A Matter of Interest: Understanding Interestingness of Math Problems in Humans and Language Models

Shubhra Mishra[1], **Yuka Machino**[2,3,4], **Gabriel Poesia**[2], **Albert Jiang**[4,5], **Joy Hsu**[2], **Adrian Weller**[4], **Challenger Mishra**[4], **David Broman**[1], **Joshua B. Tenenbaum**[3], **Mateja Jamnik**[4,†], **Cedegao E. Zhang**[3,†], **Katherine M. Collins**[3,4,†]

KTH Royal Institute of Technology[1], Stanford University[2],
Massachusetts Institute of Technology[3], University of Cambridge[4], Mistral AI[5]

`shubhram@kth.se, {cedzhang, katiemc}@mit.edu`

Abstract

The evolution of mathematics has been guided in part by *interestingness*. From researchers choosing which problems to tackle next, to students deciding which ones to engage with, people's choices are often guided by judgments about how interesting or challenging problems are likely to be. As AI systems, such as LLMs, increasingly participate in mathematics with people—whether for advanced research or education—it becomes important to understand how well their judgments align with human ones. Our work examines this alignment through two empirical studies of human and LLM assessment of mathematical interestingness and difficulty, spanning a range of mathematical experience. We study two groups: participants from a crowdsourcing platform and International Math Olympiad competitors. We show that while many LLMs appear to broadly agree with human notions of interestingness, they mostly do not capture the *distribution* observed in human judgments. Moreover, most LLMs only somewhat align with *why* humans find certain math problems interesting, showing weak correlation with human-selected interestingness rationales. Together, our findings highlight both the promises and limitations of current LLMs in capturing human interestingness judgments for mathematical AI thought partnerships.

1 Introduction

Mathematical reasoning involves not just solving problems but judging whether a problem is worth solving. In recent years, large language models (LLMs) and large reasoning models (LRMs) have substantially advanced in their ability to solve mathematics problems: they have gone from struggling to solve grade school mathematics problems to now achieving gold medal-level performance at the International Mathematical Olympiad (IMO) [2, 11, 14, 19]. While this progress is impressive, *problems are often given to models to solve*. Even computer-assisted discoveries like improved bounds on the CapSet problem (by FunSearch), better matrix multiplication algorithms (by AlphaTensor), or the improved bound on the kissing number (by AlphaEvolve) have ultimately depended on human-posed targets or carefully designed heuristics [22, 9, 18].

It is unclear whether LLMs and LRMs can adequately judge and select which problems are worth solving at all [4]. This is important, because problem selection is crucial for many potential applications of LLMs from education (e.g., proposing interesting problems and examples to students) to automated mathematical discovery (AMD) [21, 16] (involving posing interesting conjectures to explore). Answering the question *"is this problem worth solving?"* entails exploring two dimensions: whether the problem is interesting enough and whether it is challenging enough. While some prior

[†] Collaborative advising.

Preprint.

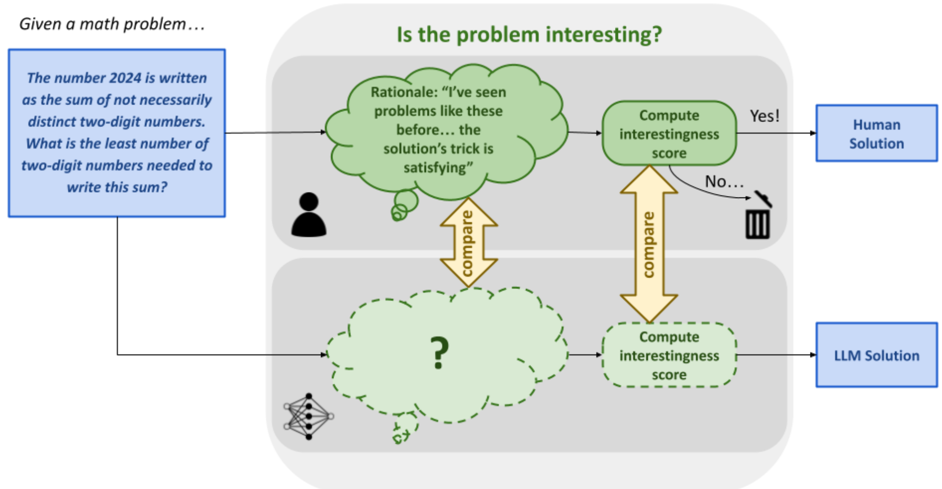


Figure 1: Human interactions with mathematics involve both judging *which* problems are worth pursuing, and then solving those problems. LLMs, in their applications today, are asked to solve problems directly. A gap remains: in settings where LLMs interact with mathematics in a human-like way, LLMs' judgments of how interesting problems are and why they are interesting remains unexplored. In our work, we study two human participant pools, exploring how language models' assessments of interestigness can be compared to people both in the final assessment of whether the problem is interesting and in the *process* by which a reasoner arrives at such a judgement (e.g., what factors are considered when judging interestingness).

work has explored the notion of interestingness in humans [12, 7, 1], in this paper, we take the first step in empirically comparing interestingness and difficulty judgments across humans and LLMs.

Prior work has attempted to model interestingness for AMD systems; however, such systems like Graffiti [5], AM [6, 13], HR [5, 6], and GT [8] rely on hundreds of hand-coded heuristics to guide conjecture generation and theorem proving, limiting scalability. Recent work has introduced complementary approaches that structure mathematical statements to yield conjectures provable through domain expertise, but lack heuristics to guide search based on human notions of interestingness [16]. The ability for LLMs and LRMs to operate more freely over language text (e.g., informal natural language problem descriptions) opens up new possibilities for flexible modeling of problem interestingness. This makes a new top-down approach possible—instead of manually specifying heuristics, we can directly measure how well LLMs capture human perceptions of interestingness, and use this to identify and address gaps.

Our work takes a step in this direction, conducting an initial study comparing LLMs and LRMs to human judgments of mathematics problem interestingness. Specifically, we contribute:

- Two new datasets of people's interestingness. 111 participants spanned a range of mathematical experience, from adults with high-school and/or college-level math experience to IMO participants, and judged problems of varied difficulty, resulting in 822 judgments.
- An evaluation of 12 language models (7 LLMs and 5 LRMs) across 5 families and compare how models' judgments of problem interestignness—and assessment of what factors ought to make a problem interesting (or uninteresting)—align with people.

2 Methods

We study human interestigness judgments in two participant pools and two different banks of problems: (1) crowdsourced participants reasoning about AMC problems, and (2) IMO participants (engaged in-person at the 2024 competition) reasoning about past IMO problems. Both studies received ethics approval by our institutional ethics review boards (the former study through MIT, and the latter through the University of Cambridge).

Crowdsourcing human interestingness judgments We recruit 63 participants from Prolific, a crowdsourcing platform common in cognitive science [20]. Each participant was assigned to one of 2 conditions, rating 10 problems each. Each participant saw the same control problems and either a problem's original version or its variant. Participants were required to think about each problem for at least 1 minute before rating its *interestingness* and *difficulty* on a scale of 0-100, and providing a 1- to 3-sentence rationale for each rating. For analysis, we filter out participants that rate the negative control math problem ("What is 28 + 13?") as having interestingness > 90, which filtered out the 6 outlier participants' responses.

We curate problems from AMC 8 and AMC 12, high school-level contests given by the Mathematical Association of America [15]. To systematically probe how dimensions of a problem impact its perception and to increase problem set diversity, we introduce the notion of a *variant*: a type of change that can be applied to a problem to transform it. Examples of variants include increasing/decreasing the sizes of the numbers in the problem, adding/removing steps, etc. For each contest problem, we hand-write a new problem based on a variant type. We also create two control problems: a negative and a positive control, which are later used to filter out any unfaithful participant ratings. This results in our final dataset for the Prolific study, which contains 18 problems (2 controls and 8 problems with one variant each). A list of all our problems and variants are provided in Appendix A3.

IMO data collection We conducted a survey of interestingness judgements made by participants at the 2024 IMO. Each of the 48 survey participants saw 4 problems. Each participant saw the same baseline: problem 1 from IMO 2024. The rest of the problems were selected randomly from IMO shortlists, with each participant survey being unique and including problems from the same area (Algebra, Combinatorics, Number Theory, and Geometry). The participants were asked if they wanted to see the solution to the problem before rating its *interestingness* and *difficulty*. They were also asked to select reasons for their interestingness and uninterestingness rating from a multiple choice list (see Appendix A4.1), plus an additional free-text box to state their own reasons. Most problems only received 1 to 2 responses; this is too few to compare human and model judgments at a per-problem level. As such, for the bulk of this paper, all IMO data comparisons are made over the interestingness *criteria* that participants selected.

LLM Experiments We evaluate 12 language models from 5 families. We sample 20 responses for each model at temperatures 0.3 and 1.0 for most models where temperature sampling is allowed. Due to computational costs, for reasoning models on our IMO dataset, we sampled 5 responses each. Additionally, note that GPT-5 and o3 do not allow temperature sampling and were thus only sampled at their default temperature (1.0).

3 Results

On average, LLM judgments of interestingness and difficulty are correlated with human judgments. For each LLM, we compute theR^2 between per-problem mean interestingness in humans and the model (see Figure 5).The human row/column reveal model–human agreement, while other row/column combinations report correlations amongst the models. On the Prolific dataset, model–human R^2 ranges from about 0.48 to 0.78, with the strongest agreements from the Mistral family. Split-half human R^2—our noise ceiling on explainable variance—is 0.71 (see Appendix A1.1). This indicates that current LLMs (especially those in the Mistral family) are able to approximate human perceptions of interestingness with surprising fidelity.

LLMs do not generally reflect the *distribution* of human judgments of interestingness While LLMs capture some aspects of human interestingness perceptions, their *distributions* of interestingness ratings usually diverge. We measure distributional similarity between human and LLM judgments using the Wasserstein Distance (WD) [23]. We also bootstrap the WD scores to build a 95% confidence interval (CI). The lowest WD is achieved by Mistral 7B (mean WD = 12.4, 95% CI = [0.3, 16.0]), whose 95% CI is the only one overlapping with the human split-half baseline (which has mean WD = 9.5, 95% CI = [7.8, 11.5]; Table 1). High distributional differences underscore the importance of careful model selection when aiming to capture the diversity of human judgments.

Elegance played a key role in human interestingness judgments of the IMO problems. For the IMO survey, each participant rated the interestingness of four problems and selected interesting-

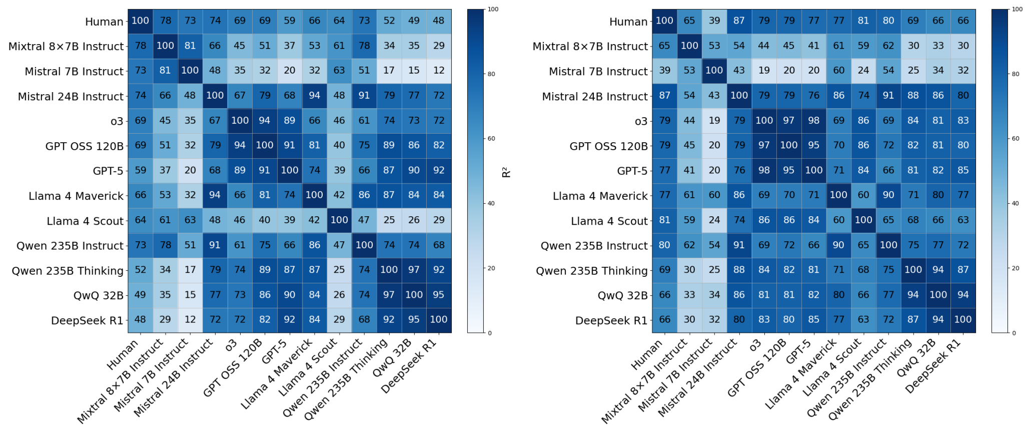


Figure 2: Agreement in humans' and LLMs' judgments about problem interestingness (left) and difficulty (right) on the Prolific dataset. Each cell shows the squared Pearson correlation (R^2 scaled by 100), between the row and column reasoners' per-problem mean ratings. The left matrix is sorted in descending order by agreement, while the right matrix follows the order of the left matrix to enable easier comparison. Darker cells indicate higher agreement (R^2). Models were sampled at temperature 1.0; additional analyses, e.g., for temperature 0.3, are included in Appendix A1.1.1.

ness/uninterestingness rationales for their rating. In Appendix Figure 16, we include a histogram of the frequency of different rationales for interestingness/uninterestingness. The three most frequently marked reasons for interestingness were "the problem statement is simple and elegant," "the solution does not require any sophisticated techniques/theorems," and "the solution is elegant." We also include a correlation matrix depicting when people chose multiple reasons for interestingness/uninterestingness for the same problem in Appendix Figure 17.

Most LLMs do not reflect *why* humans find problems interesting. To assess what factors of problems models and people find interesting, we next compared the distribution of reasons that humans and LLMs selected for problem interestingness. Since problems in our IMO survey mostly received 1-2 responses per problem, we don't focus on problem-level analysis for this. Instead, we focus on pre-survey questions which participants answered assigning importance to different interestingness reasons (e.g., "Please indicate how important this factor generally is for a problem to be interesting to you: The problem statement is simple and elegant."). The answers were collected on a four point scale of "not important" to "very important" with an option to mark if the criterion did not make sense to the participant. We replicate this experiment with all LLMs we examine, sampling 50 responses (to match the 48 participants from our IMO study). For each interestingness criterion, human participants' answers collectively spanned the allowed range of importance options. However, despite sampling each LLM 50 times at temperature 1.0, most LLMs only selected one to two importance scores for each reason. We include all distributions in Appendix A1.2. Only Mistral 7B Instruct (Figure 19) and Mistral 24B Instruct (Figure 20) reflect the human distributions of interestingness rationales well. Future work can better understand the drivers for such differences across model families.

LRMs tend to make flash judgments of uninterestingness than of interestingness for simpler problems. We next examine the resource usage LRMs engaged when reasoning about a problem, which we measure via the number of reasoning tokens used (i.e., the length of their reasoning chain, or reasoning "time"). We use this to explore the distinction in reasoning time to assess problems that LRMs judge as low- vs. high-interest (which is judged by whether a problem's interestingness score is below or above the median of interestingness scores from judgments from that LRM). See Figure 3 for results for our Prolific dataset, with results for the IMO dataset included in Appendix A1.1.3. For the Prolific dataset, all four LRMs make fast, "flash" judgments of uninterestingness while investing longer reasoning chains for problems they judged as interesting. This distinction disappears at the IMO level (see Appendix Figure 6), where judgments made under high reasoning time are no longer correlated with higher interestingness ratings. One possible explanation is that for hard Olympiad-style problems, resource usage spent on parsing and understanding the problem dominates the total resource usage spent thinking about the problem, leading to differences in interest making minimal impact on the length of the total reasoning chain.

Table 1: Wasserstein distances (WD) between human and LLM distributions (temp 1.0) of interestingness judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. Human-human split-half baseline indicates the amount of explainable variability in the human data. We report WD measures for other temperatures and for difficulty judgments in Appendix Section A1.1.2.

Model	WD	95% CI	Model	WD	95% CI
Human-Human	9.5	[7.8, 11.5]	Mixtral 8×7B Instruct	20.2	[17.5, 23.1]
Mistral 7B Instruct	12.4	[9.3, 16.0]	Llama 4 Maverick	20.7	[18.1, 23.7]
Mistral 24B Instruct	15.6	[13.5, 18.0]	Llama 4 Scout	21.1	[18.6, 23.6]
DeepSeek R1	16.4	[13.1, 19.7]	GPT-5	21.2	[18.3, 24.2]
QwQ 32B	18.1	[15.3, 21.0]	Qwen 235B Instruct	21.3	[18.5, 24.1]
GPT OSS 120B	18.3	[15.9, 21.0]	o3	21.9	[18.7, 25.5]
Qwen 235B Thinking	19.8	[16.4, 23.9]			

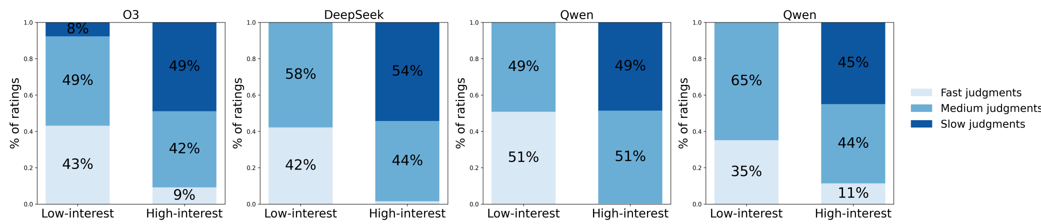


Figure 3: Judgment speed distributions across LRMs on low- vs. high-interest Prolific problems. A low-interest problem for a model is one that is given a below median interestingness score by the model. High-interest problems are those which are given higher than the median interestingness score. For each subset of problems, we label whether a judgment was slow, medium, or fast, based on the distribution of reasoning token counts for that model. Slow judgments occupy the bottom quartile and fast judgments occupy the top one, with medium-speed judgments covering the middle. We see that LRMs tend to engage in longer reasoning chains for problems that they ultimately label as being higher interest.

4 Discussion

In this work, we assess how judgments of interestingness compare between humans and LLMs, over two novel datasets of people's evaluation about math problems. We find that despite LLMs' interestingness judgments generally correlating with those of humans, their distributions do not completely match those of human judgments. However, our study is just a first step. Both the Prolific and IMO dataset use competition math problems, which are a narrow subset of the problems educators and mathematicians encounter daily. Additionally, the two surveys' populations focus on participants from a crowdsourcing website (who all had a baseline interest in math) and IMO participants, which overlooks beginners and experts, who might have differing perceptions of interestingness in mathematics. Lastly, the IMO study contained very few ratings for each problem, making it challenging to make quantitative or statistically significant conclusions about individual problems. Instead, for this survey, we focus on broader and more qualitative analyses.

More broadly, our work raises important questions: do we want models to align to the variability of human responses, which humans those responses should align to, and at what level of mathematical experience? If models are used in human-facing applications, e.g., as mathematical AI thought partners [3, 10], designing curricula for students or advising research mathematicians on problem selection, then we may want build interestingness measures that meaningfully correlate to human curriculum and the level of the learner [17]. If instead models are used to autonomously discover new mathematics—and decide what mathematics problem to pursue at all—we may set a higher standard for the interestingness judgments [12]. Overall, we hope our work motivates and informs future work in mathematically-capable AI systems that engage with subjective notions of mathematical problem interestingness. At the same time, we believe these efforts can also inform a better understanding of what drives humans to find a problem interesting in the first place.

Acknowledgments

We thank Simon Frieder, Michael Douglas, Gizem Çaylak, Xavery Davey, Arvid Eriksson, Oscar Eriksson, Lars Hummelgren, and John Wikman for helpful discussions that informed this work. For SM and DB, this work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. KMC acknowledges support from the Cambridge Trust and King's College Cambridge. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. JBT acknowledges support from work AFOSR (FA9550-22-1-0387), the ONR Science of AI program (N00014-23-1-2355), a Schmidt AI2050 Fellowship, and the Siegel Family Quest for Intelligence at MIT.

References

[1] Daniel E. Berlyne. Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology / Revue canadienne de psychologie*, 17(3):274–290, 1963.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[3] Katherine M Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, et al. Building machines that learn and think with people. *Nature human behaviour*, 8(10):1851–1863, 2024.

[4] Katherine M Collins, Cedegao E Zhang, Graham Todd, Lance Ying, Mauricio Barba da Costa, Ryan Liu, Prafull Sharma, Adrian Weller, Ionatan Kuperwajs, Lionel Wong, et al. Evaluating language models' evaluations of games. *arXiv preprint arXiv:2510.10930*, 2025.

[5] Simon Colton, Alan Bundy, and Toby Walsh. Automatic concept formation in pure mathematics. In *International Joint Conference on Artificial Intelligence*, 1999.

[6] Simon Colton, Alan Bundy, and Toby Walsh. On the notion of interestingness in automated mathematical discovery. *International Journal of Human-Computer Studies*, 53(3):351–375, 2000.

[7] Harris Day. Evaluations of subjective complexity, pleasingness and interestingness for a series of random polygons varying in complexity. *Perception & Psychophysics*, 2:281–286, 1967.

[8] Susan L. Epstein. On the discovery of mathematical theorems. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'87, page 194–197, San Francisco, CA, USA, 1987. Morgan Kaufmann Publishers Inc.

[9] Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, Demis Hassabis, and Pushmeet Kohli. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

[10] Simon Frieder, Jonas Bayer, Katherine M Collins, Julius Berner, Jacob Loader, András Juhász, Fabian Ruehle, Sean Welleck, Gabriel Poesia, Ryan-Rhys Griffiths, et al. Data for mathematical copilots: Better ways of presenting proofs for machine learning. *arXiv preprint arXiv:2412.15184*, 2024.

[11] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.

[12] Matthew Inglis and Andrew Aberdein. Beauty is not simplicity: An analysis of mathematicians' proof appraisals†. *Philosophia Mathematica*, 23(1):87–109, 07 2014.

[13] Douglas B. Lenat. Am, an artificial intelligence approach to discovery in mathematics as heuristic search. 1976.

[14] Thang Luong and Edward Lockhart. Advanced version of gemini with deep think officially achieves gold-medal standard at the international mathematical olympiad, July 2025. DeepMind blog post.

[15] Mathematical Association of America. American mathematics competitions (AMC). `https://maa.org/math-competitions`. Accessed: 2025-09-06.

[16] Challenger Mishra, Subhayan Roy Moulik, and Rahul Sarkar. Mathematical conjecture generation using machine intelligence, 2023.

[17] Shubhra Mishra, Gabriel Poesia, and Noah D. Goodman. From next-token to mathematics: The learning dynamics of mathematical reasoning in language models, 2025.

[18] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery, 2025.

[19] OpenAI. Introducing OpenAI o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`, April 2025. Accessed: 2025-08-29.

[20] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

[21] Gabriel Poesia, David Broman, Nick Haber, and Noah D. Goodman. Learning formal mathematics from intrinsic motivation, 2024.

[22] Petar Veličković, Alex Vitvitskyi, Larisa Markeeva, Borja Ibarz, Lars Buesing, Matej Balog, and Alexander Novikov. Amplifying human performance in combinatorial competitive programming, 2024.

[23] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Appendix

A1 Additional Results

A1.1 Prolific Survey Results

A1.1.1 Model Correlation

We include the agreement heatmaps for temperature 1.0 in Figure 5. We also include information about the correlation between *bootstrapped* LLM and human judgments of interestingness and difficulty in Tables 2 and 3, respectively. We also include all correlation plots for interestingness judgments in Figures 7 through 15. We sample at two temperatures and also plot the 95% confidence interval.

A1.1.2 Wasserstein Distance Tables

In Tables 4, 5, and 6, we report the Wasserstein distances between the LLM and human distributions for interestingness and difficulty.

A1.1.3 LRM Judgment Lengths

We examine the resource usage LRMs engaged when reasoning about a problem, which we measure via the amount of reasoning tokens used (i.e., the length of their reasoning chain). We use this to explore the distinction in resources used to assess problems that LRMs judge as low- vs. high-interest (which is judged by whether a problem's interestingness score is below or above the median of interestingness scores from judgments from that LRM). For the Prolific dataset, we observe that all 4 LRMs make fast, "flash" judgments of uninterestingness while investing longer reasoning chains for problems they judged to be interesting (see Figure 6). This distinction disappears at the IMO level, where judgments made under high resource usage are no longer correlated with higher interestingness ratings. One possible explanation is that for hard Olympiad-style problems, resource usage spent on parsing and understanding the problem dominates the total resource usage spent thinking about the problem, leading to differences in interest making little to no impact on the length of the total reasoning chain.

A1.1.4 Split-half R^2 between independent human raters.

To compare the LLM-human R^2 to the correlation among independent groups of humans, we calculate the split-half R^2, for which we repeatedly and randomly split human interestingness scores for each

Family	Model (tag)	Mean pairwise r (within family) temp. 0.3, 1.0	$\mathbf{R^2}$ vs. humans temp 0.3	$\mathbf{R^2}$ vs. humans temp 1.0
OpenAI	o3 (thinking)	-, 0.96	-	0.72
	GPT-5		-	0.66
	gpt-oss		0.75	0.75
Llama 4	Llama-4 Maverick	0.61, 0.65	0.73	0.75
	Llama-4 Scout		0.70	0.70
Qwen	QwQ-32B (thinking)	0.92, 0.93	0.60	0.60
	Qwen3-235B-A22B-Instruct		0.82	0.82
	Qwen3-235B-A22B-Thinking		0.60	0.62
Mistral	Mistral-7B-Instruct-v0.1	0.87, 0.82	0.86	0.81
	Mistral-Small-24B-Instruct		0.78	0.81
	Mixtral-8x7B-Instruct-v0.1		**0.88**	0.85
DeepSeek	DeepSeek-R1 (thinking)	-	0.63	0.58

Table 2: Correlation between LLM vs. human judgments of interestingness and within each model family. We include the per-model R^2 against human ratings on questions from our Prolific dataset. The "Mean pairwise r" column reports the mean pairwise Pearson correlation between models within their own family (computed over per-problem predictions). The "thinking" tag indicates reasoning models. The correlations for difficulty judgments are provided in Table 3.

Family	Model (tag)	Mean pairwise r (within family) temp. 0.3, 1.0	$\mathbf{R^2}$ vs. humans temp 0.3	$\mathbf{R^2}$ vs. humans temp 1.0
OpenAI	o3 (thinking)	-, 0.98	-	0.81
	GPT-5		-	0.79
	gpt-oss		0.80	0.81
Llama 4	Llama-4 Maverick	0.79, 0.77	**0.85**	0.83
	Llama-4 Scout		0.84	0.84
Qwen	QwQ-32B (thinking)	0.93, 0.92	0.73	0.73
	Qwen3-235B-A22B-Instruct		0.83	0.85
	Qwen3-235B-A22B-Thinking		0.71	0.74
Mistral	Mistral-7B-Instruct-v0.1	0.63, 0.71	0.52	0.53
	Mistral-Small-24B-Instruct		0.84	**0.91**
	Mixtral-8x7B-Instruct-v0.1		0.72	0.74
DeepSeek	DeepSeek-R1 (thinking)	-	0.80	0.72

Table 3: Correlation between LLM vs. human judgments of difficulty and within each model family. We include the per-model R^2 against human ratings on questions from our Prolific dataset. The "Mean pairwise r" column reports the mean pairwise Pearson correlation between models within their own family (computed over per-problem predictions). The "thinking" tag indicates reasoning models.

Table 4: Wasserstein distances (WD) between human and LLM distributions (temp 0.3) of interestingness judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. Human-human split-half baseline indicates the amount of explainable variability in the human data.

Model	WD	95% CI	Model	WD	95% CI
Human-Human	9.5	[7.8, 11.5]	Mixtral 8×7B Instruct	21.3	[18.9, 23.9]
DeepSeek R1	17.5	[14.7, 20.5]	Llama 4 Scout	21.4	[19.0, 23.9]
Mistral 24B Instruct	18.5	[16.2, 20.9]	GPT-5	21.5	[18.7, 24.6]
Mistral 7B Instruct	19.0	[16.5, 21.5]	Qwen 235B Instruct	22.0	[19.1, 24.9]
QwQ 32B	19.3	[16.6, 22.3]	Llama 4 Maverick	22.0	[19.0, 25.3]
GPT OSS 120B	19.5	[17.0, 22.2]	Qwen 235B Thinking	20.4	[16.9, 24.1]

Table 5: Wasserstein distances (WD) between human and LLM distributions (temp 0.3) of difficulty judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. Human-human split-half baseline indicates the amount of explainable variability in the human data.

Model	WD	95% CI	Model	WD	95% CI
Human–Human	9.2	[7.5, 11.0]	Mixtral 8×7B Instruct	20.1	[17.5, 23.1]
Mistral 24B Instruct	17.3	[15.2, 19.8]	Mistral 7B Instruct	20.2	[15.9, 24.7]
QwQ 32B	17.6	[15.3, 20.3]	Qwen 235B Instruct	20.5	[17.7, 23.6]
Qwen 235B Thinking	18.3	[15.2, 22.1]	GPT OSS 120B	21.3	[17.7, 25.6]
Llama 4 Maverick	19.0	[16.3, 21.9]	GPT-5	29.1	[24.7, 33.3]
Llama 4 Scout	19.2	[17.1, 21.4]	DeepSeek R1	19.2	[16.8, 21.8]

Table 6: Wasserstein distances (WD) between human and LLM distributions (temp 1.0) of difficulty judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. Human-human split-half baseline indicates the amount of explainable variability in the human data.

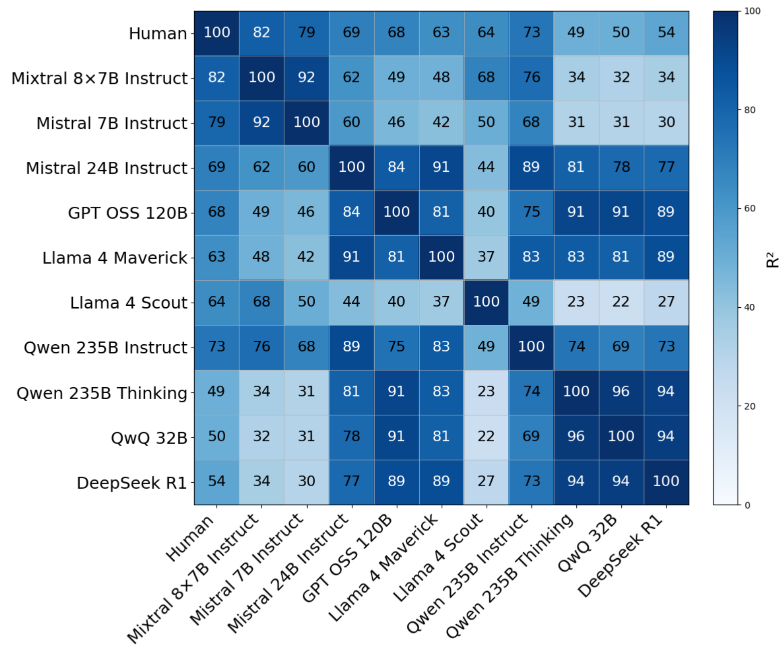
Model	WD	95% CI	Model	WD	95% CI
Human–Human	9.2	[7.5, 11.0]	Mixtral 8×7B Instruct	18.2	[15.5, 21.3]
Mistral 24B Instruct	13.3	[11.6, 15.0]	Llama 4 Scout	18.6	[16.6, 20.6]
Mistral 7B Instruct	16.0	[11.8, 20.5]	Llama 4 Maverick	19.1	[16.3, 21.7]
QwQ 32B	16.4	[13.8, 19.3]	Qwen 235B Instruct	20.2	[17.6, 22.9]
Qwen 235B Thinking	17.4	[14.5, 20.9]	GPT OSS 120B	20.7	[17.2, 25.0]
DeepSeek R1	18.0	[15.2, 21.3]	o3	23.7	[20.0, 27.8]

problem into two groups, and calculate the correlation between those two groups. The mean split-half R^2 among humans was 0.71, with a 95% confidence interval of $[0.53, 0.87]$.

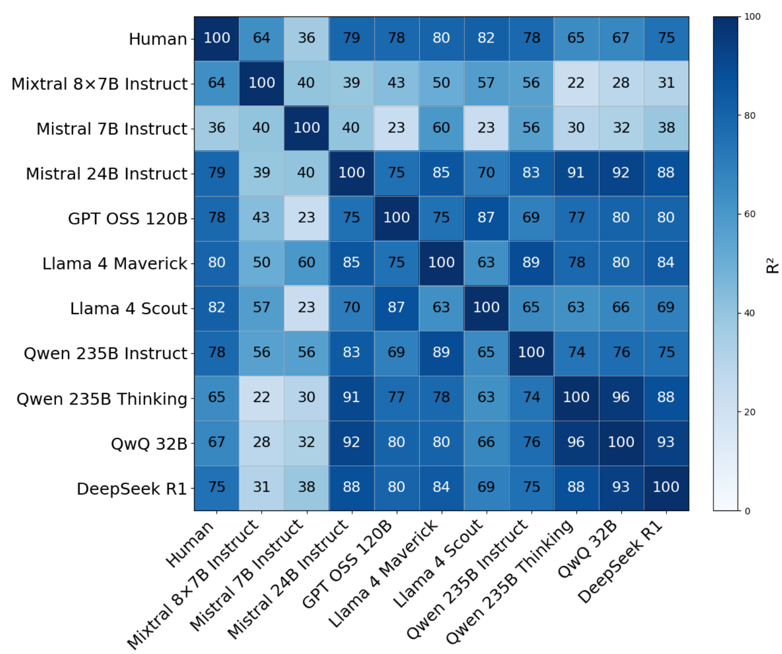
A1.2 IMO Survey Results

Elegance played a key role in human interestingness judgments of the IMO problems. For the IMO survey, each participant rated the interestingness of four problems and selected interestingness/uninterestingness rationales for their rating. In Appendix Figure 16, we include a histogram of the frequency of different rationales for interestingness/uninterestingness. The three most frequently marked reasons for interestingness were “the problem statement is simple and elegant," “the solution does not require any sophisticated techniques/theorems," and “the solution is elegant." We also include a correlation matrix depicting when people chose multiple reasons for interestingness/uninterestingness for the same problem in Appendix Figure 17.

Reasons for interestingness across LLMs. In Figure 18, we show histograms of human participants importance ratings for various interestingness criteria. Figures 19 through 30 includes this for the LLMs we examine. We sort the grid of LLM ratings by the distributions that are closest to human distributions first.

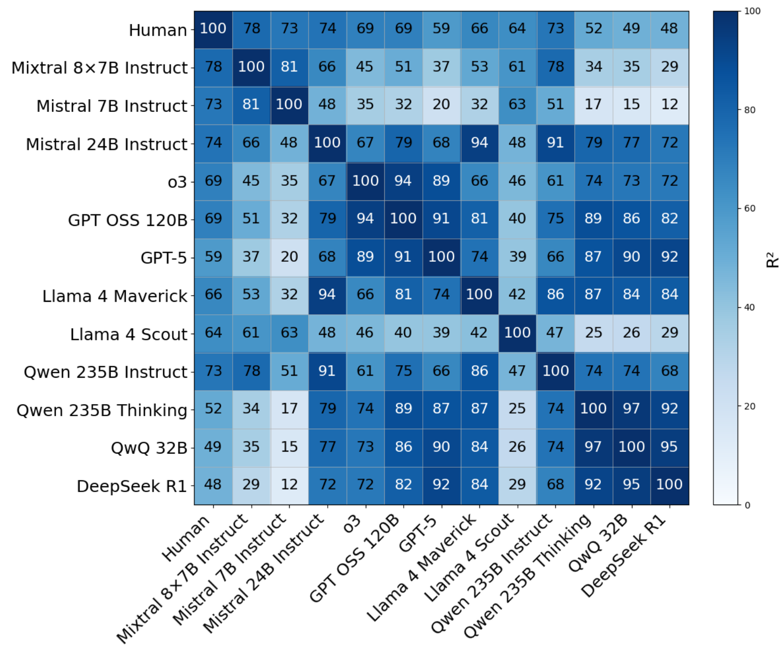


(a) Interestingness (temp = 0.3)

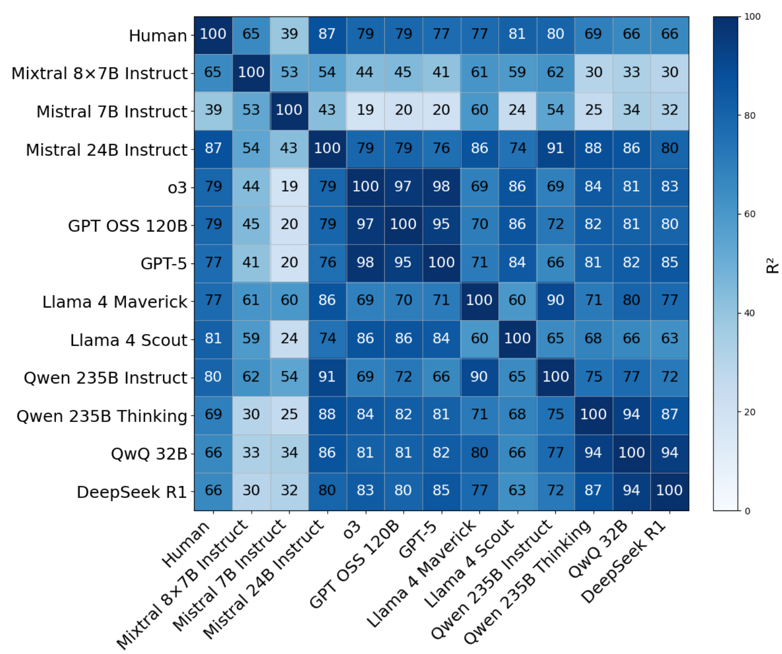


(b) Difficulty (temp = 0.3)

Figure 4: Agent–agent agreement on the **Prolific** dataset at temperature 0.30. Each cell shows the squared Pearson correlation (R^2) between the row and column agents' per-problem mean ratings. Darker cells indicate higher agreement; the diagonal is 1.00 by definition. The *Human* row/column gives model–human agreement. Top: interestingness ratings; bottom: difficulty ratings.



(a) Interestingness (temp = 1.0)



(b) Difficulty (temp = 1.0)

Figure 5: Agent–agent agreement on the **Prolific** dataset at temperature 1.00. Each cell shows the squared Pearson correlation (R^2) between the row and column agents' per-problem mean ratings. Darker cells indicate higher agreement; the diagonal is 1.00 by definition. The *Human* row/column gives model–human agreement. Top: interestingness ratings; bottom: difficulty ratings.

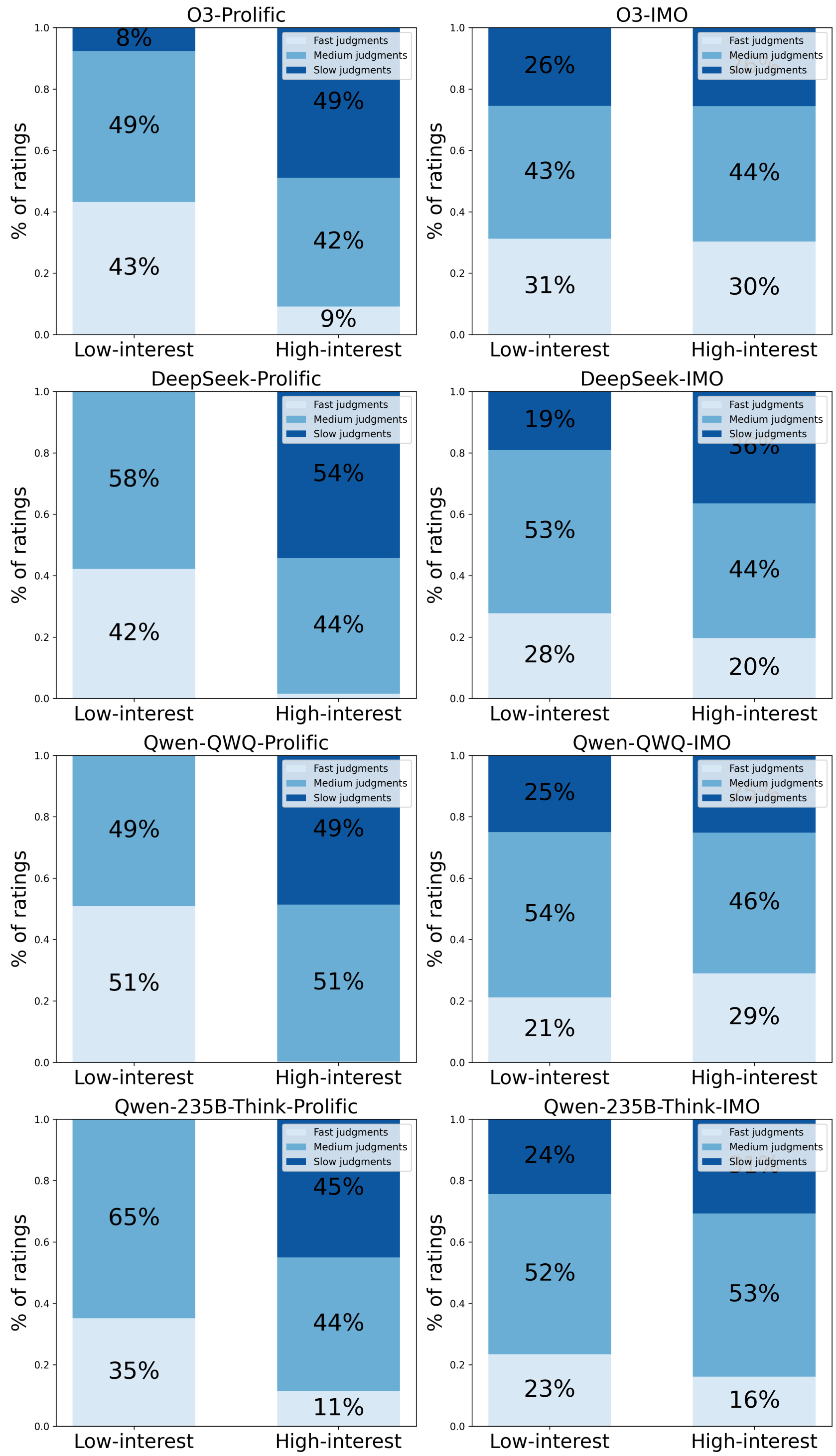


Figure 6: Judgment speed (as measured by length of reasoning chains) in LRMs across problems they judge as low- and high-interest.

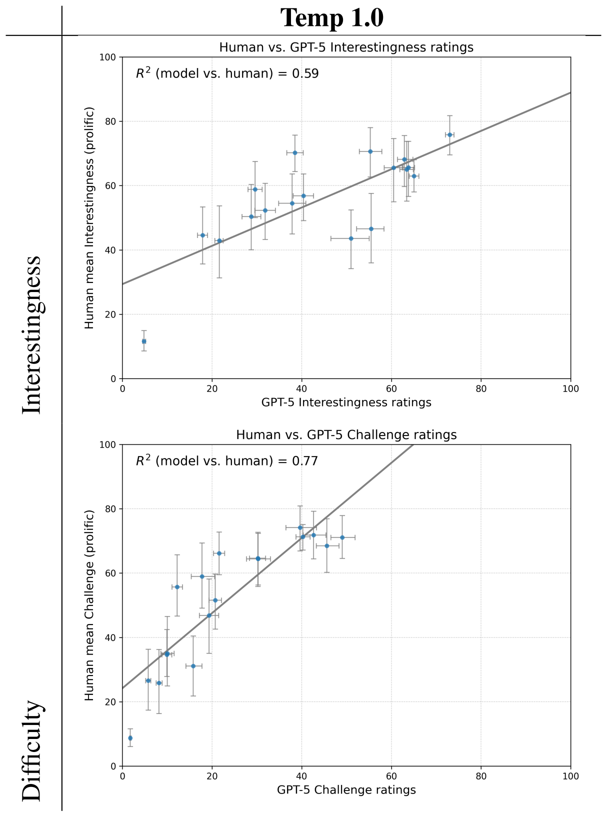


Figure 7: GPT-5: Human vs LLM ratings

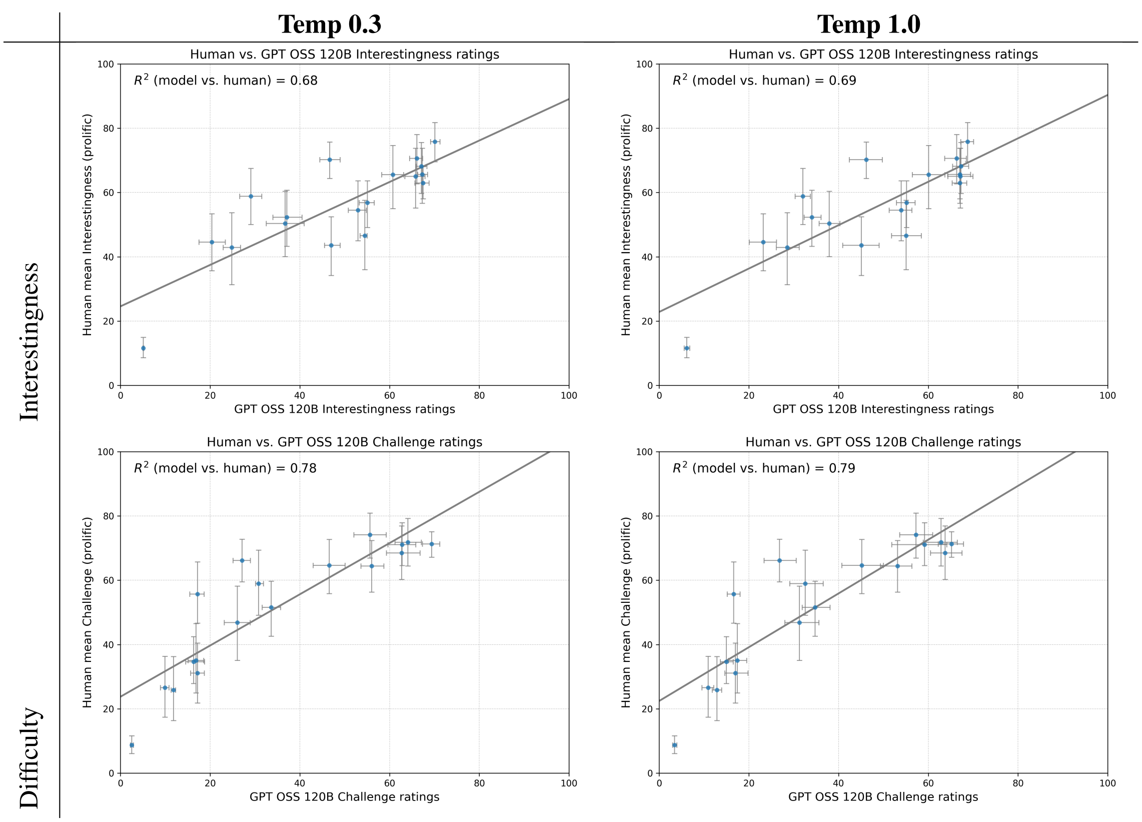


Figure 8: GPT-OSS-120B: Human vs LLM ratings

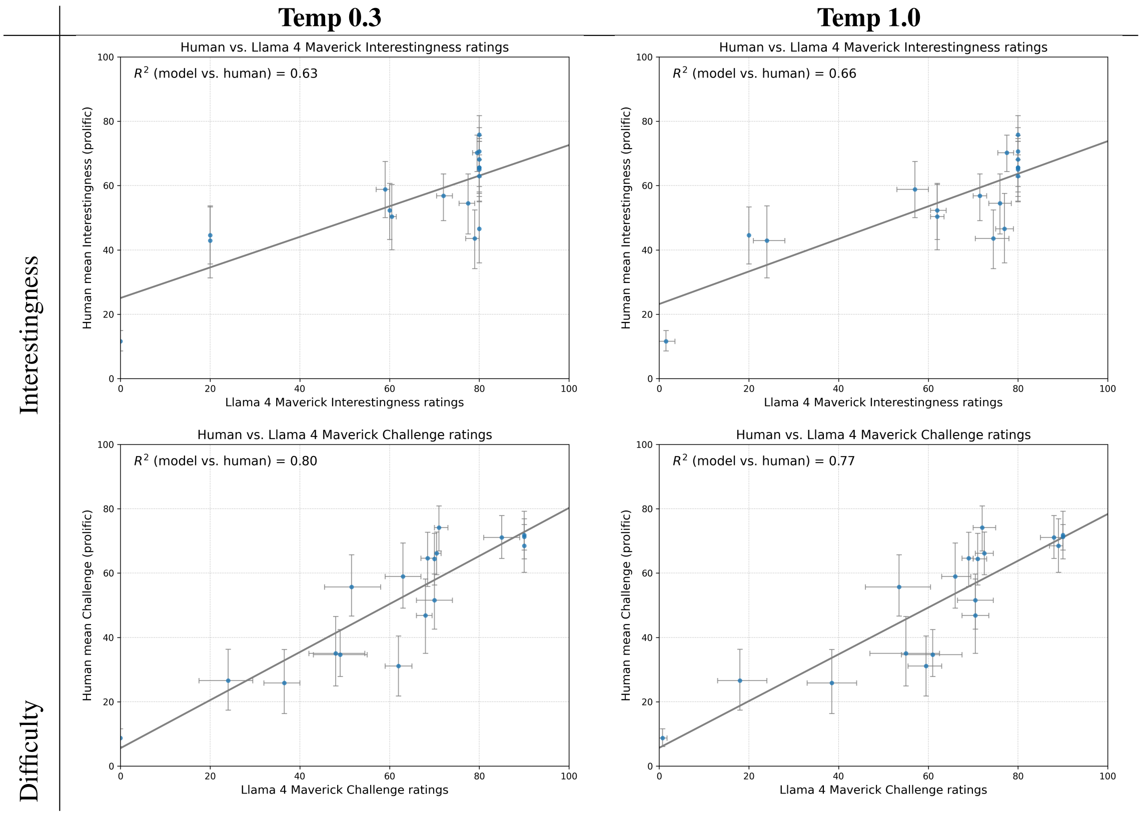


Figure 9: LLaMA-Maverick: Human vs LLM ratings

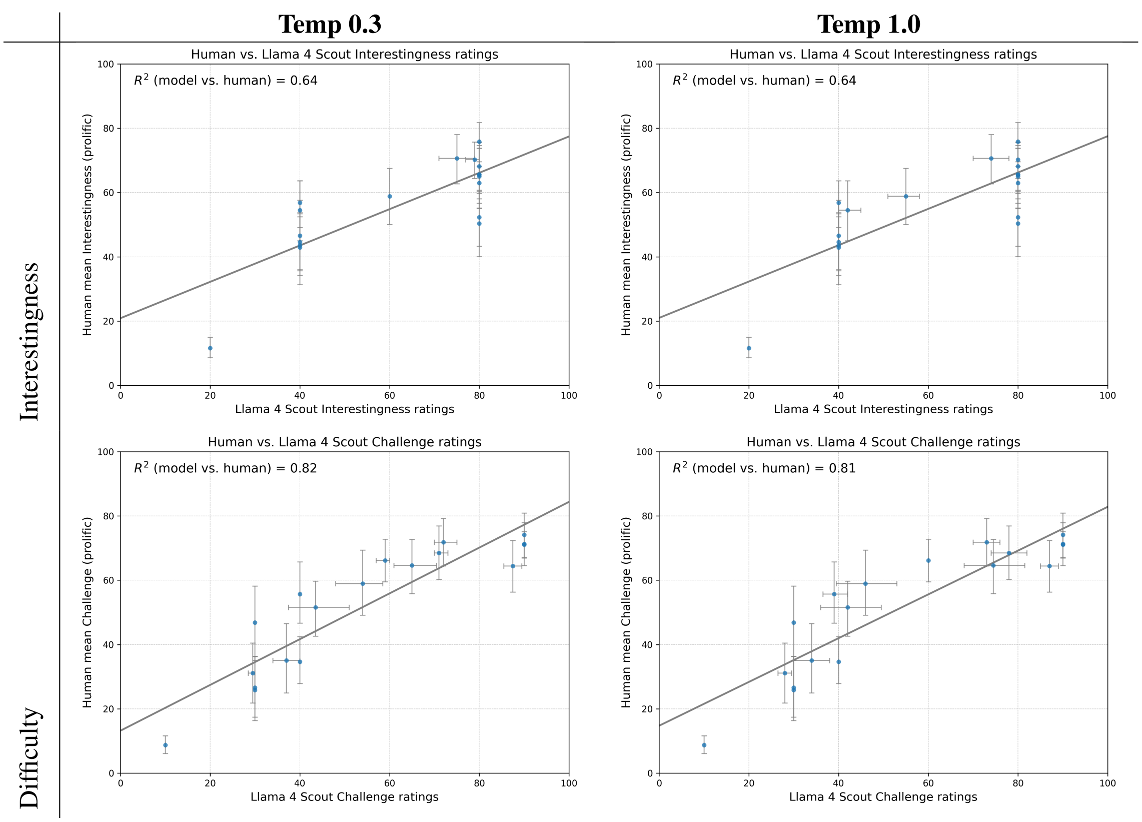


Figure 10: LLaMA-Scout: Human vs LLM ratings

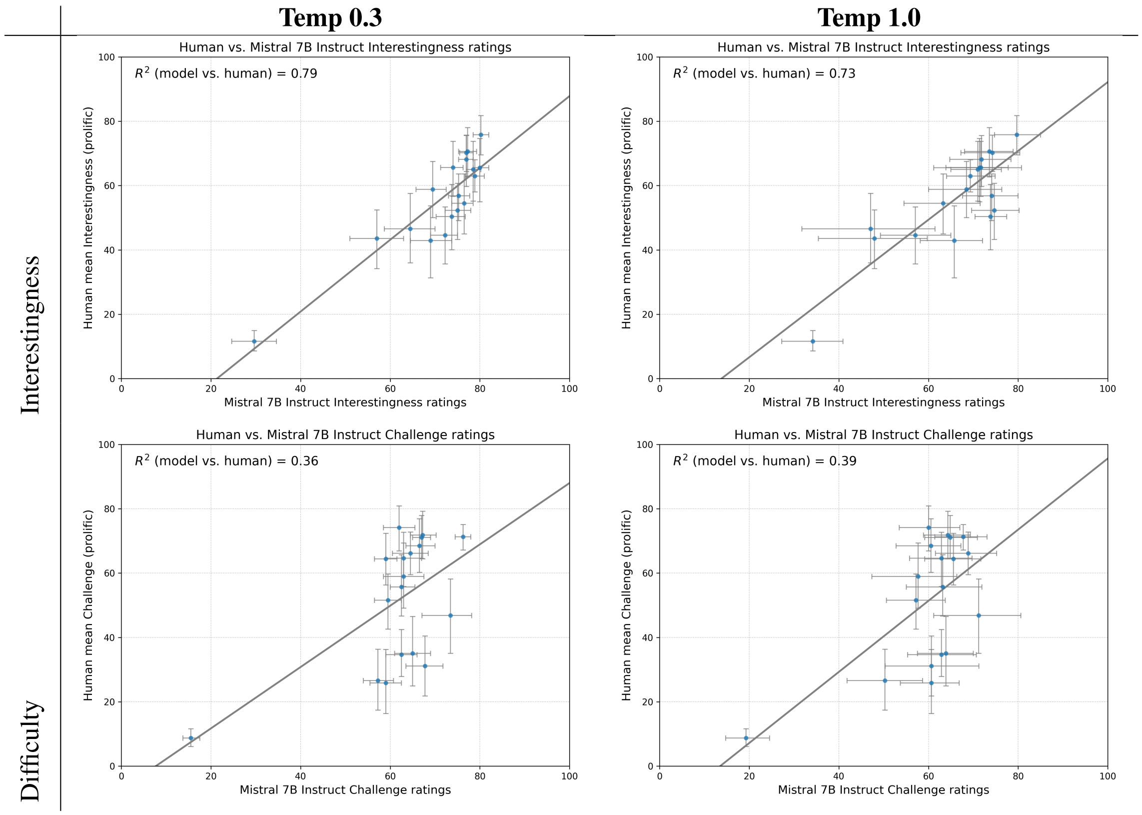


Figure 11: Mistral-7B-Instruct: Human vs LLM ratings

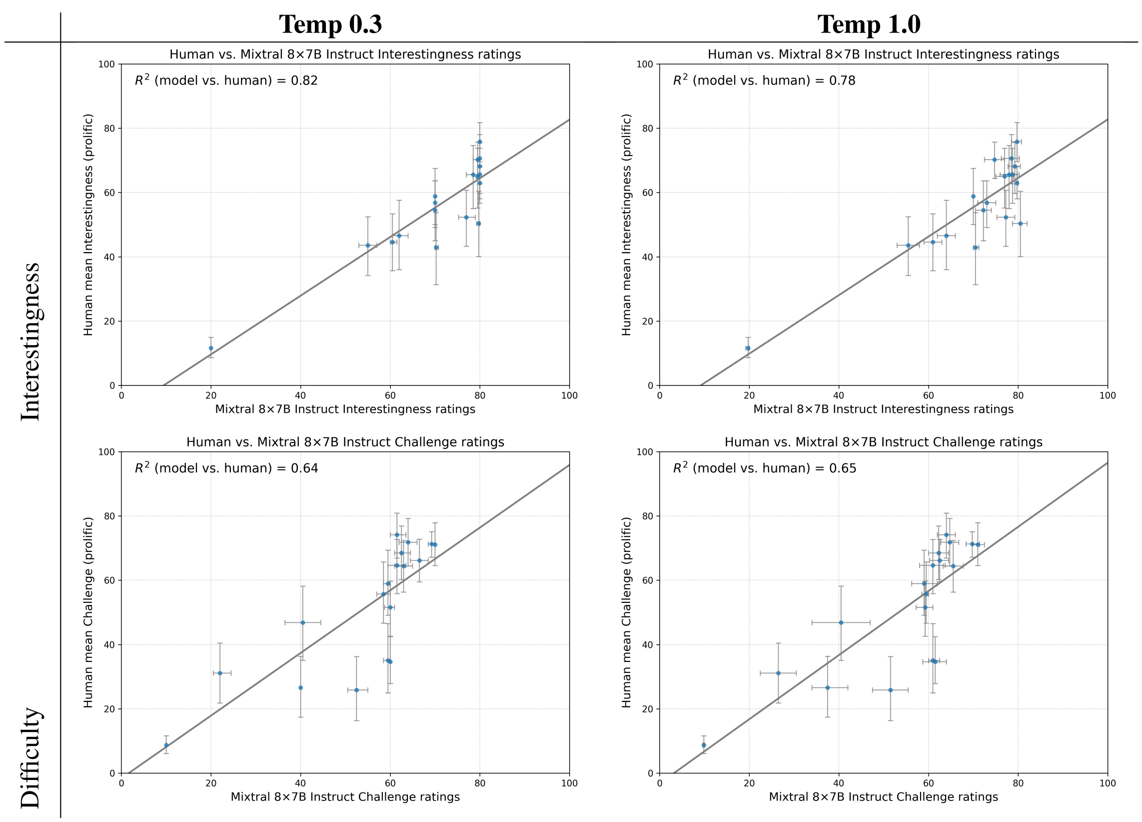


Figure 12: Mixtral-8x7B-Instruct: Human vs LLM ratings

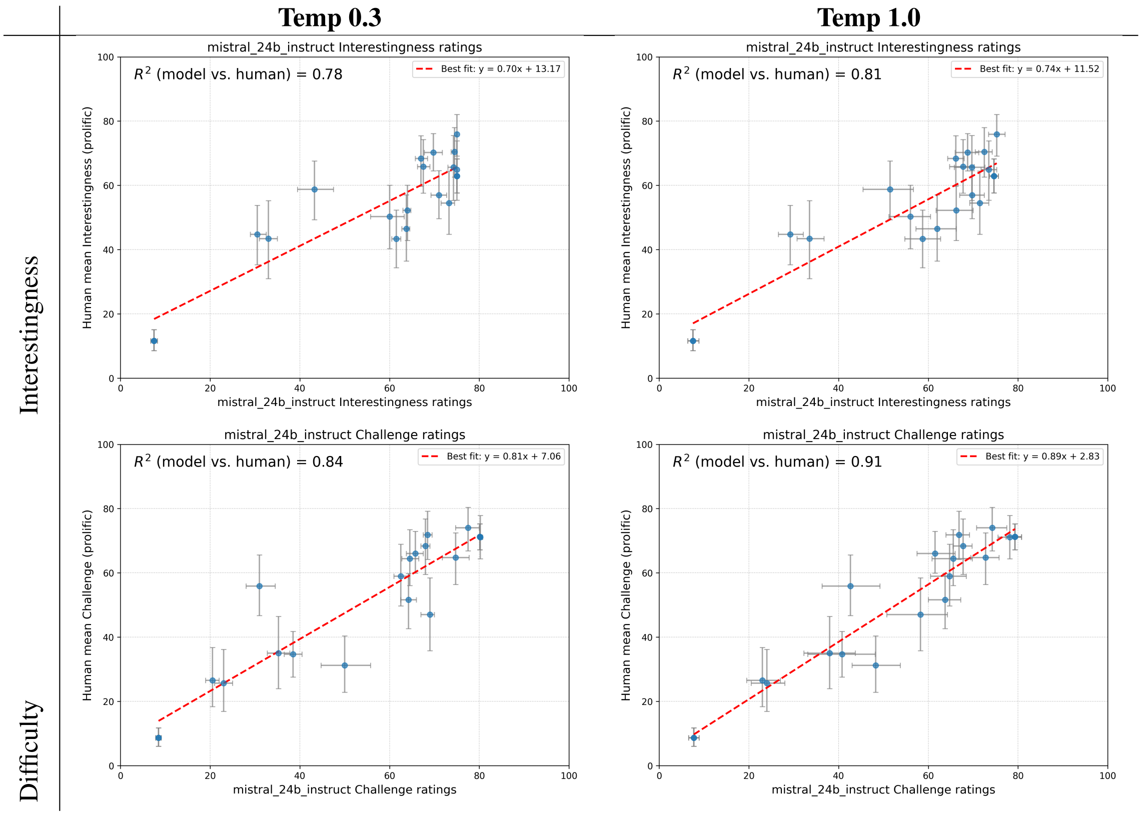


Figure 13: Mistral-24B-Instruct: Human vs LLM ratings

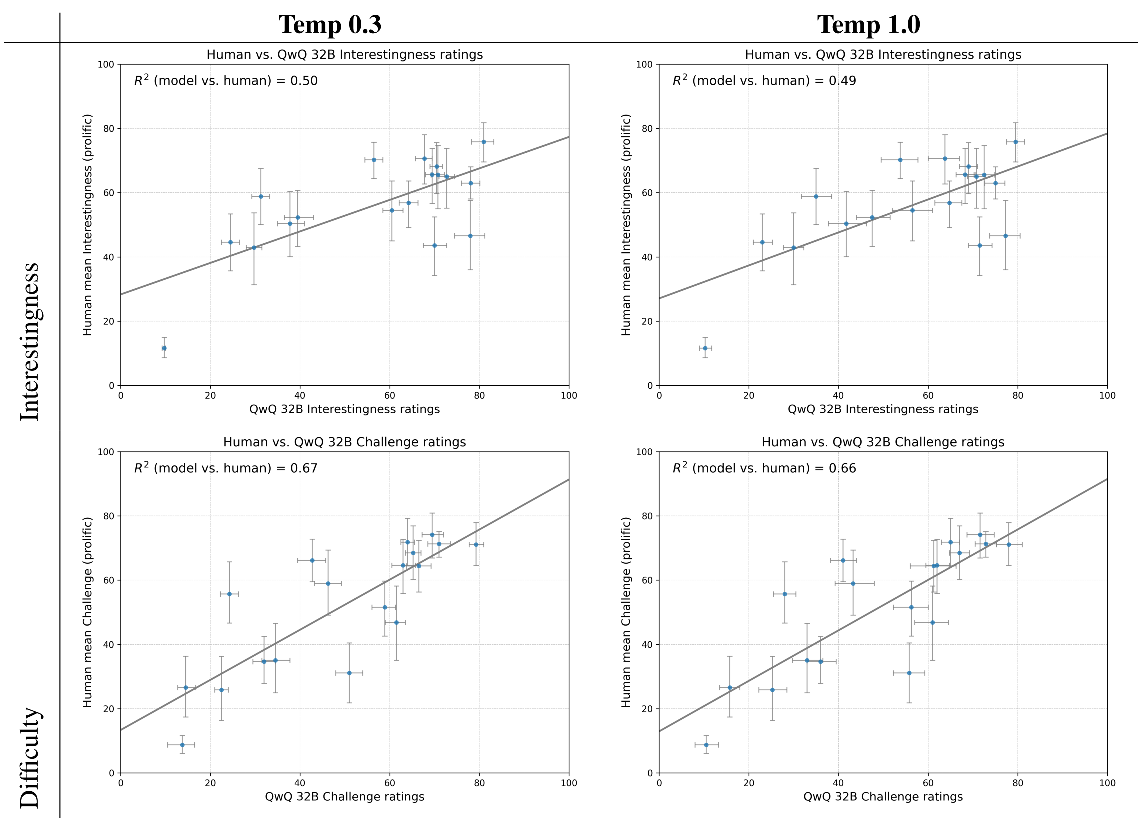


Figure 14: QwQ-32B: Human vs LLM ratings

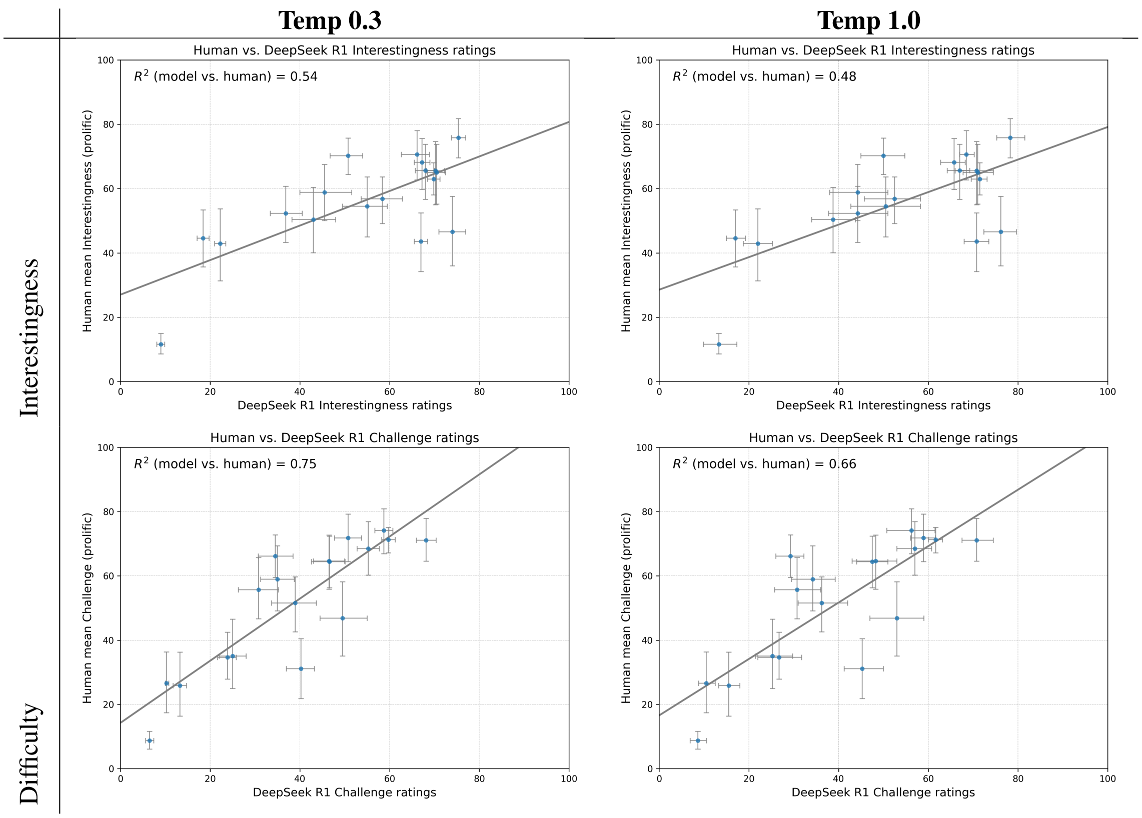


Figure 15: Deepseek R1: Human vs LLM ratings

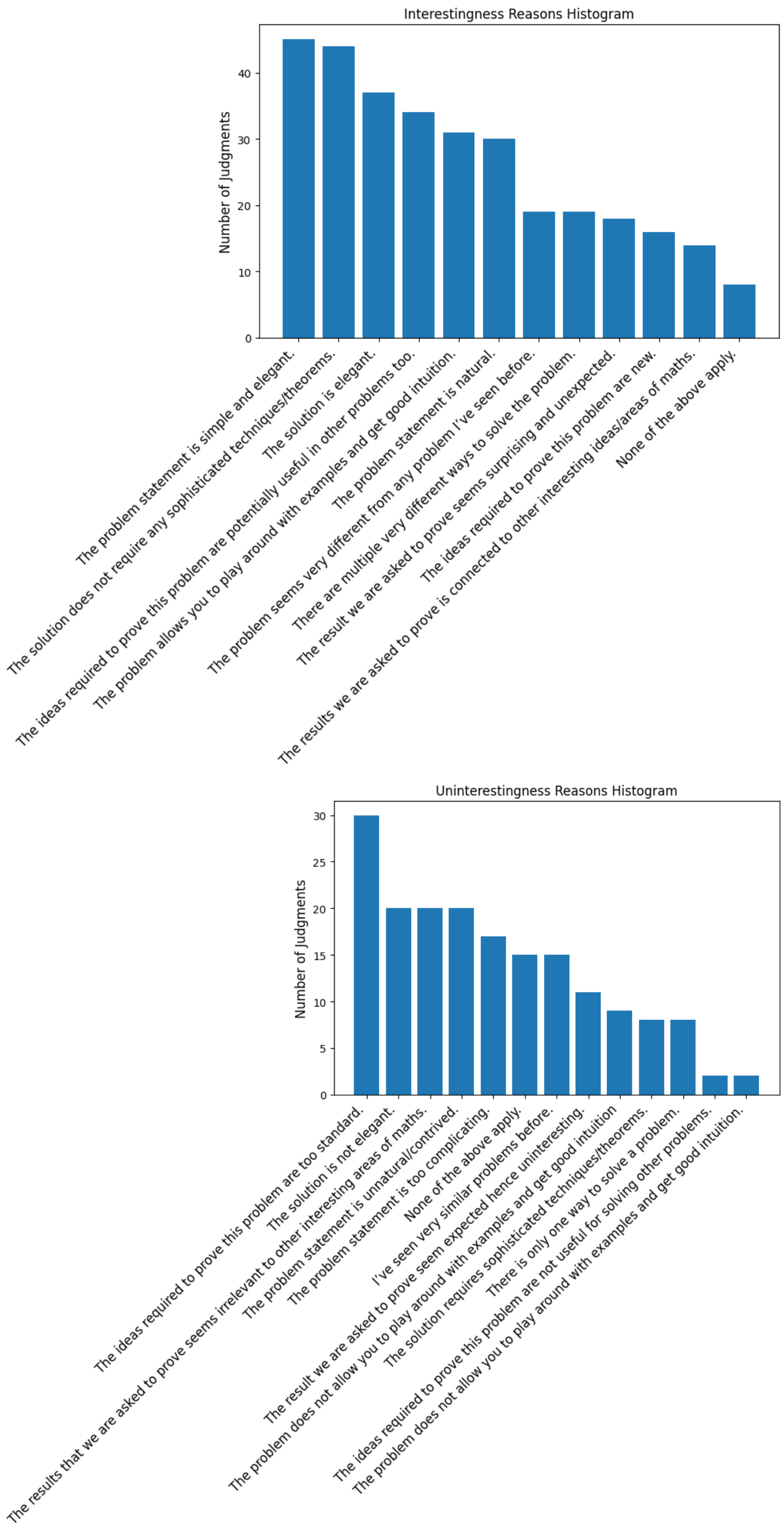
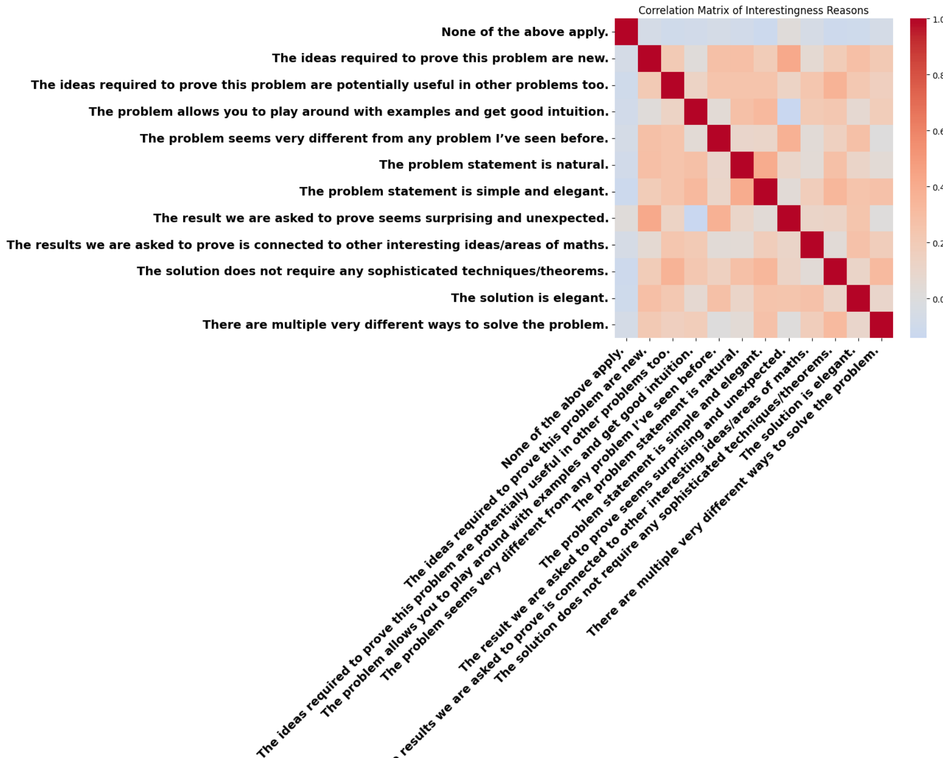


Figure 16: Histogram of how frequently interestingness (top) and uninterestingness (bottom) reasons were chosen by human participants across the survey.



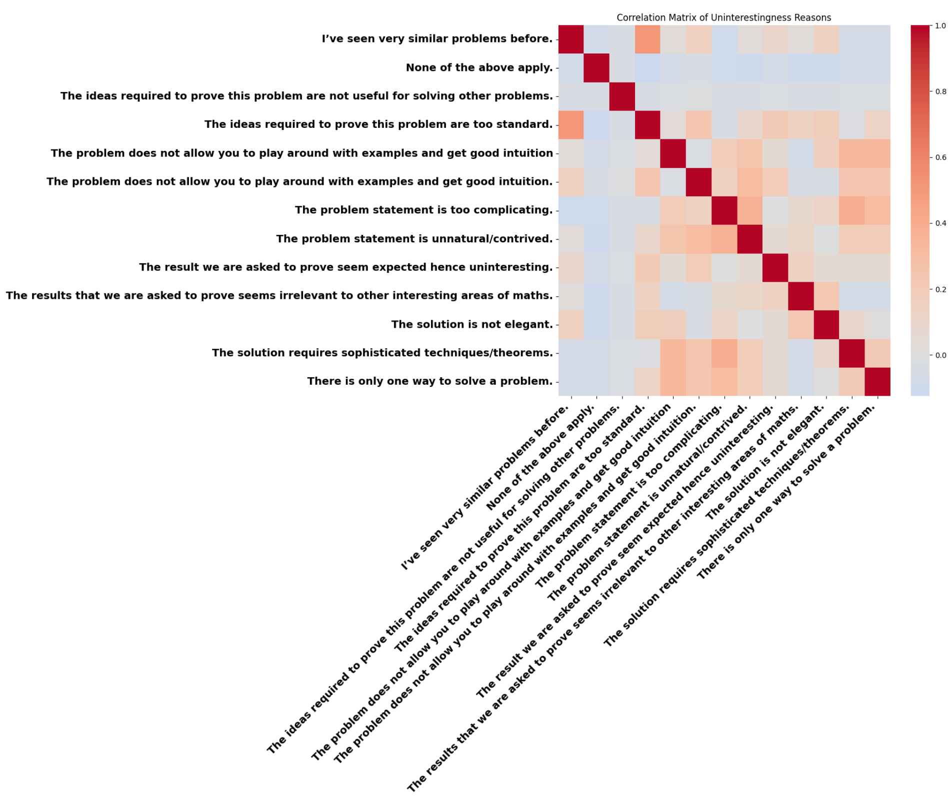


Figure 17: Correlation matrices showing when human participants in our IMO study chose multiple reasons for interestingness (top) or uninterestingness (bottom) for the same problem.

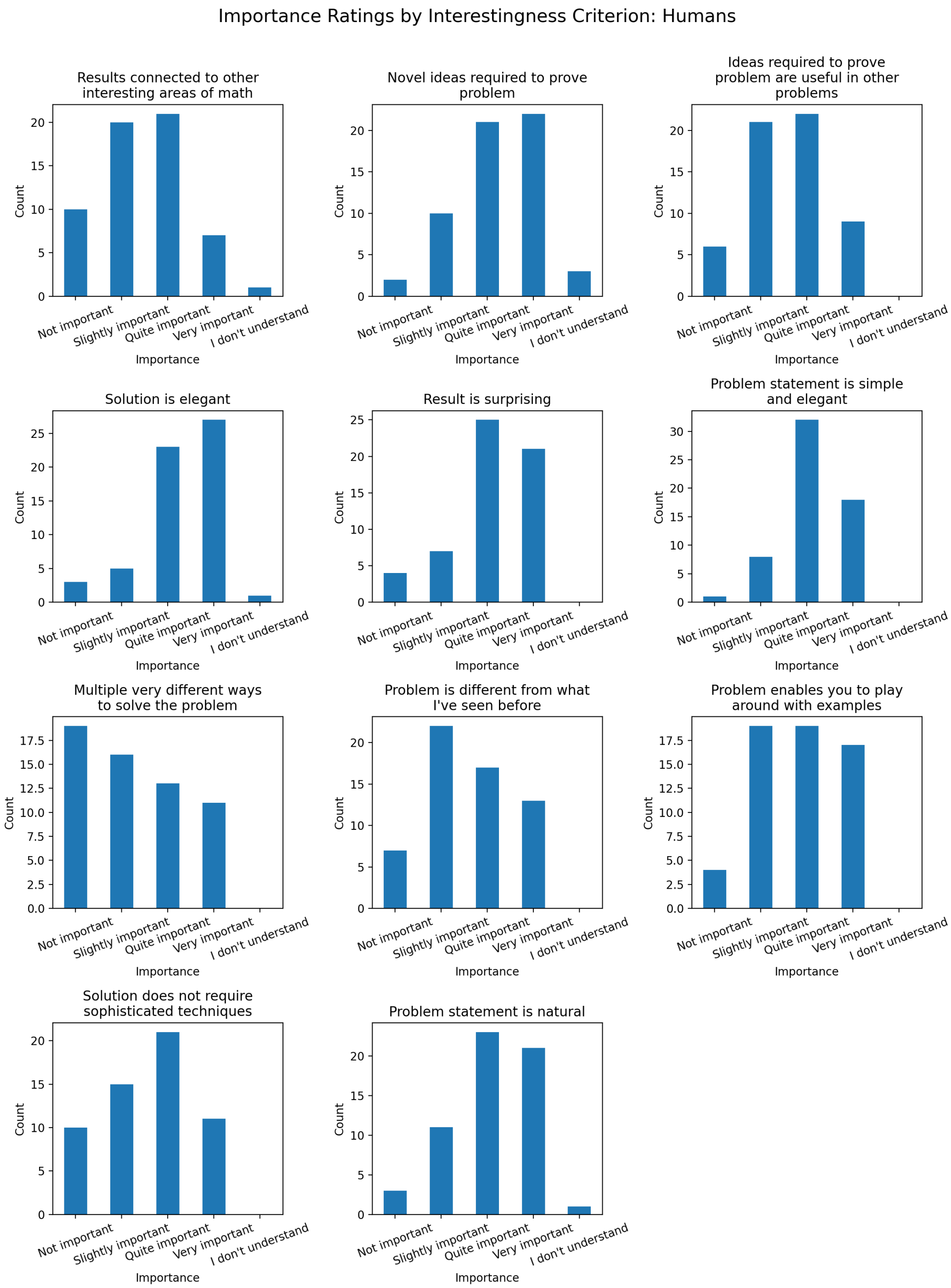


Figure 18: Distribution of importance ratings from human participants across 11 interestingness criteria.

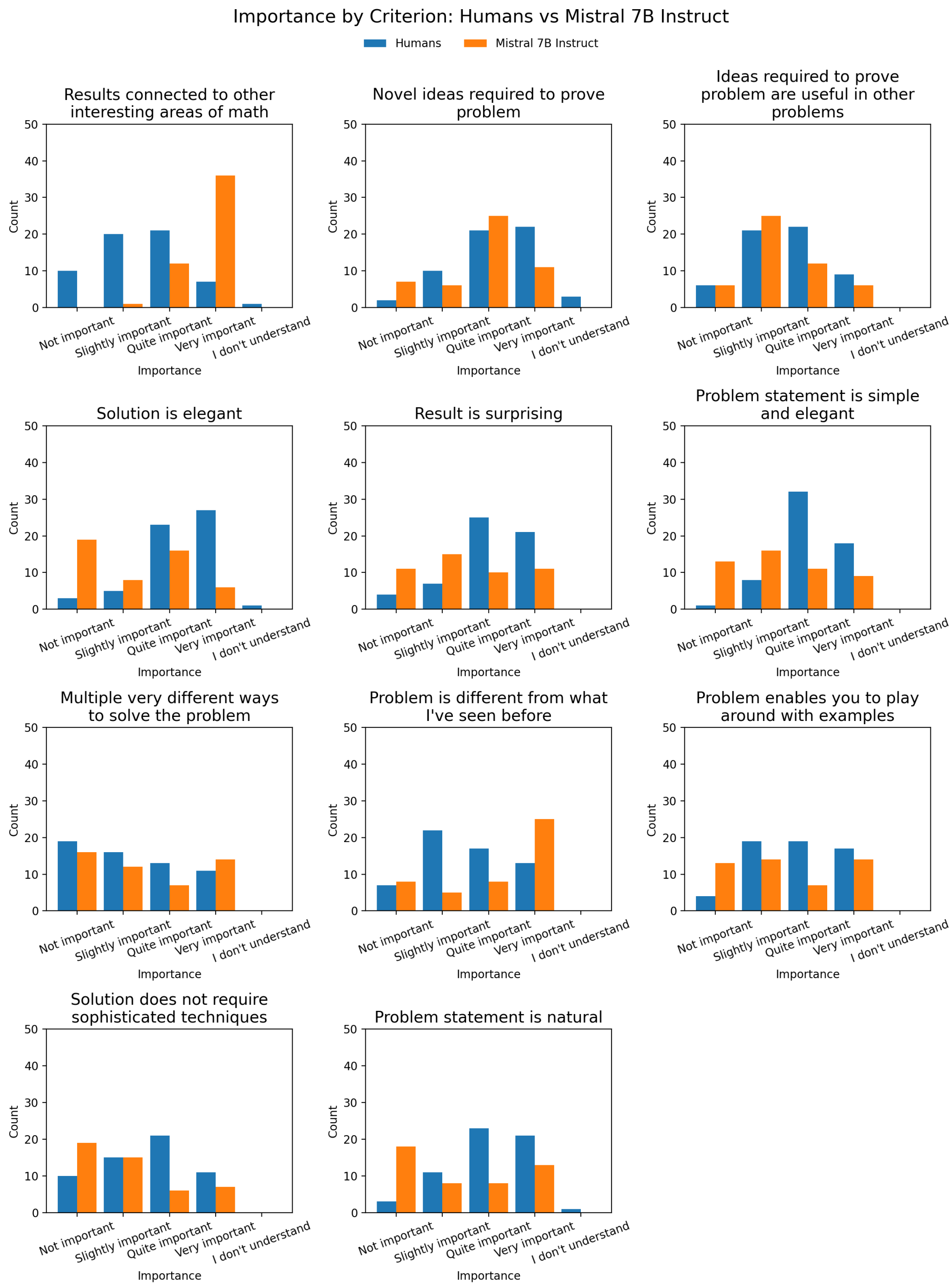


Figure 19: Importance ratings from **Mistral 7B Instruct** across 11 interestingness criteria.

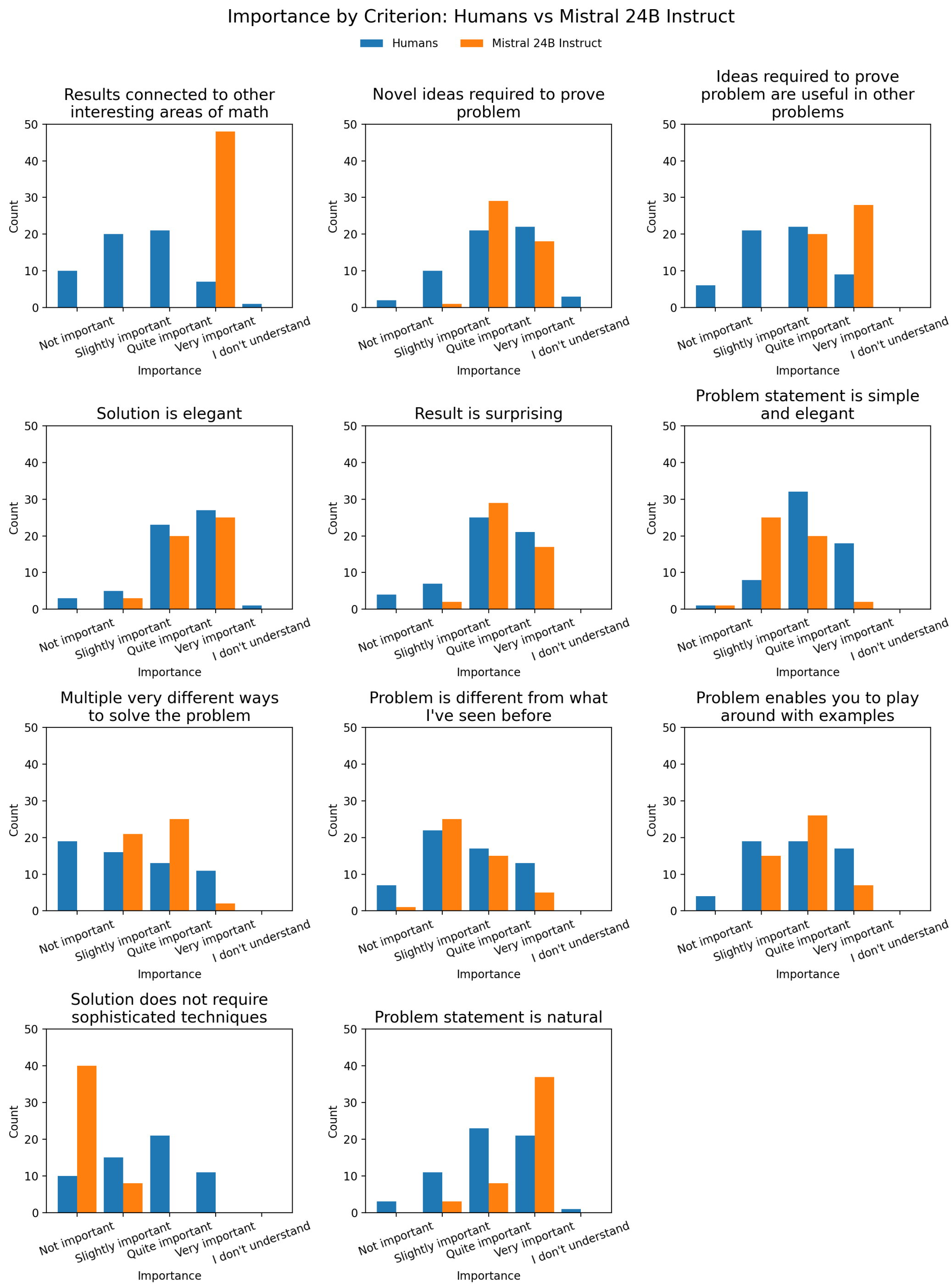


Figure 20: Importance ratings from **Mistral 24B Instruct**.

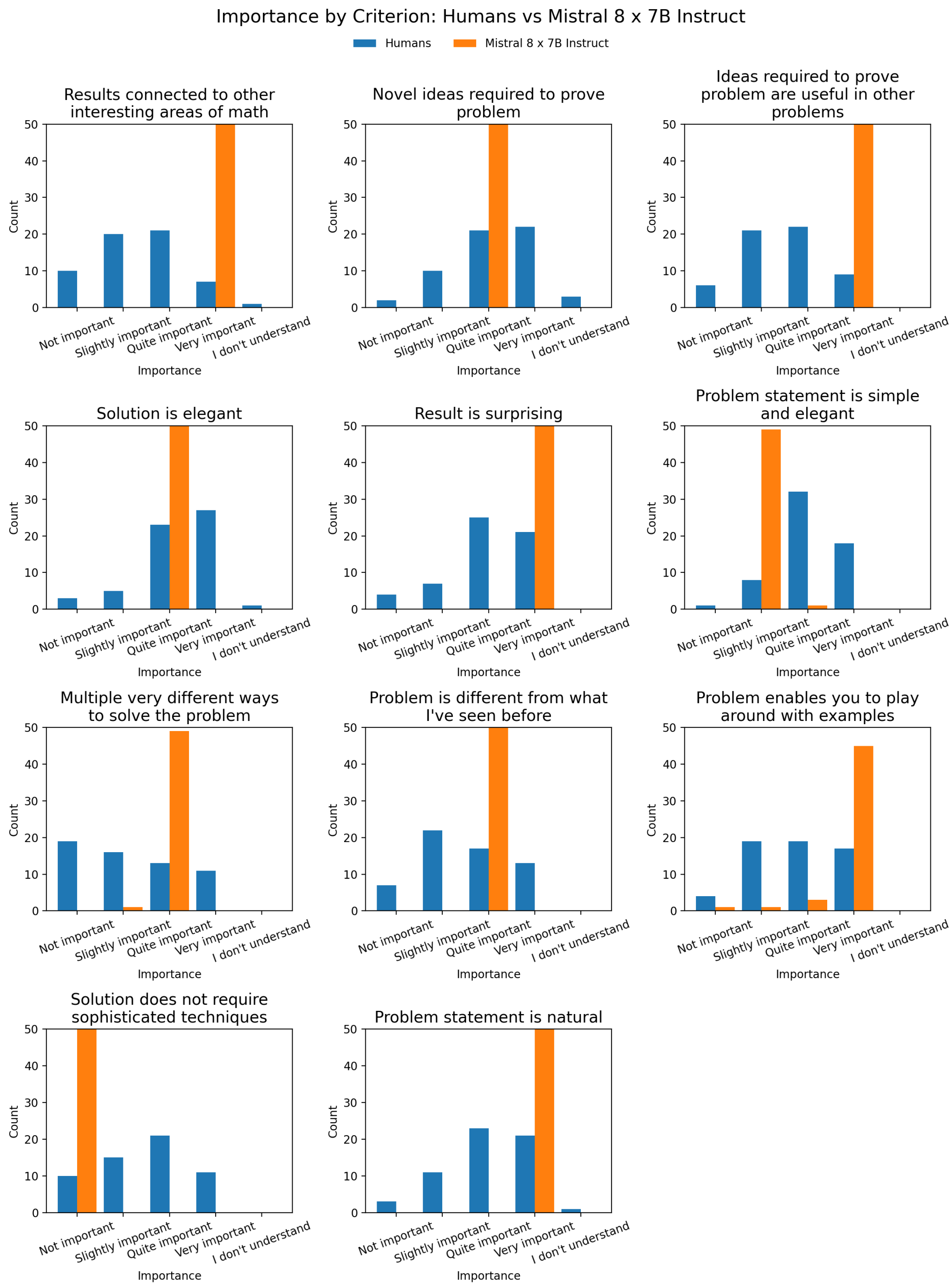


Figure 21: Importance ratings from **Mixtral 8×7B Instruct**.

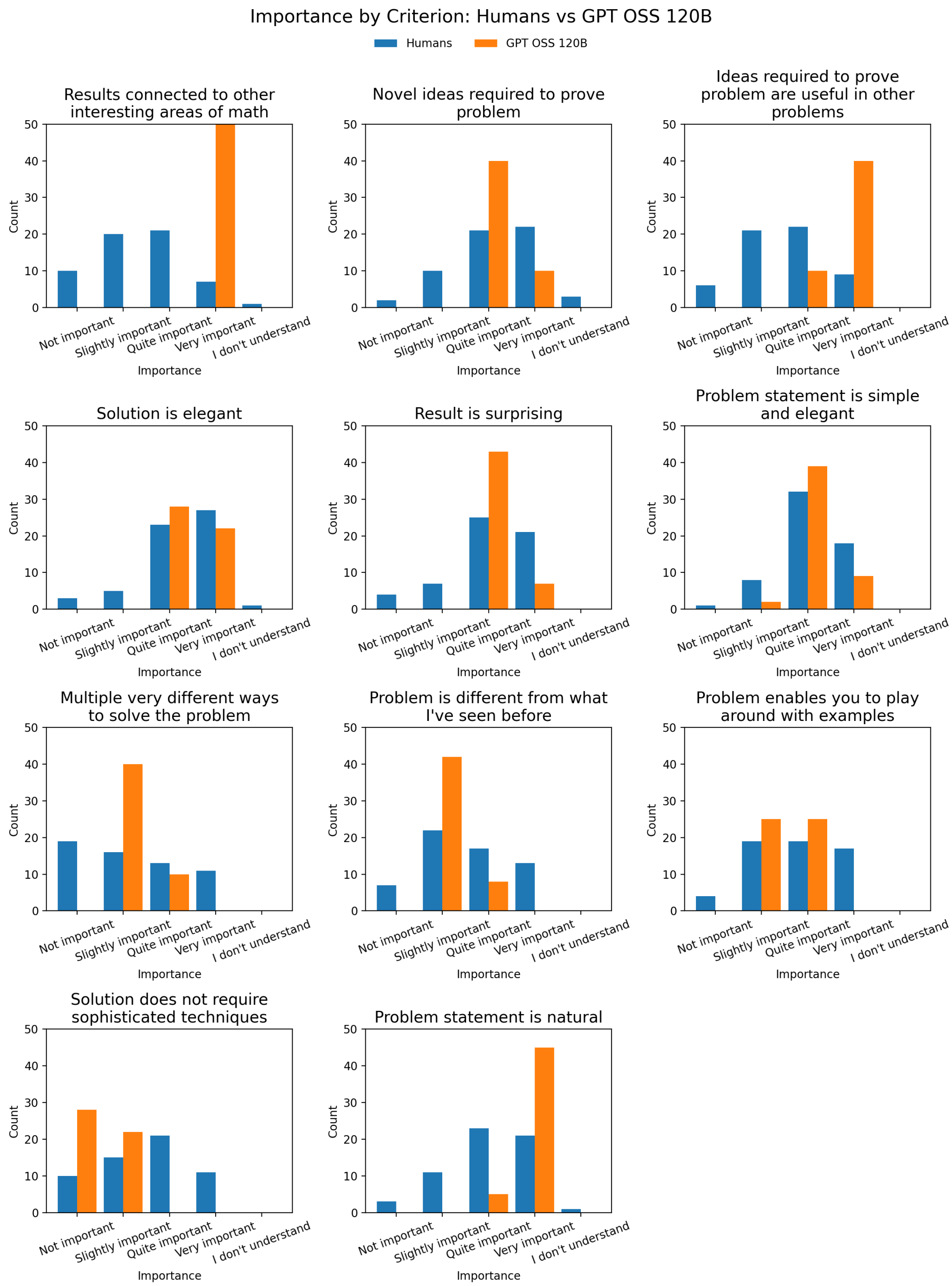


Figure 22: Importance ratings from **GPT-OSS 120B**.

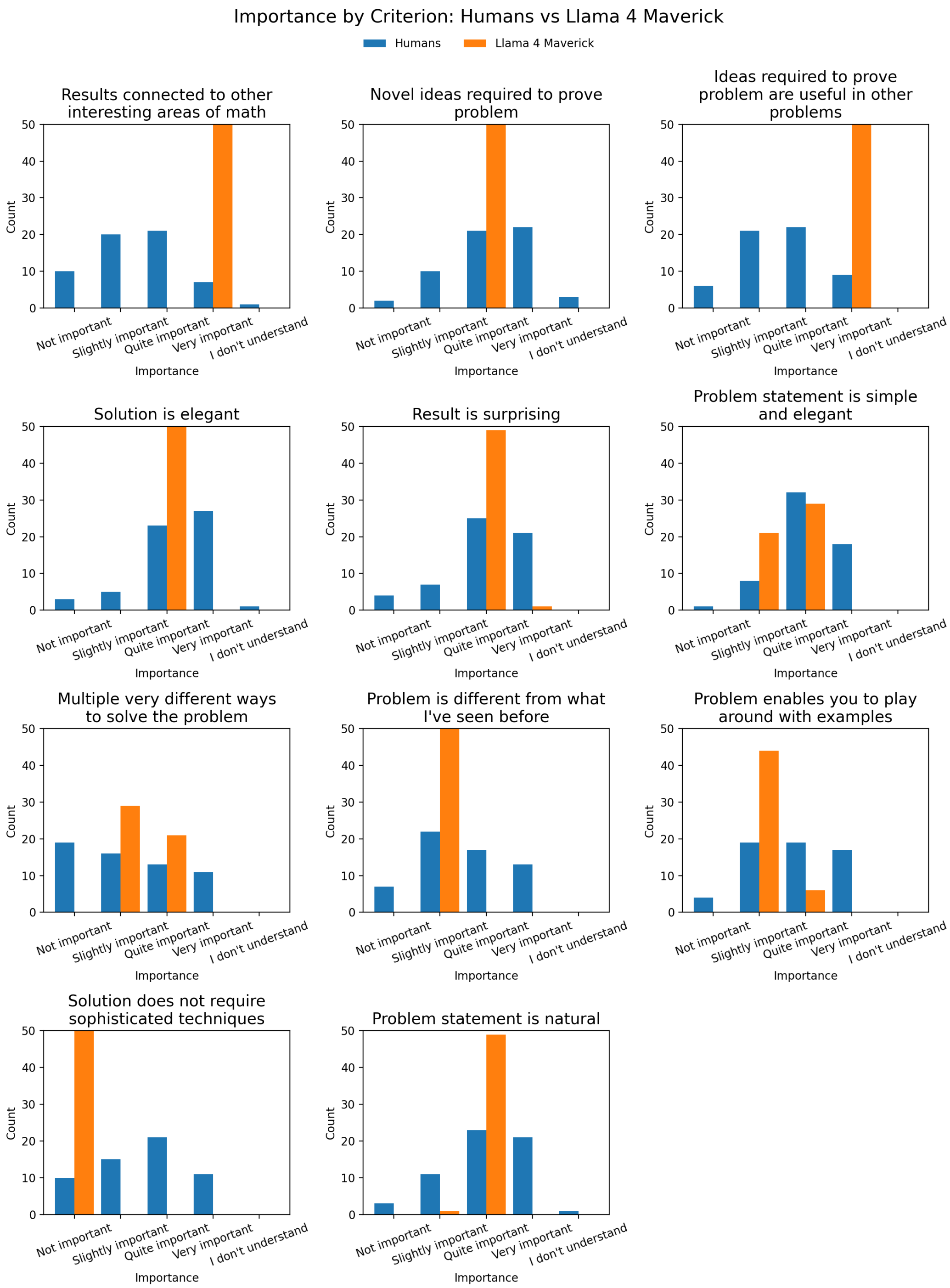


Figure 23: Importance ratings from **Llama 4 Maverick**.

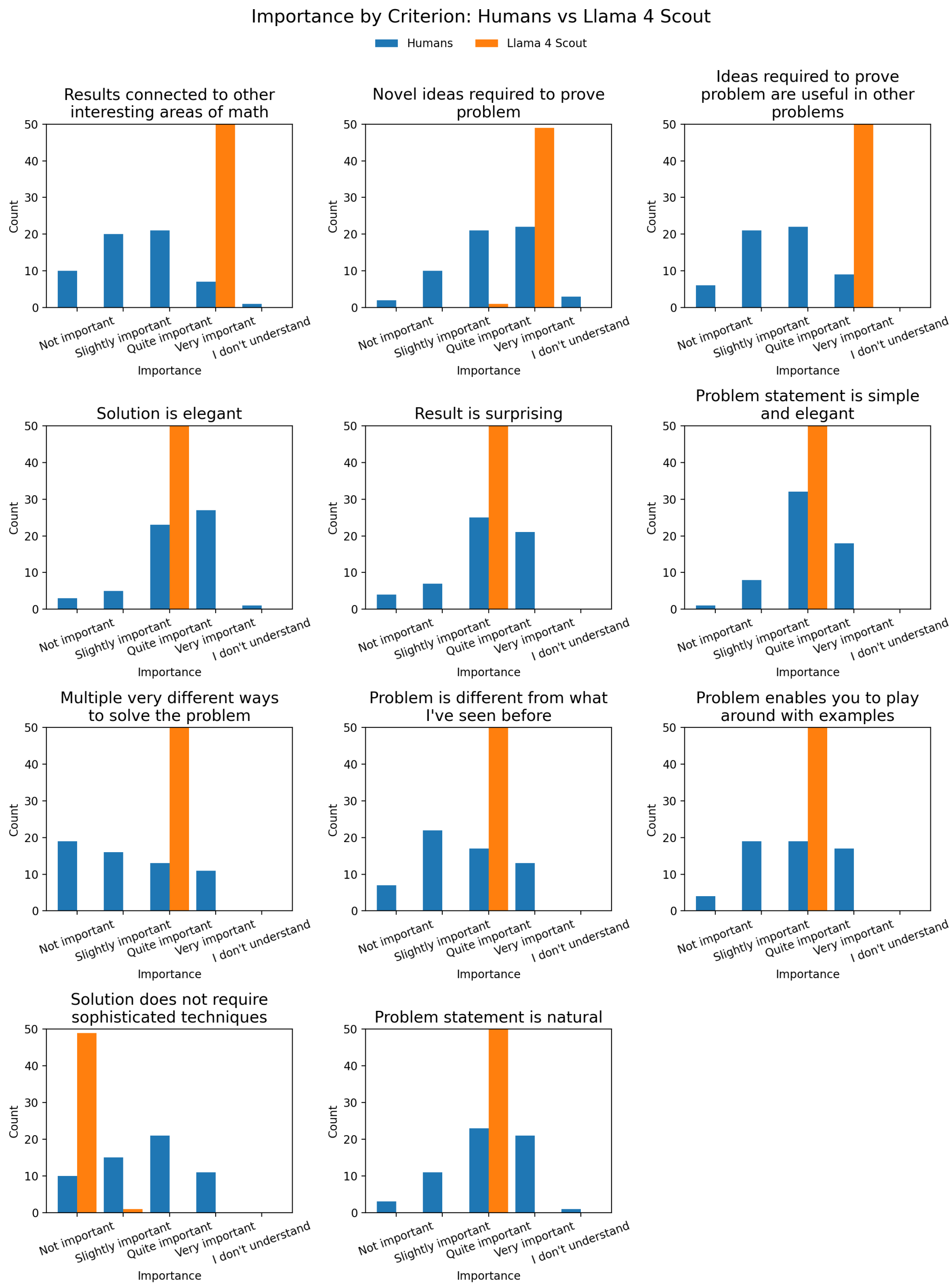


Figure 24: Importance ratings from **Llama 4 Scout**.

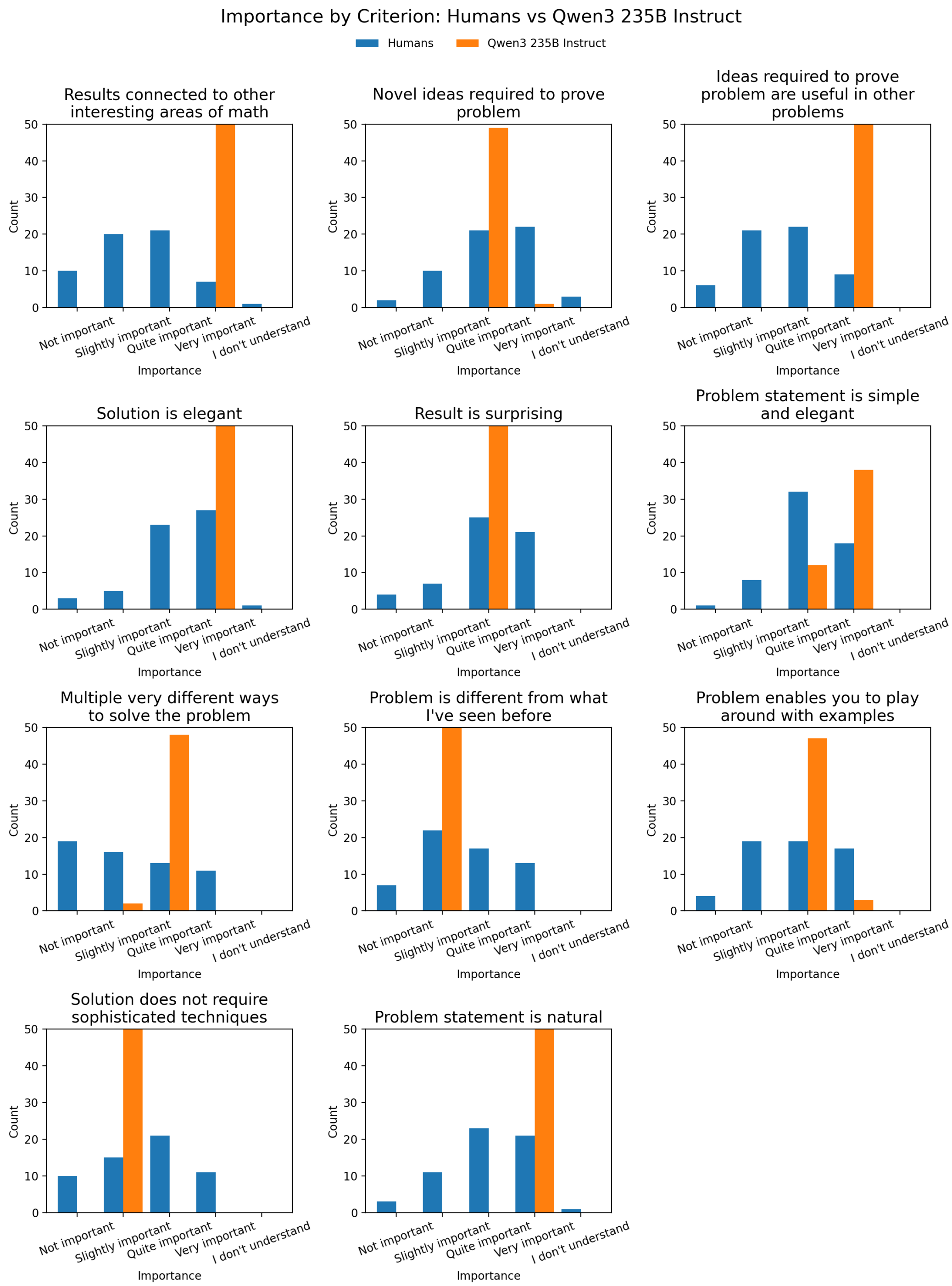


Figure 25: Importance ratings from **Qwen 235B Instruct**.

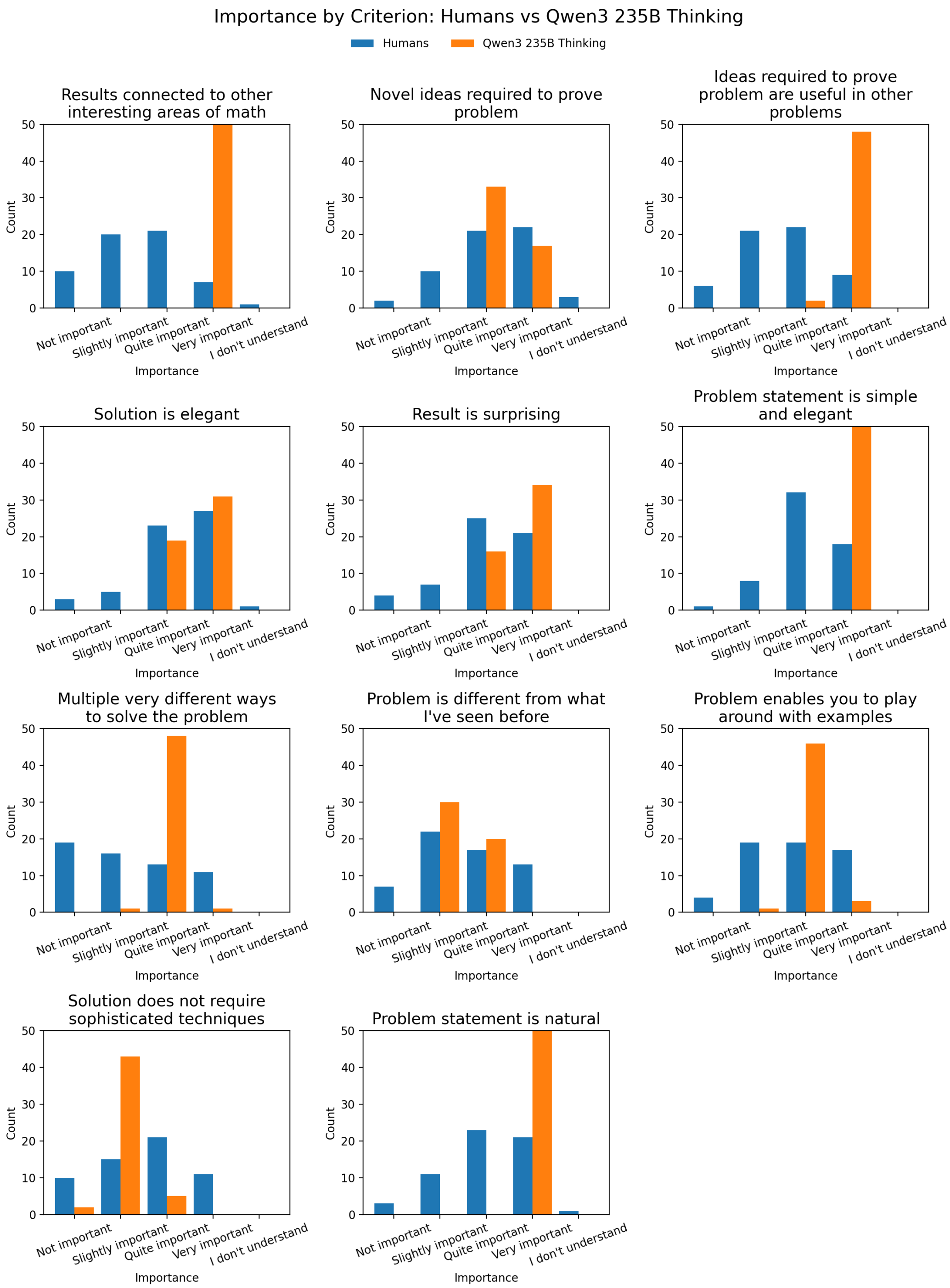


Figure 26: Importance ratings from **Qwen 235B Thinking**.

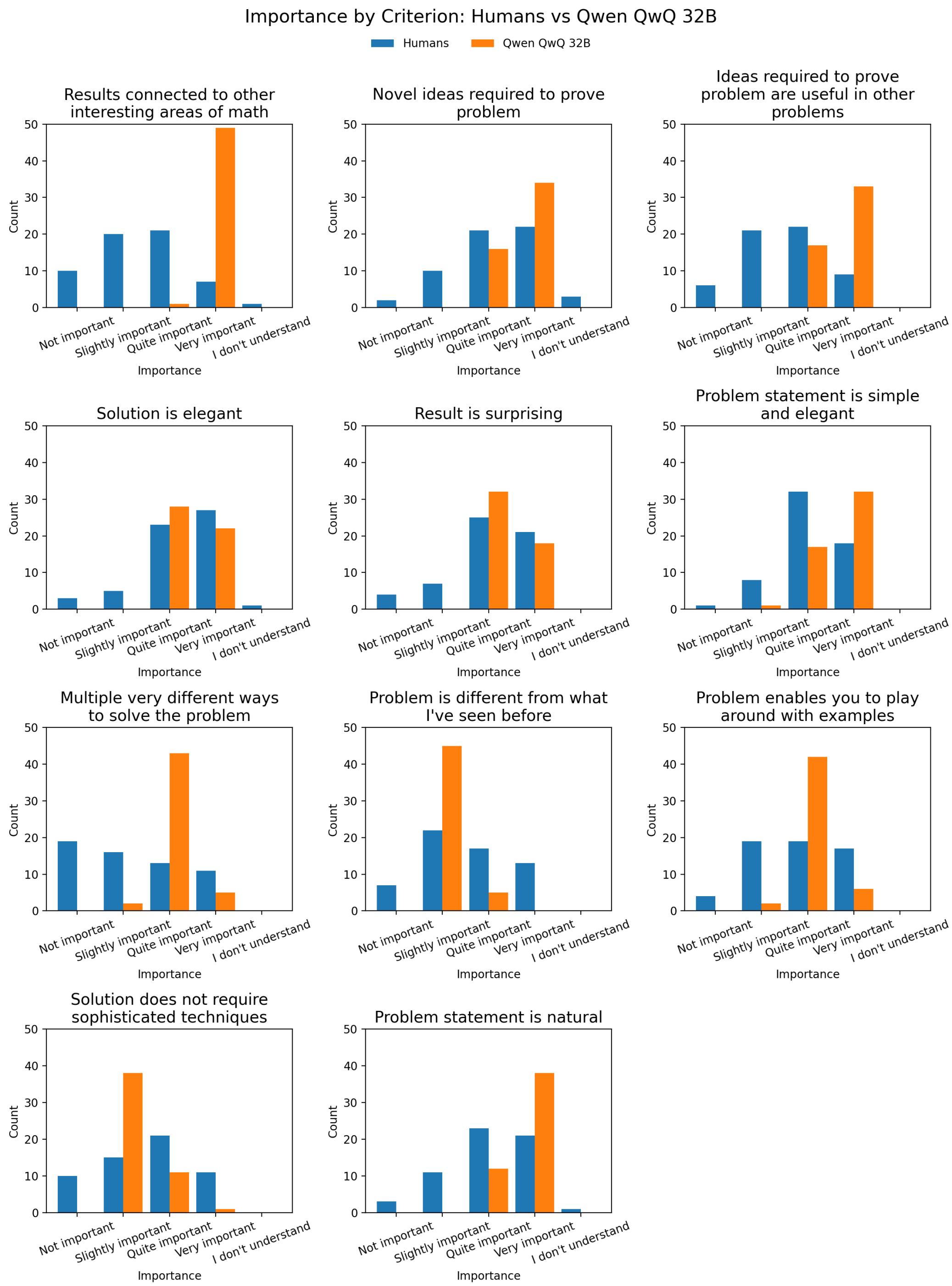


Figure 27: Importance ratings from **QwQ 32B**.

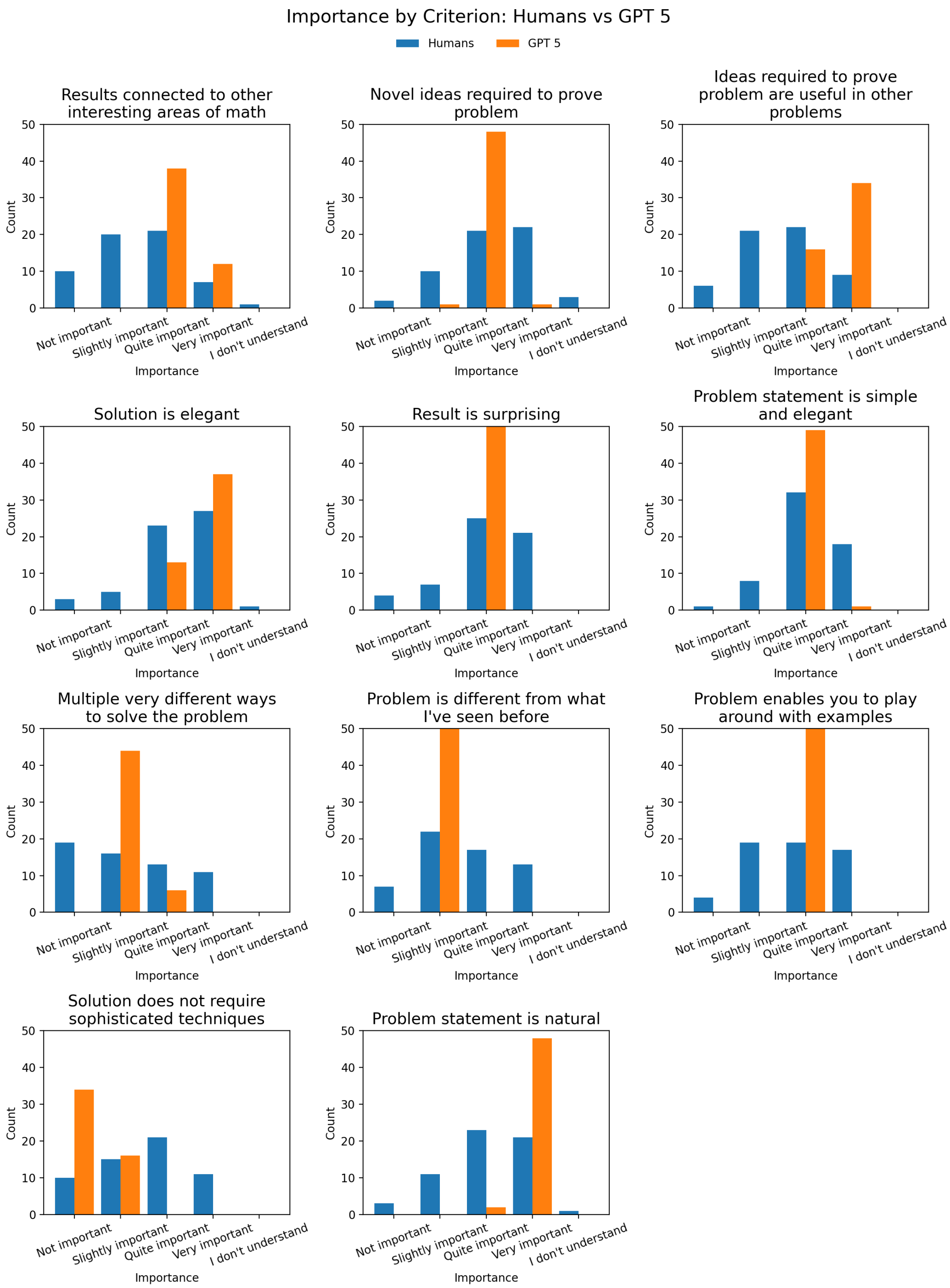


Figure 28: Importance ratings from **GPT-5**.

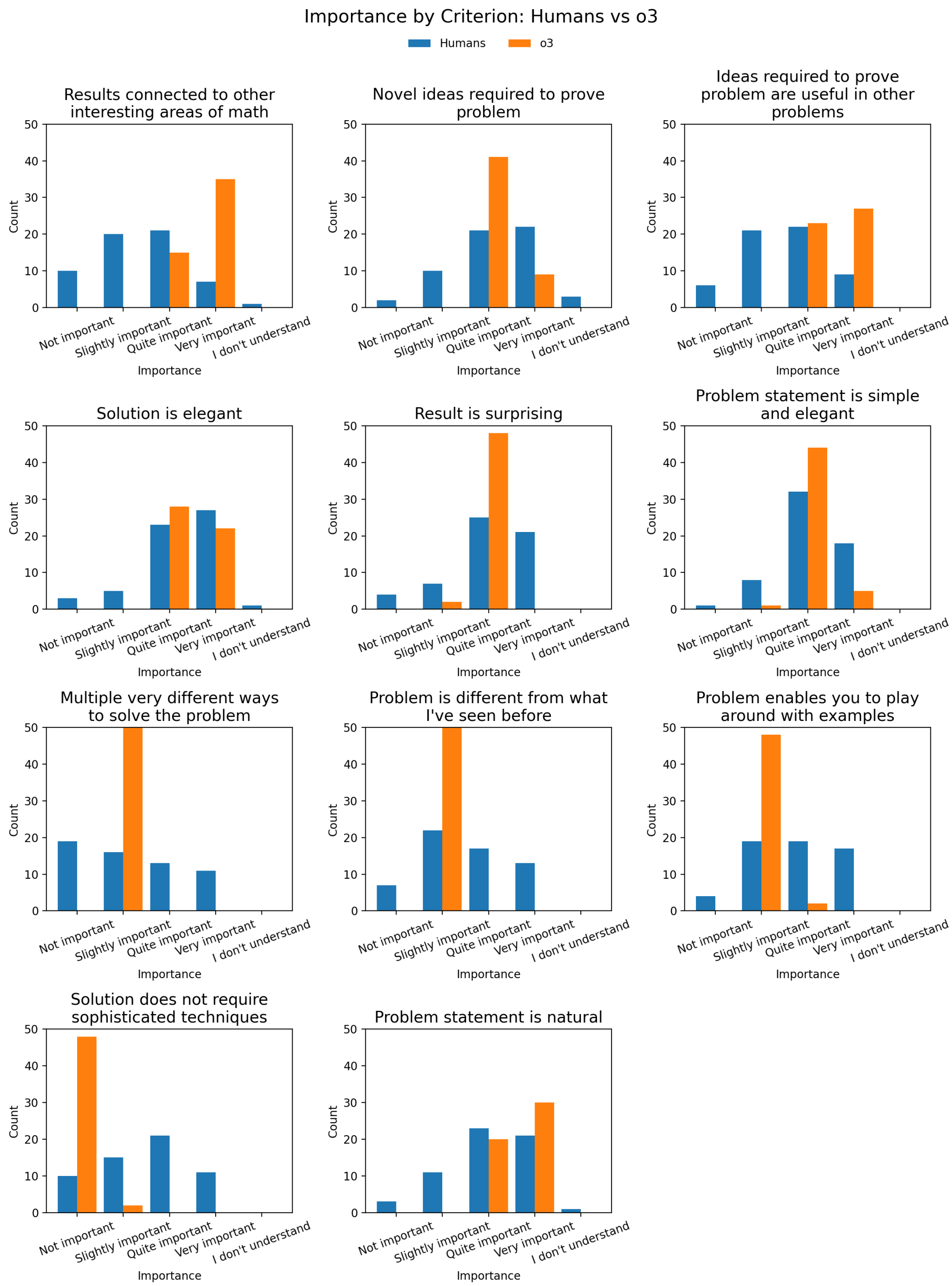


Figure 29: Importance ratings from **o3**.

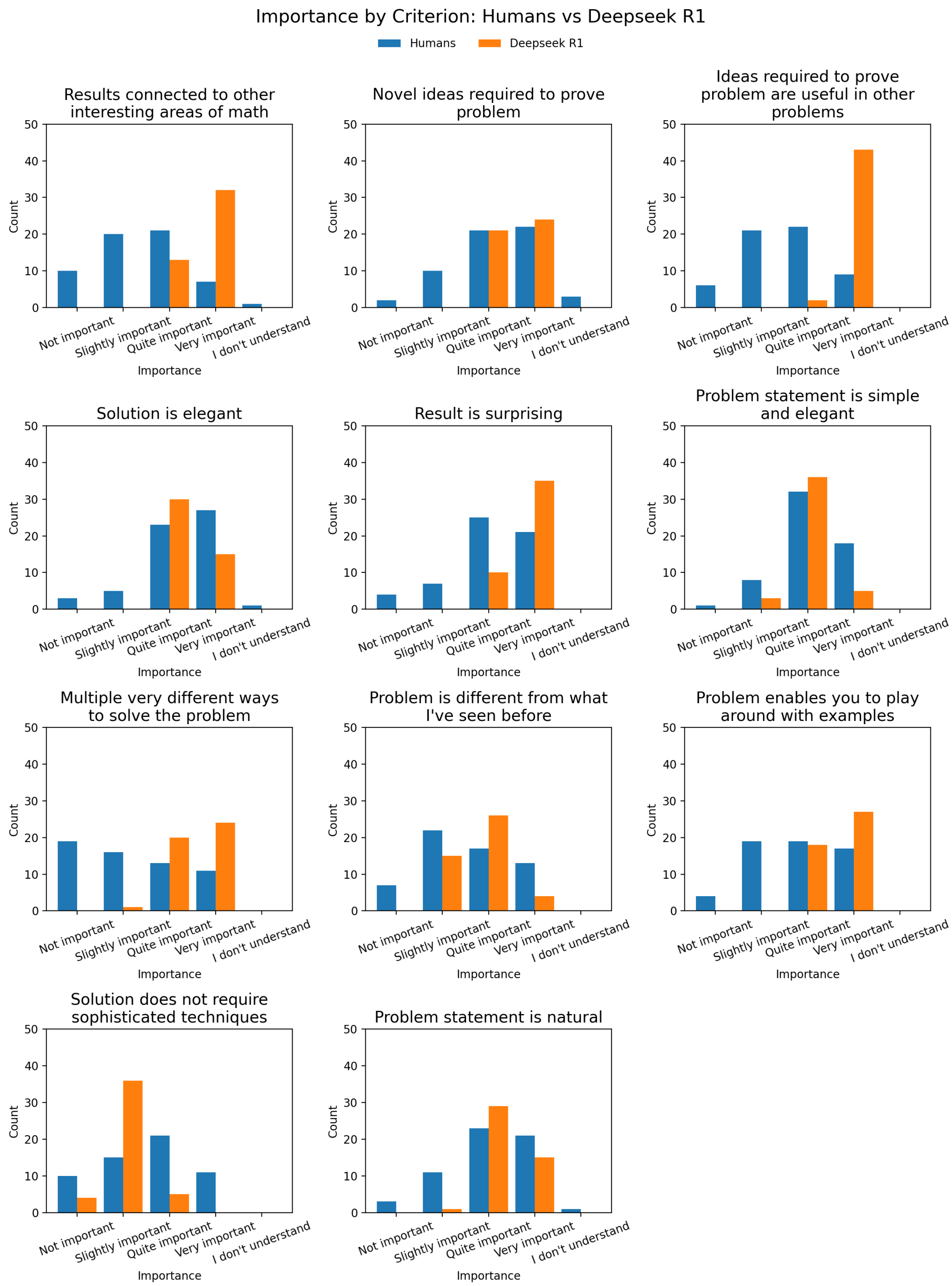


Figure 30: Importance ratings from **DeepSeek R1**.

A2 Models

Model	Full Name	Provider
llama_scout	meta-llama/Llama-4-Scout-17B-16E-Instruct	Meta (TogetherAI)
llama_maverick	meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8	Meta (TogetherAI)
deepseek_r1	deepseek-ai/DeepSeek-R1	DeepSeek (TogetherAI)
o3	o3	OpenAI
gpt_5	gpt-5	OpenAI
gpt_oss_120b	openai/gpt-oss-120b	OpenAI (TogetherAI)
mixtral_8x7b_instruct	mistralai/Mixtral-8x7B-Instruct-v0.1	Mistral (TogetherAI)
mistral_7b_instruct	mistralai/Mistral-7B-Instruct-v0.1	Mistral (TogetherAI)
mistral_24b_instruct	mistralai/Mistral-Small-24B-Instruct-2501	Mistral (TogetherAI)
qwen_235b_instruct	Qwen/Qwen3-235B-A22B-Instruct-2507-tput	Qwen (TogetherAI)
qwen_235b_thinking	Qwen/Qwen3-235B-A22B-Thinking-2507	Qwen (TogetherAI)
qwq_32b	Qwen/QwQ-32B	Qwen (TogetherAI)

Table 7: Models evaluated in this study, their full identifiers, and providers.

A3 Prolific Study

A3.1 Variant types

- **Increasing/decreasing numbers:** The numerical values in the problem are scaled up or down while keeping the structure intact.
- **Adding/removing steps:** The problem is modified to include additional intermediate steps, or simplified by removing steps.
- **Adding ambiguity:** The wording is adjusted to introduce multiple plausible interpretations.

A3.2 List of problems

We list all problems for our Prolific study, their variants, and variant types in Table 8.

Original Problem	**Variant Problem**	**Variant Type**
Real numbers x and y satisfy the following system: $x^2+y^2=25$ $(x+y+5)(-x+y+5)(x-y+5)(x+y-5)=100$ and $x+y=\sqrt(m)$. Determine m.		Positive control (no variant)
What is $28+13$?		Negative control (no variant)
What is the value of $9901 \cdot 101 - 99 \cdot 10101$?	What is the value of $999001 \cdot 1001 - 999 \cdot 1001001$?	Increase value
The number 2024 is written as the sum of not necessarily distinct two-digit numbers. What is the least number of two-digit numbers needed to write this sum?	The number 500 is written as the sum of not necessarily distinct two-digit numbers. What is the least number of two-digit numbers needed to write this sum?	Decrease value
A data set containing 20 numbers, some of which are 6, has mean 45. When all the 6s are removed, the data set has mean 66. How many 6s were in the original data set?	A data set containing 20 numbers, 7 of which are 6, has mean 45. When all the 6s are removed, what is the mean of the dataset?	Remove step
Kei draws a 6-by-6 grid. He colors 13 of the unit squares silver and the remaining squares gold. Kei then folds the grid in half vertically, forming pairs of overlapping unit squares. Let m and M equal the least and greatest possible number of gold-on-gold pairs, respectively. What is the value of $m+M$?	Kei draws a 6-by-6 grid. He colors 13 of the unit squares silver and the remaining squares gold. Kei then folds the grid in half vertically, forming pairs of overlapping unit squares. Let m equal the least possible number of gold-on-gold pairs. What is the value of m?	Remove step
In a long line of people arranged left to right, the 1013th person from the left is also the 1010th person from the right. How many people are in the line?	In a long line of people, the 1013th person from one end is also the 1010th person from the other end. How many people are in the line?	Add ambiguity
Makayla finds all the possible ways to draw a path in a 5×5 square-shaped grid. Each path starts at the bottom left of the grid and ends at the top right, always moving one unit east or north. She computes the area of the region between each path and the right side of the grid. What is the sum of the areas determined by all possible paths?	Makayla finds all the possible ways to draw a path in a 2×2 square-shaped grid. Each path starts at the bottom left of the grid and ends at the top right, always moving one unit east or north. She computes the area of the region between each path and the right side of the grid. What is the sum of the areas determined by all possible paths?	Decrease value
Lucius is counting backward by 7s. His first three numbers are 100, 93, and 86. What is his 10th number?	Lucius is counting backward by 7s. His first three numbers are 100, 93, and 86. What is his 5th number?	Decrease value
$WXYZ$ is a rectangle with $WX=4$ and $WZ=8$. Point M lies $\overline{XY}$, point A lies on $\overline{YZ}$, and $\angle WMA$ is a right angle. The areas of $\triangle WXM$ and $\triangle WAZ$ are equal. What is the area of $\triangle WMA$?	$WXYZ$ is a rectangle with $WX=4$ and $WZ=8$. Point M lies $\overline{XY}$, point A lies on $\overline{YZ}$, and $\angle WMA$ is a right angle. The areas of $\triangle WXM$ and $\triangle WAZ$ are equal. What is the sum of the areas of $\triangle WMA$ and $\triangle WAZ$?	Add step

Table 8: Prolific study problems and their variants.

A4 IMO Study

One of our authors physically attended the IMO in 2024 and asked participants to complete a survey of their mathematics judgments. The survey received prior ethics approval by our institutional ethics review board.

A4.1 List of Interestingness and Uninterestingness Reasons

Interestingness reasons:

- The results we are asked to prove is connected to other interesting ideas/areas of maths.
- The ideas required to prove this problem are new.
- The ideas required to prove this problem are potentially useful in other problems too.
- The solution is elegant.
- The result we are asked to prove seems surprising and unexpected.
- The problem statement is simple and elegant.
- There are multiple very different ways to solve the problem.
- The problem seems very different from any problem I've seen before.
- The problem allows you to play around with examples and get good intuition.
- The solution does not require any sophisticated techniques/theorems.
- The problem statement is natural.
- None of the above apply.

Uninterestingness reasons:

- The results that we are asked to prove seems irrelevant to other interesting areas of maths.
- The ideas required to prove this problem are too standard.
- The ideas required to prove this problem are not useful for solving other problems.
- The solution is not elegant.
- The result we are asked to prove seem expected hence uninteresting.
- The problem statement is too complicating.
- There is only one way to solve a problem.
- I've seen very similar problems before.
- The problem does not allow you to play around with examples and get good intuition.
- The solution requires sophisticated techniques/theorems.
- The problem statement is unnatural/contrived.
- None of the above apply.