Pro	2025-06	`gemini-2.5-pro`	API
		Gemini 2.5 Flash	2025-06	`gemini-2.5-flash`	API
	Anthropic	Claude-sonnet-4	2025-05	`claude-sonnet-4-20250514`	API
		Claude-sonnet-3.7	2025-02	`claude-sonnet-3.7-20250219`	API
Pure-text LLM	OpenAI	GPT-3.5-Turbo	2022-11	`gpt-3.5-turbo`	API
	Alibaba	Qwen-Max	2023-12	`qwen-max`	API
		Qwen-Plus	2023-12	`qwen-plus`	API
Open-source LLMs					
MLLM	Alibaba	Qwen3-VL-32B	2025-09	`Qwen3-VL-32B-Instruct`	vLLM
		Qwen3-VL-8B	2025-09	`Qwen3-VL-8B-Instruct`	vLLM
		Qwen3-VL-4B	2025-09	`Qwen3-VL-4B-Instruct`	vLLM
		Qwen2.5-VL-72B	2025-01	`Qwen2.5-VL-72B-Instruct`	vLLM
		Qwen2.5-VL-32B	2025-01	`Qwen2.5-VL-32B-Instruct`	vLLM
		Qwen2.5-VL-7B	2025-01	`Qwen2.5-VL-7B-Instruct`	vLLM
		Qwen2.5-VL-3B	2025-01	`Qwen2.5-VL-3B-Instruct`	vLLM
	Llava Transformers	LLaVA-OneVision-72B	2024-09	`llava-onevision-qwen2-72b-ov-hf`	vLLM
		LLaVA-OneVision-7B	2024-09	`llava-onevision-qwen2-7b-ov-hf`	vLLM
	OpenGVLab	InternVL3-78B	2025-04	`InternVL3-78B-Instruct`	vLLM
		InternVL3-38B	2025-04	`InternVL3-38B-Instruct`	vLLM
		InternVL3-8B	2025-04	`InternVL3-8B-Instruct`	vLLM
	OpenBMB	MiniCPM-V-2.6	2025-01	`MiniCPM-V-2.6`	vLLM
	Mistral AI	Mistral-3.2-24B	2025-06	`Mistral-Small-3.2-24B-Instruct-2506`	vLLM
Pure-text LLM	Alibaba	Qwen3-32B	2025-09	`Qwen3-32B-Instruct`	vLLM
		Qwen3-4B	2025-09	`Qwen3-4B-Instruct`	vLLM
		Qwen2.5-32B	2024-09	`Qwen2.5-32B-Instruct`	vLLM
		Qwen2.5-7B	2024-09	`Qwen2.5-7B-Instruct`	vLLM
	DeepSeek	DeepSeek-V3	2024-12	`DeepSeek-V3`	vLLM
	Meta	Llama-3.3-70B	2024-12	`Llama-3.3-70B-Instruct`	vLLM

Sender Prompt for Activity Overview Selection

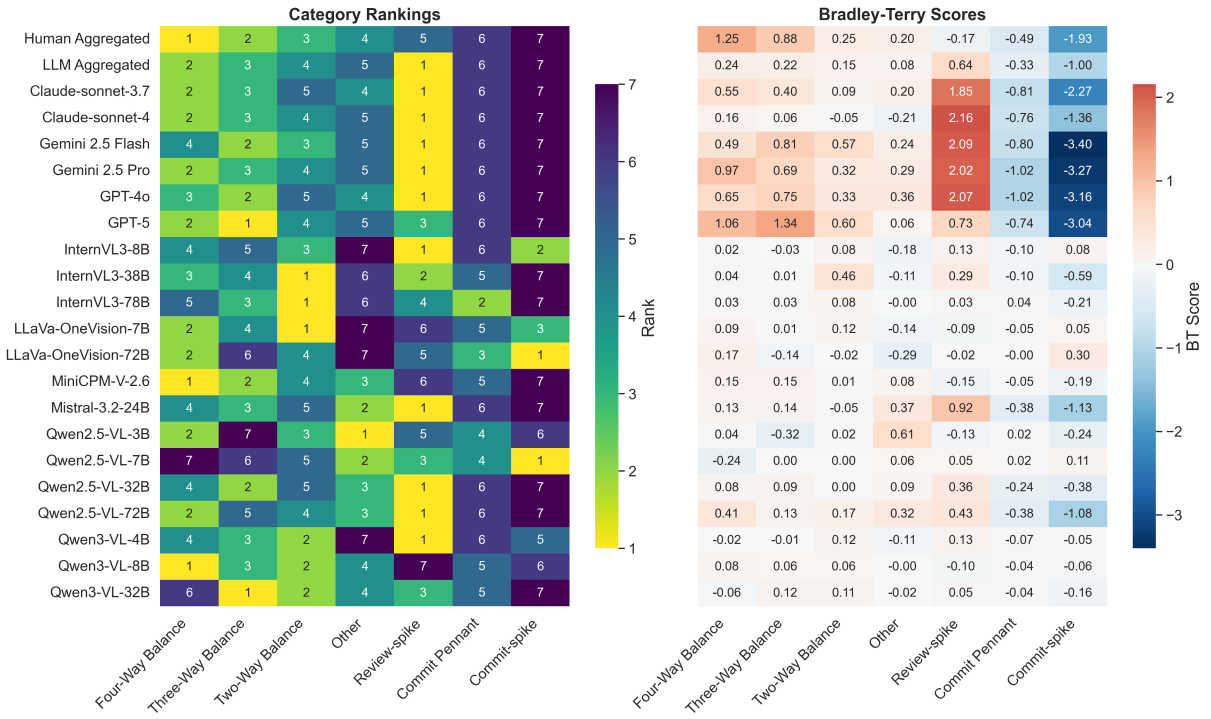
```
You are a DEVELOPER preparing to submit a pull request.
You are shown two Activity Overview (AO) visualizations that summarize how a
developer's past activity is distributed across Code Review, Pull Requests,
Commits, and Issues. Image 1 corresponds to Activity Overview A, and Image 2
corresponds to Activity Overview B. You can choose to present ONLY ONE of these
Activity Overviews as your public signal to a project maintainer.

Based only on these AOs and your first intuition, which Activity Overview would
you choose to PRESENT in order to maximize the chance that your pull request is
accepted?

Return ONLY a valid JSON object in the following schema (no markdown, no extra
text, and you must make a choice):
{
  "choice": "A" or "B"
}
```

Figure 10: Sender prompt used in the Activity Overview selection task

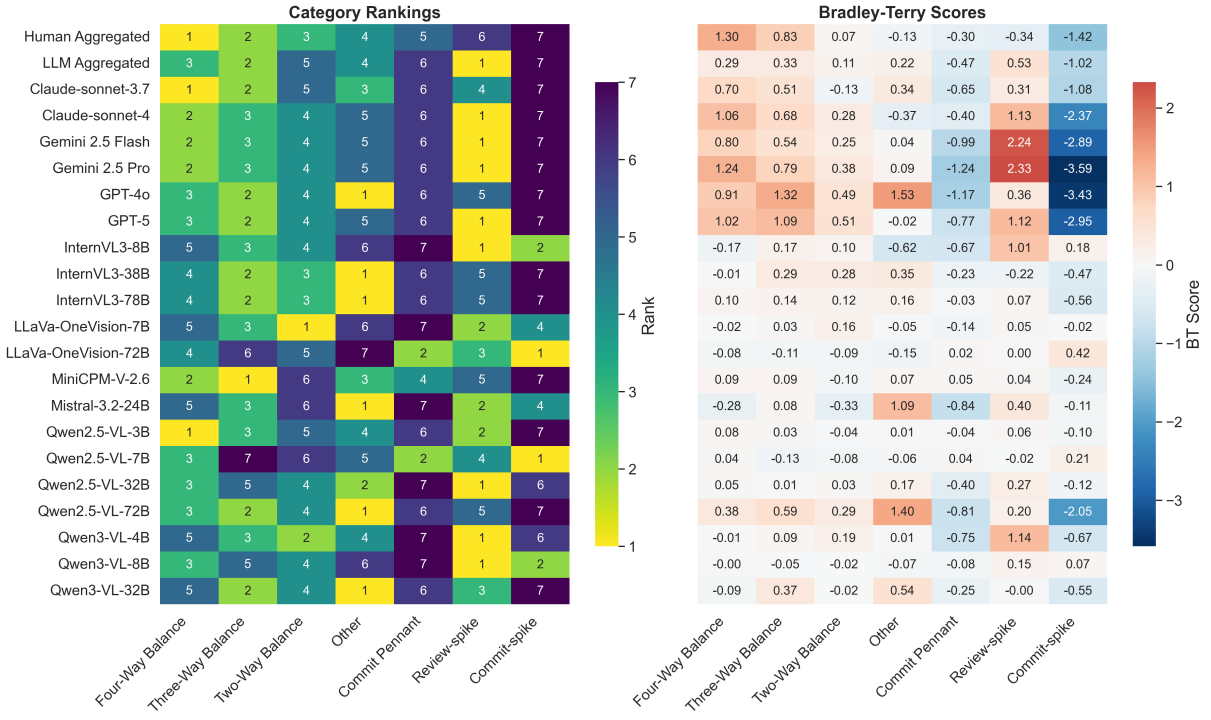
Receiver Prompt for Activity Overview Evaluation

```
You are a PROJECT MAINTAINER reviewing incoming pull requests.
You are shown two Activity Overview (AO) visualizations that summarize how each
developer's past activity is distributed across Code Review, Pull Requests,
Commits, and Issues. Image 1 corresponds to Developer A, and Image 2 corresponds
to Developer B. Each pull request is accompanied by one Activity Overview as a
signal of the contributor's past behavior.

Based only on these AOs and your first intuition, whose pull request would you be
more willing to accept?

Return ONLY a valid JSON object in the following schema (no markdown, no extra
text, and you must make a choice):
{
  "choice": "A" or "B"
}
```

Figure 11: Receiver prompt used in the Activity Overview evaluation task



(a) Rank and Bradley-Terry scores in the Sender experimental scenario.



(b) Rank and Bradley-Terry scores in the Receiver experimental scenario.

Figure 12: Comparative analysis of category rankings and Bradley–Terry (BT) scores across human and LLM in the Sender (top) and Receiver (bottom) scenarios. Each cell shows the preference rank (1 = most preferred), with color intensity indicating the BT score.

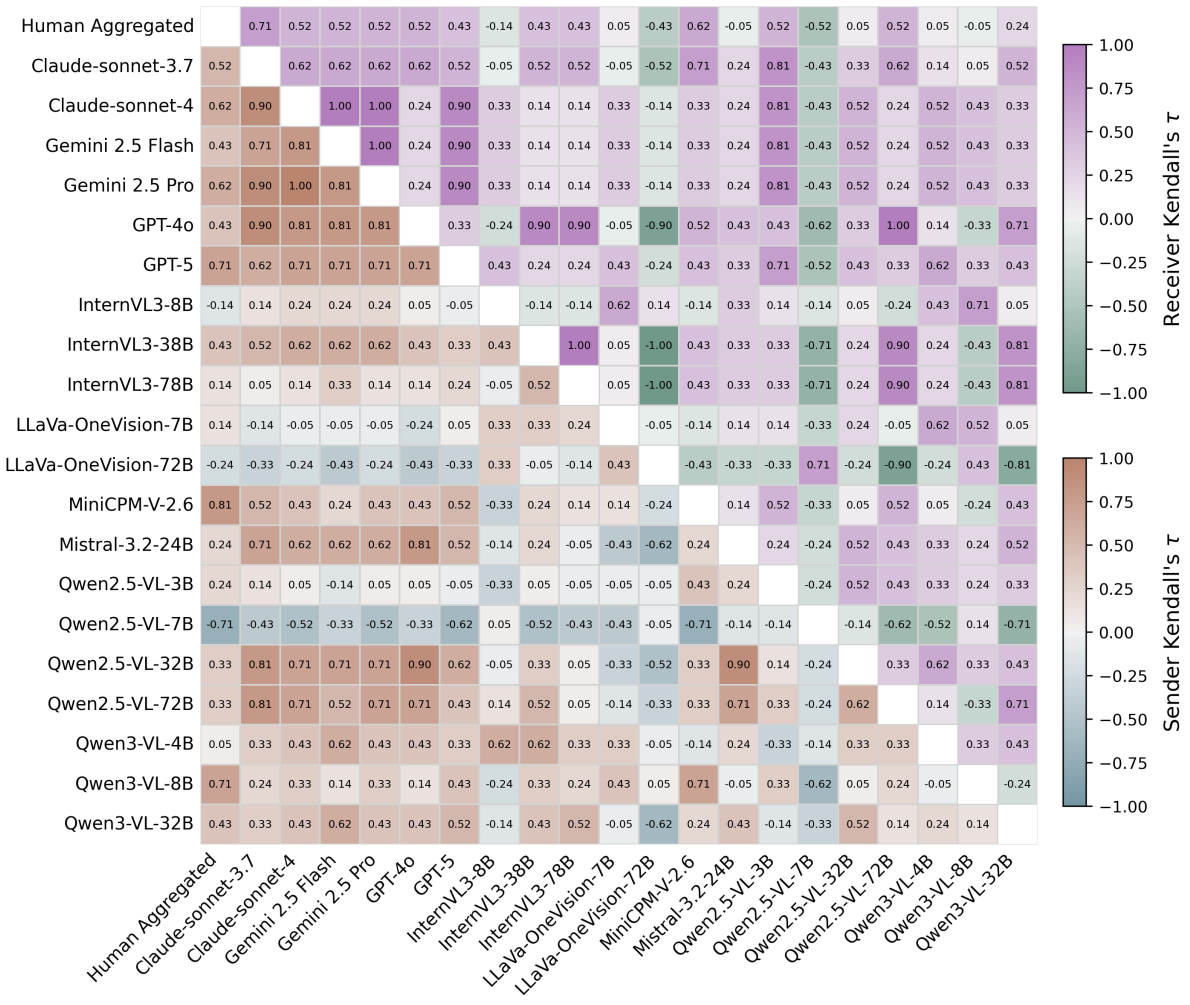


Figure 13: Similarity of preference orders across humans and all LLMs. The heatmap shows Kendall's τ correlations between human and LLM Bradley–Terry preference rankings over AO categories. The lower triangle reports correlations in the *Sender* role, and the upper triangle reports correlations in the *Receiver* role.

Table 4: Definitions and descriptions of the Activity Overview visual features extracted for regression analysis, categorized by the three conceptual dimensions: Activity Distribution, Geometric Aesthetics, and Semantic Cues.

Category	Feature Name	`Variable Name`	Description
Activity Distribution	Visual Area	`feat_area`	Represents how large the activity shape appears overall. Larger area indicates higher overall activity and more balanced participation across different types of actions.
	Visual Entropy	`feat_entropy`	Measures how evenly activity is distributed across categories. High entropy means activity is spread across many types, while low entropy indicates concentration on a few.
Geometric Aesthetics	Long/Width Ratio	`feat_geo_aspect_ratio`	Captures whether the activity shape is vertically or horizontally elongated. Vertical elongation reflects emphasis on reviews and pull requests, while horizontal elongation reflects emphasis on commits and issues.
	Vertical Symmetry	`feat_sym_vertical`	Measures the balance between pull requests and code reviews. Higher symmetry indicates similar levels of code contribution and review activity.
	Horizontal Symmetry	`feat_sym_horizontal`	Measures the balance between commits and issues. Higher symmetry reflects a more even mix of implementation work and coordination or issue tracking.
Semantic Cues	Maintainer Status	`feat_maintainer_status`	Represents the combined share of code review and issue activity ($P_{cr} + P_{is}$). Higher values indicate stronger involvement in activities of maintainers.
	Cosine Similarity	`feat_pair_similarity`	Measures how similar the two compared activity shapes are. Higher similarity indicates more similar activity profiles and typically makes the comparison harder. These feature serves as a control variable for decision difficulty.

Table 5: Comprehensive Descriptive Statistics. Detailed distributional analysis of the seven visual features. We report the Mean (Standard Deviation), Median [25th, 75th Percentile], and Range [Min, Max] to provide a complete view of feature dispersion for both Sender and Receiver datasets.

	Sender Dataset			**Receiver Dataset**		
Features	**Mean (SD)**	**Median [Q1, Q3]**	**Range**	**Mean (SD)**	**Median [Q1, Q3]**	**Range**
Activity Distribution						
Visual Area	0.04 (0.64)	-0.01 [-0.34, 0.40]	[-1.98, 2.76]	-0.01 (0.65)	-0.01 [-0.41, 0.36]	[-1.89, 2.94]
Visual Entropy	-0.01 (1.92)	0.05 [-1.22, 1.25]	[-5.32, 5.01]	0.04 (1.91)	0.12 [-1.16, 1.28]	[-5.56, 4.88]
Geometric Aesthetics						
L/W Ratio	0.05 (1.19)	0.00 [-0.68, 0.77]	[-3.45, 3.82]	0.02 (1.20)	0.00 [-0.69, 0.75]	[-3.67, 3.55]
Vertical Symmetry	-0.01 (0.45)	-0.01 [-0.27, 0.23]	[-1.00, 1.00]	0.01 (0.45)	0.00 [-0.23, 0.28]	[-1.00, 1.00]
Horizontal Symmetry	-0.01 (0.43)	-0.01 [-0.24, 0.22]	[-1.00, 1.00]	0.00 (0.43)	0.00 [-0.23, 0.23]	[-1.00, 1.00]
Semantic Cues						
Maintainer Status	0.00 (0.48)	0.00 [-0.29, 0.30]	[-1.00, 1.00]	-0.01 (0.48)	0.00 [-0.32, 0.26]	[-1.00, 1.00]
Cosine Similarity	0.70 (0.17)	0.72 [0.59, 0.83]	[0.21, 1.00]	0.70 (0.17)	0.72 [0.59, 0.83]	[0.19, 1.00]

Table 6: Visual Feature Regression (Sender). We report the standardized logistic regression coefficients (β) for the dataset. Shaded rows denote section groupings for Reference, Proprietary, and Open-Source models. Significance levels: $^{*}p < 0.05$,$^{**}p < 0.01$,$^{***}p < 0.001$.

	Activity Distribution		**Geometric Aesthetics**			**Semantic Cues**	
	Area	**Entropy**	**L/W Ratio**	**V.Sym**	**H.Sym**	**Maintainer**	**Similarity**
Reference							
HUMAN	0.80***	0.47	0.82***	-0.64***	-0.83**	-0.40***	0.09
Proprietary LLMs							
Claude-sonnet-3.7	-0.25	1.36***	2.03***	-0.10	1.06***	1.68***	-0.06
Claude-sonnet-4	-0.31	0.30	1.34***	-0.10	1.18***	2.33***	-0.60***
GPT-4o	2.19***	-0.65	3.31***	-0.64***	-0.59	1.53***	-0.03
GPT-5	2.18***	2.42***	5.20***	-1.10***	1.57***	0.95***	0.33**
Gemini 2.5 Flash	1.08***	3.60***	5.56***	0.29	3.28***	1.71***	-0.06
Gemini 2.5 Pro	0.33	4.87***	4.89***	0.42*	2.86***	2.43***	0.32*
Open-Source LLMs							
InternVL3-38B	0.57***	-0.52*	0.40***	-0.22	-0.12	0.18*	0.09
InternVL3-78B	0.11	0.27	-0.02	0.17	0.16	-0.03	0.05
InternVL3-8B	0.18	-0.35	-0.08	-0.09	-0.09	0.24**	-0.04
LLaVa-OneVision-72B	0.12	-0.06	-0.20*	-0.08	0.09	0.12	-0.02
LLaVa-OneVision-7B	0.10	-0.11	-0.07	-0.07	-0.14	-0.04	-0.01
MiniCPM-V-2.6	0.13	-0.27	0.07	-0.32*	-0.47*	-0.04	-0.03
Mistral-3.2-24B	0.50**	-1.11***	0.32**	-0.44***	-1.02***	0.64***	0.27***
Qwen2.5-VL-32B	0.50***	-0.84**	0.00	-0.27*	-0.71***	0.23**	-0.15**
Qwen2.5-VL-3B	0.25	-0.42	-0.11	-0.15	-0.93***	-0.40***	-0.01
Qwen2.5-VL-72B	0.76***	-0.88**	0.13	-0.38**	-1.02***	0.43***	-0.31***
Qwen2.5-VL-7B	0.27	-0.50	-0.02	-0.00	-0.10	0.09	0.02
Qwen3-VL-32B	0.08	0.45	1.84***	0.19	1.52***	0.03	0.17
Qwen3-VL-4B	0.73*	-0.34	-0.39	0.11	1.00	1.96***	1.24***
Qwen3-VL-8B	2.35***	-3.05**	-1.16**	-0.87*	-2.84***	0.41	-0.15

Table 7: Visual Feature Regression (Receiver). We report the standardized logistic regression coefficients (β). Shaded rows denote section groupings for Reference, Proprietary, and Open-Source LLMs. Significance: $^{*}p < 0.05$,$^{**}p < 0.01$,$^{***}p < 0.001$.

	Activity Intensity		**Geometric Aesthetics**			**Semantic Cues**	
	Area	**Entropy**	**L/W Ratio**	**V.Sym**	**H.Sym**	**Maintainer**	**Similarity**
Reference							
HUMAN	0.76***	0.48	0.29*	-0.79***	-0.86***	-0.39***	0.01
Proprietary LLMs							
Claude-sonnet-3.7	0.08	-0.00	0.82***	-0.73***	-0.72**	0.02	-0.23***
Claude-sonnet-4	-0.19	3.42***	1.78***	0.03	2.72***	1.90***	-0.45***
GPT-4o	4.14***	-1.73***	5.49***	-2.01***	-3.94***	-1.78***	-0.29**
GPT-5	1.38***	2.10***	3.97***	-1.13***	1.94***	1.80***	0.04
Gemini 2.5 Flash	0.17	2.35***	2.71***	0.04	1.92***	2.10***	0.01
Gemini 2.5 Pro	1.13**	4.92***	4.56***	0.17	2.30***	3.83***	0.92***
Open-Source LLMs							
InternVL3-38B	0.50***	-0.47	0.64***	-0.31*	-0.60**	-0.61***	-0.04
InternVL3-78B	0.64***	-0.58*	0.05	-0.25*	-0.57**	0.01	0.07
InternVL3-8B	1.48***	-3.59***	0.20	-1.25***	-1.23***	1.76***	-0.19***
LLaVa-OneVision-72B	-0.08	-0.12	-0.17	0.02	0.10	0.10	0.01
LLaVa-OneVision-7B	-0.06	0.05	0.03	0.08	0.04	-0.02	-0.03
MiniCPM-V-2.6	0.37*	-0.47	0.26*	-0.18	-0.43*	-0.21*	0.16**
Mistral-3.2-24B	-0.02	-1.96***	1.05***	-0.72***	-2.40***	-0.85***	0.59***
Qwen2.5-VL-32B	1.07***	-2.29***	0.18	-0.80***	-1.90***	-0.01	-0.35***
Qwen2.5-VL-3B	0.10	-0.13	0.04	-0.03	-0.16	-0.03	-0.03
Qwen2.5-VL-72B	1.24***	-0.91**	1.44***	-0.53***	-2.40***	-1.24***	-0.47***
Qwen2.5-VL-7B	0.08	-0.18	-0.14	-0.06	-0.26	-0.17*	0.13*
Qwen3-VL-32B	0.62***	-0.89**	1.50***	-0.82***	-0.99***	-0.93***	0.21**
Qwen3-VL-4B	0.67***	-1.04**	0.70***	-0.43**	-0.46	1.04***	0.17*
Qwen3-VL-8B	2.83***	-3.46***	0.03	-1.26**	-2.63***	0.17	-0.78***

Table 8: Impact of Input Modality on Cue Regression. Standardized coefficients (β) for a representative set of human participants and LLMs, categorized by proprietary and open-source models. Darker shaded rows denote section groupings, while light blue rows marked with (T) indicate text-only inputs. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Activity Distribution		**Geometric Aesthetics**			**Semantic Cues**	
Human&LLMs	**Area**	**Entropy**	**L/W Ratio**	**V.Sym**	**H.Sym**	**Maintainer**	**Similarity**
Sender							
HUMAN	0.80***	0.47	0.82***	-0.64***	-0.83**	-0.40***	0.09
Proprietary LLMs							
GPT-4o	2.19***	-0.65	3.31***	-0.64***	-0.59	1.53***	-0.03
GPT-4o (T)	0.04	2.42***	2.97***	-0.12	2.04***	1.13***	-0.30**
Gemini 2.5 Flash	1.08***	3.60***	5.56***	0.29	3.28***	1.71***	-0.06
Gemini 2.5 Flash (T)	-0.83**	5.30***	4.08***	0.81***	4.14***	2.02***	-0.01
Claude-3.7-Sonnet	-0.25	1.36***	2.03***	-0.10	1.06***	1.68***	-0.06
Claude-3.7-Sonnet (T)	0.13	4.05***	4.11***	0.42*	3.61***	3.81***	1.29***
Open-Source LLMs							
InternVL3-8B	0.18	-0.35	-0.08	-0.09	-0.09	0.24**	-0.04
InternVL3-8B (T)	-0.15	0.58*	0.30**	0.21	0.53*	0.31***	-0.08
LLaVa-OV-72B	0.12	-0.06	-0.20*	-0.08	0.09	0.12	-0.02
LLaVa-OV-72B (T)	0.09	0.60	1.55***	0.02	1.39***	0.90***	-0.23***
Mistral-3.2-24B	0.50**	-1.11***	0.32**	-0.44***	-1.02***	0.64***	0.27***
Mistral-3.2-24B (T)	-0.15	2.30***	1.51***	0.16	1.50***	1.18***	0.44***
Qwen2.5-VL-7B	0.27	-0.50	-0.02	-0.00	-0.10	0.09	0.02
Qwen2.5-VL-7B (T)	-0.20	0.39	0.12	0.15	0.26	0.04	0.15**
Qwen3-VL-4B	0.73*	-0.34	-0.39	0.11	1.00	1.96***	1.24***
Qwen3-VL-4B (T)	-0.09	0.44	0.60***	-0.09	2.02***	1.80***	0.24***
Qwen3-VL-8B	2.35***	-3.05**	-1.16**	-0.87*	-2.84***	0.41	-0.15
Qwen3-VL-8B (T)	0.50	0.62	-0.22	-0.47*	-0.41	-0.08	-0.10
Qwen3-VL-32B	0.08	0.45	1.84***	0.19	1.52***	0.03	0.17
Qwen3-VL-32B (T)	-1.31***	3.97***	4.61***	1.28***	1.99***	-0.08	0.65***
Receiver							
HUMAN	0.76***	0.48	0.29*	-0.79***	-0.86***	-0.39***	0.01
Proprietary LLMs							
GPT-4o	4.14***	-1.73***	5.49***	-2.01***	-3.94***	-1.78***	-0.29**
GPT-4o (T)	0.07	2.80***	3.23***	-0.45*	1.89***	1.19***	-0.44***
Gemini 2.5 Flash	0.17	2.35***	2.71***	0.04	1.92***	2.10***	0.01
Gemini 2.5 Flash (T)	-0.95**	6.12***	4.32***	0.60**	4.73***	3.11***	-0.16
Claude-3.7-Sonnet	0.08	-0.00	0.82***	-0.73***	-0.72**	0.02	-0.23***
Claude-3.7-Sonnet (T)	0.44	5.50***	4.71***	0.13	6.44***	5.39***	-1.35***
Open-Source LLMs							
InternVL3-8B	1.48***	-3.59***	0.20	-1.25***	-1.23***	1.76***	-0.19***
InternVL3-8B (T)	0.12	0.06	0.37***	-0.08	-0.06	0.22**	-0.05
LLaVa-OV-72B	-0.08	-0.12	-0.17	0.02	0.10	0.10	0.01
LLaVa-OV-72B (T)	-0.09	0.83**	0.42***	-0.09	0.70**	0.29***	-0.17***
Mistral-3.2-24B	-0.02	-1.96***	1.05***	-0.72***	-2.40***	-0.85***	0.59***
Mistral-3.2-24B (T)	-0.34	3.09***	1.44***	0.16	2.19***	1.80***	0.20*
Qwen2.5-VL-7B	0.08	-0.18	-0.14	-0.06	-0.26	-0.17*	0.13*
Qwen2.5-VL-7B (T)	-0.24	0.50*	0.26**	0.26*	0.16	-0.28***	0.08
Qwen3-VL-4B	0.67***	-1.04**	0.70***	-0.43**	-0.46	1.04***	0.17*
Qwen3-VL-4B (T)	-0.66***	1.57***	0.69***	0.24	2.80***	2.28***	-0.41***
Qwen3-VL-8B	2.83***	-3.46***	0.03	-1.26**	-2.63***	0.17	-0.78***
Qwen3-VL-8B (T)	0.29	0.89**	-0.15	-0.31*	0.86***	0.14	-0.07
Qwen3-VL-32B	0.62***	-0.89**	1.50***	-0.82***	-0.99***	-0.93***	0.21**
Qwen3-VL-32B (T)	-0.13	1.37***	3.06***	-0.08	-0.25	0.16	-0.88***

Table 9: Pure-Text LLM Regression. Standardized coefficients (β) for human and pure text large language models, categorized by proprietary and open-source models. Shaded rows denote section groupings. Significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Activity Distribution		**Geometric Aesthetics**			**Semantic Cues**	
Agent	**Area**	**Entropy**	**L/W Ratio**	**V.Sym**	**H.Sym**	**Maintainer**	**Similarity**
Sender							
HUMAN	0.80***	0.47	0.82***	-0.64***	-0.83**	-0.40***	0.09
Proprietary LLMs							
GPT-3.5-Turbo	0.49*	-0.55	1.78***	-0.27	0.49	1.99***	0.16
Qwen-Max	-0.36	2.53***	0.86***	0.28	0.47	0.03	-0.15*
Qwen-Plus	-1.00***	2.67***	3.30***	0.54**	4.62***	3.14***	-0.24*
Open-Source LLMs							
Qwen2.5-7B	-0.09	0.66*	1.08***	0.22	0.83***	0.41***	0.54***
Qwen2.5-32B	-0.72***	2.70***	0.57***	1.13***	3.00***	0.55***	0.15*
Qwen3-4B	-0.08	0.65*	-0.17	0.38**	1.38***	1.05***	0.47***
Qwen3-32B	0.04	1.70***	1.07***	-0.61***	-1.28***	-0.66***	0.10
DeepSeek-V3	-0.60*	3.89***	4.45***	0.55*	5.02***	3.95***	-0.83***
Llama-3.3-70B	0.02	2.98***	3.45***	-0.52*	3.28***	2.53***	-0.38**
Receiver							
HUMAN	0.76***	0.48	0.29*	-0.79***	-0.86***	-0.39***	0.01
Proprietary LLMs							
GPT-3.5-Turbo	0.56**	-0.14	2.50***	-0.29	0.83**	1.70***	0.18*
Qwen-Max	-0.50*	3.83***	0.71***	0.09	1.45***	1.50***	-0.35***
Qwen-Plus	-0.81***	2.67***	1.67***	0.02	3.13***	1.36***	-0.04
Open-Source LLMs							
Qwen2.5-7B	-0.17	0.49	0.49***	0.00	-0.14	-0.75***	0.42***
Qwen2.5-32B	-0.75***	2.49***	0.05	0.59***	2.67***	1.17***	-0.38***
Qwen3-4B	-0.62***	1.23***	0.23	0.45**	2.11***	1.13***	0.39***
Qwen3-32B	-0.96***	2.26***	-0.26*	-0.21	-1.16***	-0.93***	0.05
DeepSeek-V3	-0.43	3.33***	2.49***	0.12	3.77***	3.63***	-0.32**
Llama-3.3-70B	0.27	3.76***	2.67***	-0.57*	3.50***	3.13***	-1.43***

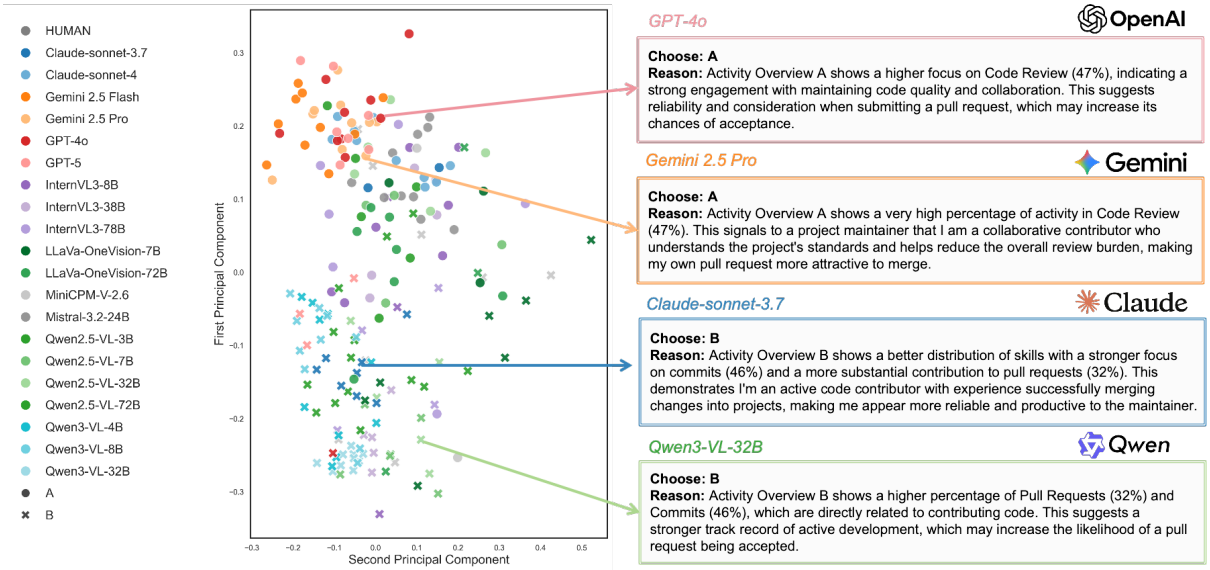


Figure 14: Semantic embedding projection of reasoning (temperature=1) for the *Review-Spike* vs. *Commit-Pennant* pair (Sender).

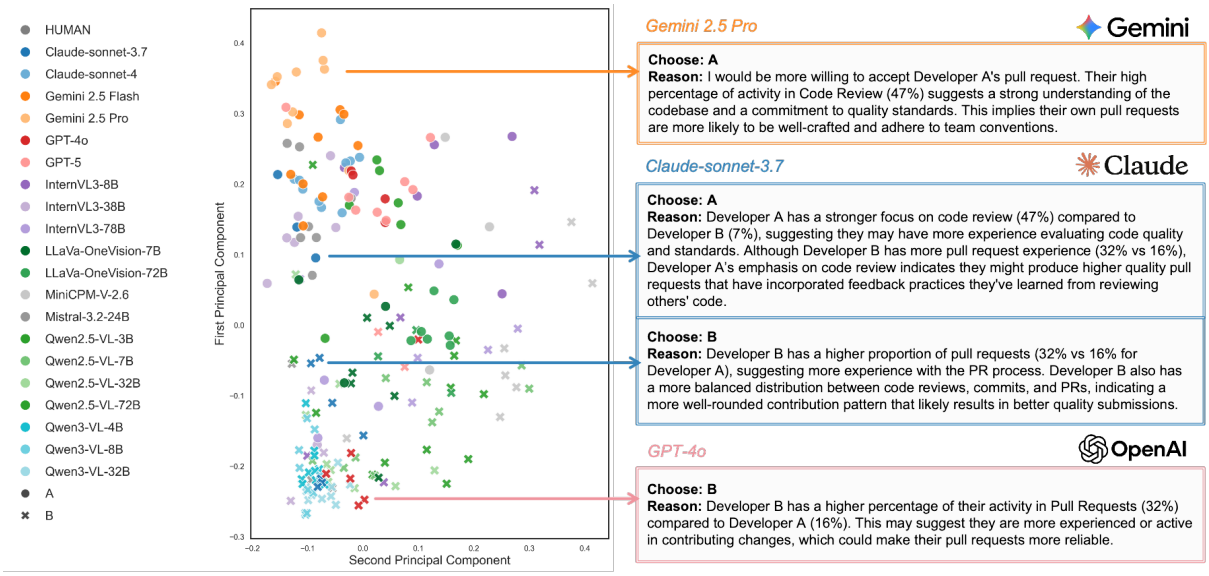


Figure 15: Semantic embedding projection of reasoning (temperature=1) for the *Review-Spike* vs. *Commit-Pennant* pair (Receiver).