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

Shubhra Mishra[1], **Yuka Machino**[2,3,4], **Gabriel Poesia**[5], **Albert Jiang**[4,6], **Joy Hsu**[2], **Adrian Weller**[4], **Challenger Mishra**[4], **Joshua B. Tenenbaum**[3], **David Broman**[1,*], **Mateja Jamnik**[4,*], **Cedegao E. Zhang**[3,*], **Katherine M. Collins**[3,4,7*]
KTH Royal Institute of Technology[1], Stanford University[2], Massachusetts Institute of Technology[3], University of Cambridge[4], University of Michigan, Ann Arbor[5], Mistral AI[6], Princeton AI Lab[7]
{shubhram, dbro}@kth.se, mateja.jamnik@cl.cam.ac.uk, {cedzhang, katiemc}@mit.edu

Abstract

The evolution of mathematics is shaped importantly by *interestingness*: researchers choose which problems to pursue, and students choose which problems to engage with, based on expectations of interest and challenge. As AI systems, particularly large language models (LLMs) that operate flexibly over natural language and formal mathematics, are increasingly used in mathematics research and education, it becomes crucial to characterize how closely their judgments align with people from different mathematical backgrounds. We study whether LLMs align with human interestingness judgments by comparing LLM ratings with those of two populations, crowdsourced participants with college math experience and International Math Olympiad competitors. Although many LLMs broadly agree with human notions of interestingness, they largely fail to match the *distribution* of human judgments. They also weakly align with *why* humans find problems interesting, with low correlation to human-selected rationales. Finally, we evaluate LLMs' ability to *generate* interesting problems and find that, after filtering for validity, LLMs are able to generate engaging problems. We conclude with takeaways, including the need for multi-LLM human-AI collaborative systems, that highlight both the promise and current limits of LLMs as partners in mathematical reasoning. We make all data and code available here.

1 Introduction

Mathematical reasoning involves not just solving problems but also judging whether a problem is worth solving. 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) [7, 25, 34, 41, 1]. While this progress is impressive, *problems are often given to models by humans*. Even computer-assisted discoveries like improved bounds on the CapSet problem (by FunSearch [50]), better matrix multiplication algorithms (by AlphaTensor [18]), or the improved bound on the kissing number (by AlphaEvolve [40]) have ultimately depended on human-posed targets or carefully designed heuristics.

It is unclear whether LLMs and LRMs can adequately judge and select which problems are worth solving at all [8]. This is important because problem selection is crucial for many potential appli-

*Co-senior authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

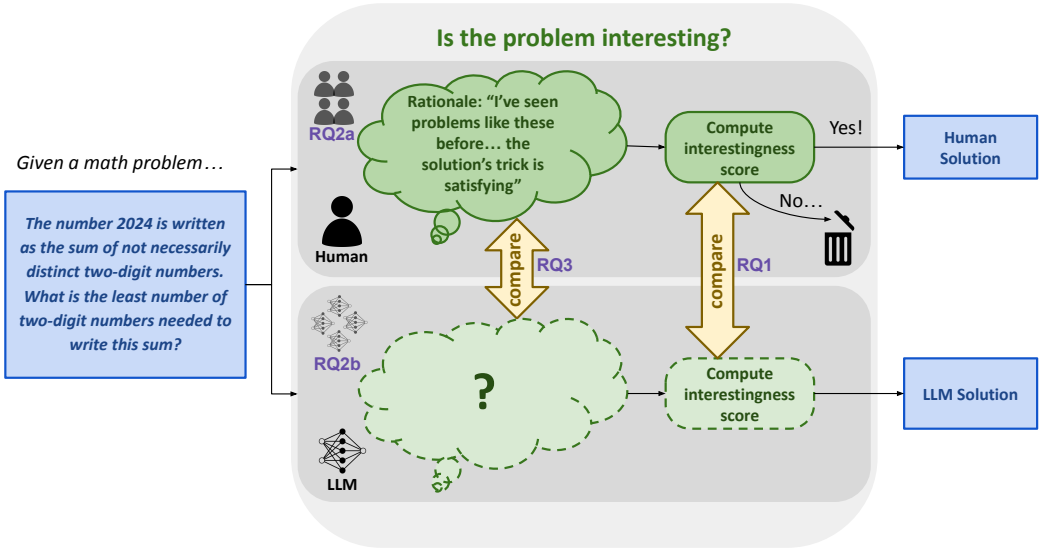


Figure 1: Our experimental pipeline for studying human-LLM alignment on interestingness is depicted above. For any problem, human participants rate *how* interesting it is and *why* they find it interesting (top panel). We also evaluate LLMs' ratings for these (bottom panel), comparing human-LLM alignment on how interesting a problem is (RQ1) and their rationales for interestingness (RQ3). We explore whether the *distributions* of judgments between humans and LLMs align (RQ2a) and if model pooling can improve this alignment (RQ2b). While not depicted in the figure, we also study how reasoning models reason about problem interestingness (RQ4) and explore whether LLMs are capable of generating mathematical problems that interest humans (RQ5).

cations of LLMs from education (e.g., proposing interesting problems and examples to students) to automated mathematical discovery (AMD; posing interesting conjectures to explore) [43, 37, 2]. Answering the question *"is this problem worth solving?"* often involves estimating properties of the problem before any experience e.g., whether the problem is likely to be challenging, or generally interesting. While some prior work has explored the notion of interestingness in humans [26, 14, 3], in this paper, we take **the first step** in empirically comparing interestingness and difficulty judgments across humans and LLMs.

Classical work in AMD modeled interestingness using hand-coded heuristics to guide conjecture generation and theorem proving—often hundreds of them, limiting scalability [12, 13, 31, 12, 17] . 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 [37]. Other work has also explored evolutionary search to find symbolic interestingness evaluators that lead to human-written statements being rediscovered [48]. 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 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 initial studies comparing LLMs and LRMs to human judgments of mathematical problem interestingness and understanding whether these models can actually *generate* interesting problems. We run two controlled studies, one with crowdsourced participants and one with IMO participants rating math problems from the American Math Competition and International Math Olympiad, respectively. As we see in Figure 1, participants are asked to rate how interesting they think a problem is and also why they find it interesting. We then evaluate LLMs on the same task of problem evaluation and compare the human and LLM judgments for all problems. Specifically, we make the following contributions:

- A systematic evaluation of whether a battery of LLMs judge problem interestingness in human-aligned ways.
- Insights into how LLMs evaluate the interestingness and difficulty of math problems, and how they do (and don't) align with those of humans. Importantly, we make a recommendation for the

use of multiple LLMs in discovery and human-AI collaboration systems and highlight a proxy for interestingness in LRMs.
- An initial exploration of LLMs' ability to generate problems that humans find interesting.

2 Related work

Our work lies at the intersection of (i) AI for mathematical problem solving, (ii) evaluating AI evaluations, and (iii) human judgments of interestingness.

AI for mathematical problem solving. The use of large language models for math problem solving began with grade- and high-school level mathematics with benchmarks like GSM8K and MATH [7, 25]. LLMs quickly saturated these benchmarks due to increased problem-solving abilities [4]. To address benchmark saturation, more challenging benchmarks like Putnam-AXIOM, PutnamBench, FrontierMath, and the AIMO prizes containing PhD and competition mathematics emerged [24, 47, 22, 27, 20]. Large Reasoning Models (LRMs), trained with Reinforcement Learning techniques to utilize additional compute at inference-time, quickly rose to this challenge. Closed LRMs from providers like DeepMind [34], Harmonic [1], and OpenAI [41] achieved Gold medal-level performance at the 2025 International Mathematics Olympiad, highlighting the gap between proprietary and open models. However, these LLMs were only evaluated on their ability to *solve* mathematical problems. These accomplishments complement a parallel line of computer-assisted discovery systems (e.g., FunSearch [50], AlphaTensor [18], and AlphaEvolve [40]), which demonstrated that search, learning, and symbolic tools can together yield novel or improved results. Despite the achievements, a gap remains: humans pose the problems that models solve and make discoveries for, and thus models are not evaluated at judging problems or generating interesting ones [cf. 54].

Evaluating AI Evaluations. LLMs have mostly interacted with human-posed problems so far; there is comparatively less understanding of how well LLMs *evaluate* whether problems are worth solving. This research direction goes beyond evaluating solution correctness or proof validity, and asks whether models can evaluate *tasks themselves* (e.g., rating difficulty, quality, or broader subjective properties) and whether these evaluations align with people. For example, recent work has examined whether LLMs' evaluations of board games align with those of humans [8]. This is in contrast to the "LLM-as-Judge" approach, which usually uses LLMs to evaluate *solutions*, for example, to evaluate the quality of writing, code, and LLM-generated answers to math problems [55, 28, 45, 23]. Recent work has used LLM judgments to filter synthetic mathematical problems and conjectures at scale [53, 2]: here, we systematically study how such judgments compare with those of humans.

Human interestingness judgments. Previous work in cognitive science has attempted to build models of what humans find interesting, and ultimately, what drives human curiosity [32, 15]. Some experiments studied students' reactions to irregular patterns [3] and complex shapes [14], linking how they judged interestingness to complexity, novelty, and incongruity, and how these factors shape exploratory behavior. Studies done within mathematics specifically explore expert appraisals of what makes a proof "beautiful" [26]. Attempts to build artificially intelligent mathematical systems, including AMD systems such as Graffiti [12], AM [13, 31], HR [12, 13], and GT [17], operationalized "interestingness" via large sets of hand-coded heuristics, which enabled early progress but limited scalability and adaptability across domains. While research in cognitive science has tried to explain how humans judge interestingness, work involving AMD systems often manually imbues these systems with problem features humans deem interesting. LLMs, which can be probed using natural language, reveal a new possibility. Drawing on the spaces of cognitive science and AMD, in this work, we evaluate whether LLMs are capable of evaluating interestingness in math problems and outline suggestions for how AI systems can better be aligned with how humans judge mathematical problems. While not the primary focus in this work, our data collection and analyses also raise new questions about how people decide what problems are even worth solving [21, 39, 6, 11, 52].

3 Methods

We design and run a suite of human and model experiments to assess interestingness judgments (Figure 1). Specifically, we study six research questions to characterize LLMs' ability to evaluate and generate interesting math problems.

RQ1: How do LLMs' judgments of interestingness align with those of humans?
RQ2a: Are the *distributions* of interestingness judgments aligned between humans and LLMs?

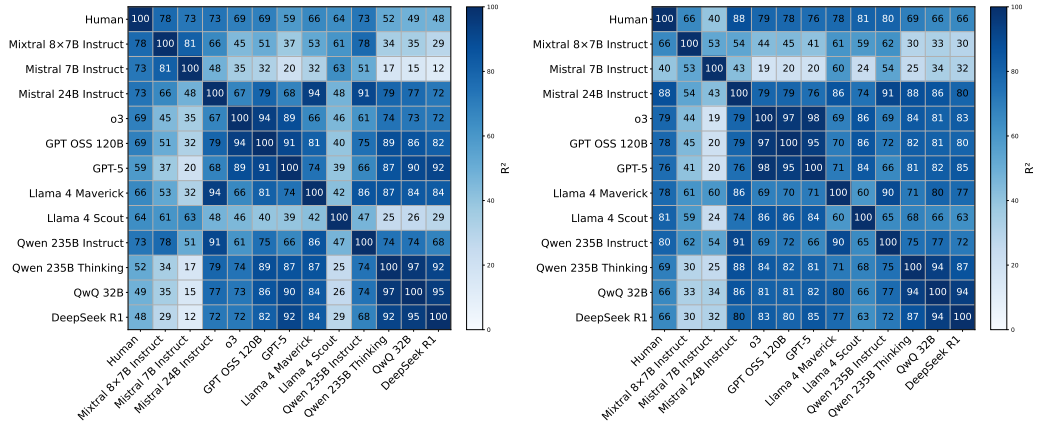


Figure 2: Agreement between humans' and LLMs' judgments about problem interestingness (left) and difficulty (right) on the Prolific dataset. Each cell shows the squared Pearson correlation (R^2 scaled by 100), between the row and column reasoners' per-problem mean ratings. The left matrix includes models from the same family together and 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 B.1.1. We see that different LLMs align with human judgments of interestingness and difficulty to different extents. While models from the Mistral family better approximate human judgments of interestingness, the other families excel at aligning with human judgments of difficulty.

RQ2b: If not aligned, how can we improve alignment of LLMs judgments of interestingness with humans?

RQ3: Do humans and LLMs align on *why* they find problems interesting?

RQ4: How much do LRMs reason when they judge problem interestingness?

RQ5: Are LLMs capable of generating math problems humans find interesting?

To take steps to address **RQ1-RQ4**, we study human interestingness 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. To address **RQ5**, we take initial steps to test whether LLMs can generate interesting problems by crowdsourcing human ratings for LLM-generated math problems. All studies received ethics approval by our institutional ethics review boards.

Prolific dataset: crowdsourcing human interestingness judgments. We recruit 63 participants from Prolific, a crowdsourcing platform common in cognitive science [42]. Our sample size aligns with standard practice for related human participant empirical studies [5]. Each participant was assigned to one of two conditions, rating 10 problems each. Each participant saw the same control problems and either a problem's original version or its variant (see below). Participants were required to think about each problem for at least one minute before rating its *interestingness* and *difficulty* on a scale of 0-100, and providing a one- to three-sentence rationale for each rating.

We curate problems from AMC 8 and AMC 12, high school-level contests given by the Mathematical Association of America [35]. Our goal is not to construct a large-scale benchmark, but to create a controlled stimulus set for the human-model study. We therefore deliberately select a set of contest problems and design hand-written variants of base problems to systematically probe dimensions of a problem to increase problem set diversity. For each contest problem, we hand-write a new problem based on a variant type. Examples of variants include increasing/decreasing the sizes of the numbers in the problem, adding/removing steps, etc. We also create two control problems: a negative and a positive control, which are later used to filter out any unfaithful participant ratings. That is, we filter out participants that rate the negative control math problem ("What is 28 + 13?") as having interestingness > 90. The result is 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 E.

IMO data collection. We conducted a survey of interestingness judgments made by 48 participants at the 2024 IMO. As part of the pre-survey, participants answered questions about *factors of interestingness*, where participants marked how important they found dimensions of problems (e.g. "The problem statement is simple and elegant") to making a problem interesting. Our analysis for RQ3 utilizes this portion of the pre-survey, with additional details about the survey provided in Appendix F.

Comparing LLM and Human Judgments. We evaluate 12 language models from five 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 10 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).

Understanding whether LLMs can generate interesting problems. Finally, we conduct an initial exploration into whether LLMs can generate competition math problems that humans find interesting. We first generate 90 math problems using three models, Mistral 7B Instruct, Qwen 235B Thinking, and OpenAI o3. The LLMs are prompted specifically to generate problems that are interesting high school-level competition math problems. We then filter out invalid problems manually. Problem validity is manually determined (e.g., problems with inconsistent information, no correct answer, etc. are discarded). The set of 90 problems is split into 5 batches. 30 total participants evaluate problems from two of these batches, resulting in 360 judgments we evaluate.

4 Results

RQ1: How do LLMs' judgments of interestingness align with those of humans? For each LLM, we compute the R^2 between per-problem mean interestingness in humans and the model (see Figure 2). 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, which is our noise ceiling on explainable variance, is 0.72 with 95% CI: [0.53, 0.87] (see Appendix B.1.5). This indicates that current LLMs (especially those in the Mistral family) are able to approximate human perceptions of interestingness with surprising fidelity. We include power analyses to support our claims in Appendix B.1.6.

For ratings of difficulty, we see the opposite: most models within the Mistral family struggle to align with human judgments of problem difficulty, with other families aligning more closely with humans. We also see that the Deepseek and Qwen reasoning models struggle in aligning with both the human judgments of interestingness and difficulty.

Beyond human-LLM agreement on interestingness and difficulty, respectively, we next test whether any one agent (human or model) is correlated with respect to its own ratings of interestingness and difficulty. That is, is an agent's judgment of difficulty related to its judgment of interestingness? In humans, interestingness and difficulty are not highly correlated. We define the per-participant interestingness/difficulty correlation as the R^2 between a participant's interestingness and difficulty scores over the problem set, and find a mean correlation of 0.47 (95% CI: [0.39, 0.55], standard deviation 0.32). In reasoning models, the correlation is consistently at or above 0.9. Other LLMs range from 0.41 to 0.89, with only Llama Scout and Mistral 7B Instruct below 0.80 (Appendix B.1). Prior work has used proxies like difficulty [43] and usefulness [30] for interestingness; given the low human correlation between interestingness and difficulty on the same problems, we encourage future work to consider treating interestingness as its own dimension in discovery systems and understanding the factors that drive peoples' judgments, as we begin to explore later in RQ3.

> **Takeaway.** LLMs broadly align with peoples' ratings of interestingness and difficulty in math problems. However, people show moderate correlation between their *own* ratings of interestingness and difficulty for any given problem. LLM ratings for interestingness and difficulty for a problem, on the other hand, are typically highly correlated.

RQ2a: Are the *distributions* of interestingness judgments aligned between humans and LLMs? While LLMs capture some aspects of human interestingness judgments, in many instances, their *distributions* of interestingness ratings diverge. We measure distributional similarity between human and LLM judgments using the Wasserstein-1 Distance (WD) [51], based on the L1 norm. For each problem, we calculate the WD between the human distribution of judgments and the distribution

Table 1: Distributional alignment in interestingness judgments, as measured by Wasserstein distances (WD) between human and LLM distributions of interestingness judgments on the Prolific dataset. We use the Wasserstein-1 distance with the L1 norm, treating each individual rating with equal weighting. Lower values indicate closer alignment to human distributions. The human-human split-half baseline indicates the amount of explainable variability in the human data. LLMs are sampled at a temperature of 1.0. WD measures for other temperatures and for difficulty judgments are reported in Appendix B.1.3. We find that only one model, Mistral 7B Instruct, comes close to reflecting the distribution of human judgments observed in our study.

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]			

of one model's judgments. We compute the bootstrapped average WD score over our dataset of problems for each model. We report these values in Table 1.

For a human comparison baseline, we calculate the bootstrapped split-halves WD by repeatedly halving the human judgment set, converting each set into a distribution, and and calculating the WD between them. 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). We also demonstrate that our claim about distributional divergence is robust to our stimulus set, and include details in Appendix B.1.7.

Takeaway. LLM's interestingness judgments generally do not well-capture the diversity across the population of human judgments. This highlights the importance of carefully considering which models we integrate in AMD and human-AI collaborative systems.

RQ2b: How can we improve the alignment of LLMs' interestingness judgments to that of humans? We next explore whether multiple models, when pooled, can better capture the distribution of human judgments than any one model alone. We construct the full space of combinatorial subsets S of the 12 models we experiment with ($|S| \geq 2$). We calculate the WD for the human distribution of judgments against the distribution of judgments derived from each pooled set. We compute 95% CIs via 1000 bootstrapping iterations. In each iteration, we resample problems with replacement and draw a fresh subsample of 30 model ratings from the pooled distribution (i.e., if there are M models in a pool, where each model produced N ratings for any one problem, we repeatedly sample 30 judgments with replacement from the set of M $\times$ N judgments). Thus, this reflects both problem-level sampling variability and uncertainty from the rating subsampling. We present the top- and bottom-3 distributionally aligned subsets in Table 2. The best-performing minimal subset is {Mistral-24B, Mistral-7B, Mixtral-8x7B, OpenAI o3}. Adding more models beyond this point tends to worsen alignment, likely because intra-family similarities reduce rather than increase judgment diversity [29]. We also conduct 5-fold cross-validation over problems, selecting the best subset of k models on the training folds and evaluate WD on the held-out fold. While the set of models that achieves the lowest mean held-out WD changes, the WD estimate stays approximately the same (WD = 9.13 [7.65, 10.61]).

Takeaway. Pooled sets of LLMs can generally better reflect the human distribution of judgments compared to individual LLMs. This discrepancy highlights the importance for future work to explore AI-mathematics systems that rely on multiple LLMs, since most such work currently only relies on one LLM [43, 16, 9].

RQ3: Do humans and LLMs align on *why* they 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. We focus on pre-survey questions which IMO

Table 2: Wasserstein Distance (WD) between the human and pooled model judgments, averaged over our dataset (lower is better). The human-human WD represents a split-half baseline calculated by repeatedly halving the human judgments, calculating the WD between the resulting distributions, and bootstrapping. The top 3 and bottom 3 pooled LLM combinations ranked by pooled per-problem Wasserstein distance to human rating distributions are reported here. We see that pooling models considerably improves their alignment with the human distribution of judgments.

Pooled model set	WD ↓	95% CI
Most aligned pooled sets (lowest WD)		
{Mistral-24B, Mistral-7B, Mixtral-8x7B, o3}	9.07	[8.34, 10.9]
{Mistral-7B, o3, Qwen Instruct}	9.27	[8.30, 10.8]
{Mistral-7B, Mixtral-8x7B, o3, Qwen Instruct }	9.28	[8.26, 11.1]
Human baseline		
Human–Human (split-half baseline)	9.5	[7.8, 11.5]
Least aligned pooled sets (highest WD)		
{GPT-OSS-120B, o3}	19.3	[16.6, 22.3]
{GPT-5, GPT OSS 120B, o3}	19.6	[16.5, 22.5]
{GPT-5, o3}	20.6	[17.8, 24.0]

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 spanned the allowed range of importance options from not important to very important. 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 comparisons between the human and model distributions for all models we examine in Appendix B.2. Only Mistral 7B Instruct (Appendix Figure 20) and Mistral 24B Instruct (Appendix Figure 21) reflect the human distributions of interestingness rationales well. While this section focused on the IMO survey, Appendix B.1.8 includes rationale misalignment results for our Prolific study as well. Future work can better understand the drivers for such differences across model families.

Takeaway. People and LLMs are generally misaligned on *why* they find problems interesting. This encourages future work towards a better understanding of the drivers behind this divergence.

RQ4: How much do LRMs reason when they judge problem interestingness?

We examine the resource usage of LRMs when reasoning about a problem via the number of reasoning tokens used (i.e., the length of their reasoning chain). 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). In Figure 3, we see that 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, 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.

Takeaway. There is a correlation between how long LRMs think for and how they rate problem interestingness for the Prolific problems in our dataset. However, the relation between thinking time and interestingness breaks down for more challenging IMO-level problems.

RQ5: Are LLMs capable of generating math problems humans find interesting? So far, we explored whether LLMs' interestingness judgments align with those of humans. But ultimately, in human-AI collaborative applications, it matters whether LLMs are also capable of *generating*

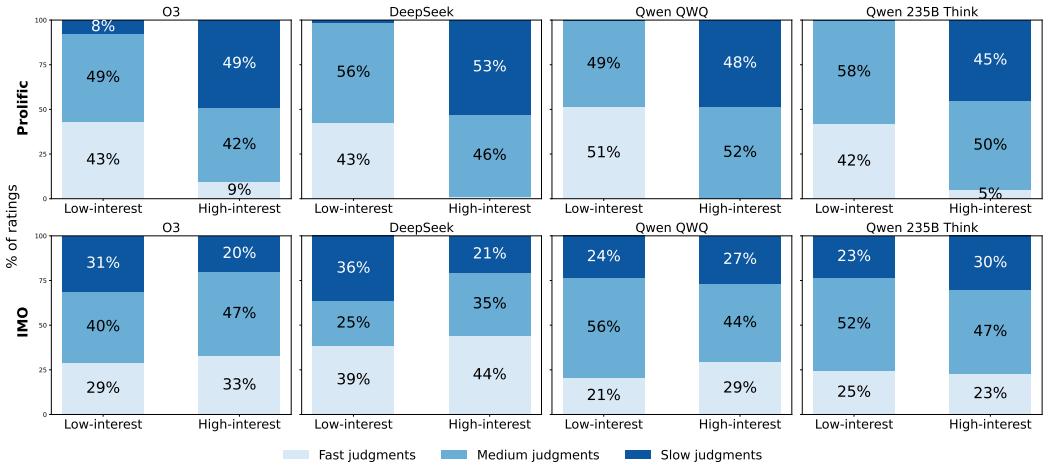


Figure 3: Judgment speed distributions across LRMs on low- vs. high-interest Prolific problems. A low-interest problem for a model is one that is given a below median interestingness score by the model. High-interest problems are those which are given higher than the median interestingness score. For 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. We include the judgment speed distributions across LRMs for difficulty judgments in Appendix B.1.

interesting problems; this could either be to pique students' interest as an AI tutor or to complement the problems a researcher thinks about as an AI co-scientist [33, 44]. As an initial case study, we begin to explore whether LLMs are capable of generating competition math problems that humans find interesting.

First, we sample 30 "interesting, high-school level competition math problems" from the 12 models in our experiments. To understand how LLM-written problems compare to those written by humans, we embed the 360 generated problems alongside all AMC problems (which, to the best of our knowledge, are human-authored) and the problems from our Prolific study (which are taken or adapted from the AMC). All problems are embedded with tSNE [49] (see details in Appendix section G.1). We see in Figure 4 that human- and LLM-written problems tend to cluster in different regions of the semantic space, highlighting the need for a better understanding of the gaps between the problems humans vs. LLMs write.

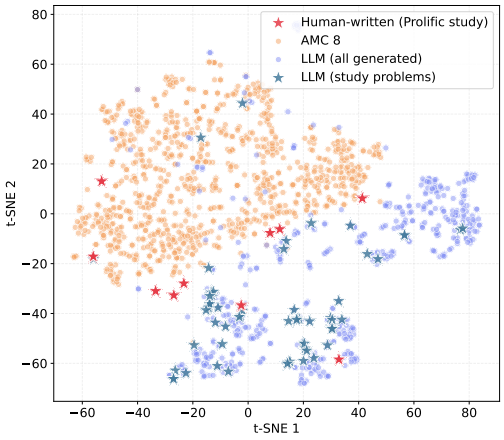


Figure 4: A t-SNE visualization of human- vs. LLM-written math problem embeddings. The LLM-generated problems from our study (purple circles) are plotted against all AMC problems (orange circles). Prolific problems (red stars), taken or adapted from the AMC, and LLM-generated problems we collect human judgments for (blue stars) are also included. We see that human- and LLM-written problems cluster separately in the semantic space, suggesting systematic differences between the two.

But do people find these generated problems interesting? We take initial steps towards this question via a pilot problem evaluation study with a new group of Prolific participants. The study follows our previous experimental design (to collect the judgments in RQ1-2), wherein participants now judge LLM-generated problems. Participants are only shown the problems and do not know the problems are LLMs generated to avoid bias. We choose LLMs along the range of human-alignment (as measured by WD in RQ2a): we use Mistral 7B instruct (most human aligned), Qwen 235B Thinking (somewhat human aligned), and o3 (least hu-

man aligned) as problem generators. This also allows us to assess whether models that are more aligned to people in their evaluations of problem interestingness are also better generators, or vice versa. From these 90 problems, we manually filter invalid problems (e.g., problems with inconsistent information, no correct answer, etc.). Additional information about the filtering process and ablations addressing automated filtering are provided in Appendix G.

We find substantial variation in human ratings of how interesting the LLM-generated problems are. In Appendix table 8, we present the top- and bottom-3 problems and their generator models from both batches we run. We see no significant difference in human ratings for problems across generators. Mistral-generated problems receive a mean rating (with 95% confidence intervals) of 59.36 [52.35, 68.09], while o3 and Qwen 235B Thinking receive scores of 62.46 [56.76, 69.00] and 62.84 [52.25, 69.83]. We also observe that no individual participant's top 3 problems originate from the same generator model. These results set up future work to explore the role of multiple problem generators, diverging from the current norm of using one base model in AI tutor and AI research systems [36, 33].

Since human-aligned LLM judges are not also necessarily better generators (potentially due to distribution shift in the problems we see in Figure 4), we next explore whether it is possible to build systems where an LLM *both generates and selects* interesting problems to present to humans. To begin to assess, we compare whether a model's ability to align with human interestingness on the human-written problems also transfers to LLM-written ones. We find no significant relationship in either direction ($r = -0.450, R^2 = 0.203, p = 0.165$) and include the plot in Appendix Figure 32. Even the LLM that is most aligned with human judgments on LLM-generated problems (GPT-OSS-120B), achieves a Pearson r correlation with human ratings of 0.336 ($p = 0.109$), lagging vastly behind the correlations we see between humans and LLMs for human-written (Prolific) problems. A possible reason for this lack of relationship could be the divergence in the types of problems written by humans vs. LLMs, as we see in Figure 4. This highlights an area where much additional future work is necessary. If the goal is to build AI systems that can independently teach students or assist mathematicians, they will need to be capable of both generating problems, and judging which ones are best to suggest to the user. We encourage future work to explore such generator-selector systems end-to-end.

Takeaway. LLMs, after filtering for problem validity, are capable of generating interesting problems. However, models that are aligned evaluators of interestingness on human-written problems are not necessarily good problem generators or evaluators of interestingness on LLM-generated problems, raising new open work. Because no singular model dominated participants' favorite problems, incorporating multiple LLMs in AMD and human-AI collaboration systems should be explored.

5 Conclusion and Limitations

In this work, we rigorously assess how judgments of interestingness compare between humans and LLMs, over two new datasets of people's evaluation about math problems. These datasets are collected over a varying difficulty of problems (from AMC- and IMO-level competition math), and from participants with a varying level of expertise (crowdsourced participants and IMO competitors). While LLMs' interestingness judgments generally correlate with the average judgment of people, they do not necessarily match the *distribution* of human judgments. We find that model pooling can better align LLMs with humans and propose future directions in this space. We also find that resource usage in LRMs during problem judgment may be correlated with problem interestingness, but only for less difficult competition problems, highlighting thinking patterns that can serve as proxies for interestingness in the future (and raising additional questions about resource usage in models' reasoning). As an initial exploration, we also evaluate whether humans find LLM-generated math problems engaging.

However, our study is just a first step. Future work should expand on the number and type of math problems, possibly in collaboration with educators and mathematicians. 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.

More broadly, our work raises many important questions, in particular, whether models should be aligned to the variability of human responses, which humans should those responses align to, and at what level of mathematical experience? If models are used in human-facing applications, for example, as mathematical AI thought partners [10, 19] designing curricula for students, then we may want to build interestingness measures that meaningfully correlate to human curriculum and the level of the learner [38, 46]. If instead models work alongside research mathematicians or autonomously discover new mathematics—and decide what mathematics problems to pursue at all—we may set a higher standard for the interestingness judgments [26]. Overall, we hope our work motivates and informs future work on 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, Tom Griffiths, Phoebe Zeng, Gizem Çaylak, Xaver 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 an NSF SBE SPRF, 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] Tudor Achim, Alex Best, Alberto Bietti, Kevin Der, Mathïs Fédérico, Sergei Gukov, Daniel Halpern-Leistner, Kirsten Henningsgard, Yury Kudryashov, Alexander Meiburg, Martin Michelsen, Riley Patterson, Eric Rodriguez, Laura Scharff, Vikram Shanker, Vladmir Sicca, Hari Sowrirajan, Aidan Swope, Matyas Tamas, Vlad Tenev, Jonathan Thomm, Harold Williams, and Lawrence Wu. Aristotle: Imo-level automated theorem proving, 2025.

[2] Luke Bailey, Kaiyue Wen, Kefan Dong, Tatsunori Hashimoto, and Tengyu Ma. Scaling self-play with self-guidance. *arXiv preprint arXiv:2604.20209*, 2026.

[3] 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.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[5] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA, 2016. Association for Computing Machinery.

[6] Junyi Chu, Kristine Zheng, and Judith E Fan. What makes people think a puzzle is fun to solve? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[7] 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.

[8] Katherine Collins, Cedegao Zhang, Graham Todd, Lance Ying, Mauricio da Costa, Ryan Liu, Prafull Sharma, Adrian Weller, Ionatan Kuperwajs, Lio Wong, et al. Evaluating language models' evaluations of games. In *International Conference on Learning Representations*, volume 2026, pages 157702–157740, 2026.

[9] Katherine M. Collins, Albert Q. Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller, and Mateja Jamnik. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024.

[10] 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.

[11] Katherine M Collins, Cedegao E Zhang, Lionel Wong, Mauricio Barba da Costa, Graham Todd, Adrian Weller, Samuel J Cheyette, Thomas L Griffiths, and Joshua B Tenenbaum. People use fast and flat simulation to reason about new games. *Nature*, 655(8123):598–607, 2026.

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

[13] 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.

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

[15] Hy Day. Role of specific curiosity in school achievement. *Journal of Educational Psychology*, 59:37–43, 02 1968.

[16] Kefan Dong and Tengyu Ma. Stp: Self-play llm theorem provers with iterative conjecturing and proving, 2025.

[17] 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.

[18] 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.

[19] 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.

[20] Simon Frieder, Sam Bealing, Arsenii Nikolaiev, Geoff C. Smith, Kevin Buzzard, Timothy Gowers, Peter J. Liu, Po-Shen Loh, Lester Mackey, Leonardo de Moura, Dan Roberts, D. Sculley, Terence Tao, David Balduzzi, Simon Coyle, Alex Gerko, Ryan Holbrook, Addison Howard, and XTX Markets. Ai mathematical olympiad - progress prize 2. `https://kaggle.com/competitions/ai-mathematical-olympiad-progress-prize-2`, 2024. Kaggle.

[21] Jacob W Getzels. The problem of the problem. *New directions for methodology of social and behavioral science: Question framing and response consistency*, 11:37–49, 1982.

[22] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2025.

[23] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[24] Aryan Gulati, Brando Miranda, Eric Chen, Emily Xia, Kai Fronsdal, Bruno Dumont, Elyas Obbad, and Sanmi Koyejo. Putnam-axiom: A functional and static benchmark for measuring higher level mathematical reasoning in llms, 2025.

[25] 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.

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

[27] XTX Investments. Ai mathematical olympiad - progress prize 1. `https://kaggle.com/competitions/ai-mathematical-olympiad-prize`, 2024. Kaggle.

[28] Hongchao Jiang, Yiming Chen, Yushi Cao, Hung yi Lee, and Robby T. Tan. Codejudgebench: Benchmarking llm-as-a-judge for coding tasks, 2025.

[29] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025.

[30] Timothe Kasriel, Thomas Lu, Qinghua Ding, Jingxuan He, and Dawn Song. Usefulness-driven learning of formal mathematics. In *NeurIPS 2025 Workshop on MATH-AI*, Oct 2025. Poster.

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

[32] George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116(1):75–98, 1994.

[33] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024.

[34] 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.

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

[36] Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

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

[38] Shubhra Mishra, Gabriel Poesia, and Noah D. Goodman. From next-token to mathematics: The learning dynamics of mathematical reasoning in language models. In *Proceedings of the Conference on Language Modeling (COLM)*, Montreal, Canada, 2025. arXiv:2407.00900.

[39] Thomas Nickles. What is a problem that we may solve it? *Synthese*, pages 85–118, 1981.

[40] 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.

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

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

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

[44] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[45] Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks, 2025.

[46] Pavan Kalyan Tankala, Shubhra Mishra, Satya Lokam, and Navin Goyal. CurLL: A developmental framework to evaluate continual learning in language models. In Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams, editors, *Proceedings of the First BabyLM Workshop*, pages 256–278, Suzhou, China, November 2025. Association for Computational Linguistics.

[47] George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition, 2024.

[48] George Tsoukalas, Rahul Saha, Amitayush Thakur, Sabrina Reguyal, and Swarat Chaudhuri. Learning interestingness in automated mathematical theory formation, 2025.

[49] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

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

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

[52] Lionel Wong, Tracey Mills, Ionatan Kuperwajs, Katherine M Collins, and Tom Griffiths. Meta-reasoning: Deciding which game to play, which problem to solve, and when to quit. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025.

[53] Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

[54] Cedegao E Zhang, Katherine M Collins, Adrian Weller, and Joshua B Tenenbaum. Ai for mathematics: A cognitive science perspective. *arXiv preprint arXiv:2310.13021*, 2023.

[55] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Appendix

A Impact Statement **15**

B Additional Results **15**

B.1 Prolific Survey Results 15

B.1.1 Model Family Correlation 15

B.1.2 Interestingness/Difficulty Correlations 15

B.1.3 Wasserstein Distance Tables 15

B.1.4 LRM Judgment Lengths 16

B.1.5 Split-half R^2 between independent human raters. 17

B.1.6 Power Analyses 17

B.1.7 RQ2 Robustness Analyses 18

B.1.8 RQ3 Rationale Misalignment 18

B.2 IMO Survey Results 20

B.3 LLM Generated Problems Results 21

C Models **43**

D Compute Information **43**

E Methods: Prolific Study **43**

E.1 Participant Information 43

E.2 Variant types 43

E.3 List of problems 44

F Methods: IMO Study **45**

F.1 List of Interestingness and Uninterestingness Reasons 45

G Methods: Problem Generation Study **46**

G.1 t-SNE Visualization 46

H Methods: Prompts to Humans and LLMs **46**

H.1 Study Interface 46

H.2 Survey Text - Prolific Study 46

H.3 Survey Text - IMO Study 48

H.4 LLM Prompt - Prolific Study 51

H.5 LLM Prompt - IMO Study 51

H.6 LLM Prompt - Interestingness Reasons 53

A Impact Statement

The goal of this paper is to drive a better understanding of *interestingness* in mathematical problems and whether language models are capable of understanding how humans judge problem interestingness. AI systems are becoming an increasingly larger part of students' and researchers' interactions with mathematics. AI tutor chatbots decide which problems are shown to students, and AI co-scientists often suggest directions for exploration to mathematicians. If people increasingly rely on AI systems to guide problem choice, these systems may shift the landscape of research and education. It is also an open question whether the kinds of selections models endorse for what problem to solve then also influence people's own future evaluations of problems.

B Additional Results

B.1 Prolific Survey Results

We next include additional details on the Prolific crowdsourced participant survey and related model predictions.

B.1.1 Model Family Correlation

We include the R^2 heatmaps for temperatures 0.3 and 1.0 in Figures 5 and 6. We also include information about the correlation between interestingness and difficulty scores between LLM families in Table 3. All correlation plots for interestingness judgments are included in Figures 8 through 16. We sample at two temperatures and also plot the 95% confidence interval.

B.1.2 Interestingness/Difficulty Correlations

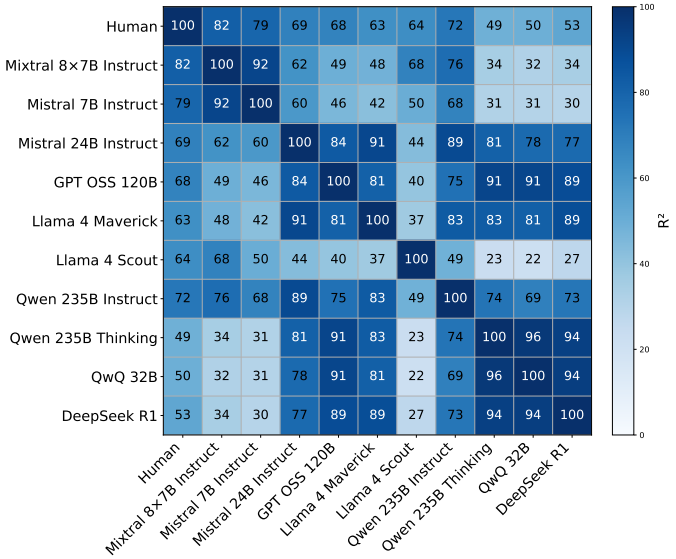
In Table 4, we report the per-problem correlation between interestingness and difficulty scores for each LLM.

B.1.3 Wasserstein Distance Tables

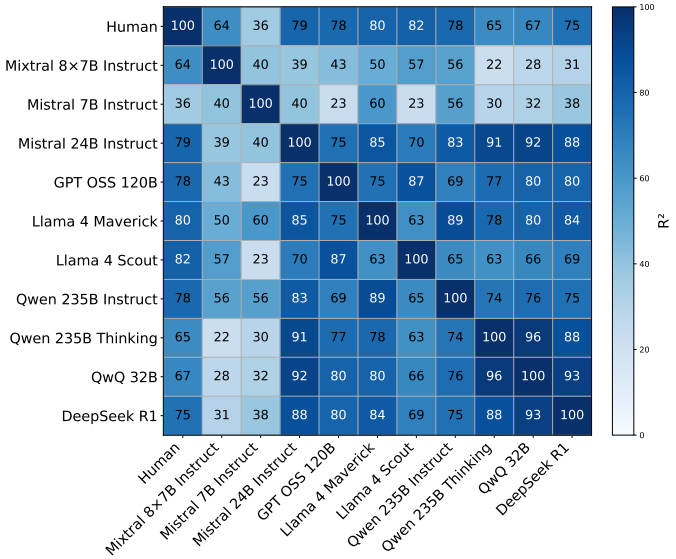
In Tables 5, 6, and 7, we report the Wasserstein distances between the LLM and human distributions for interestingness and difficulty ate temperatures 0.3 and 1.0.

Table 3: Correlation among each model family for interestingness and difficulty judgments. 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.

Family	**Model (tag)**	**Mean pairwise r (Interestingness) temp. 0.3, 1.0**	**Mean pairwise r (Difficulty) temp. 0.3, 1.0**
OpenAI	o3 (thinking) GPT-5 gpt-oss	-, 0.95	-, 0.98
Llama 4	Llama-4 Maverick Llama-4 Scout	0.61, 0.65	0.79, 0.77
Qwen	QwQ-32B (thinking) Qwen3-235B-A22B-Instruct Qwen3-235B-A22B-Thinking	0.89, 0.90	0.90, 0.90
Mistral	Mistral-7B-Instruct-v0.1 Mistral-Small-24B-Instruct Mixtral-8x7B-Instruct-v0.1	0.96, 0.90	0.63, 0.73
DeepSeek	DeepSeek-R1 (thinking)	-	-



(a) Interestingness (temp = 0.3)

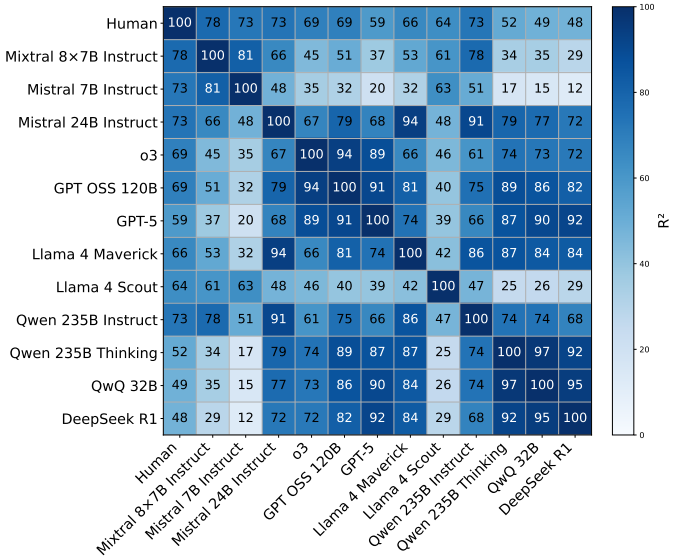


(b) Difficulty (temp = 0.3)

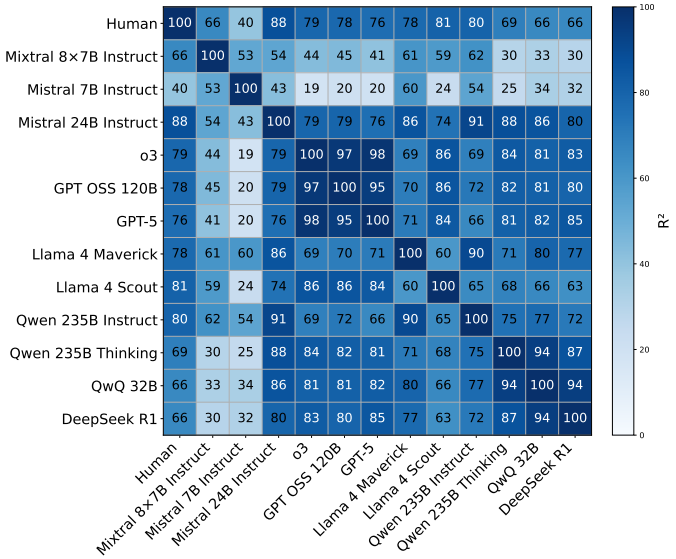
Figure 5: Agent–agent agreement on the **Prolific** dataset at temperature 0.30. Each cell shows the squared Pearson correlation (R^2) between the row and column agents' per-problem mean ratings. Darker cells indicate higher agreement; the diagonal is 1.00 by definition. The *Human* row/column gives model–human agreement. Top: interestingness ratings; bottom: difficulty ratings.

B.1.4 LRM Judgment Lengths

In Figure 7, 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-difficulty (which is judged by whether a problem's difficulty score is below or above the median of difficulty scores from judgments from that LRM).



(a) Interestingness (temp = 1.0)



(b) Difficulty (temp = 1.0)

Figure 6: Agent–agent agreement on the **Prolific** dataset at temperature 1.0. 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 *Human* row/column gives model–human agreement. (Top) interestingness ratings; (bottom)difficulty ratings.

B.1.5 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 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 $[0.53, 0.87]$.

B.1.6 Power Analyses

The unit of analysis for our human-model correlations is the problem ($n = 18$). At $n = 18$, the minimum detectable Pearson correlation is $r = 0.62 (R = 0.38)$ at 80% power and $r = 0.68$

Table 4: LLM interestingness–difficulty correlation (Pearson R^2) by model and temperature.

Model name	Temperature	R^2
Llama-Scout	0.30	0.58
	1.00	0.62
Llama-Maverick	0.30	0.84
	1.00	0.88
DeepSeek-R1	0.30	0.91
	1.00	0.90
o3	1.00	0.96
GPT-5	1.00	0.86
GPT-OSS-120B	0.30	0.84
	1.00	0.87
Mixtral-8x7B-Instruct	0.30	0.87
	1.00	0.89
Mistral-7B-Instruct	0.30	0.61
	1.00	0.41
Mistral-24B-Instruct	0.30	0.83
	1.00	0.87
QwQ-32B	0.30	0.93
	1.00	0.93
Qwen-235B-Instruct	0.30	0.84
	1.00	0.86
Qwen-235B-Thinking	0.30	0.92
	1.00	0.90

Table 5: Distributional alignment (WD) over **interestingness** judgments between human and LLM distributions of interestingness judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. The human-human split-half baseline indicates the amount of explainable variability in the human data. LLMs are sampled with a temperature of **0.3**.

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]

($R = 0.46$) at 90% power ($\alpha = .05$, two-sided). Our observed model-human agreements are $R = 0.48 - 0.78 (r = 0.69 \tilde{} 0.88)$, i.e., above the detection floor, so our interestingness evaluation experiments are adequately powered.

B.1.7 RQ2 Robustness Analyses

We run a leave-one-problem-out jackknife on the Table 1 ranking by removing each of the 18 problems in turn and re-computing every model's mean Wasserstein distance. We find that the ranking is highly stable. The best-matching model (Mistral-7B) remains rank-1 in all 18 folds, every model's mean WD moves by ≤ 1 point across folds, and each fold's ranking correlates with the full ranking at Spearman $\rho \geq 0.87$ (mean 0.98). Thus, we see that no single problem drives the distributional conclusions.

B.1.8 RQ3 Rationale Misalignment

In the Prolific experiment, participants were asked to provide free-text rationales for why they did or didn't find a problem interesting. Here, we explore whether humans and LLMs are misaligned

Table 6: Distributional alignment (WD) over **difficulty** judgments between human and LLM distributions of difficulty judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. The human-human split-half baseline indicates the amount of explainable variability in the human data. LLMs are sampled with a temperature of **0.3**.

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 7: Distributional alignment (WD) over **difficulty** judgments between human and LLM distributions of difficulty judgments on the Prolific dataset. Lower values indicate closer alignment to human distributions. The human-human split-half baseline indicates the amount of explainable variability in the human data. LLMs are sampled at a temperature of **1.0**.

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]

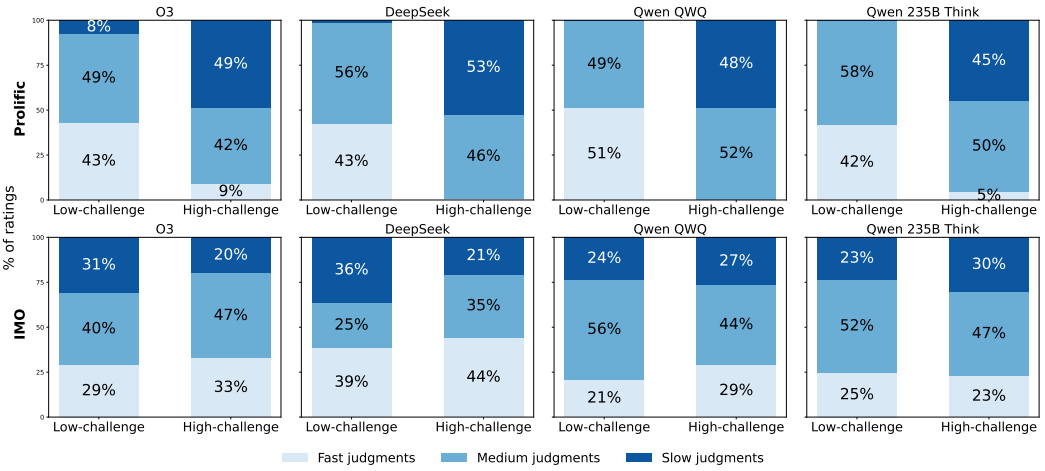


Figure 7: Judgment speed distributions across LRMs on low- vs. high-interest Prolific problems. A low-challenge problem for a model is one that is given a below median challenge score by the model. High-interest problems are those which are given higher than the median challenge score. For each subset 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 more difficult.

on their interestingness rationales for the Prolific experiment as well (as they were for the IMO experiment, as demonstrated in the main text in RQ3).

We conduct manual coding of 100 interestingness rationales from the Prolific study, which results in 60 labels of subquotes from a participant-given text rationale mapped to an interestingness reason.

Note that not every human rationale necessarily included something that could be flagged as an interestingness reason, and some participant rationales included multiple interestingness reasons. This free coding was used to come up with a codebook of 15 "interestingness rationales" that were grouped into 6 broader categories.

Then, Claude Sonnet 5 was used to classify the remaining rationales into one of the 15 interestingness reasons. The prompt to Claude includes the 15 interestingness reasons and a golden human-labeled example quote from the human participant rationales which were manually coded. Claude then classified the rest of the human rationales, and all the LLM-generated rationales. We validated this coding against the 60 manual labels, obtaining 75.9% agreement at the 6-category level.

We then used the LLM-labeled classifications and compared the distribution of reasons humans vs. models cite for problem interestingness. We use the total variation distance (TVD) to measure how closely the two distributions of reasons match, which is similar to Wasserstein distance used in RQ3, but for categorical distributions. We build a human-human noise ceiling by repeatedly splitting the human rationales into halves and computing the TVD between them. This yields a baseline TVD of $0.06 (95\%$ CI $[0.03, 0.11])$. We then compute the TVD between the human distribution and each model's distribution. We find that all 12 models diverge from humans in the reasons they cite. All models' TVD to humans ranges from 0.17 to 0.50, and we find that their confidence intervals do not overlap with those of humans either.

Below, we provide the codebook of LLM interestingness reasons we developed. The 15 fine-grained categories, grouped into the 6 broader buckets, are:

1. **Structure and aesthetics:**
   (a) Patterns: the problem exhibits or hints at an underlying numeric/structural pattern
   (b) Symmetry: notices symmetry, near-equal values, or a pleasing layout of the numbers
   (c) Simple/satisfying numbers: interest driven by clean, easy-to-compute values
   (d) Elegant solution: the solution simplifies to a neat result or reveals a hidden pattern
2. **Solving process:**
   (a) Mental simulation/brain games: puzzle-like; invites strategy or play rather than pure calculation
   (b) Clever trick/creative thinking: an unexpected trick, or creativity forced by the problem's constraints
   (c) Casework: interest from working through several individual cases
   (d) Decomposition: the problem must be broken into parts; not a one-step answer
3. **Difficulty and simulability:**
   (a) Challenging at the right level: engaging difficulty, neither trivial nor too complex
   (b) Mentally simulate the solution: one can reason toward the solution in one's head
4. **Novelty:**
   (a) Unusual/novel problem setup: an unusual construction or phrasing, different from the norm
5. **Content and scope:**
   (a) Certain class of problems: enjoys this kind of problem specifically (e.g. decimals, grids)
   (b) Addresses multiple concepts: combines several mathematical ideas at once
   (c) Real-world applications: interest tied to real-world applicability
6. **Ambiguity:**
   (a) Ambiguity: missing information is left open to interpretation

B.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 17, 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 18.

Reasons for interestingness across LLMs. In Figure 19, we show histograms of human participants importance ratings for various interestingness criteria. Figures 20 through 31 includes this for the LLMs we examine.

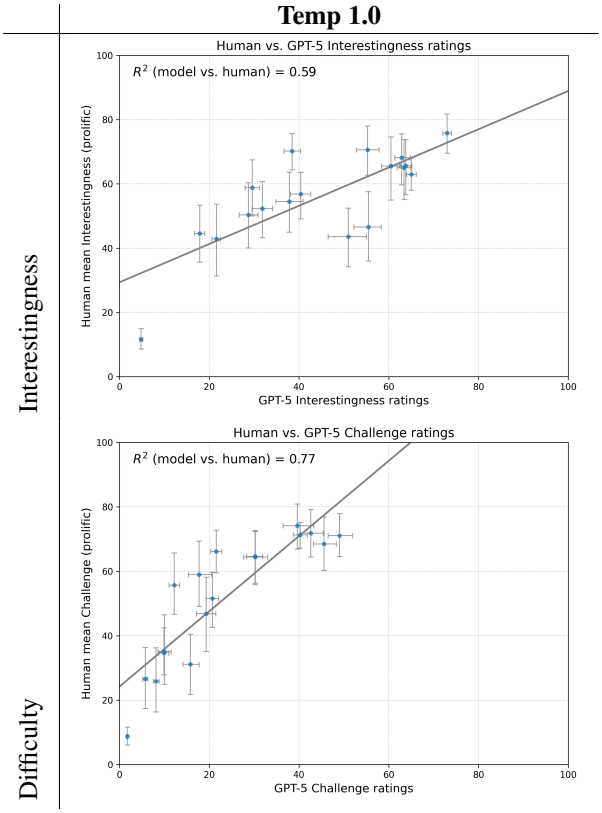


Figure 8: GPT-5: Human vs LLM ratings

B.3 LLM Generated Problems Results

In Figure 32, we plot the correlation between LLMs' performance on aligning with humans on the interestingness of human-written problems vs. on the interestingness of those generated by LLMs. In table 8, we include the top- and bottom-3 generated problems for the two batches of problems we run.

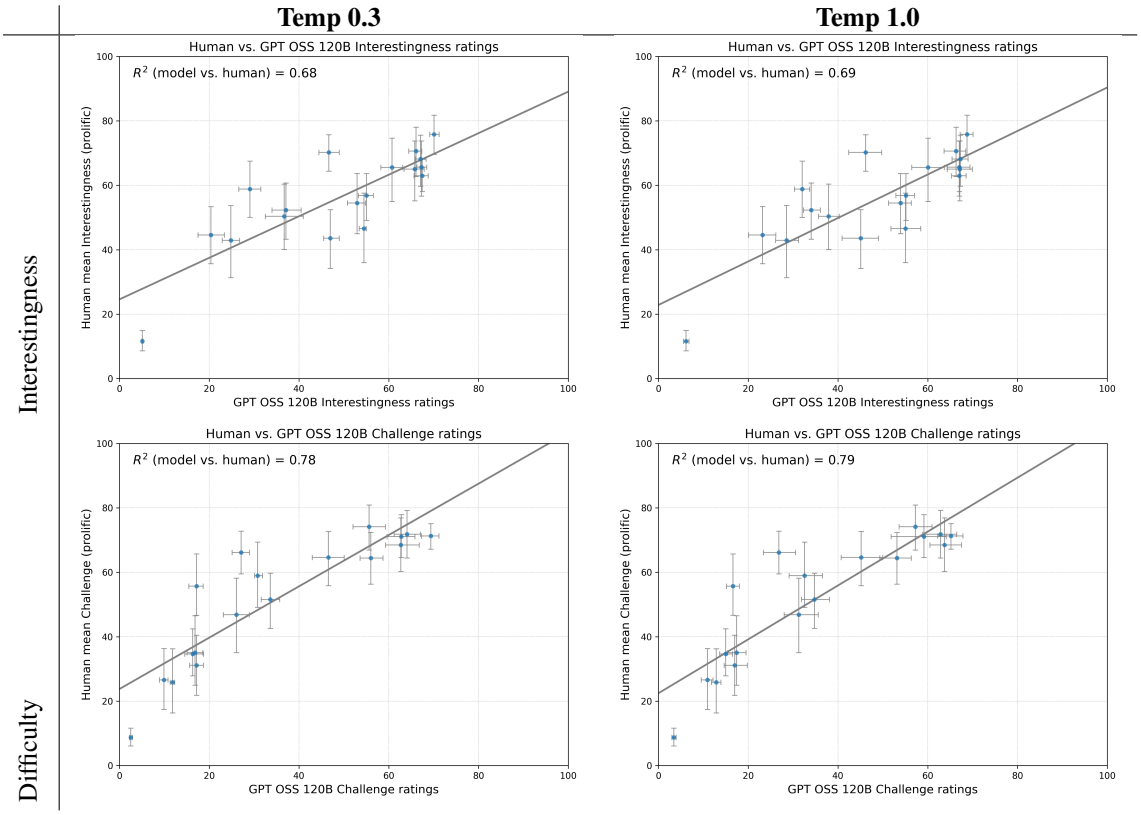


Figure 9: GPT-OSS-120B: Human vs LLM ratings

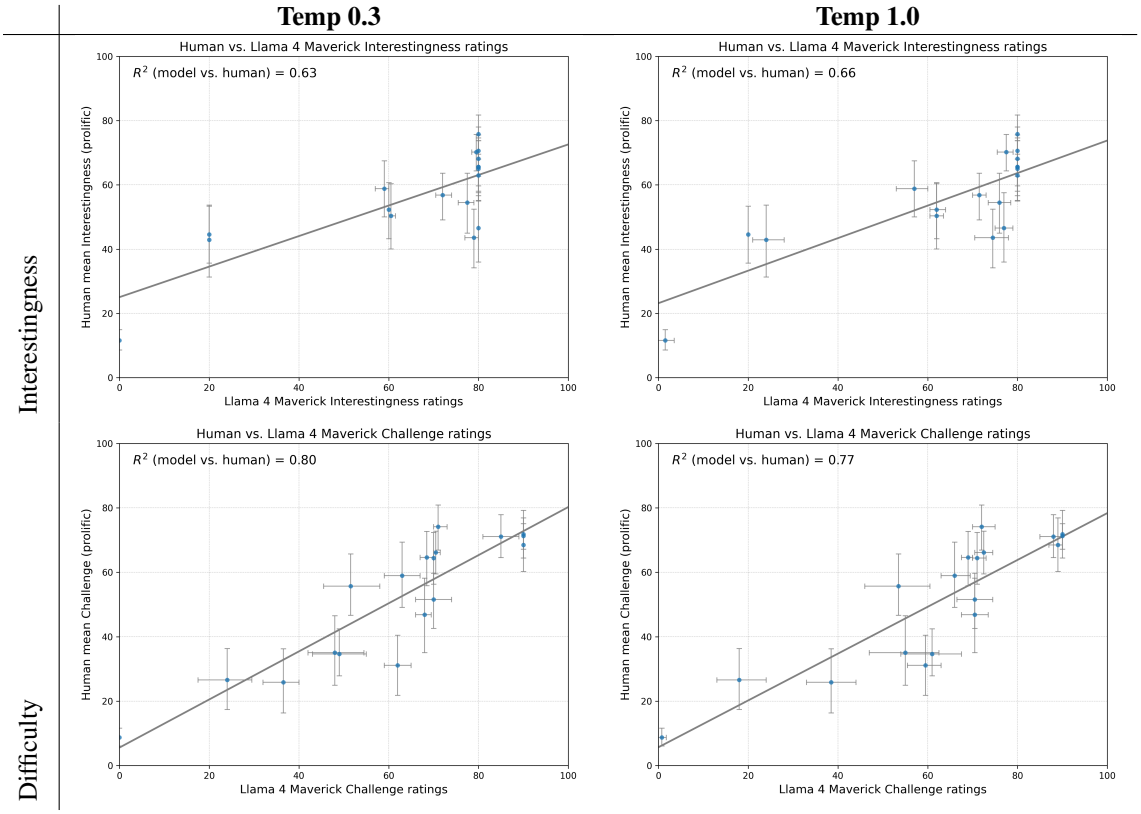


Figure 10: LLaMA-Maverick: Human vs LLM ratings

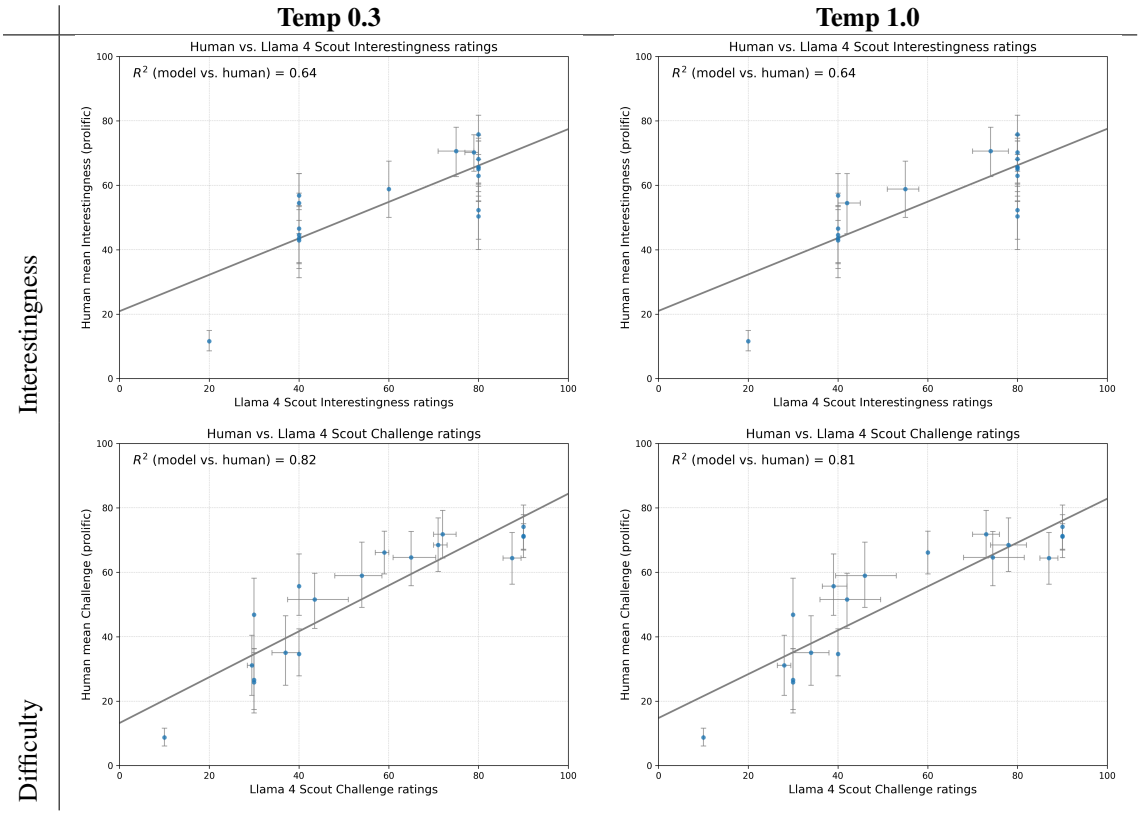


Figure 11: LLaMA-Scout: Human vs LLM ratings

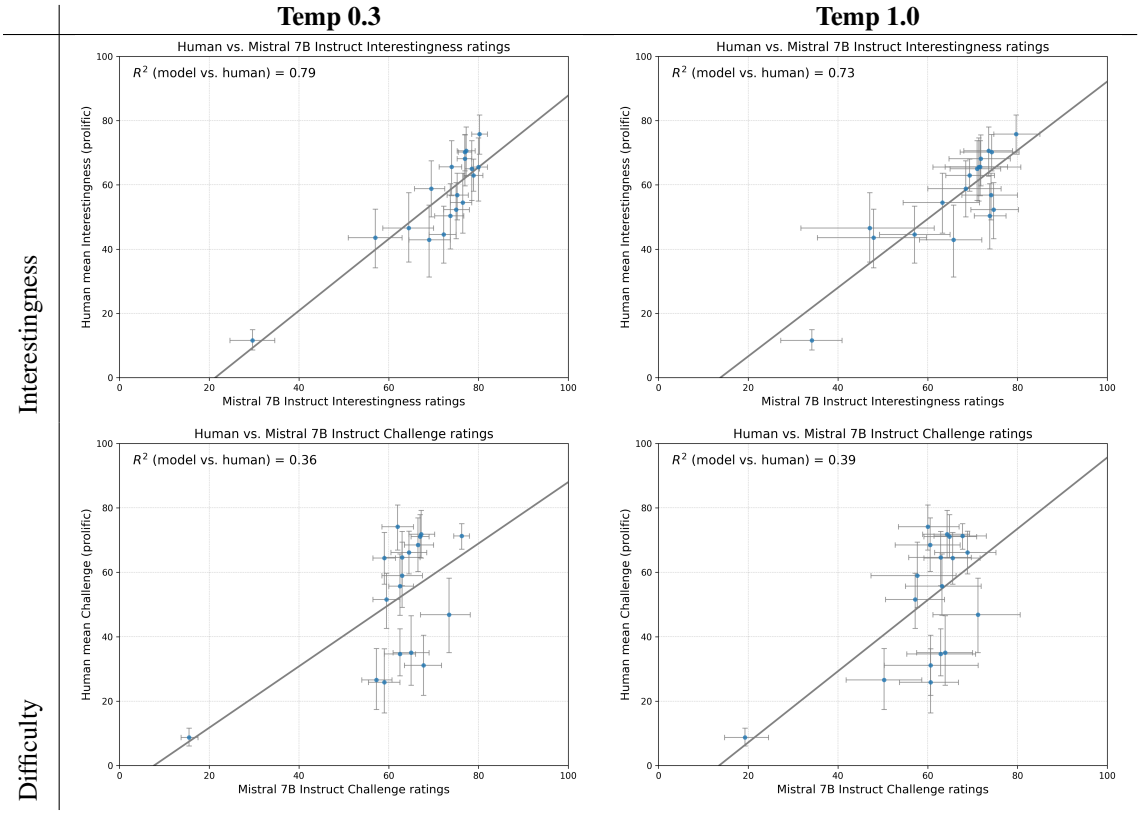


Figure 12: Mistral-7B-Instruct: Human vs LLM ratings

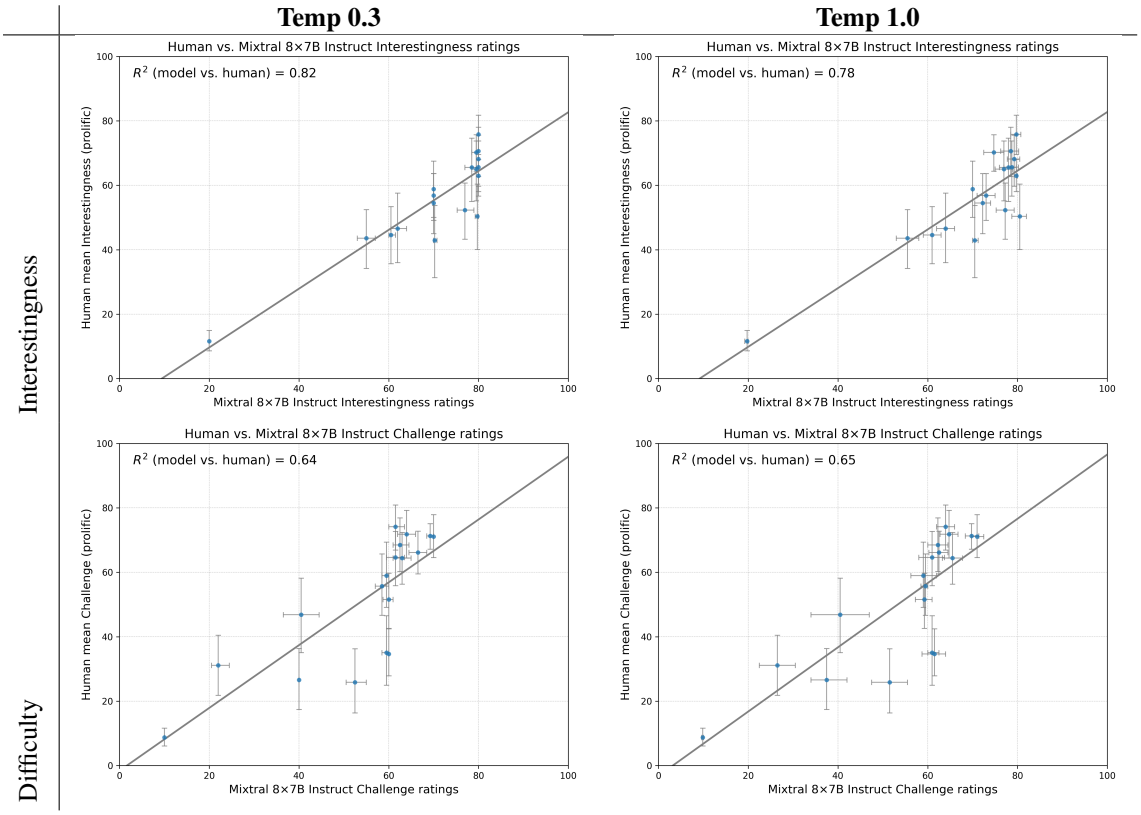


Figure 13: Mixtral-8x7B-Instruct: Human vs LLM ratings

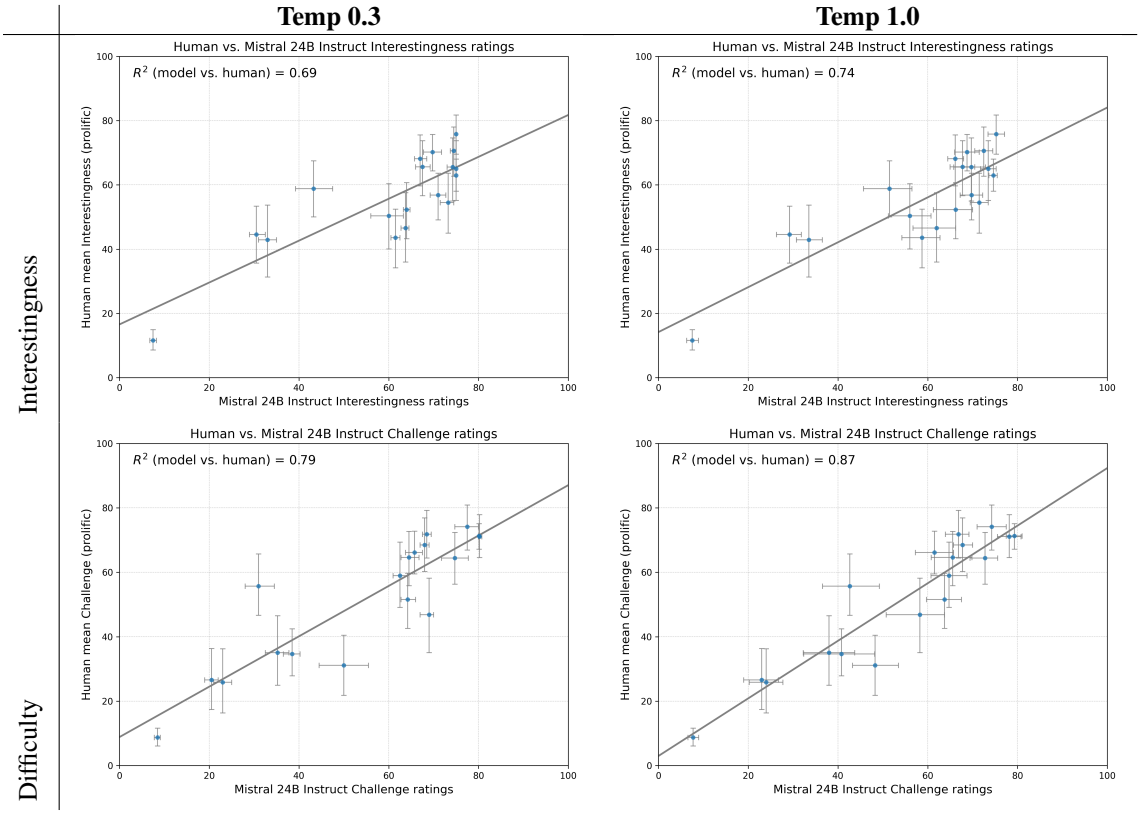


Figure 14: Mistral-24B-Instruct: Human vs LLM ratings

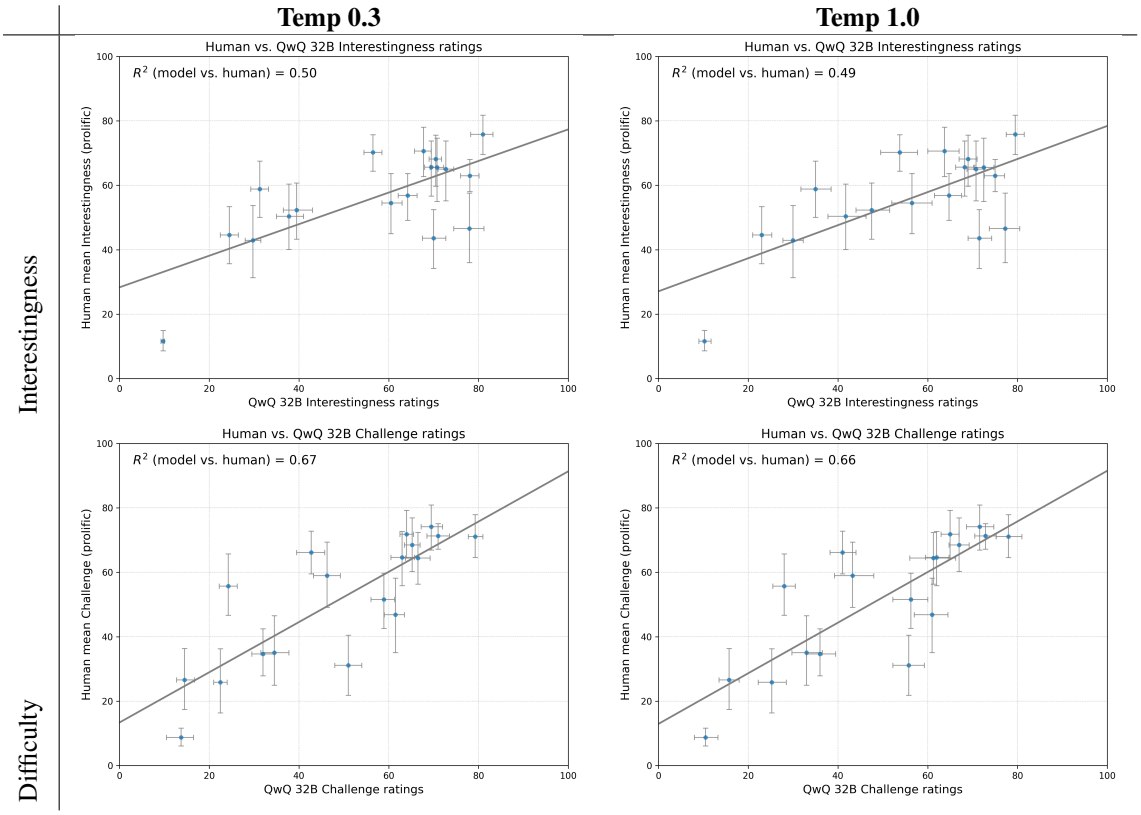


Figure 15: QwQ-32B: Human vs LLM ratings

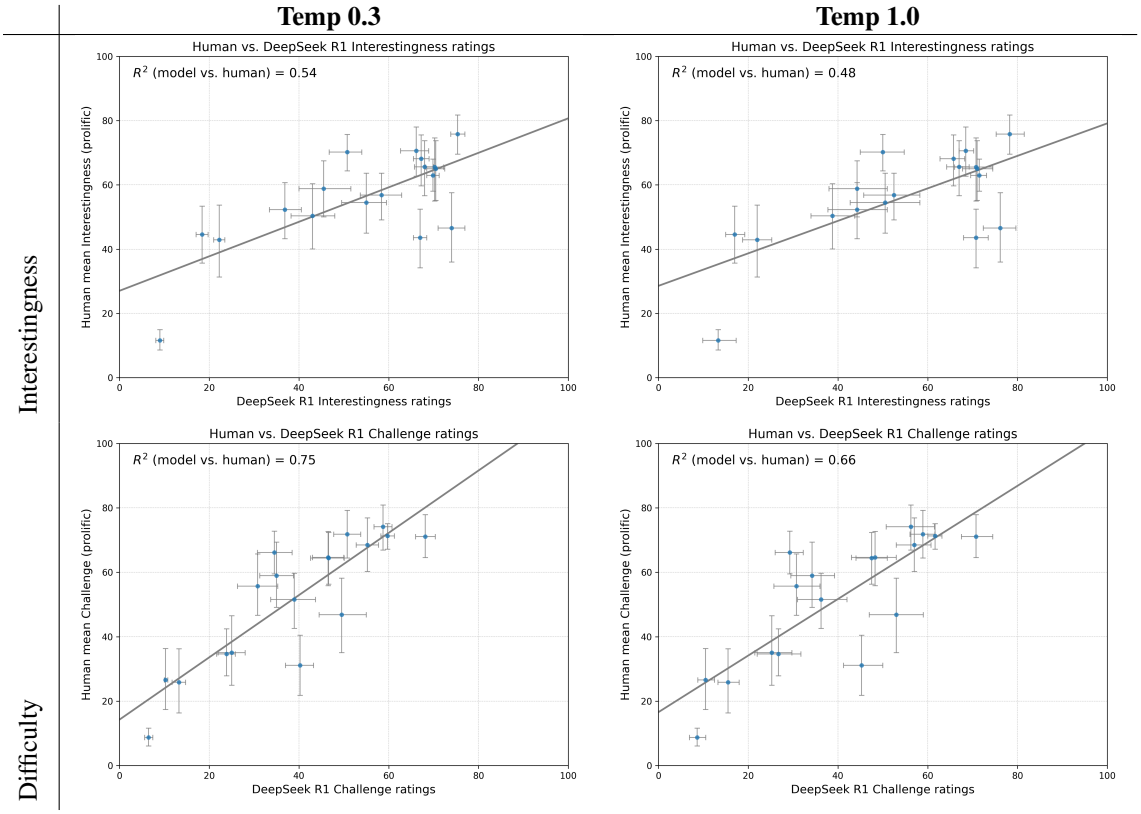


Figure 16: Deepseek R1: Human vs LLM ratings

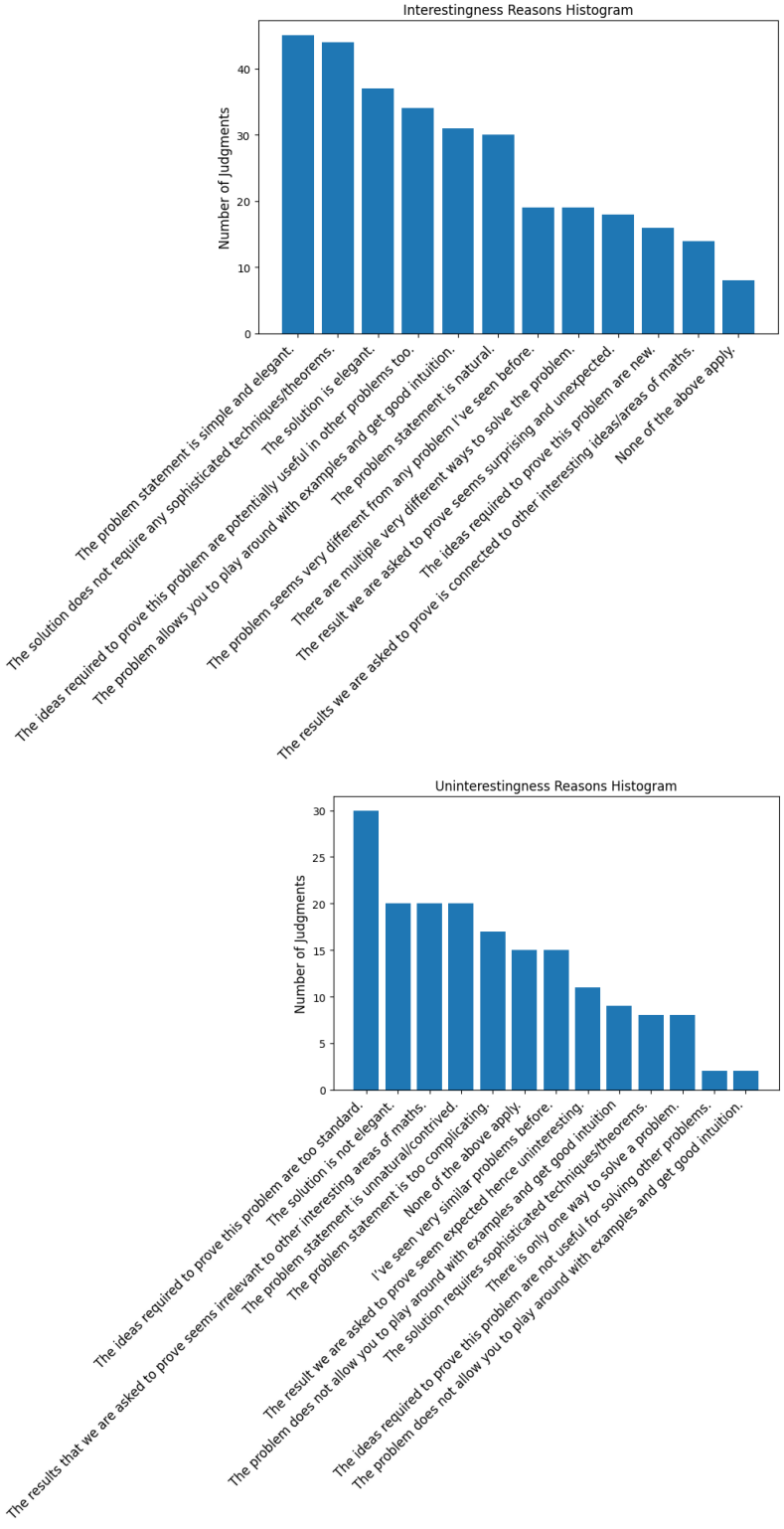


Figure 17: Histogram of how frequently interestingness (top) and uninterestingness (bottom) reasons were chosen by human participants across the survey.

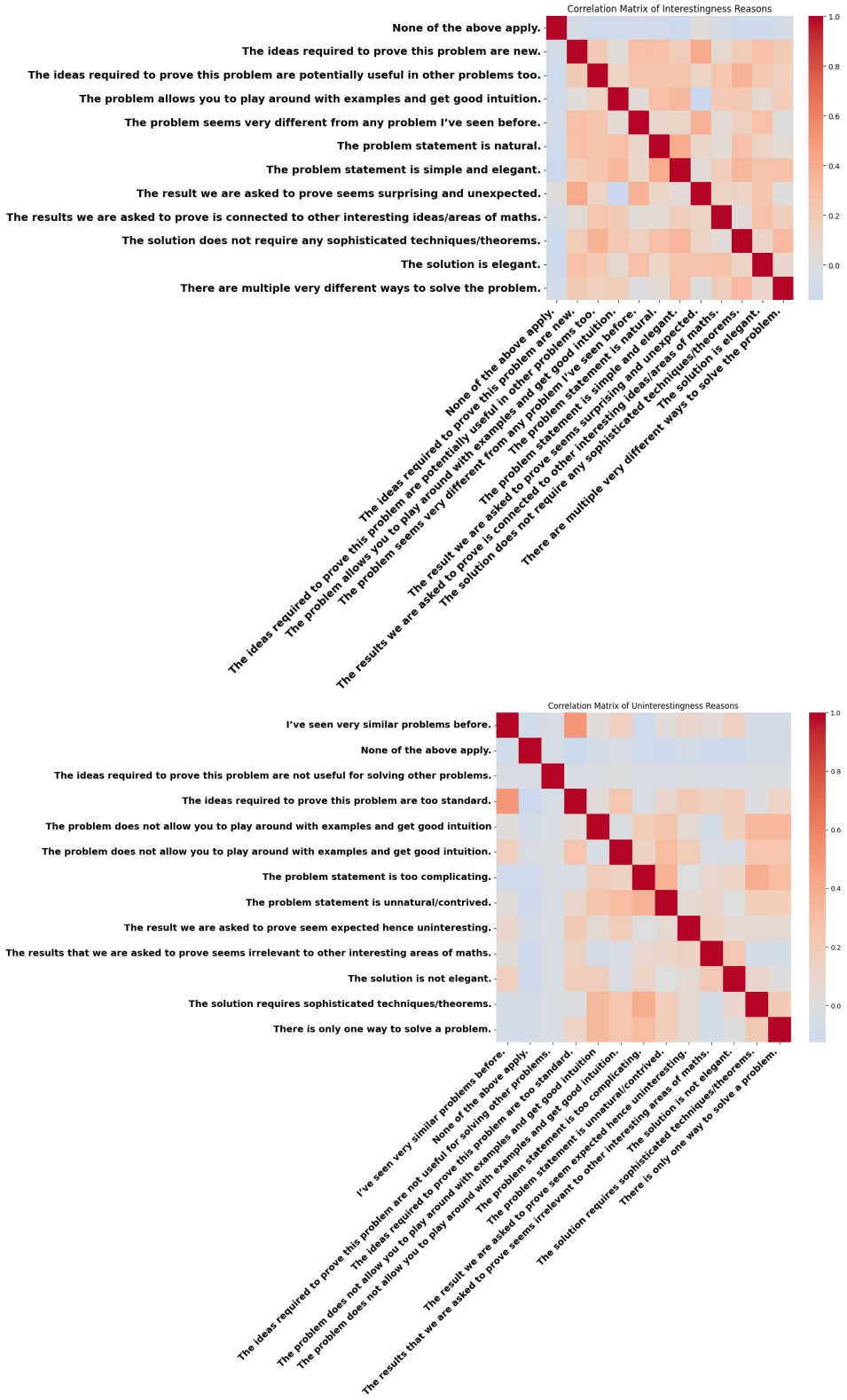


Figure 18: Correlation matrices showing when human participants in our IMO study chose multiple reasons for interestingness (top) or uninterestingness (bottom) for the same problem.

Table 8: Top- and bottom-rated generated problems by batch, according to human interestingness ratings. Scores are means with bootstrapped 95% confidence intervals.

Rank	Problem	Generator	Mean	95% CI
Batch 3: Top-rated problems				
1	Find the sum of all integer solutions to the equation $(x^2 - 5x + 5)^{(x^2-11x+30)} = 1$.	Qwen3-235B	77.8	[68.1, 86.9]
2	Suppose $d = -2w + 17$, $2d - 6w - 6 = 0$. Suppose $-4i = 2u + 2$, $-10 = -i + u$. What is the remainder when d is divided by i?	Mistral-7B	75.3	[61.3, 87.0]
3	How many 5-digit positive integers, whose five digits are all different and whose digits add up to 10, are divisible by 11?	o3	70.7	[54.1, 84.9]
Batch 3: Bottom-rated problems				
10	Let a_n be the nth positive integer in increasing order whose digits sum to 10. Find a_{15}.	Qwen3-235B	43.5	[29.6, 57.5]
9	An arithmetic sequence of positive integers has first term 3 and common difference d. Knowing that both the first term and the 10th term are divisors of 720, find the sum of all possible positive integers d.	o3	55.5	[40.0, 70.1]
8	Find the number of ordered pairs of positive integers (x, y) such that $\frac{1}{x} + \frac{1}{y} = \frac{1}{2023}$.	Qwen3-235B	56.3	[42.5, 70.5]
Batch 4: Top-rated problems				
1	A 4-digit number is called balanced if the sum of its digits in odd positions equals the sum of its digits in even positions. How many balanced 4-digit numbers exist?	Qwen3-235B	82.6	[74.1, 89.7]
2	During one complete 12-hour cycle, a digital clock displays every minute from 12:00 through 11:59. Removing the colon forms an integer, e.g., 7:05 becomes 705. How many of the 720 resulting integers are divisible by 3?	o3	78.3	[65.8, 88.3]
3	Find the positive integer k such that the equation $\left\|4x^2 - 20x\right\| = k$ has exactly three distinct real solutions.	Qwen3-235B	64.2	[50.1, 77.2]
Batch 4: Bottom-rated problems				
10	Let $n(y) = 12y - 109$. What are the prime factors of $n(-8)$?	Mistral-7B	49.5	[35.6, 64.7]
9	How many integers n between 1 and 2023, inclusive, make $n^3 + 2023$ divisible by $n + 2023$?	o3	52.0	[39.1, 65.9]
8	A sequence is defined such that the first term is 1, and each subsequent term is the smallest positive integer not already in the sequence for which the sum of the current term and the previous term is a perfect square. What is the fifth term of this sequence?	Qwen3-235B	53.3	[37.9, 67.4]

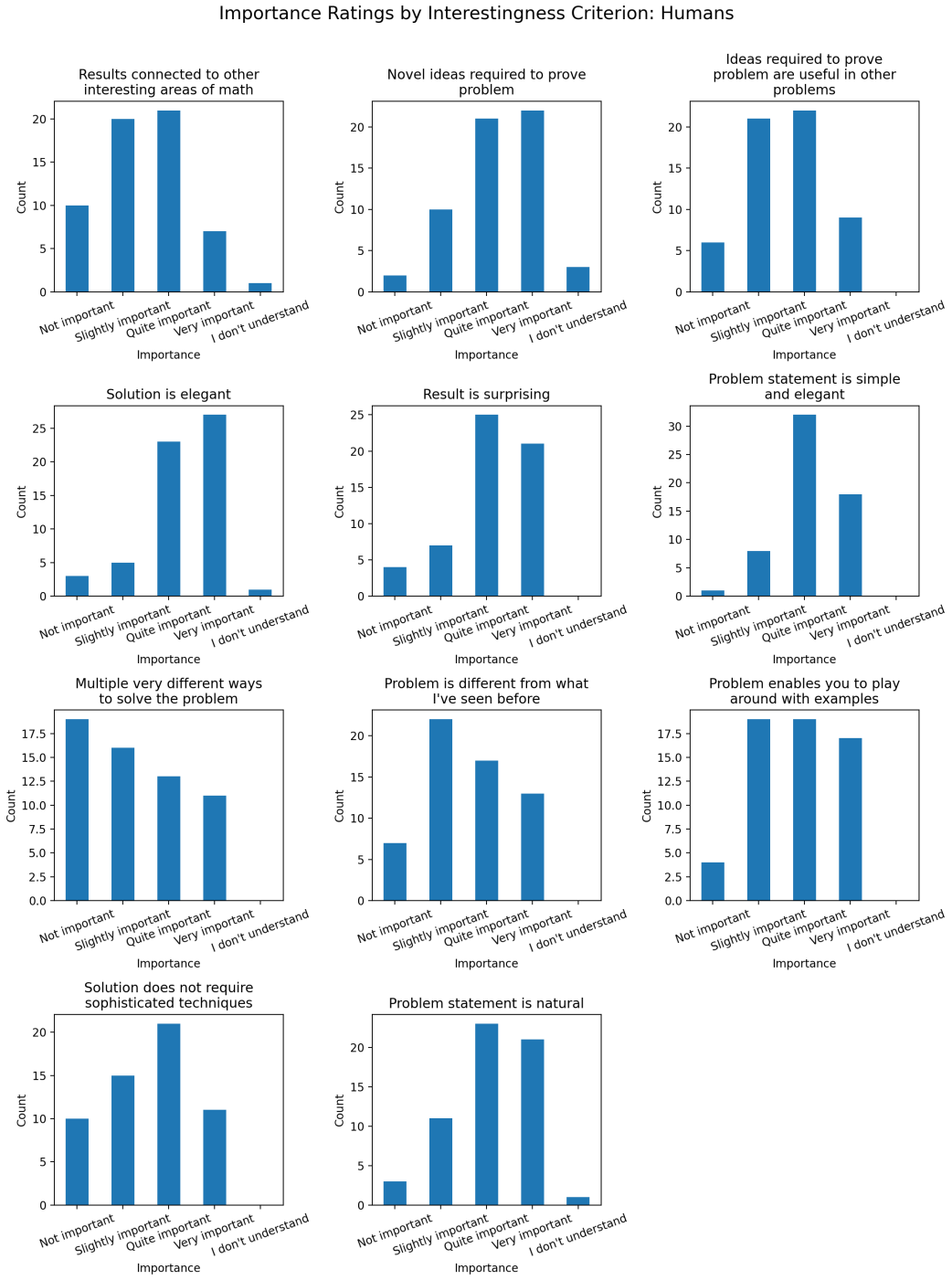


Figure 19: Distribution of importance ratings from human participants across 11 interestingness criteria.

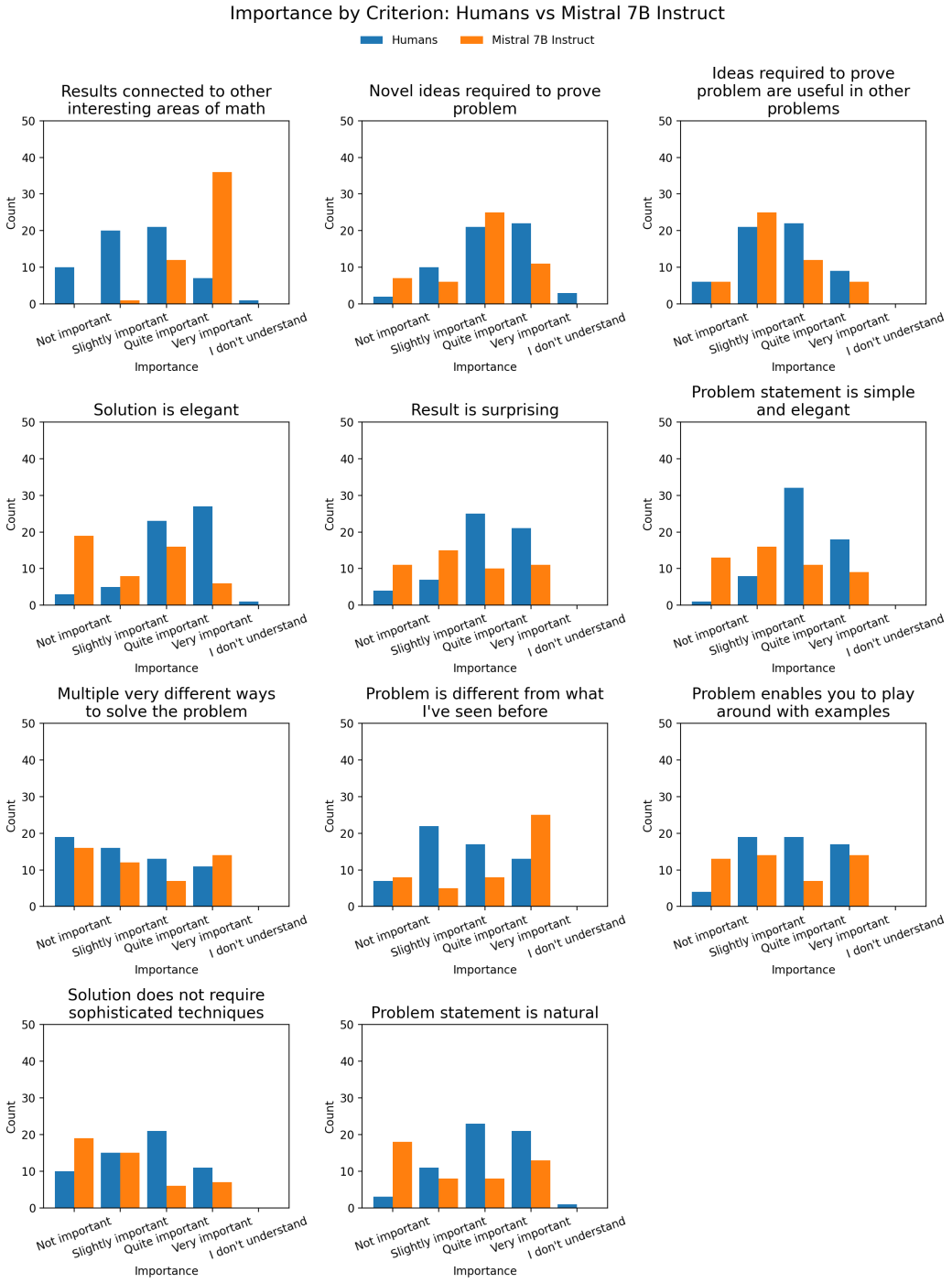


Figure 20: Importance ratings from Mistral 7B Instruct (orange) across 11 interestingness criteria compared to averaged human importance ratings (blue).

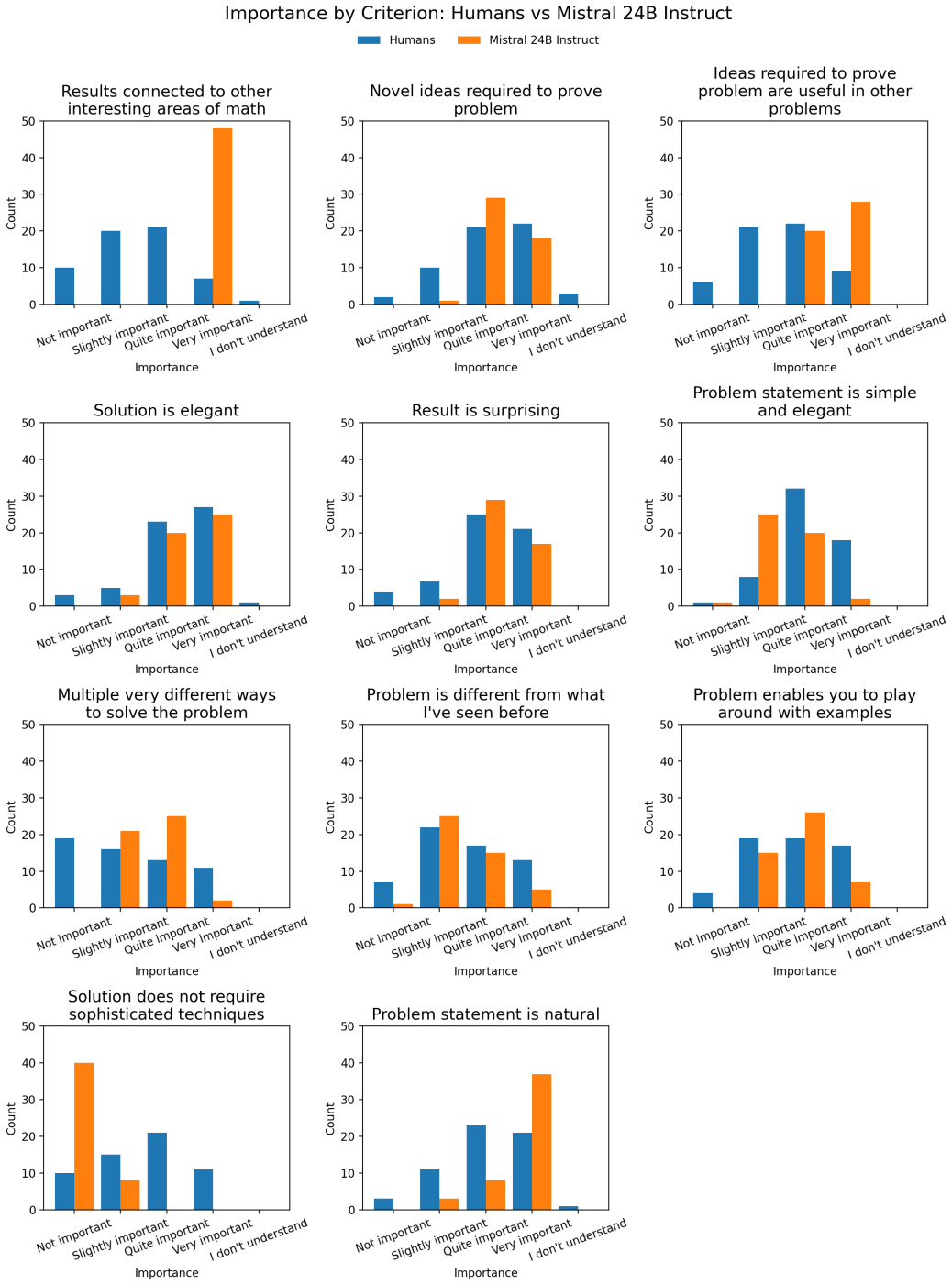


Figure 21: Importance ratings from **Mistral 24B Instruct**.

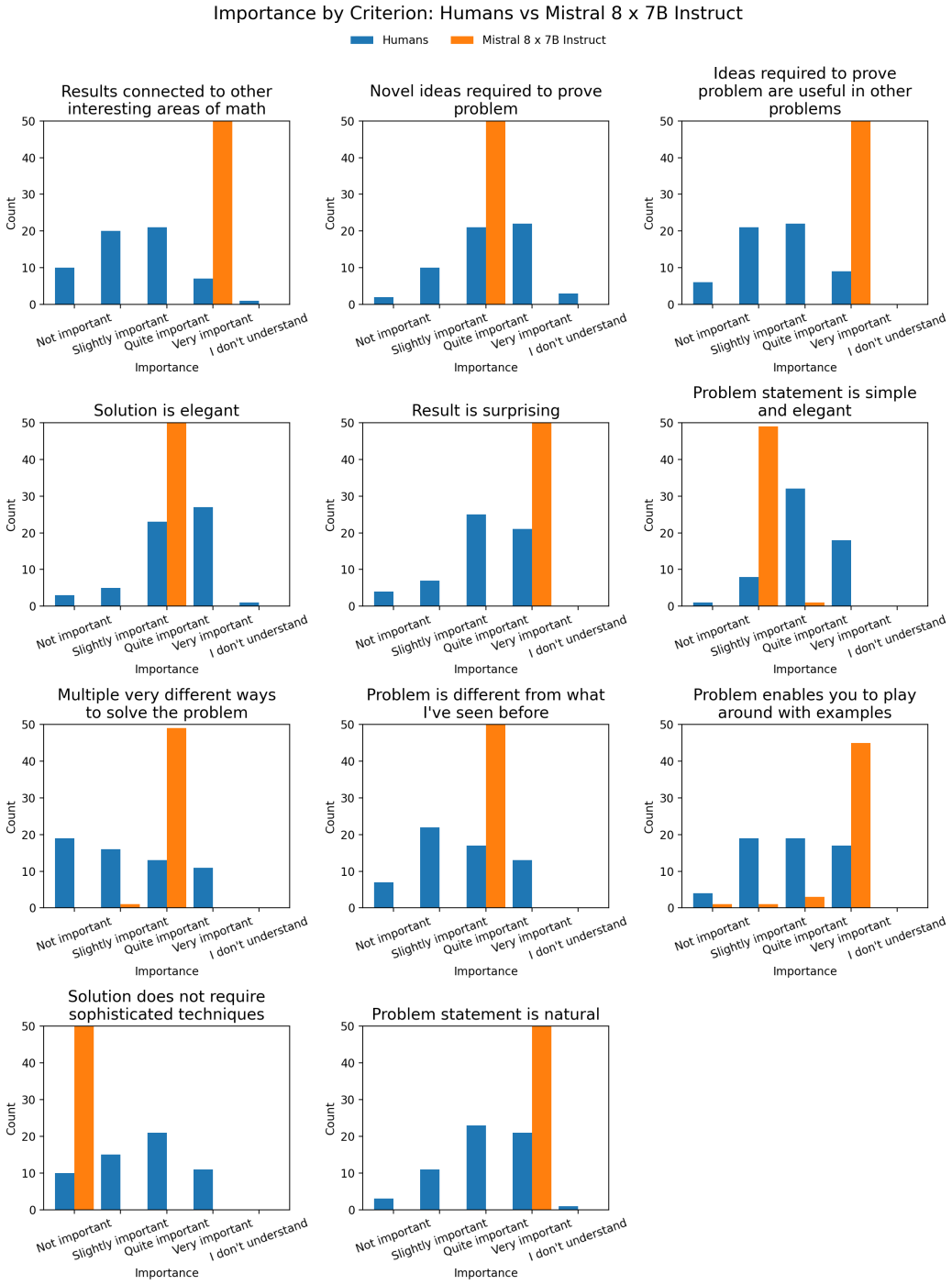


Figure 22: Importance ratings from **Mixtral 8×7B Instruct**.

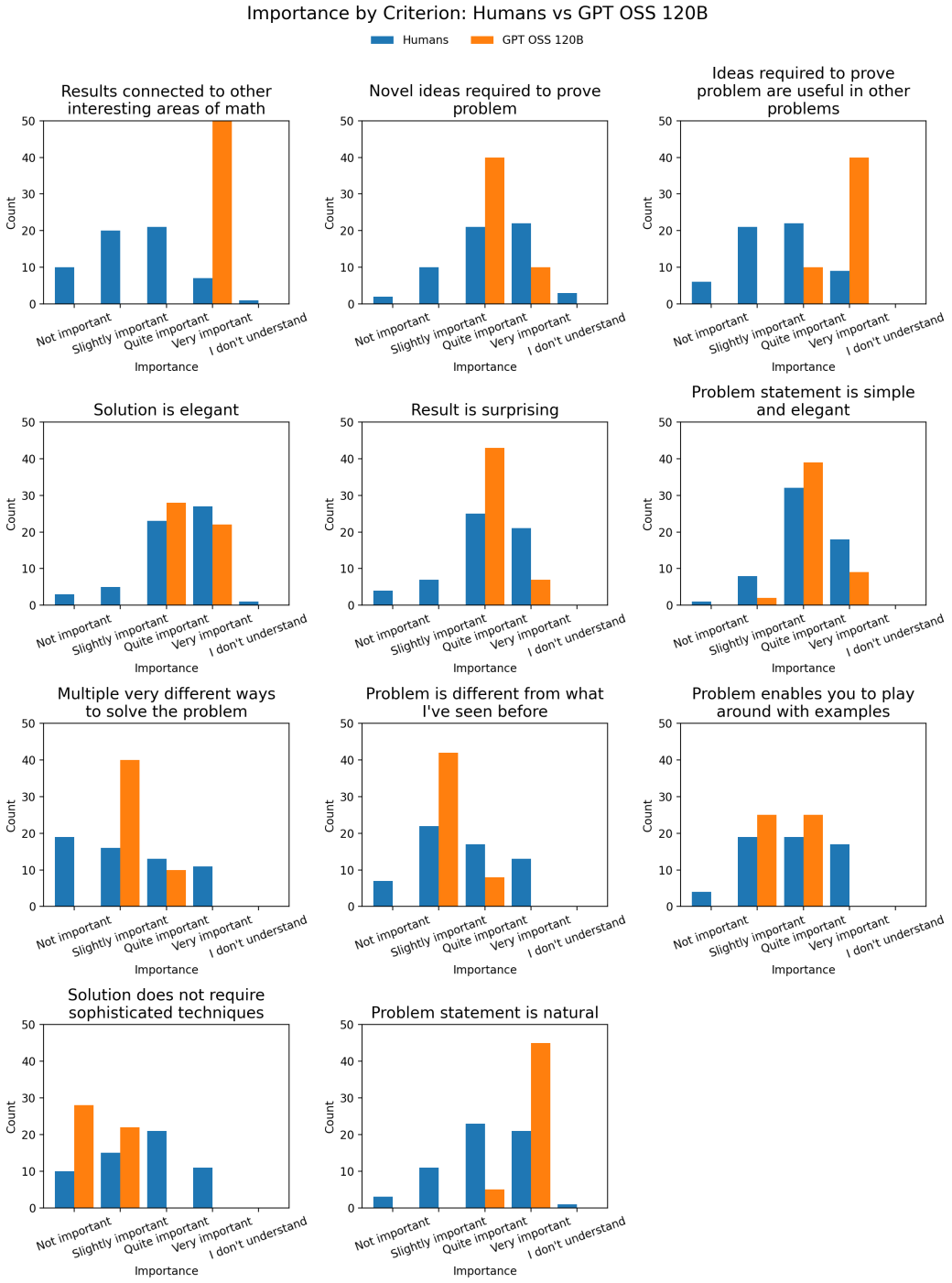


Figure 23: Importance ratings from **GPT-OSS 120B**.

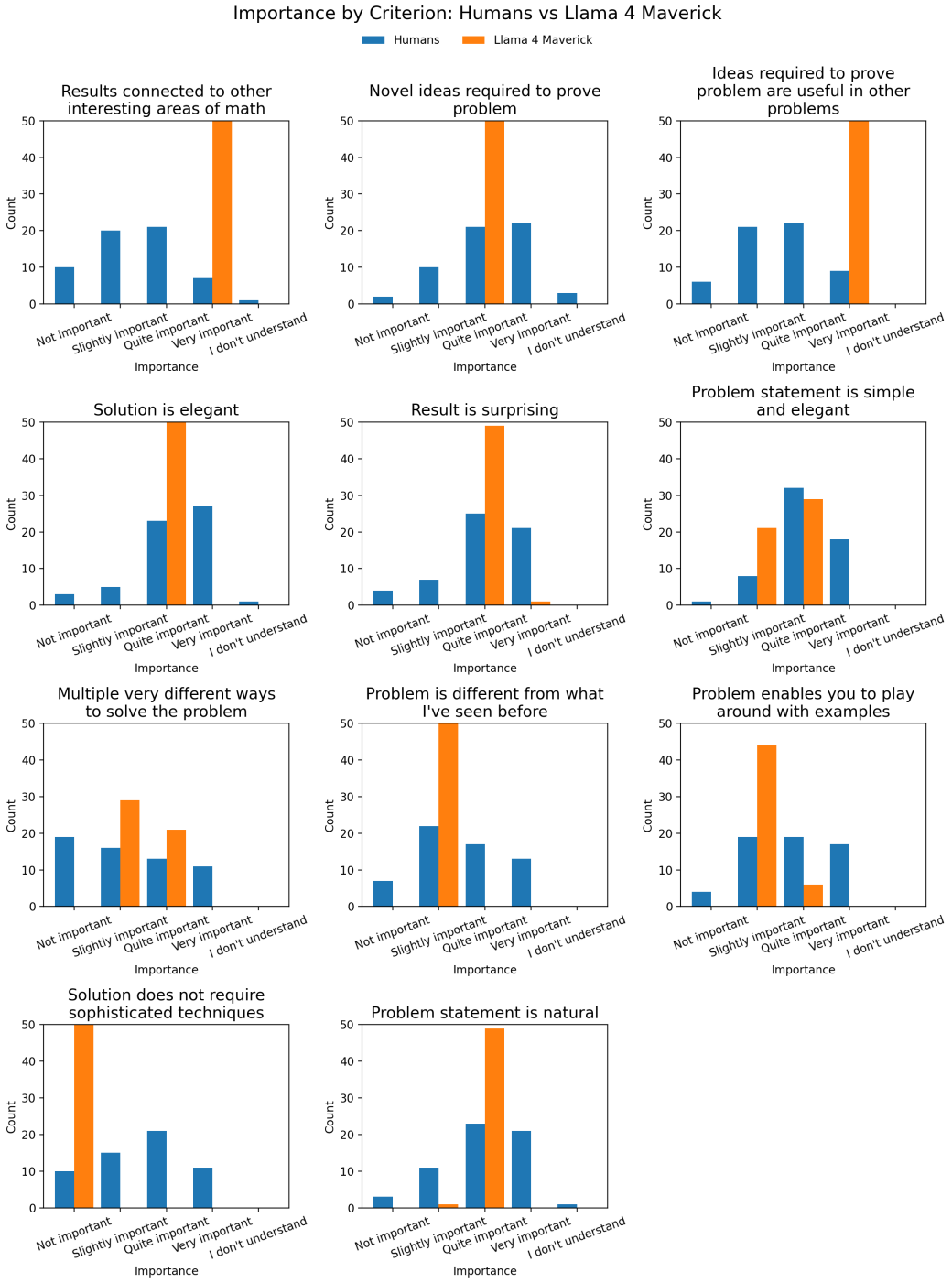


Figure 24: Importance ratings from **Llama 4 Maverick**.

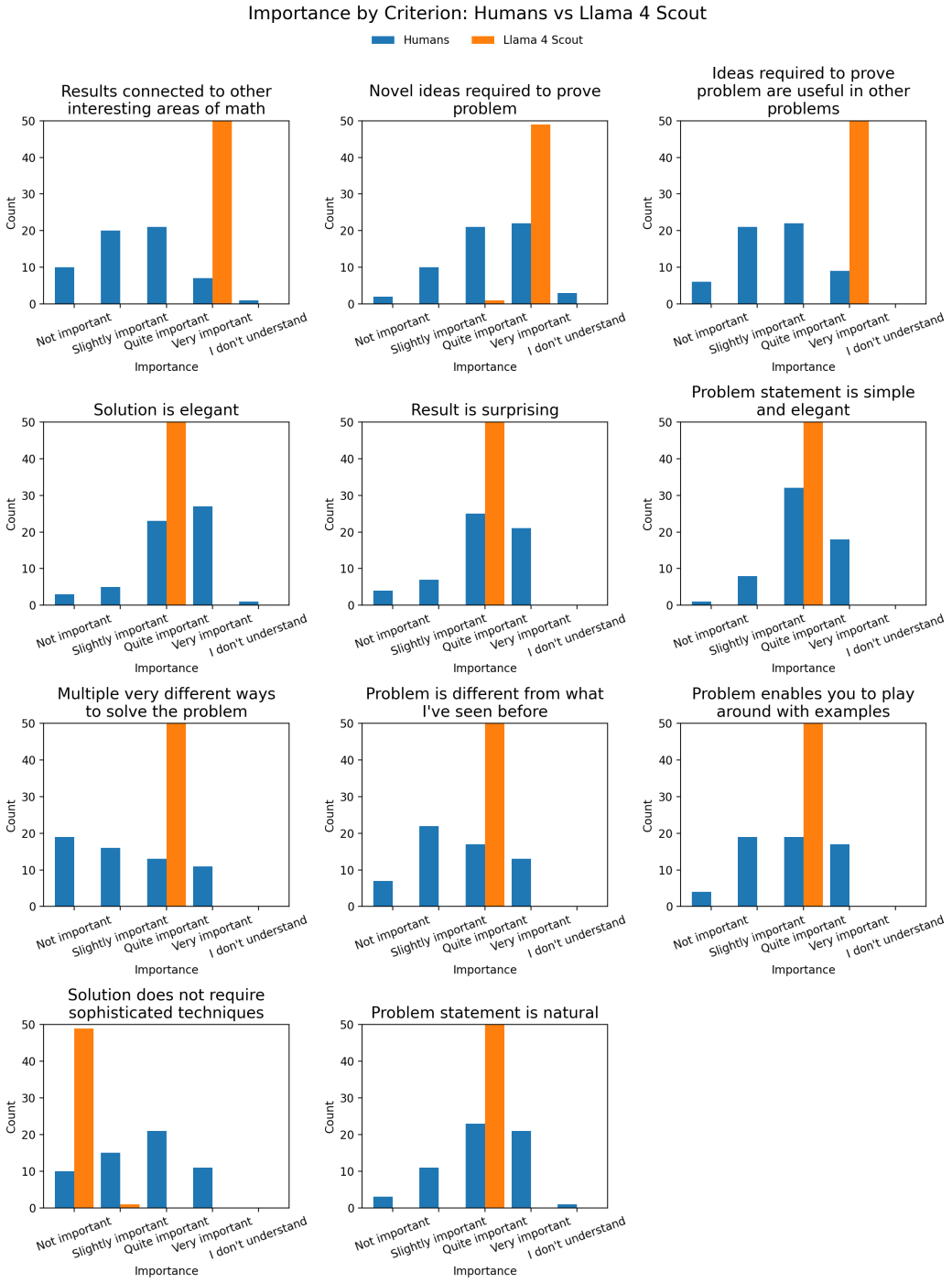


Figure 25: Importance ratings from **Llama 4 Scout**.

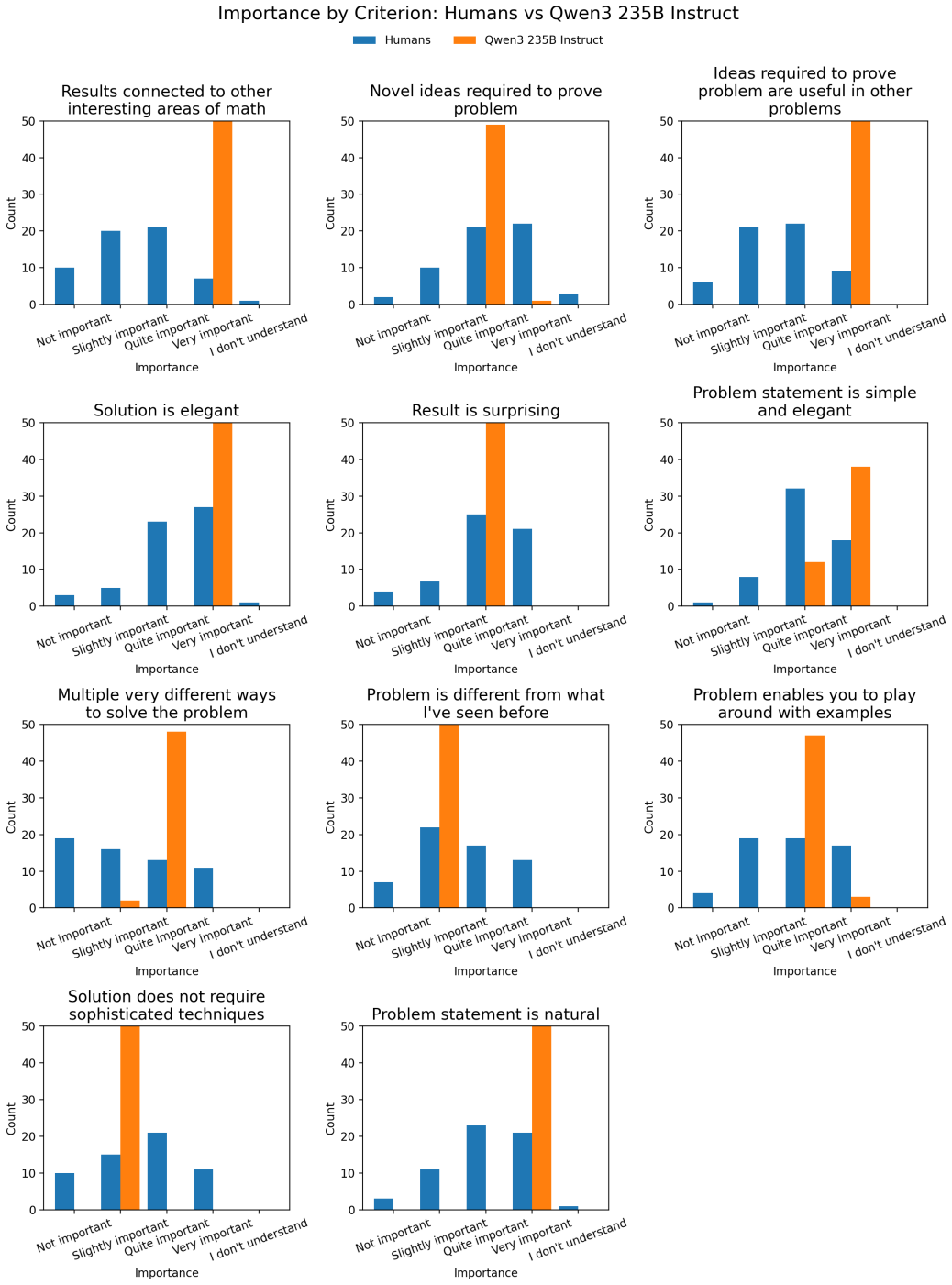


Figure 26: Importance ratings from **Qwen 235B Instruct**.

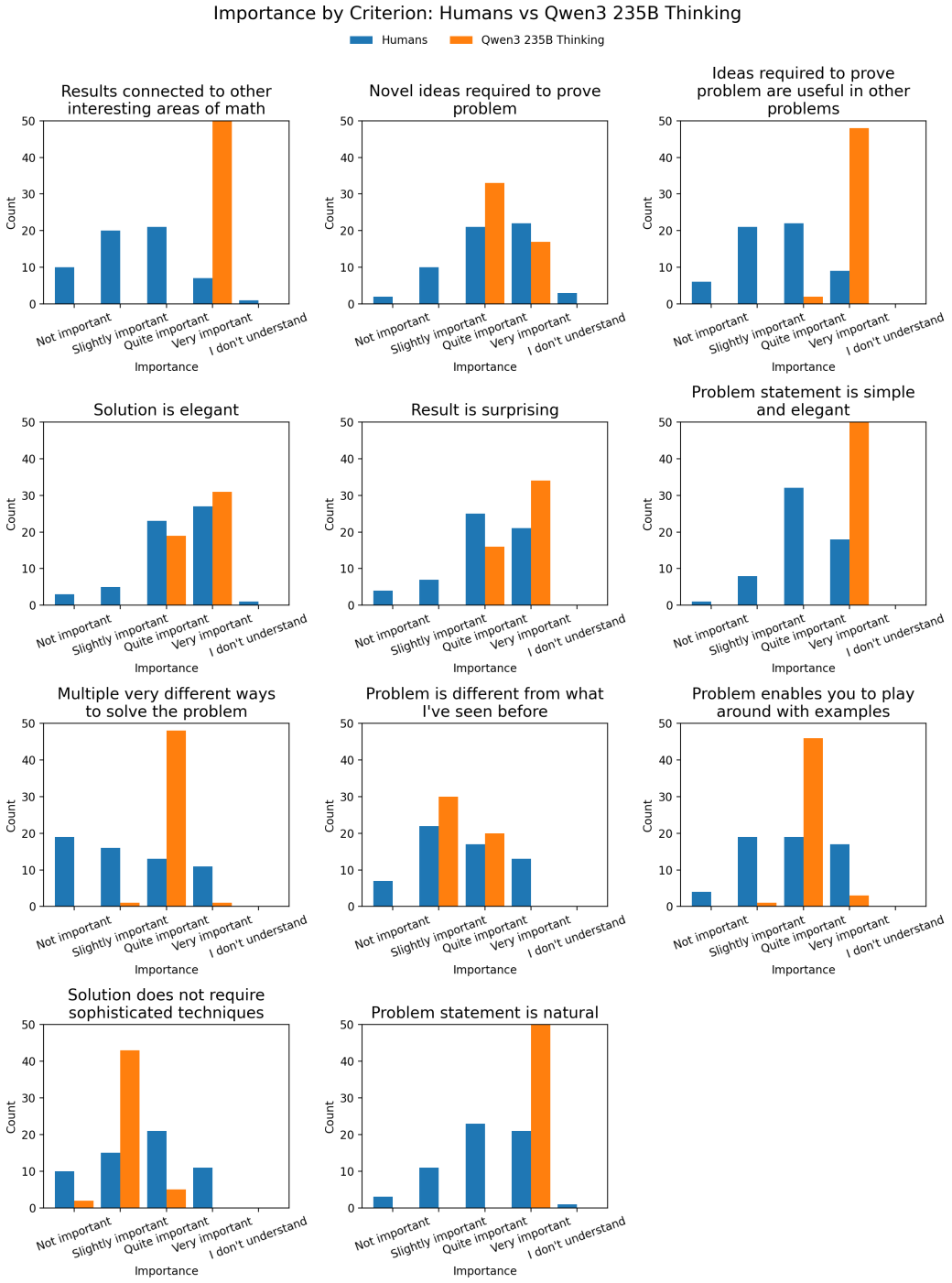


Figure 27: Importance ratings from **Qwen 235B Thinking**.

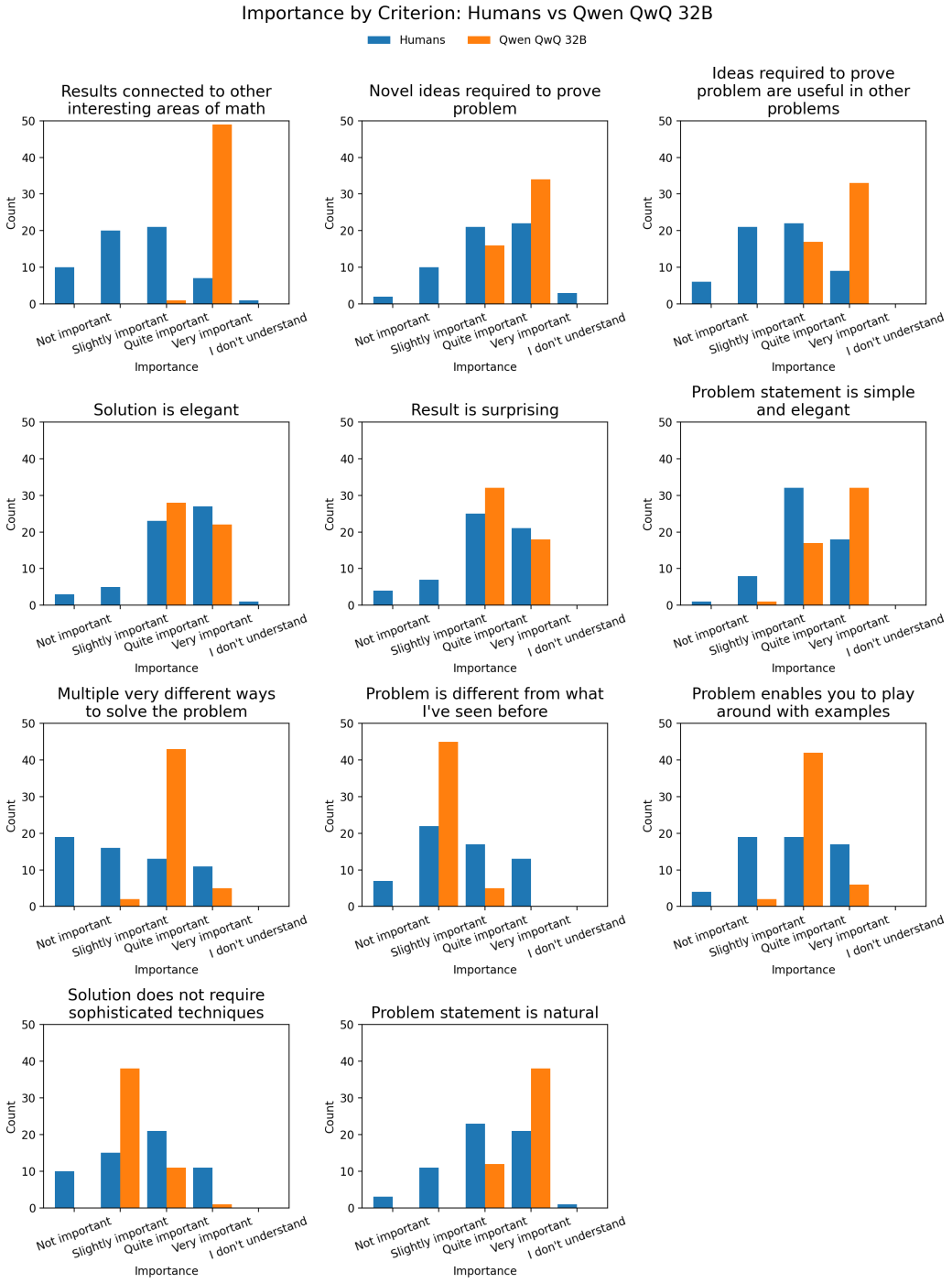


Figure 28: Importance ratings from **QwQ 32B**.

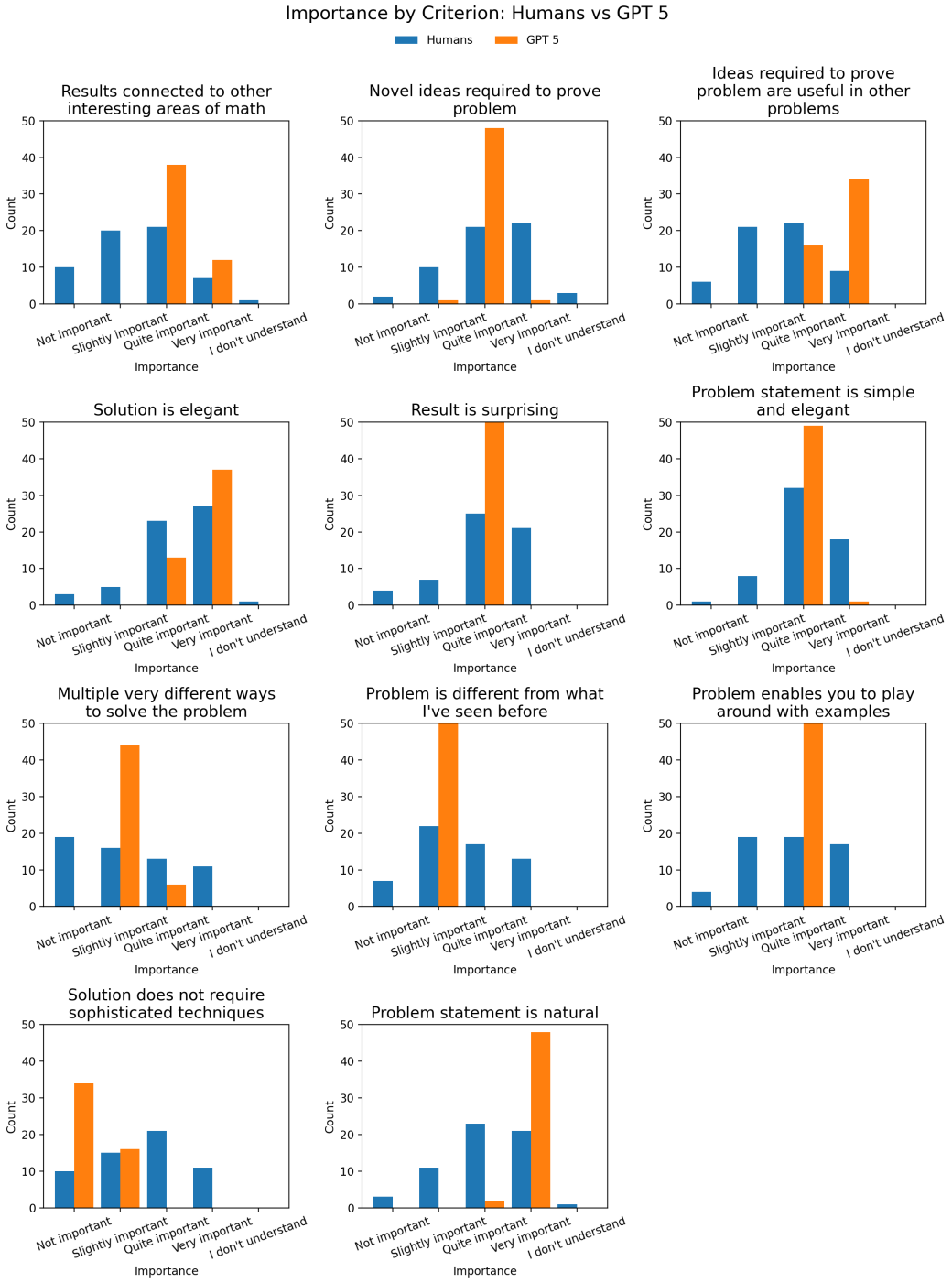


Figure 29: Importance ratings from **GPT-5**.

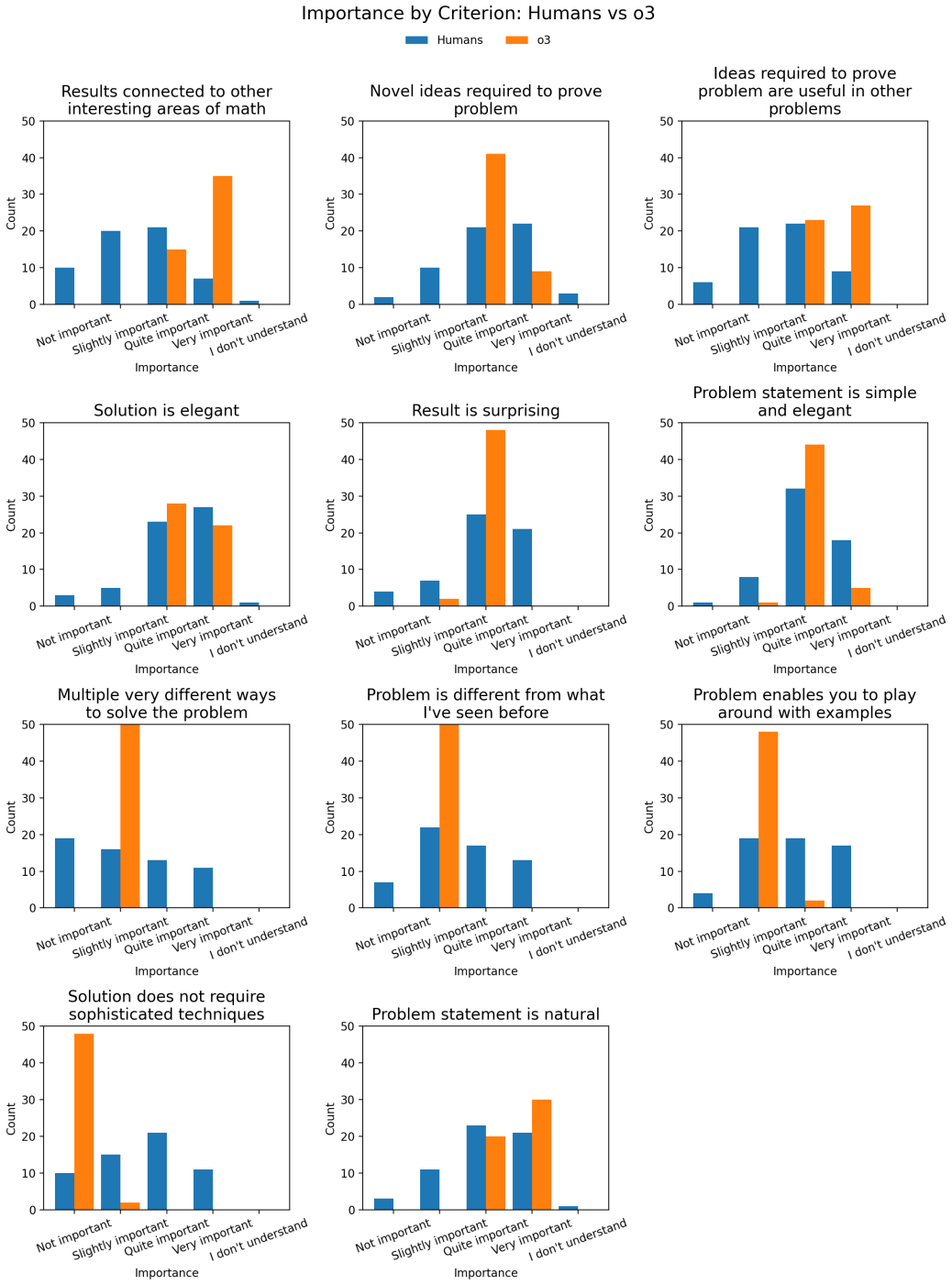


Figure 30: Importance ratings from **o3**.

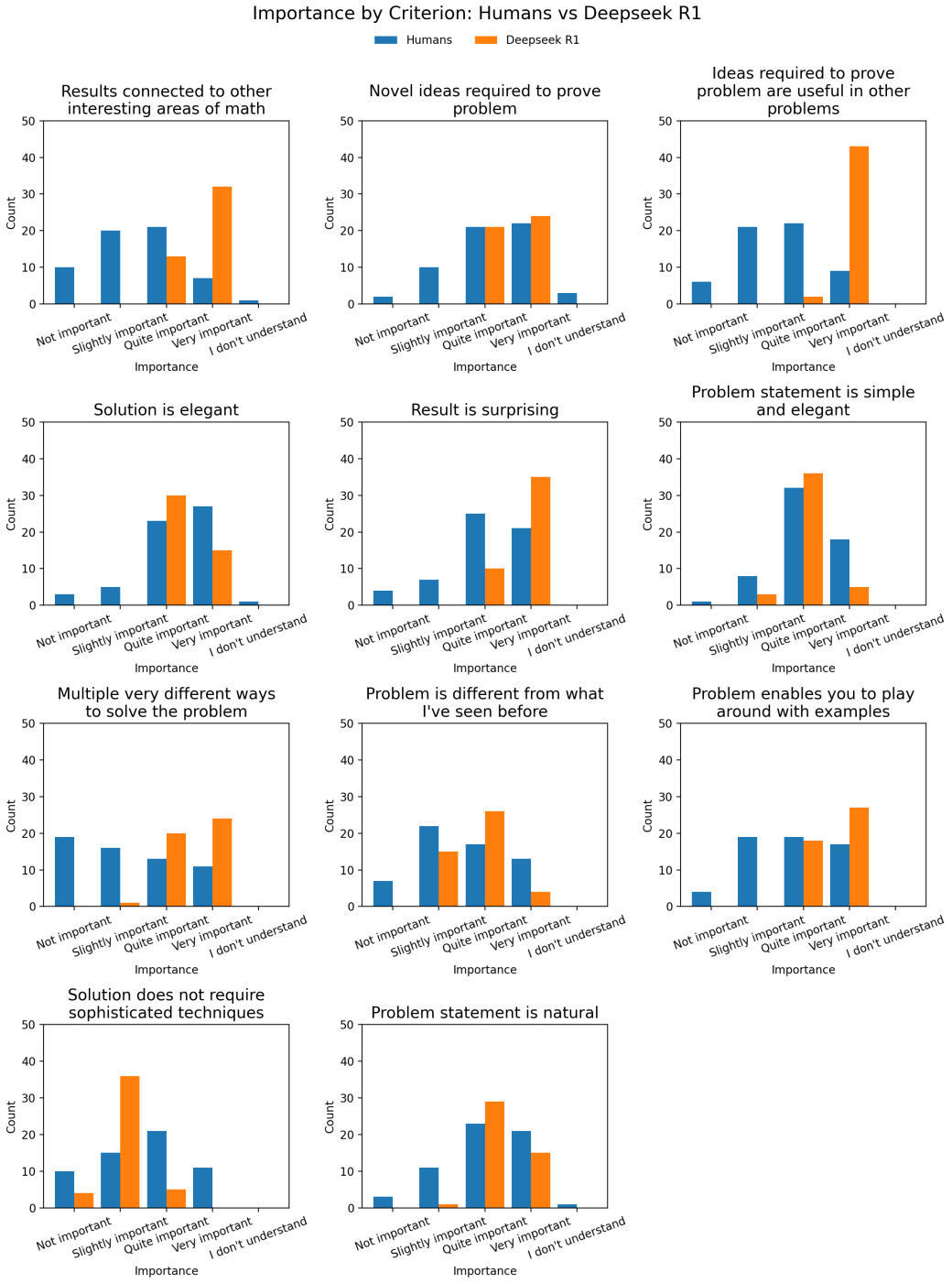


Figure 31: Importance ratings from **DeepSeek R1**.

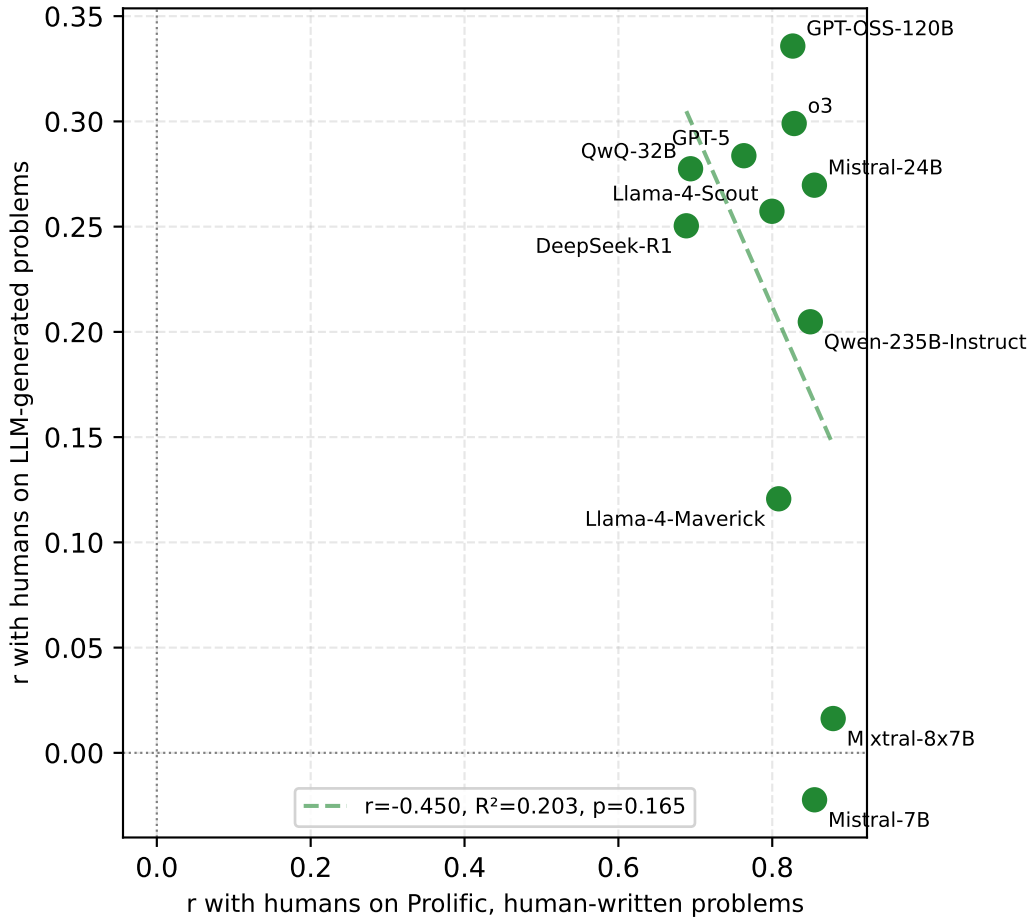


Figure 32: Comparing LLM-human alignment on interestingness judgments for human-written (Prolific) problems to LLM-human alignment on LLM-generated problems (i.e. are the models that are most aligned with humans on human-written problems also the most aligned with humans on LLM-generated problems?). We see no significant relationship between whether a model is more human-aligned on judging human-written problems and whether it is aligned on judging LLM-generated ones.

C Models

In table 9, we list all models and the providers we use.

Table 9: Models evaluated in this study, their full identifiers, and providers.

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)

D Compute Information

Our analyses do not require GPU compute, and all LLM inference was done using TogetherAI and OpenAI's API. The device used to run the analyses relies on the Apple M3 Pro chip.

E Methods: Prolific Study

Below, we describe additional information used to collect and adapt problems for the Prolific study. The majority of the problems were taken from the AMC contests, which are property of the Mathematical Association of America.

E.1 Participant Information

For this study, we recruited 63 participants across 2 conditions. The first condition included 33 participants, 4 of whose responses we chose to discard for the study since they failed the negative control. The second condition included 30 participants, 2 of whose responses were discarded for the same reason, resulting in 57 total participants whose responses we analyze.

E.2 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.

E.3 List of problems

We list all problems for our Prolific study, their variants, and variant types in Table 10.

Table 10: Prolific study problems and their variants.

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 $\overline{WZ} = 8$. 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 $\overline{WZ} = 8$. 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

F Methods: IMO Study

We conducted a survey of interestingness judgments made by participants at the 2024 IMO. Each of the 48 survey participants saw four problems. Each participant saw the same baseline problem. 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). Participants were given the option to view the solution before rating the problem's *interestingness* and *difficulty*. They were also asked to select reasons for their interestingness and uninterestingness ratings from a multiple choice list (see Appendix F.1), plus an additional free-text box to state their own reasons. Most problems only received one to two responses; this is too few to compare human and model judgments at a per-problem level. As such, IMO data comparisons are made over the interestingness *criteria* that participants selected.

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.

F.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.

G Methods: Problem Generation Study

To begin to assess whether people find these problems interesting, we run an initial pilot problem evaluation study with Prolific participants, following our previous experimental design (to collect the judgments in RQ1-2), judging LLM-generated problems. Participants are only shown the problems and do not know the problems are LLMs generated to avoid bias. We choose LLMs along the range of human-alignment (as measured by WD in RQ2a): we use Mistral 7B instruct (the most human aligned), Qwen 235B Thinking (somewhat human aligned), and o3 (the least human aligned) as problem generators. This also allows us to assess whether models that are more aligned to people in their evaluations of problem interestingness are also better generators, or vice versa. As we sample 30 problems from each LLM, this creates a set of 90 problems, from which we manually filter invalid problems (e.g., problems with inconsistent information, no correct answer, etc.). This step removes 12 problems, all of which were generated by Mistral 7B Instruct. From these problems, we create 2 batches of 12 problems rated by 15 participants each (30 total participants and 24 total problems), resulting in 360 judgments we evaluate.

We also ablate whether an automated problem verifier would also discard the problems as in our manual filtering, and find that the LLM we use (OpenAI's O3) filters out the same cases. Specifically, we provide the following prompt:

```
You are a meticulous math contest grader.

Task:
1) Decide whether the problem statement is well-posed and solvable as
written, with a determinate answer.
2) If it is NOT well-posed/solvable, set valid=false and give a concise
reason (missing diagram/data, ambiguity, contradiction, non-unique, etc.).
Do not attempt to guess missing info.
3) If it IS well-posed, solve it correctly and set valid=true and provide
the final answer only in 'answer'.

Output must match the provided schema exactly.
```

G.1 t-SNE Visualization

To create the t-SNE visualization in fig. 4, we use text-embedding-3-small for embeddings, a perplexity of 15.0, and plot using cosine distance.

H Methods: Prompts to Humans and LLMs

H.1 Study Interface

In Figure 33, we include pictures of the interface where the Prolific study was conducted.

H.2 Survey Text - Prolific Study

First, we include the pre-survey text seen by Prolific participants.

```
Welcome! We are conducting an experiment to understand how people think about
math problems. Your answers will be used to inform cognitive science,
mathematics, and AI research.

PLEASE ONLY PARTICIPATE IF YOU FIND MATH PROBLEMS INTERESTING. OTHERWISE
THIS STUDY WILL NOT BE FUN!

This experiment should take approximately 45 minutes.

You will be compensated at a base rate of $13.5/hour, with an optional bonus
to bring the total up to a rate of $15/hr if you try your best throughout
```

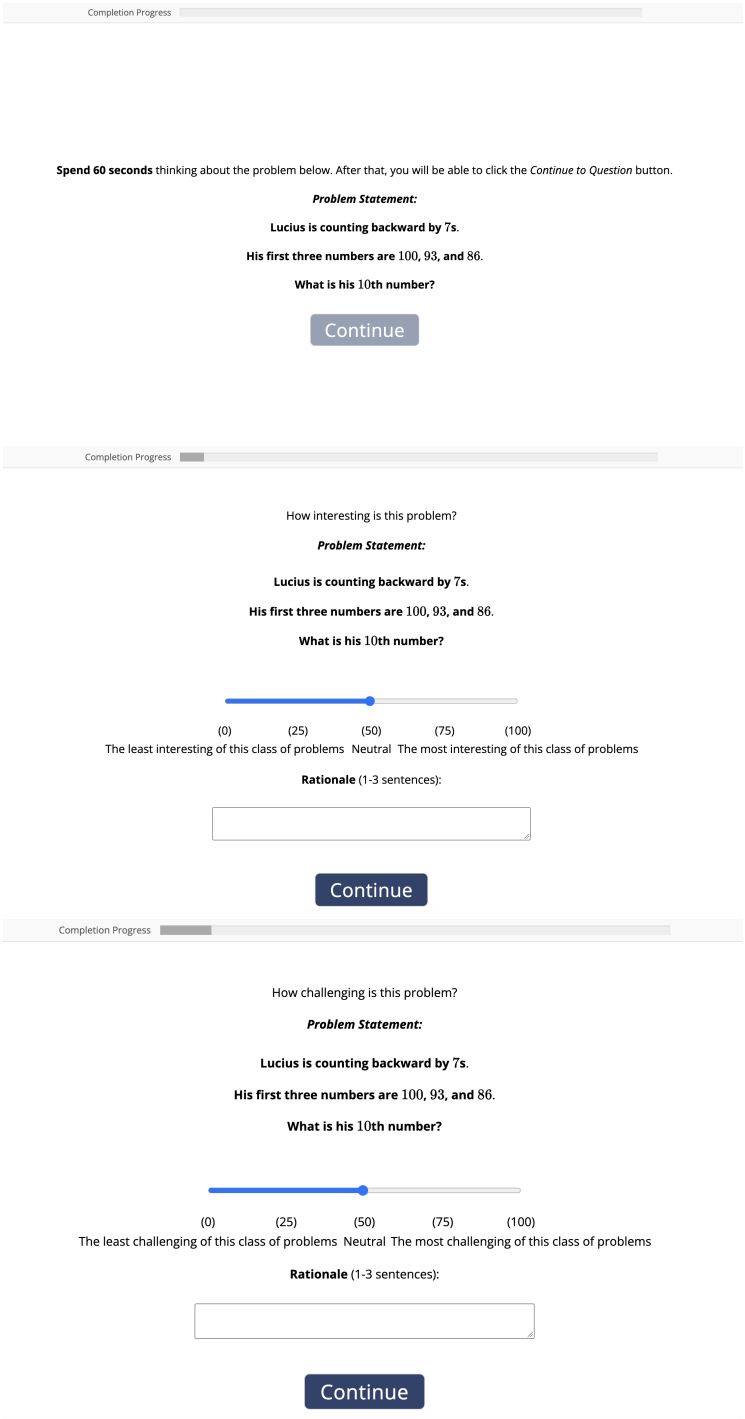


Figure 33: This is the interface where the participants view math problems in our Prolific study. The button to continue is grayed out for one minute, disallowing participants from moving on to scoring the problems and writing their rationales as they reason about the problem. Once participants have reasoned about the problem for at least one minute, they are asked to score the interestingness on a 1-100 scale, and then give a rationale for their score. Participants then repeat this to also rate the difficulty of the problem.

the experiment to answer each question

Please set the experiment to full screen.

We take your compensation and time seriously! The email for the main experimenter is [removed for anonymity].

Please write this down now, and email us with your Prolific ID and the subject line Human experiment compensation if you have problems submitting this task, or if it takes much more time than expected.

Please do not use any other Internet source or aid, including internet search, calculators, or chatbots - this experiment is designed to measure how people think. Unfortunately, if we find evidence that you did this, you may not receive the payment for this task.

There are two parts to this experiment. We present the instructions for Part I next. Instructions for a brief Part II will follow after you complete Part I.

In this experiment, you will be reading math problem statements and answering a few questions about each problem.

For each problem, your task is to answer two questions.

You will be asked to assess how interesting that problem is and how challenging the problem is.

You can interpet interestigness however you wish.

You will answer the question by dragging a slider.

We also ask that you type your rationale for your judgment per problem (1-3 sentences) in a text box.

Before you answer the question for each problem, you will have as much time as you want to think about the problem and its possible solutions. You must spend at least 60 seconds thinking about each problem. Please use scratch paper and focus while doing so.

After you feel like you understand the problem, please press the CONTINUE TO QUESTION to indicate that you are ready to answer the question.

We encourage you to take your time and carefully analyze the problem before providing your answer.

You will see a total of 10 problems.

When you are ready, please click "Next" to complete a quick comprehension check, before moving on to the experiment.

Please make sure to window size is in full screen to properly view the questions.

H.3 Survey Text - IMO Study

Your Task

The purpose of this survey is to study what mathematical olympiad problems people

find interesting. In order to do so, we will ask you to judge the interestingness of various problems to get as detailed information as possible of what kinds of problem statements are interesting.

In the first part, you will identify some key notions which you think are most important for a maths olympiad problem to be interesting in general.

In the second part, you will rate the interestingness of four past problems from various IMO shortlists. You will evaluate how these specific problems are interesting or not interesting.

Finally, you will be asked to self-evaluate what kind of a mathematician you are, which will provide more context for your interestingness judgement.

The rating process for each problem will be as follows:
* First, you will read the problem statement, and decide whether you want to read the solution to the problem. This is to accommodate for people who want to get a gist of the solution before making a judgement of interestingness. You are welcome to skip this step if you like.
* You will then be asked to rate the interestingness of the problem, and elaborate on why the problem was interesting or not interesting.
* We will also ask how well you understand the problem, as well as how difficult you think the problem is, as this might have an effect on your judgement of interestingness.

The survey should take approximately 15 min. If you feel that you are taking too long, feel free to judge interestingness of problems just based on the problem statement without considering the solution.

Privacy and Data Usage

We very much value your privacy! We will not save any identifying information, beyond your self-reported level of mathematical olympiad expertise and your participation at the IMO. We plan to release these ratings anonymously open-source for other researchers. Please do not participate if you are not comfortable with those data sharing procedures. Please note that you are welcome to leave the study at any time, and that your participation in this research is voluntary. Please only proceed to the study if you are comfortable with the above, and acknowledge that you wish to participate in this research.

If you are comfortable with all of the above -- please read the instructions closely! -- then we welcome your participation in survey.

In this survey, you will be asked to judge the interestingness of the following four problems from various IMO shortlists. Before this, please tell us what you thinks is generally important for a maths Olympiad problem to be interesting.

Here are some possible notions of interestingness. Please indicate how important these factors generally are for a problem to be interesting to you.

[The following reasons each had a multiple choice selector that let participants choose one of the following: Very important, Quite important, Slightly important, Not important, and I don't understand what this criteria means.]

1. The results we are asked to prove is connected to other interesting ideas/areas of mathematics.
2. The ideas required to prove this problem are novel.
3. The ideas required to prove this problem are potentially useful in other problems too.

4. The solution is elegant.
5. The result we are asked to prove seems surprising and unexpected.
6. The problem statement is simple and elegant.
7. There are multiple very different ways to solve the problem.
8. The problem seems very different from any problem I've seen before.
9. The problem allows you to play around with examples and get good intuition.
10. The solution does not require any sophisticated techniques/theorems.
11. The problem statement is natural.

[This was followed by all four problems being shown at once, and then each problem being shown one at a time, with the participant having the option to either view the solution or not.]

Would you like to see the solution to this problem before making a judgement? For geometry problems, choosing to see the solution will allow you to see the diagram too.

[0 - 7 scale of Least interesting problem I've ever seen - Most interesting problem I've ever seen]

What made this problem interesting? If this problem was not interesting at all, or if you think none of these reasons capture why this problem was interesting, tick the last option and move on.

- 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.

Were there any other factors which made this problem interesting?
[open-ended response space]

What made this problem uninteresting for you? Please tick the last option if none of the other options apply.

- The results 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 the 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.

```
Were there any other factors which made this problem uninteresting?
[open-ended response space]

Do you know the solution to this problem? [multiple choice]
No. My judgement is based purely on the problem statement.
Yes. I've solved the problem before.
Other:

How difficult is this problem?
0: too easy for an IMO
1: an easy IMO problem 1
2: a hard IMO problem 1
3: an easy IMO problem 2
4: a hard IMO problem 2
5: an easy IMO problem 3
6: a hard IMO problem 3
7: too hard for an IMO
Other:
```

H.4 LLM Prompt - Prolific Study

Below, we include the prompt given to LLMs while rating problems from the Prolific study. The majority of it is similar to what the human participants saw.

```
In this experiment, you will be reading math problem statements
and answering a few questions about each problem.

For each problem, your task is to answer two questions.

You will be asked to assess how interesting that problem is and how
challenging the problem is. The scale for both these ratings is between 0-100.

You can interpet interestigness however you wish.

We also ask that you type your rationale for your judgment per problem (1-3
sentences).

We encourage you to take your time and carefully analyze the problem before
providing your answer.

Your response should be in the following format.

### Interestingness: rating_between_0_and_100_here
### Int_rationale: 1-3 sentence rationale here
### Challenge: rating_between_0_and_100_here
### Chall_rationale: 1-3 sentence rationale here

{problem}
```

H.5 LLM Prompt - IMO Study

Next, we include the prompt used to elicit interestingness and difficulty judgments for the IMO problems from LLMs.

```
In this experiment, you will be reading math problem statements and answering
several questions about each problem.

For each problem, you will be asked to:
```

```
1. Rate how interesting the problem is on a scale of 0-7.
    (0 = least interesting problem you've ever seen,
    7 = most interesting problem you've ever seen)

2. Select reasons why you found the problem interesting
(or select "None of the above" if none apply).
   Options:
   - 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.

3. Select reasons why you found the problem uninteresting
(or select "None of the above" if none apply).
   Options:
   - The results 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 the 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.

4. Rate how difficult the problem is on a scale of 0-7.
   (0 = too easy for an IMO, 7 = too hard for an IMO)

We encourage you to take your time and carefully analyze the problem before
providing your answer.

Your response should be in the following format:

### Interestingness: rating_between_0_and_7_here
### Int_rationale: 1-3 sentence rationale here
### Interesting_reasons: list only the reasons that apply (ONLY from the provided
options, exactly as they are written). Do not list reasons you don't agree with.
### Uninteresting_reasons: list only the reasons that apply (ONLY from the provided
options, exactly as they are written). Do not list reasons you don't agree with.
### Challenge: rating_between_0_and_7_here
### Chall_rationale: 1-3 sentence rationale here

{problem}
```

H.6 LLM Prompt - Interestingness Reasons

Below, we include the prompt used to prompt LLMs about the importance they assign to different reasons for interestingness.

```
Here are some possible notions of interestingness.
For each criterion below, please indicate how important it generally is for a
problem to be interesting.

Choose one of the following values for each criterion:
- Very important
- Quite important
- Slightly important
- Not important
- I don't understand what this criteria means

Criteria:
1. The results we are asked to prove is connected to other interesting ideas/areas
of mathematics.
2. The ideas required to prove this problem are novel.
3. The ideas required to prove this problem are potentially useful in other
problems too.
4. The solution is elegant.
5. The result we are asked to prove seems surprising and unexpected.
6. The problem statement is simple and elegant.
7. There are multiple very different ways to solve the problem.
8. The problem seems very different from any problem I've seen before.
9. The problem allows you to play around with examples and get good intuition.
10. The solution does not require any sophisticated techniques/theorems.
11. The problem statement is natural.

Please respond in the following JSON format. Respond only from the options provided,
nothing else.
Don't return anything besides the JSON.

{
"criterion_1": "...",
"criterion_2": "...",
"criterion_3": "...",
"criterion_4": "...",
"criterion_5": "...",
"criterion_6": "...",
"criterion_7": "...",
"criterion_8": "...",
"criterion_9": "...",
"criterion_10": "...",
"criterion_11": "..."
}
```

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Yes, the methods for the study described in the abstract and introduction are explained in section 3, and the results we see are included in section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the limitations can be found in section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include any theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We release all our code and data here. We disclose all the information necessary to replicate exactly the studies we conduct. We also provide all experimental details necessary to replicate our experiments to analyze similarly-collected data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code can be accessed here.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper describes all necessary information to replicate the analyses we conduct in section 3, section 4, and the appendices.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All our analyses include bootstrapped confidence intervals, either in section 4 or in the corresponding appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include compute information in Appendix section D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, our research conforms, in all aspects, with the NeurIPS Code of Ethics. Participants in all studies provided informed consent. Participants in all Prolific studies were paid at fair wages; IMO participants elected voluntarily to participate in the study.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we include an impact statement discussion both possible positive and negative societal impacts in Appendix section A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we properly credit the necessary organizations in Appendix section E.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, all assets introduced in the paper are stored and documented here.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes, all instructions and compensation instructions are provided in Appendix section H.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We mention in both section 3 and Appendix section F that the studies received approvals from our institutions' review boards. The studies posed no potential risks to the participants, and we have included all instructions given to the participants are described in Appendix section H.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We do not use LLMs for any important, original, or non-standard components, only as a programming aid, as mentioned by the NeurIPS handbook.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.