

NLP 245 | Final Project

Judging LLM-as-a-Judge in an All-to-All Scenario

Kiara LaRocca | Jack St. Clair | Shriya Sravani Y | Sam Silver

klarocca | jrstclai | sy4 | ssilver1@ucsc.edu

17 March 2025

Abstract

The fast evolving world of Large Language Models (LLMs) has spawned a plethora of options for users. While this suite of options continues to grow, there is still a space for better evaluations of LLMs. In this paper, we will take a critical eye to one such option: "LLM-as-a-Judge". We use LLMs to create generations and then use the models and some humans to judge the generations on a set of criteria. We do this to compare and contrast the efficacy of particular models as judges, as well as compare judge models and human judges. We find that while more recent models are stricter judges than older models, the correlation with even stricter human judgements is not strong.

1 Introduction

LLMs have rapidly advanced in recent years, demonstrating impressive capabilities in generating human-like, fluent, and contextually relevant text. However, evaluating the quality of these outputs remains a difficult problem. Traditional metrics such as BLEU, ROUGE, and METEOR rely on n-gram overlap and fail to capture deeper semantic and pragmatic qualities. While human evaluation is more nuanced, it is expensive, time-consuming, and prone to subjectivity.

As a result, there is growing interest in using LLMs as evaluators in an approach known as LLM-as-a-Judge. These models can assess outputs based on criteria like coherence, factuality, and fluency, offering a scalable alternative to human annotation. Prior work has shown promising results, with some LLM judgements aligning well with human preferences.

In our project, we propose a novel all-to-all evaluation framework where multiple LLMs serve as both response generators and judges. We use subsets of two datasets, one spanning open-domain QA (ELI5) and the other based on creative writing (WritingPrompts) to enable diverse, analyzable

outputs. These subsets are evaluated on qualitative dimensions (factuality, coherence, style) and each model assesses the outputs of its peers, producing a cross-evaluation matrix that captures both generation quality and judgement behavior. To further validate these judgements, we conduct human annotation on a subset of responses, allowing us to compare LLM evaluations with human preferences. This cross-evaluation setup offers insights into how LLMs critique language and highlight areas for improving automated evaluation methods.

2 Related Work

There exists a long history of evaluation of language model outputs. Traditional automated metrics such as BLEU ([Papineni et al., 2002](#)) and ROUGE ([Lin, 2004](#)) have been widely used in machine translation and text generation tasks, primarily focusing on n-gram overlap between generated and reference texts. However, these metrics often fail to capture semantic similarity and nuanced aspects of language quality. More recent learned metrics like BERTScore ([Zhang et al., 2020](#)) leverage contextual embeddings to better capture semantic similarity, while BLEURT ([Sellam et al., 2020](#)) uses BERT-based models fine-tuned on human judgements. Despite these advances, such automated metrics still struggle to evaluate open-ended dialogue responses where multiple valid outputs are possible and where subtle qualities like coherence and appropriateness are crucial.

Recent work has explored various approaches to automated evaluation of LLM output using generative language models themselves.

Work in this subfield is commonly divided into two broad, related areas of research: reward models and judge models.

2.1 Reward Models

Reward models have served as the backbone of preference-based learning systems.

Reward models (RMs) are perhaps the most crucial component in Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) pipelines. These reward models are trained to predict human preferences between different text completions for a given prompt, learning to assign scalar reward values that correlate with human judgements of quality. Formally, reward models were initially trained using a Bradley-Terry (BT) preference model:

$$p^*(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \quad (1)$$

While the Bradley-Terry model is historically used for skill assessment and matchmaking (via ELO scores) in games like Chess, it naturally finds a home in the context of human preference data collection, where humans typically amass a collection of pairwise preferences over generations, which the BT model helps turn into scalar preference/reward scores.

Training the BT model minimizes the maximum likelihood loss:

$$L(\theta, D) = \mathbb{E}_D[\log(1 + e^{r_\theta(x, y_r) - r_\theta(x, y_c)})] \quad (2)$$

More recent models like Starling (Zhu et al., 2023) have expanded data collection to other formulations, including the Plackett-Luce model for K-wise comparisons, enabling rich learning from fully ranked response sets rather than just pairwise preferences:

$$P(r) = \prod_{i=1}^{n-1} \frac{\pi_{r_i}}{\sum_{j=i}^n \pi_{r_j}} \quad (3)$$

It's notable too that recent work has explored replacing human preference data with model-generated preferences, in a process known as Reinforcement Learning from AI Feedback (RLAIF) (Zhu et al., 2023).

Architecturally, reward models are usually implemented by adapting pretrained language models by removing the final decoding components and replacing them with reward prediction heads (Ouyang et al., 2022) before finetuning on curated human preference datasets like Ultrafeedback (Cui et al., 2024) or other similar more domain-appropriate datasets.

Several variants of reward modeling have emerged in recent literature. Standard outcome-based approaches directly output scalar rewards for the entire generation (Zhu et al., 2023), while process reward models (PRMs) (Lightman et al., 2023) provide more granular stepwise feedback for complex reasoning tasks. Only recently has a popular benchmark for the more popular outcome reward models (ORMs) been developed (Lambert et al., 2024) which bodes well for the development of reward models; the research community can generally only improve what it can measure.

Regardless of the training methodology, reward models typically assign pointwise scalar rewards to quantify either stepwise or overall response quality.

2.2 Judge Models

Unlike reward models, which primarily assign scalar values to responses, judge models are typically finetuned language models that can flexibly provide more nuanced, criteria-specific evaluations of text completions.

Research in this area varies in terms of how much context (e.g. scoring rubric, high-and-low-quality exemplars) language model judges are provided during grading. It also varies in how feedback is generated and optionally incorporated into a larger learning system. Language model judges typically provide a pre-hoc textual rationalization for their ultimate scalar feedback, though some models, typically referred to as critic models, generate entirely text-based feedback. This is in contrast to reward models, which typically have their language-model classification head replaced with a scalar prediction head, erasing any opportunity to use Chain of Thought (CoT) reasoning (Wei et al., 2023), a technique known to significantly improve language model performance across a variety of tasks. The importance of CoT reasoning has only grown, with recent models (Erziev, 2025; Arrieta et al., 2025) undergoing reinforcement learning-based "reasoning" post-training for the purpose of learning to roll out vast and exploratory chains of thought for better performance.

Literature refers to the development of both pointwise and pairwise judges. Pointwise judges evaluate a single completion for a given prompt, producing a score or assessment:

$$J_\theta(x, y) \rightarrow s \quad (4)$$

In contrast, pairwise judges compare two completions for the same prompt, producing either a

probability that one response is preferred over the other or a discrete choice between the responses.

$$J_\theta(x, y_1, y_2) \rightarrow p(y_1 \succ y_2 | x) \quad (5)$$

With respect to architecture and training, judge models are typically trained under a simple supervised finetuning (SFT) regime on top of a base instruction-tuned language model, with training objectives usually involving traditional classification losses for binary decisions or regression losses for scalar scoring.

GPT-Score (Fu et al., 2023) was an early exploration of LM-as-a-Judge. The premise is elegantly simple: high-quality text will have a higher generation probability from an LLM when provided with appropriate instructions. An LLM told to generate text with specific desirable qualities will assign higher probabilities to examples that exemplify those qualities. This flexible framework allowed for multi-dimensional assessment of aspects like fluency, relevance, informativeness, and more via appropriate prompting.

G-Eval (Liu et al., 2023) used GPT-4 (OpenAI and et al., 2024) for automatic evaluation, demonstrating that the use of CoT significantly improves evaluation performance by providing more context and guidance for assessment. Authors noted that the recently-released GPT-4 significantly outperformed GPT-3.5's evaluate performance, especially on nuanced dimensions like consistency and relevance, and noticed that LLM-based evaluators appeared to have an inherent bias towards preferring LLM-generated text over human-written text, even when humans would judge the human-written generations to be superior.

AlpacaEval (Dubois et al., 2024) introduced an LLM-based pairwise evaluation metric in which GPT-4 Turbo acts in a judge capacity, outputting a "probability" of preferring model A's response over model B's response. Authors showed that LLM-based evaluators significantly favored models that generated longer outputs, creating an exploitable bias. Authors proposed a regression-based correction that estimates preferences if both the evaluated model and the strong baseline model (gpt4_1106_preview) had the same output length. Length-controlled AlpacaEval increased Spearman correlation with human preferences on Chatbot Arena higher than any other automated evaluation metric. Additional ablations suggested that the self-annotator bias of AlpacaEval was insignificant,

with model ranking remaining relatively stable regardless of which model was judging.

In LLM-Eval (Lin and Chen, 2023), authors perform multi-criteria LLM-as-a-Judge evaluations of language model generations, but focus on single-prompt efficiency; rather than making separate judge calls for each evaluation dimension, they combine the evaluation criteria, dialogue context, and response into a single prompt, with the judge returning a consistent JSON structure.

In Zheng et al. (2023), authors construct 80 multi-turn conversations across categories like writing, roleplay, math, and question answering. Second "user" turns often introduce constraints or modifications that create challenging scenarios requiring model adaptation (e.g. "Rewrite your previous response, starting every sentence with the letter A"). MT Bench then uses GPT-4 to grade responses on a scale of 1-10. While not groundbreaking with respect to its construction of a judge system, MT Bench, along with AlpacaEval, reigned for nearly a year as one of the preferred automatic evaluations for frontier language models. This can be attributed to their excellent curation of data and the fact that their judges aligned highly with human preferences.

2.3 Model Biases

Despite their importance, both reward models and judge models face several challenges. Ye et al. (2024) identify twelve types of biases that judge models (and reward models) can exhibit: Position bias (favoring responses based on their position in the prompt, e.g. preferring the first answer over the second), Verbosity Bias (preferring longer responses, even if they don't contain uniformly higher-quality content), Compassion Fade Bias (judging differently when given model names), Bandwagon Bias (giving preference to majority opinions, regardless of their correctness), Distraction Bias (paying undue attention to irrelevant details), Fallacy Oversight Bias (ignoring logical errors in reasoning steps, focusing only on candidates' final answers), Authority Bias (assigning more credibility to statements attributed to authority figures), Sentiment Bias (showing bias towards certain demographic groups), Chain of Thought Bias (evaluations varying with and without CoT reasoning), Self-Enhancement Bias (favoring answers they generated themselves), and Refinement-Aware Bias (scoring differently when told that a

response is a refined version of an initial response).

With respect to their political bias, Bang et al. (2024) performed experiments indicating that western models hold progressive beliefs on hot-button topics like immigration, reproductive rights, gun control, climate change, and healthcare reform. While it seems likely that modern jailbreaking techniques could be used to elicit political "opinions" of any variety from language models. Current properties of western companies' pre-training and post-training datasets seem to have "biased" models towards the progressive end of the political spectrum, impacting their impartiality as judges in certain scenarios.

2.4 Relation to Our Work

The vast majority of the related works cited above give explicit instructions or rubrics regarding what high-quality performance looks like in a specific domain. In contrast, our experiment has models generate responses in the creative writing and question answering domains. They are then prompted in a *zero-shot* manner to generating judgement rationales and judgement scores using *minimal descriptions* of domain-relevant criteria (often referred to as aspects or dimensions in literature). Our intuition is that providing detailed definitions and rubrics of various grading criteria will narrow the expressed diversity of judgement reasoning and scoring. This alternate method serves as more of a test of model instruction-following capabilities than as an open-minded evaluation into the role of intrinsic model characteristics in LLM-as-a-Judge evaluation.

Prior work discussed in this section primarily used the then-recently-released GPT-4 as a judge model (which represented a leap in capabilities), we've seen since early 2023 a cambrian explosion of even more capable models of varying sizes and origins. An increased focus on model post-training and on model "character" development has resulted in a wide array of capable models offering a diverse set of experiences to users. How might the specialized training modern language models undergo influence their performance as a judge?

3 Datasets

There are a variety of datasets that we explored during our project. After trawling HuggingFace for hundreds of interesting question answering datasets, our most promising candidates included

SQuAD v2.0 (Rajpurkar et al., 2018), Natural Questions (Kwiatkowski et al., 2019), Reddit ELI5 (Fan et al., 2019), and a Quora Question-Pairs dataset (Quora, 2017). The topical diversity of these datasets allows us to evaluate models' generation and judgement capabilities across a variety of question types, response formats, and assessment criteria.

Because of our limited computational budget, selecting realistic, high-quality, diverse datasets is of the utmost importance. While datasets can be selectively sampled to improve average quality, it's a more difficult task to *add* diversity to a dataset that doesn't have enough in the first place. Given these criteria, we ultimately selected the Reddit Explain Like I'm 5 (ELI5) dataset (Fan et al., 2019) and added the Reddit WritingPrompts dataset (Fan et al., 2018) for task diversity.

While multi-turn conversational datasets were considered, they pose a number of challenges related to experimental control and cost that we ultimately decided were untenable. In single-turn datasets, user turns are fixed and models generate only one assistant turn to be evaluated, meaning model responses can be compared on an "apples-to-apples" basis. When doing multi-turn evaluation of a collection of models, conversation trajectories can diverge wildly, as different assistant responses lead to divergent second user turns. This makes straightforward comparison of model conversation trajectories more difficult and requires a method for generation of coherent user turns.

Alternatively, some studies fix a number of user and assistant turns as a prefix, having the model under evaluation generate the latest assistant turn in a conversation, thereby turning a multi-turn evaluation into a single-turn evaluation. This forces models into completing significantly off-policy trajectories, an unrealistic scenario that intuitively degrades model performance.

Furthermore, napkin math put our projected experiment (detailed in the experiment section) cost at approximately \$150 USD for all the single-turn scenarios we describe in this paper, making multi-turn evaluation financially unreasonable. As a result, we moved forward with the single-turn question answering and creative writing datasets described below.

3.1 ELI5 Dataset

The ELI5 dataset (Fan et al., 2019) provides diverse user-generated information-seeking questions such as "what is a MAC address," "why does rain make people tired," and "why are mice associated with cheese?" The simply-stated questions allow for somewhat unconstrained response generation, opening the door for models to not only generate highly factual and relevant responses, but also be *creative* in the way these concepts are communicated.

The dataset is a large-scale corpus, containing titles from 272,000 threads on the Reddit forum `r/explainlikeimfive`. These titles contain open-ended questions and community-generated answers meant (somewhat hyperbolically) to be understandable to a five-year old. The dataset is divided into 237,000 training, 10,000 validation, and 25,000 test examples, carefully split to prevent near-duplicate questions across sets using TF-IDF similarity.

We choose to evaluate generations on this question answering task along the following criteria: relevancy (does this response answer the question?), factuality (is this response factually correct and verifiable?), clarity (is this response easy to understand?), creativity (does this response provide a creative explanation?), coherence (is the response well-structured?), and conciseness (is this response to the point?). We provide only minimal descriptions of each criteria, so as to maximize diversity of intrinsic interpretations among models, as described earlier. See Table 3 for full criteria information.

3.2 WritingPrompts Dataset

The Reddit WritingPrompts dataset (Fan et al., 2018) provides creative writing prompts such as "in a world of TechnoMagic, you are a Hacker," and "while putting your favorite condiment on a sandwich, you accidentally draw a symbol that summons demons." The diverse set of open-ended prompts allow for authors to take stories in any number of creative directions, testing whether models can create engaging plots, believable characters, snappy dialogue, and coherent stories.

The dataset is a large-scale corpus designed for open-ended creative story generation. It has over 300,000 human-written stories scraped from Reddit's `r/WritingPrompts` forum, each paired with a corresponding prompt. The dataset is split into 272,600 training, 15,620 validation, and 15,138

test examples.

We choose to evaluate generations on this creative writing task along the following criteria: plot (how compelling and well-structured is the narrative?), character (how well-developed and believable are the characters?), style (how distinctive and effective is the writing style?), emotion (how effectively does the story evoke feelings?), flow (how smooth and well-paced is the story-telling?), and dialogue (how natural and purposeful is the conversation and narration?). Again, we provide only minimal descriptions of each criteria, in an attempt to maximize diversity of intrinsic interpretations among models. See Table 3 for full criteria information.

3.3 Dataset Filtering

Given the community-sourced, user-generated nature of the datasets, it's no surprise that manual examination of our data revealed a significant number of duplicate and low-quality examples. Considered together with our computational/budgetary constraints, this meant that we needed to filter our datasets to a high-quality subset that maximally maintained diversity of data.

We performed this manually and heuristically, individually examining thousands of records from each dataset to select 50 representative, well-formed examples from each. This curation process prioritized prompts that were clear, diverse in topic coverage, and well-suited for both generation and evaluation.

4 Models

For both generators and judges, we evaluate a collection of frontier open-weight language models across a variety of model sizes and architectures. OpenRouter provides a uniform interface across models, making integration to evaluate a variety of models straightforward. OpenRouter provides access to models from more than 40 inference providers, with over 300 models available through the API. Models evaluated include Claude 3.7 Sonnet (Erziev, 2025), Gemini 2.0 Flash (Team and et al., 2024a), Gemma 2 27B Instruct (Team and et al., 2024b), GPT 4o (OpenAI and et al., 2024), Llama 3.3 70B Instruct (et al., 2024), Mistral NeMo 12B Instruct (Mistral AI, 2024), o3-mini (Arrieta et al., 2025), Phi-4 Instruct (Abdin et al., 2024), Qwen 2.5 72B Instruct (Qwen et al., 2025), and GPT 3.5 Turbo Instruct (Ye et al., 2023). For more

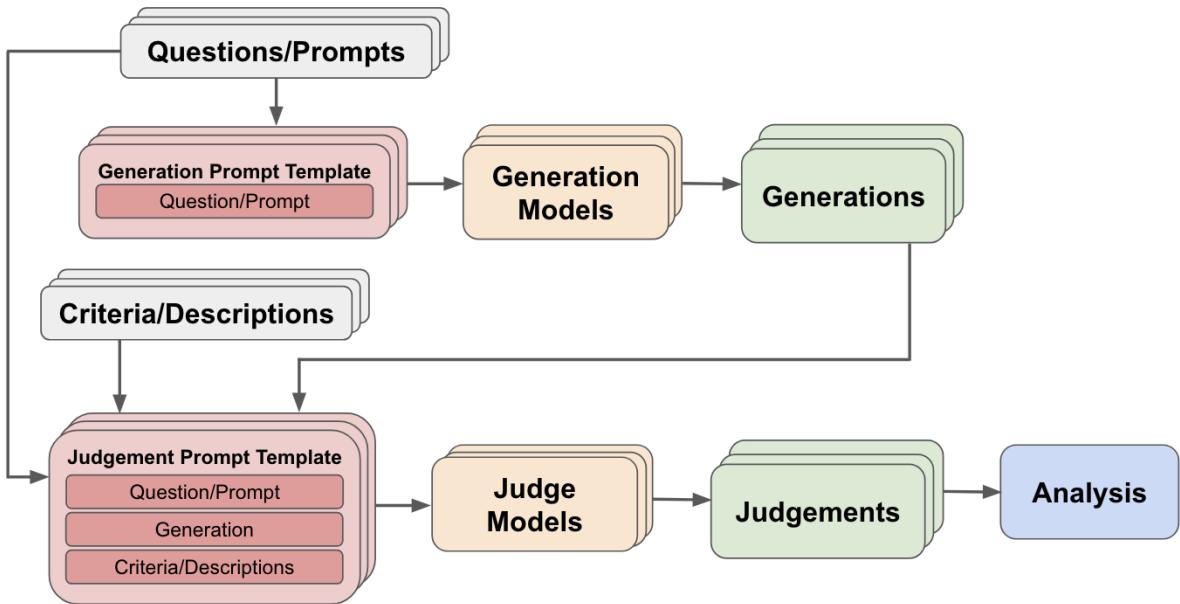


Figure 1: Diagram showing the two experiment phases: response generation and judgement generation.

information, please refer to Table 1 in Appendix A.

To run experiments with these models, we used a group member’s spare OpenRouter compute credits that were nearing expiration. Pricing for models varied, with our cheapest model, Phi-4, being provided at \$0.07/mTok input and \$0.14/mTok output, and our most expensive model, Claude Sonnet 3.7 (Thinking), being provided at \$3.00/mTok input and \$15.00/mTok output (a staggering >100x price increase over our cheapest model, for output tokens). Given Claude 3.7 Sonnet’s high cost per token, we omit it from the longer-winded creative writing experiments for budgetary reasons.

Evaluated models are all instruct-tuned and preference-tuned so as to enable preference-aligned instruction-following behavior in a helpful, harmless, and honest (HHH) manner. Evaluated models are accessed in their original precision, avoiding the loss in intelligence (and boost in inference profitability) that comes with model quantization. With respect to model diversity, we select an even balance of five open-weight models and five closed-source models. Models sizes (while not officially known for closed models) range from 12B parameters (i.e. Mistral NeMo 12B, an open-weight model that could run in a heavily-quantized manner on today’s high-end mobile devices) to >100B parameters. We select primarily recent frontier-level models that broadly lead the field in capability for their size and training FLOPS.

Both o3-mini and Claude 3.7 Sonnet (Thinking)

are examples of recent “reasoning” models, trained in reinforcement learning to generate long, meandering chain-of-thought reasoning before generating a user-presented response, scaling “inference-time compute” to generate meaningfully-better (and meaningfully more expensive) responses.

5 Experiments

At a high level, our experimental setup, shown pictorially in Figure 1, involves all models under evaluation first generating responses to all QA questions or CR prompts before rating the responses of every model (including itself) for each question/prompt, according to 6 domain-specific criteria shown in Table 3. Generation and judgement prompts are shown in Appendix Section C.

For our generation prompts, we kept the format as lightweight as possible in an effort to mimic the manner in which assistants would encounter users in real-world scenarios (e.g. the first user turn may only contain the question “How does carbon dating work,” with no preamble or specific instructions). In the question answering scenario, user ELI5 questions suffice as the entire user turn, whereas for the creative writing scenario, models are instructed to generate a creative writing story based on the provided writing prompt.

Our judgement prompt, shown in Figure 6, was meaningfully more involved, including a description of the model’s identity and goals, XML-tag-enclosed injected content (question, response,

judgement criterion, and criterion description), instructions, criteria, and output format requirements. The diligent separation of data, instruction, and criteria is seen as a best practice in language model prompting regardless of model architecture or origin. Models are instructed to first generate a thoughtful analysis of how well the provided response lives up to the given criterion before finally providing a numeric score in the 1-10 range indicating their ultimate judgement, with 1 indicating extremely poor performance and 10 indicating exceptional performance.

Each model completes a response to each question/prompt from each dataset before generating judgements for all available generations, leading to:

$$N = 50M + 50M^2C$$

Where N = number of language model calls, M = number of models, and C = number of criteria. This implies 30,500 language model calls for question answering and 24,750 calls for creative writing. It's an astounding feat of modern infrastructure engineering, language model development, and the venture capital funding environment that 55,250 calls to frontier models is affordable on a student budget.

5.1 Human Annotation

In addition to our selected language models acting as judges, we performed limited human judge annotation on both the creative writing and question answering generations. Due to time and sanity constraints, we were unable to perform human annotation for all 100 prompts. We evenly distributed 20 CR prompts and 20 QA among our four group members for annotation, with no annotator overlap. This meant that each groupmember annotated 600 question answering criteria-judgements and 570 creative writing criteria-judgements.

These human annotations served as a baseline against which we compared model judgements.

5.2 Decoding Parameters

When generating initial responses to questions/prompts, we use Nucleus Sampling with temperature=0.2 and P=0.8. The goal of this was to emulate a realistic scenario balancing diversity of generations while still maintaining high generation quality; such a decoding strategy is commonly used in literature and industry.

When generating judgement reasoning and judgement scores, we use greedy decoding, since the assigning of an ultimate score in a highly detail-oriented task demanding precision rather than creativity.

6 Results & Analysis

6.1 Expected Results

Before going over the results of our experiments, we will go over the expectations that we had for said results. Based on previous work on similar tasks and our own understandings of LLMs, we believed that the models would judge their own generations more highly than those of other language models. We also expected that our human annotators would be more critical of the generations than the language models were. We further expected that older models and, to a lesser extent, smaller models would receive lower scores from judges and be more lenient as judges themselves. We know from our own experiences as language model users and as of observed of performances on Chatbot Arena ([Chiang et al., 2024](#)) that models have improved with respect to human preference scoring over time, so we expect the newest, largest models to have the highest scores with respect to our human annotators.

6.2 Question Answering

For question answering, we used all 10 models to generate responses. The overall judgement scores given by the different models, excluding human judgement, was 8.22, as seen in Figure 37. On average, Relevancy, Clarity, and Coherence had scores that were mostly high (greater than 9.0) with one to three scores that were almost as high (between 8.0 and 9.0), which is evident in Figure 39. Creativity and Conciseness have an average lower score, with a majority of scores for Creativity being 5.0 or below, and Conciseness being 7.0 and below. The highest average score given by a judge model was 9.09, which was GPT 3.5 Turbo. Gemma 2 27B gave the lowest average scores, at 7.82. The full breakdown of scores given by each model can be seen in Table 8 of the Appendix.

Breaking this down further, we compared the highest and lowest received scores for each criteria. Gemini 2.0 Flash received the highest score in 4 out of 6 criteria, being outperformed in Conciseness and Factuality. The model received the lowest Conciseness score of all the models, and

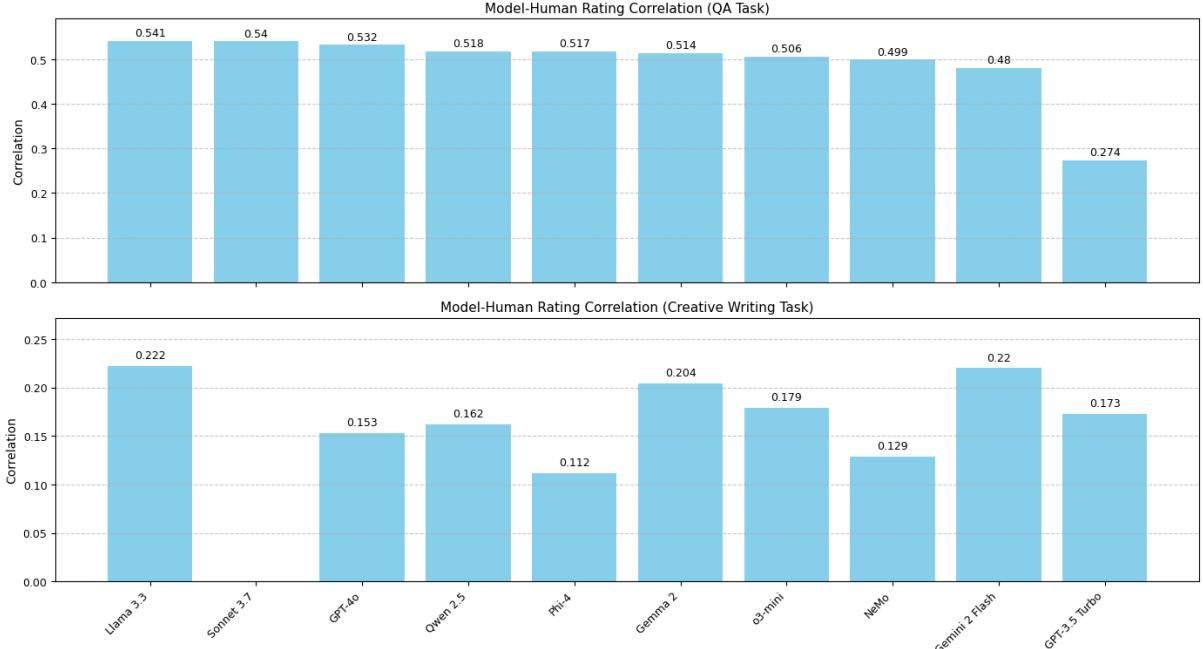


Figure 2: Overall correlation with human annotations for Question Answering and Creative Writing.

received an overall score of 8.25. This is highly impressive, given how cost-efficient the distilled model is; if our judges' ratings are strong predictors of model performance, an engineer in industry would be hard-pressed to justify another option. GPT 3.5 Turbo received the lowest score in 5 out of 6 criteria, and the model's average score across all criteria was also the lowest, at 7.93. This was highly predictable, given that this was the "oldest" model tested, having been introduced in November 2022. GPT 3.5 was an early pioneer in modern consumer-facing language models, building on InstructGPT's initial RLHF formulation for language models. Although Claude 3.7 Sonnet did not receive the highest average score for a single criterion, the model had the highest average score of 8.49, edging out Gemini 2.0 Flash by 0.20. Sonnet's strong performance comes as no surprise, as it's the most recent "full-size" (non-distilled, e.g. Gemini "Flash") model in our lineup that also takes advantage of recent RL-based "reasoning" post-training techniques. The full breakdown of scores received by each model can be found in Table 9 of the Appendix.

When it comes to self-rating, GPT 3.5 Turbo gave itself the highest scores, for an average of 8.82. While this could be significant, the model rated everyone highly, and this does not necessarily mean the model thought that it's own generations were actually the best. Despite having the high-

est average self-rating, GPT 3.5 Turbo only actually gave itself the highest score for two criteria - Conciseness and Creativity, which are the two categories with the lowest overall scores. Llama 3.3 70B gave itself more bad scores, averaging 7.82. Notably, Claude 3.7 Sonnet received the highest overall average scores, and it rated itself second best, with an average score of 8.56. These results are shown in Table 7.

As mentioned in the paragraph above, GPT 3.5 Turbo gave the most generous scores across all categories and models. The average score given by this model is 9.09. Gemma 2 27B gave the lowest overall scores at an average of 7.82. Notably, while GPT 3.5 Turbo was the most generous judge model, it never gave any model, itself included, a score of 10.0, but Gemini 2.0 Flash gave itself a 10.0 for Relevancy. Qwen 2.5 72B and o3-mini fall within 0.02 of the mean with averages of 8.24 and 8.20 respectively.

For most criteria, the distribution of scores lies in the range of 8.0 to 10.0, which is quite high. As mentioned before, Creativity and Conciseness have lower scores and wider ranges. Comparing the model evaluation to human evaluation, the scores given by human annotators are lower on average: 7.44. The combined score is 0.78 lower than the average score of models without human annotation. The distribution when the judge model is a human is wider, from 5.0 to 10.0, and conciseness and

Model	Clarity	Coherence	Conciseness	Creativity	Factuality	Relevancy
Claude 3.7 Sonnet	9.36	9.51	6.78	3.94	9.28	9.75
Gemini 2.0 Flash	9.80	9.69	7.56	3.24	9.74	9.96
Gemma 2 27B	9.42	9.37	6.38	3.01	9.25	9.48
GPT-4o	9.61	9.68	6.11	5.43	9.53	9.72
Llama 3.3 70B	8.92	9.35	5.88	4.70	9.31	9.68
Mistral NeMo	8.79	8.81	6.85	5.25	9.14	9.46
o3-mini	9.63	9.78	7.17	3.39	9.40	9.84
Phi-4	9.41	9.40	6.33	4.06	9.61	9.61
Qwen 2.5 72B	8.93	9.23	7.03	5.58	9.30	9.40
GPT-3.5 Turbo	9.46	9.42	9.07	8.05	9.37	9.18

Figure 3: Average scores given by different models for each evaluation criteria in a QA setting.

creativity scores exist within the full range 1.0 to 10.0.

The overall correlation with human judgements can be found in Figure 53. Llama 3.3 70B has the highest overall correlation between model and human ratings, with Claude 3.7 Sonnet just 0.001 worse. The range of correlation scores are mainly between 0.4800 and 0.5410. GPT 3.5 Turbo is an outlier, with the correlation at 0.2741, which is 0.2060 worse than the next best model, Gemini 2.0 Flash.

The average correlation between human and model judges is positive in general, but when analyzing correlation for Conciseness in Figure 56, GPT 3.5 Turbo actually has a negative correlation.

In general, the scores between humans and models may differ due to multiple things, such as the lack of guidelines or the subjectivity of the criteria. For example, conciseness and factuality may mean different things to different models. When annotating model generations, we saw responses where the scores might have differed according to just how inaccurate a response was. Getting an event right but the location wrong or completely hallucinating are two different kinds of inaccuracies that we, along with the models, had to consider when rating generations for Factuality. Without guidelines, models can decide how arbitrarily to score these two different things, leading to different scores, albeit similar reasoning.

6.3 Creative Writing

For the creative writing generations, we used 9 models – only excluding Claude 3.7 Sonnet for cost reasons – to generate our stories. The overall judgements given by the models were fairly high, with the average score of all criteria from all gen-

eration models being an 8.34 (see Figure 9). As can be seen in Table 4, the spread of judgements by criteria is quite uniform with the models scoring between a 6.41 and 9.44 on average for all criteria (and the vast majority lying in the 7.50 to 9.00 range). The only outliers are Character and Dialogue with only slightly lower scores overall. As a judge, GPT 3.5 Turbo is the most favorable, giving the highest score on average with a mean score of 9.24 given to the generations. This is an outlier when compared to the other models, but this lines up with our assumptions that the older models would give higher scores as judges and be more favorable towards the generations more than newer models. When we investigate the smaller subset of the data including human judgements we see that the human judges are much more critical of the model generations than the models are of themselves. The mean human judgement score is 1.21 points lower than the next judge model (Gemma 2 27B) as seen in Figure 10. Additionally, the distribution of human judge score varies much more than the judge models scores. As seen in Figure 12, the models are very prone to specific judgements, with the vast majority being between an 8 and a 10. On the contrary, the human judges utilize a broader spectrum of judgement ratings and are less monotonous. This invariability in judge model judgements shows that the models find that the majority of generations to be positive. Something that does not fully line up with human opinion.

From the perspective of received scores, the model with the lowest overall scores is GPT 3.5 Turbo with an average score of 7.54 (see Figure 7). On all individual criteria, GPT 3.5 Turbo received the lowest scores by at least 0.41 points when compared to the next lowest model for each criteria

(see Figure 8). The most liked model overall is Gemini 2.0 Flash which had an average rating just 0.07 points higher than the next model. Overall, the spread of the judgements received by models is very small with the mean score for the generator models being within 1.20 points of each other.

When judging themselves, models like their own judgements more. As can be seen in Figure 13, the same judge models are giving the lowest and highest average scores as seen previously (Gemma 2 27B and GPT 3.5 Turbo). However the average self judgement score of 8.54 is 0.2 points higher than the average judge model score when considering all generation models. This significant bump seems to show the bias that these models have when judging their own generations. While this is partially true, Figure 14 shows a slightly different story. The only model with very significant differences in overly positive self ratings is GPT 3.5 Turbo. Because the other models see less of this undesirable behavior it points to the possibility that newer models continue to improve judge models’ abilities to be subjective judges.

After further investigation of results by individual criteria, some expected results arise. The Character and Plot criteria are positively correlated with length, this makes sense as if the story is longer, these characteristics are given more time to develop. Overall the most widely distributed criteria was Character, with judge models ranking generations very highly but other very low as well. This points to the character development being varied in quality across the stories.

When comparing our human judgements with the model judgements for the creative writing task, we see positive correlations across all models and criteria. As can be seen in Figure 2, the model with the highest correlation with humans annotators is Llama 3.3. The correlations in this domain are much lower for any given model than they are in QA. We believe this is due to the more subjective domain of creative writing. It is more difficult as humans to assess the criteria that we chose and even more so for a language model.

7 Conclusion and Future Work

Our judge models were largely enthusiastic about the generations, with newer models like Gemma 2 27B Instruct being more critical and older models like GPT 3.5 Turbo being more lenient.

We found that the human annotations were more

diverse in scoring and used a broader range of values. This contrast highlights a potential over-optimism bias in model judges and underscores the need for calibration if LLMs are to replace or augment human evaluation at scale. The results also reveal that models tend to prefer their own generations, a bias observed in both pairwise and scalar judgments. Moreover, correlations between output length and specific criteria scores (e.g., plot and character) suggest that some models may implicitly associate verbosity with quality.

In the future, we would like to continue this investigation by broadening the amount of data that we used for generation and judging to increase the statistical significance of any findings. Additionally, future work comparing different decoding parameters used for generation and judgement would be interesting.

8 Discussion of Limitations

Limited Samples: Due to our limited budget, we were only able to generate responses and judgements for curated 50 - question subsets of each of our original datasets. While this resulted ultimately in >50k criteria-judgements, for more diversity in prompts/questions, it would have been nice to be able to afford to generate responses for more questions.

Limited Models: We accessed our models through OpenRouter, which is a proxy to inference providers like TogetherAI, Novita, DeepInfra, and others. Inference provision is a red-hot area of activity and is highly competitive, with each provider endeavoring to maximize utilization of their available compute. The downstream consequence of this is that only highly-sought-after models are generally available for inference. It’s likely that these frontier labs are optimizing for similar popular general-purpose benchmarks, which likely narrows the design space of model architectures, datasets, and post-training techniques to one that produces largely similarly-charactered models. Furthermore, OpenRouter inference providers are largely biased towards western audiences and rarely serve capable models from other countries with different value systems, such as excellent models from China’s Bytedance, Baidu, 01-AI, and DeepSeek. DeepSeek R1 and Qwen QwQ 32B are extremely interesting and capable recent Chinese reasoning models that we were unable to include in our experiment because OpenRouter inference providers

weren't able to reliably serve them at the scale we needed, often timing out and returning error status codes regardless of provider.

Limited Domains: While our choice of creative writing and question-answering were grounded in a belief that they offered fertile ground to understand differences in judge characteristics, it's likely that an exploration of more datasets and more domains would yield interesting results. Both general creative writing and question-answering are highly "in-domain" from the perspective of the majority of Chatbot Arena traffic, which model developers are known to highly optimize for. Perhaps it would be interesting to test models on more "reasoning"-oriented datasets like ZebraLogic from AI2, alignment-oriented datasets like Anthropic's HH dataset, or on specialized medical or science datasets for which there are objectively correct answers but room to express reasoning in more or less interesting and effective ways.

Statistical Significance: As noted during our paper presentation, we present interesting observations and descriptive statistics showing the supposed difference (or lack of difference, in many cases) of various LM judges, but do not (yet) support these observation with strong claims backed by measures of statistical significance like the non-parameteric Wilcox Signed-Rank test for hypothesis testing. Due to the limited time for this project and brief treatment we received in 203 on statistical significance testing (primarily focused on parametric statistical testing like Student's t-test), we feel it reasonable to leave this as future work for this project, but acknowledge the importance of such testing for future work.

Human Annotations: For expediency, our human annotators annotated disjoint segments of our datasets, with judgements being wrapped under a single "human" judge model. We didn't expect that our human annotators opinions of LM-generated text would vary as much as they did; we saw significantly more inter-human disagreement than we saw inter-model disagreement, making comparison via correlation between the singular "human" annotator and any given judge language model *directionally* correct but fuzzy at the detail-level.

9 Ethical Considerations

Our research on LM-as-a-Judge is primarily an evaluation of existing capabilities, but development of judge systems and reward models touches on

several important ethical implications:

Language Model Bias: As mentioned in the introduction and related works sections, language models (whether employed as generators or as judges) carry with them the biases implicit in their training datasets. Some biases, like the "compassion fade" bias (judging differently when given the name of the generating model) seem relatively unlikely to harm society, while others like the "bandwagon bias" (giving preference to majority opinions) or "sentiment bias" (showing preference towards certain demographic groups) portend social harm if not better understood and rectified or designed around.

Reliability and Trust: LM Judge systems that unroll Chain of Thought reasoning before providing a judgement are seen as interpretable and highly-aligned with human preferences, but the research community needs more research to understand to what degree generated rationales represent robust first-principles reasoning rather than fluent pre-rationnalizations of latently-held priors based on biases and heuristics.

Dual Use Considerations: Language models are a dual-use technology, as useful for therapy and other pro-social applications as they are for digital propaganda and other nefarious applications. Development of any assistive tools in the LM system toolchain that improve capabilities need to soberly reckon with possible negative uses of such technology.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. [Phi-4 technical report](#).

Aitor Arrieta, Miriam Ugarte, Pablo Valle, José Antonio Parejo, and Sergio Segura. 2025. [Early external safety testing of openai's o3-mini: Insights from the pre-deployment evaluation](#).

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. [Measuring political bias in large language models: What is said and how it is said](#).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E.

- Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback.
- DeepSeek-AI and Aixin Liu et al. 2024. Deepseek-v3 technical report.
- DeepSeek-AI and Daya Guo et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.
- Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpaca-farm: A simulation framework for methods that learn from human feedback.
- K. O. T. Erziev. 2025. à la recherche du sens perdu: your favourite llm might have more to say than you can understand.
- Aaron Grattafiori et al. 2024. The llama 3 herd of models.
- Yuntao Bai et al. 2022. Constitutional ai: Harmlessness from ai feedback.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.
- Sarah E. Finch, Ellie S. Paek, and Jinho D. Choi. 2023. Leveraging large language models for automated dialogue analysis.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-ku Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillem Lampe, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling.
- Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harry Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment.
- Mistral AI. 2024. Mistral nemo: our new best small model. <https://mistral.ai/en/news/mistral-nemo>. Accessed July 2024.
- OpenAI and Josh Achiam et al. 2024. Gpt-4 technical report.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John

- Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Quora. 2017. Quora question pairs dataset. <https://www.kaggle.com/datasets/quora/question-pairs-dataset>. Accessed February 2024.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don't know: Unanswerable questions for squad](#).
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. [Bleurt: Learning robust metrics for text generation](#).
- Gemini Team and Rohan Anil et al. 2024a. [Gemini: A family of highly capable multimodal models](#).
- Gemma Team and Morgane Riviere et al. 2024b. [Gemma 2: Improving open language models at a practical size](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#).
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. [Justice or prejudice? quantifying biases in llm-as-a-judge](#).
- Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhua Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. [A comprehensive capability analysis of gpt-3 and gpt-3.5 series models](#).
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. [Self-rewarding language models](#).
- Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. [Textgrad: Automatic "differentiation" via text](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#).
- 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. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#).
- Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. [Starling-7b: Increasing llm helpfulness & harmlessness with rlaif](#). Accessed April 7, 2025.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. [Judgelm: Fine-tuned large language models are scalable judges](#).

Appendices

A Dataset and Model Details

Model	Make	Weight Type	Parameters	Tokens	Release Date
Gemini 2.0 Flash	Google	Closed	?	?	February 2025
Sonnet 3.7 (Thinking)	Anthropic	Closed	?	?	February 2025
o3-mini	OpenAI	Closed	?	?	January 2025
Qwen 2.5 72B	Alibaba	Open	72B	18T	December 2024
Llama 3.3 70B	Meta	Open	70B	15T	December 2024
Phi-4	Microsoft	Open	14B	9.8T	December 2024
Mistral NeMo 12B	Mistral	Open	12B	?	July 2024
Gemma 2 27B	Google	Open	27B	13T	June 2024
GPT 4o	OpenAI	Closed	?	?	May 2024
GPT 3.5 Turbo	OpenAI	Closed	Est. 175B	Est. 400B	September 2023

Table 1: Comparison of language models by source availability, parameter size, training tokens, and release date. Models are ordered by release date, descending. All models are accessed via OpenRouter interfaces in their original precision.

Table 2: Dataset Statistics

Dataset	Train Size	Validation Size	Test Size	Total Size
WritingPrompts	273k	15.6k	15.1k	303.7k
ELI5	325k	0	0	325k

B Judgement Criteria

QA Criteria	Creative Writing Criteria
Relevancy: Does this response answer the question?	Plot: How compelling and well-structured is the narrative?
Factuality: Is this response factually correct and verifiable?	Character: How well-developed and believable are the characters?
Clarity: Is this response clear and easy to understand?	Style: How distinctive and effective is the writing style?
Creativity: Is this a creative response?	Emotion: How effectively does the story evoke feelings?
Coherence: Is this response coherent and well-structured?	Flow: How smooth and well-paced is the storytelling?
Conciseness: Is this response concise and to the point?	Dialogue: How natural and purposeful is the conversation and narration?

Table 3: Criteria and Criteria Descriptions for question answering and creative writing domains.

C Prompts

```
{ question }
```

Figure 4: Prompt template used for response generation in the question-answering domain.

```
Given the following creative writing prompt, generate a short story on the topic :
```

```
<prompt>
{ question }
</prompt>
```

Figure 5: Prompt template used for response generation in the creative writing domain.

You are an AI language model tasked with evaluating responses to open-domain questions. Your goal is to assess the quality of a given response based on a specific criterion. Please carefully read the question, response, and evaluation criterion before providing your assessment.

Question to be answered:

```
<question>
{ question }
</question>
```

Response to be evaluated :

```
<response>
{ response }
</response>
```

You will evaluate this response based on the following criterion :

```
<criterion>
{ criterion }: { criterion_description }
</criterion>
```

Please consider the response carefully in light of this criterion. Provide a thoughtful analysis of how well the response lives up to the given criterion. Be objective and fair in your assessment, considering both strengths and weaknesses of the response.

After your analysis, assign a score on a scale from 1 to 10, where:

1 = Extremely poor performance on this criterion
5 = Average performance on this criterion
10 = Exceptional performance on this criterion

Use the full range of the scale as appropriate. Provide your answer rounded to the nearest tenths place. Don't hesitate to give high or low scores if they are warranted, but also don't be afraid to use middle range scores if that best reflects the quality of the response.

Present your evaluation in the following format:

```
<evaluation>
<reasoning>
[Your detailed reasoning here]
</reasoning>
<score>
[Your numerical score here]
</score>
</evaluation>
```

Remember, your task is to evaluate this specific response based solely on the given criterion. Do not compare it to other potential responses or speculate about the capabilities of the model that generated it. Focus on providing a fair, objective assessment based on the content of the response and how well it meets the specified criterion.

Figure 6: Prompt template used for per-criteria judgement of generations.

D Creative Writing Figures

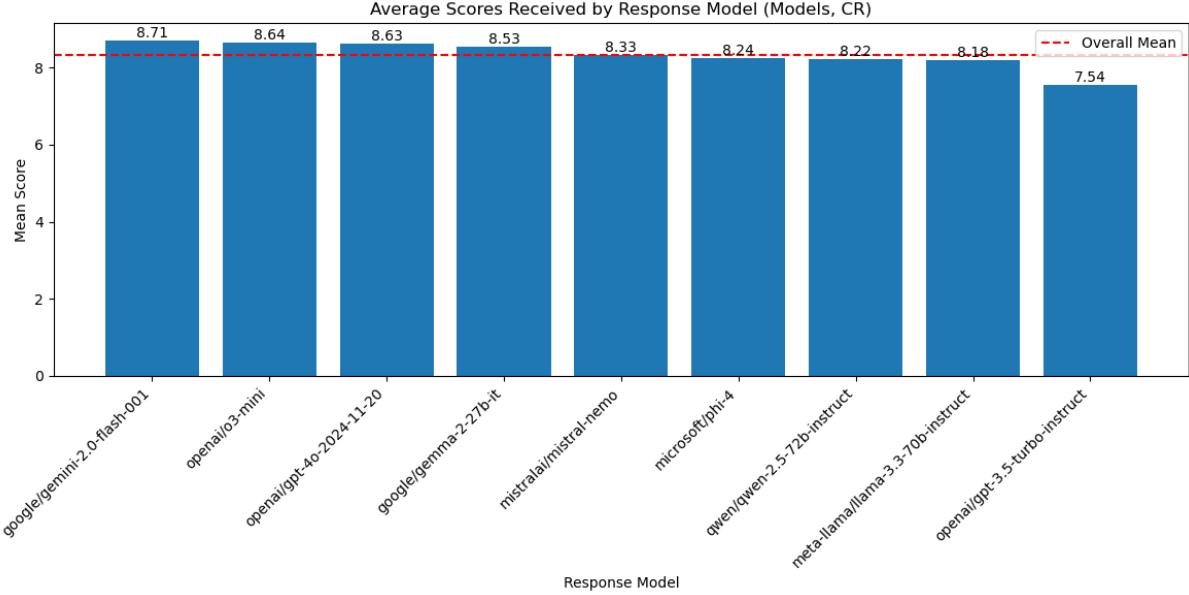


Figure 7: Average Judgements received by Generated Models (Full 50 Questions)

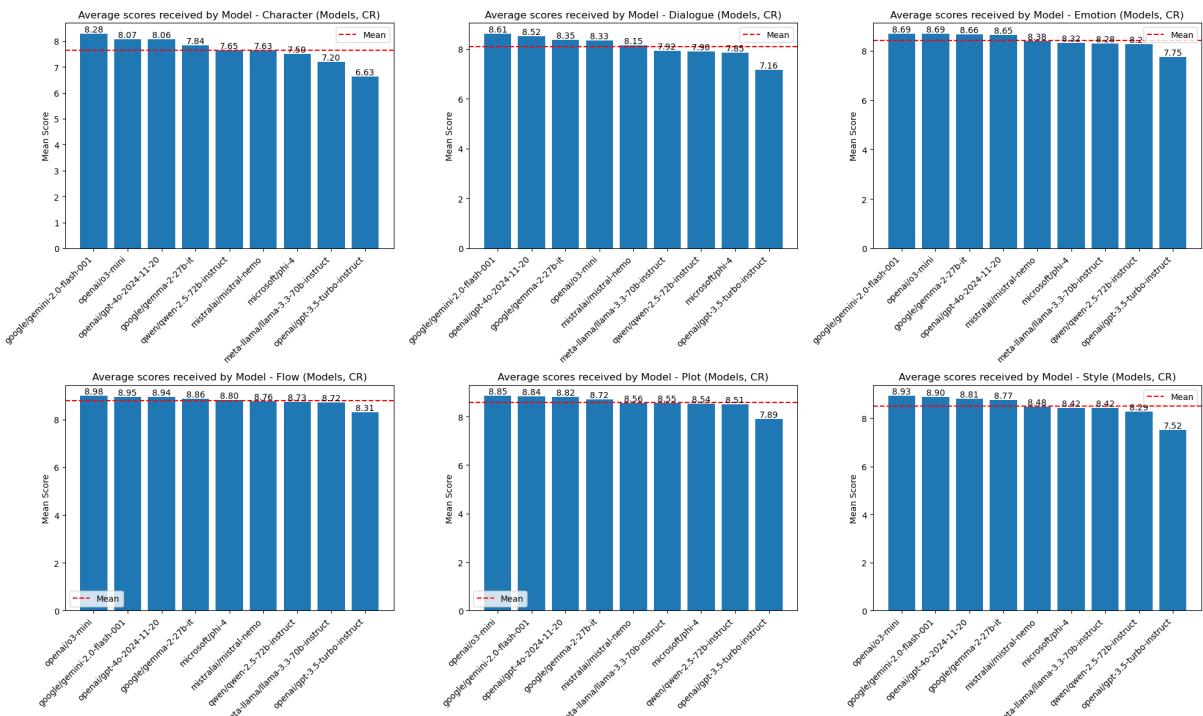


Figure 8: Average Judgements received by Generated Models by criteria (Full 50 Questions). Domain: **Creative Writing**

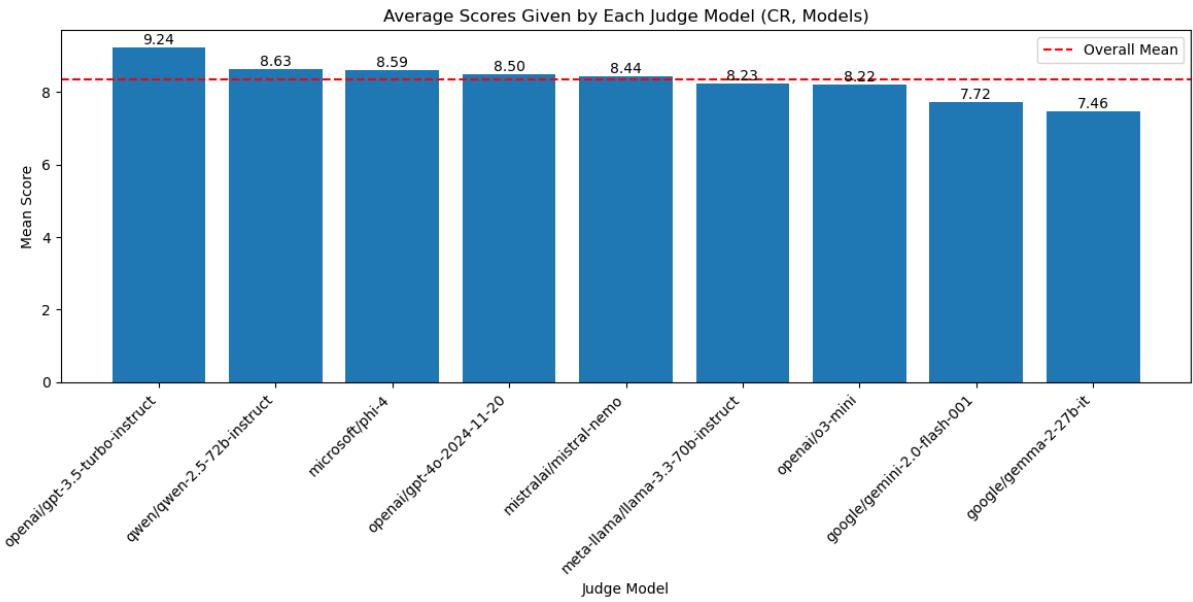


Figure 9: Average Judgements given by Judge Models (Full 50 Questions. Domain: **Creative Writing**)

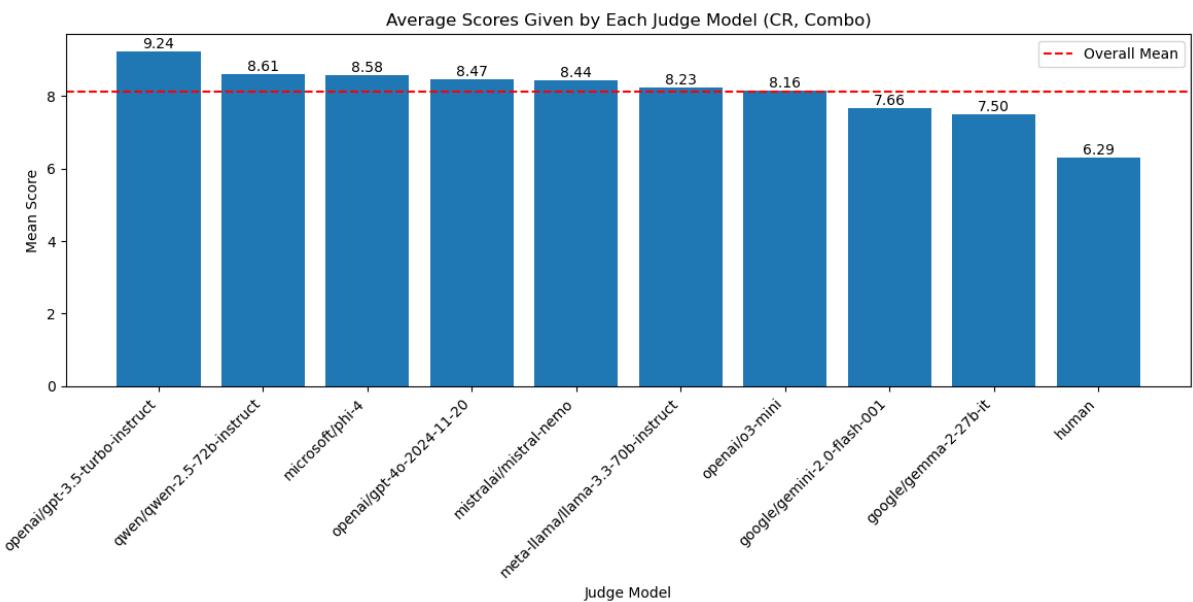


Figure 10: Average Judgements given by Judge Models and Human Judges (20-Question Annotated Subset). Domain: **Creative Writing**

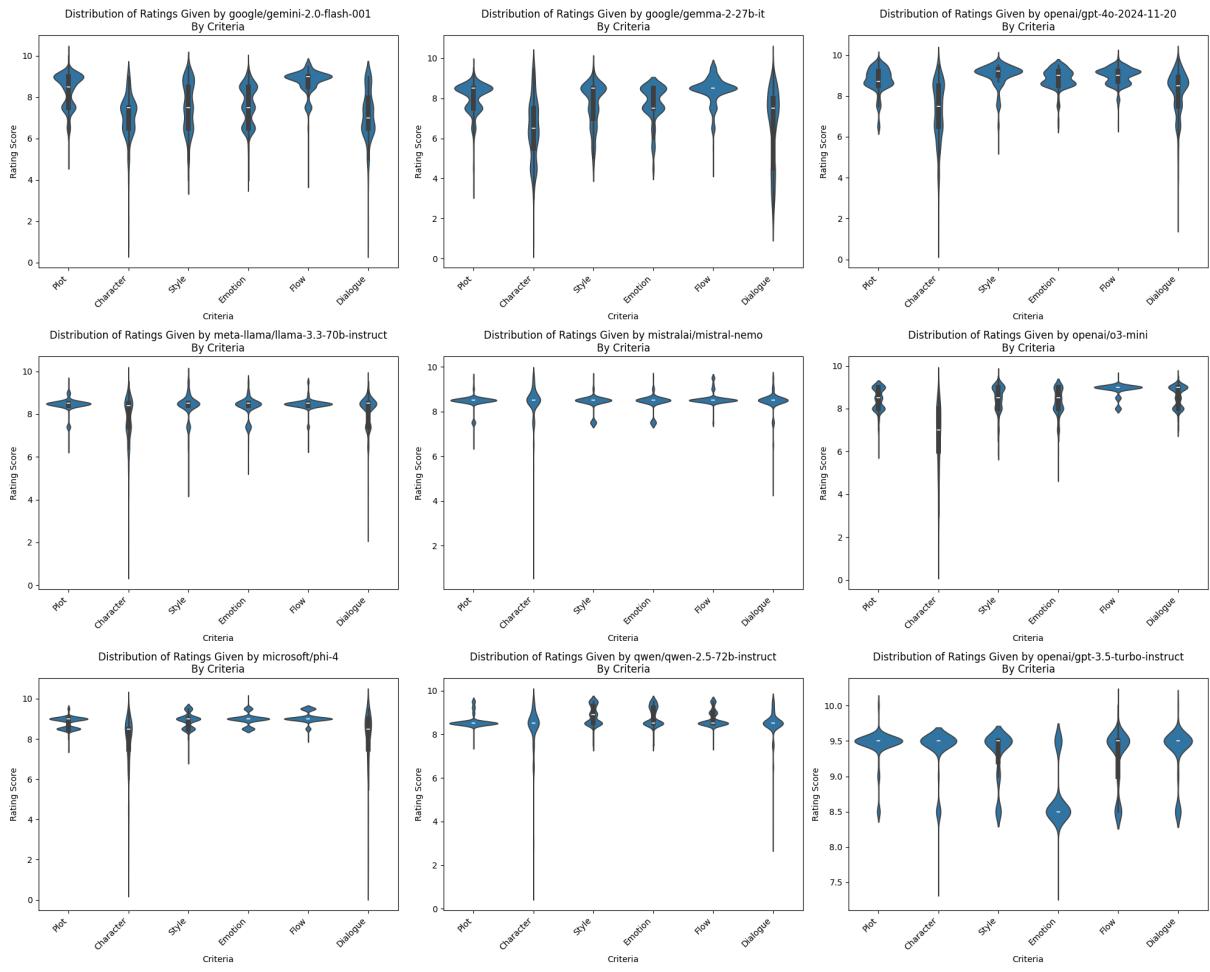


Figure 11: Distribution of Judgements by Criteria. Domain: **Creative Writing**

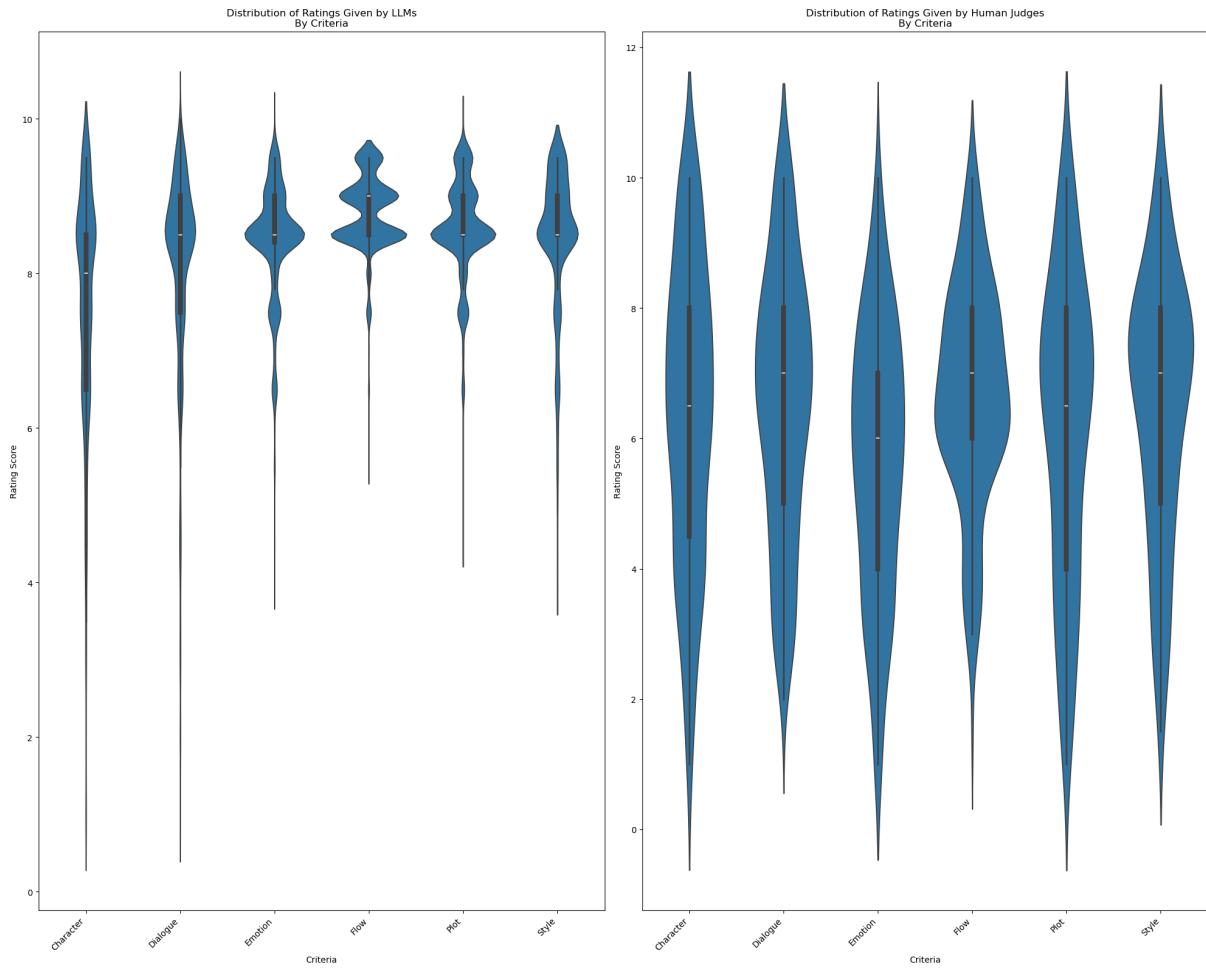


Figure 12: Distribution of Judgements for Judge Models and Human Judges. Domain: **Creative Writing**

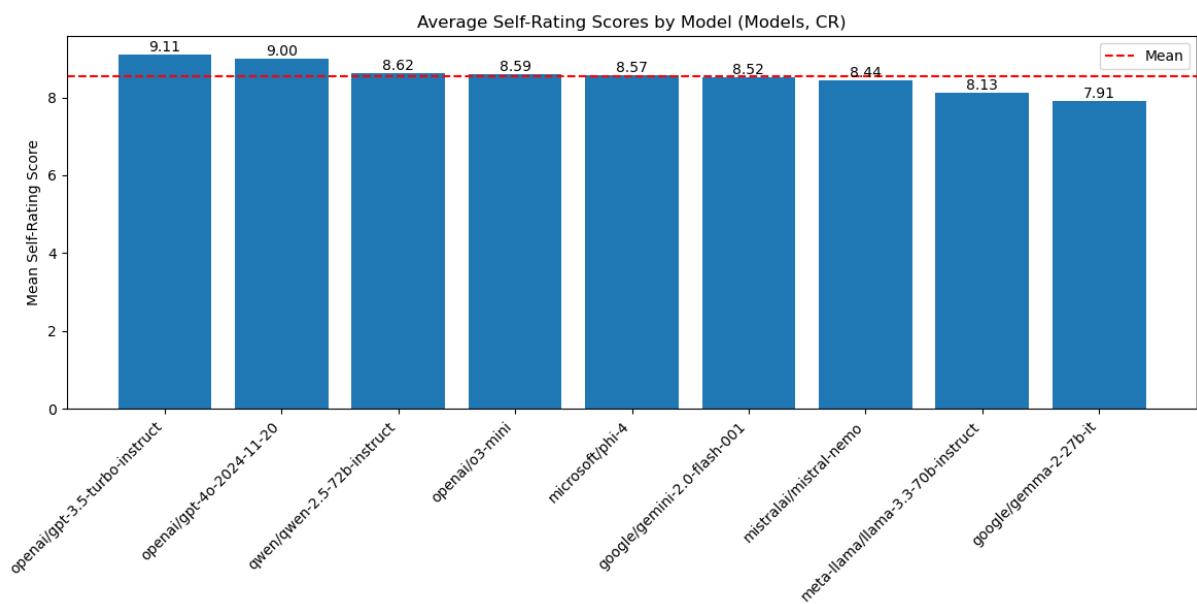


Figure 13: Judgements given by Judge models to their own generations. Domain: **Creative Writing**



Figure 14: Average Ratings given to a Model's own Generations vs Other Model's Generations Domain: **Creative Writing**

Model	Character	Dialogue	Emotion	Flow	Plot	Style
Gemini 2.0 Flash	6.96	7.11	7.63	8.74	8.34	7.56
Gemma 2 27B	6.41	6.60	7.70	8.39	7.99	7.68
GPT 4o	7.24	8.16	8.83	8.98	8.77	9.00
Llama 3 70B	7.68	8.13	8.32	8.45	8.43	8.36
Mistral NeMo 12B	8.39	8.45	8.40	8.58	8.45	8.39
o3-mini	6.69	8.57	8.28	8.86	8.44	8.46
Phi-4	7.81	7.94	8.97	9.07	8.82	8.91
Qwen 2.5 72B	8.30	8.48	8.82	8.72	8.60	8.86
GPT 3.5 Turbo	9.38	9.33	8.72	9.26	9.44	9.30

Table 4: Average ratings **given** by different models across six creative writing (CR) criteria. Domain: **Creative Writing**

Model	Character	Dialogue	Emotion	Flow	Plot	Style
Gemini 2.0 Flash	8.28	8.61	8.69	8.95	8.84	8.90
Gemma 2 27B	7.84	8.35	8.66	8.86	8.72	8.77
GPT 4o	8.06	8.52	8.65	8.94	8.82	8.81
Llama 3.3 70B	7.20	7.92	8.28	8.72	8.55	8.42
Mistral NeMo 12B	7.63	8.15	8.38	8.76	8.56	8.48
o3-mini	8.07	8.33	8.69	8.98	8.85	8.93
Phi-4	7.50	7.85	8.32	8.80	8.54	8.42
Qwen 2.5 72B	7.65	7.90	8.27	8.73	8.51	8.29
GPT 3.5 Turbo	6.63	7.16	7.75	8.31	7.89	7.52

Table 5: Average ratings received by different models across six creative writing (CR) criteria. Domain: **Creative Writing**

Model	Character	Dialogue	Emotion	Flow	Plot	Style
Gemini 2.0 Flash	7.94	8.16	8.39	9.09	8.90	8.66
Gemma 2 27B	6.86	7.26	8.26	8.53	8.29	8.24
Llama 3.3 70B	7.29	7.96	8.26	8.44	8.45	8.36
Phi-4	7.73	7.78	9.00	9.18	8.78	8.95
Mistral NeMo 12B	8.38	8.47	8.41	8.57	8.47	8.32
GPT 3.5 Turbo	9.21	9.21	8.67	9.22	9.25	9.11
GPT 4o	8.04	8.98	9.14	9.28	9.27	9.29
o3-mini	7.30	8.80	8.74	9.00	8.77	8.94
Qwen 2.5 72B	8.42	8.48	8.70	8.67	8.60	8.84

Table 6: Average self-rating scores given by each model for creative writing (CR) evaluation criteria. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	8.69	8.43	7.78	8.30	8.36	7.60	8.51	8.58	8.42	8.30
Dialogue	8.70	8.60	8.29	8.51	8.56	7.97	8.65	8.60	8.48	8.48
Emotion	9.09	9.03	8.71	8.77	8.67	8.56	8.95	8.94	8.70	8.82
Flow	8.83	8.68	8.62	8.70	8.68	8.46	8.93	8.96	8.67	8.72
Plot	8.74	8.66	8.56	8.55	8.56	8.50	8.68	8.59	8.60	8.60
Style	9.13	9.06	8.78	8.84	8.74	8.40	9.08	8.90	8.84	8.86
Average	8.86	8.74	8.46	8.61	8.60	8.25	8.80	8.76	8.62	8.63

Figure 15: This table shows how **Qwen 2.5 72B** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	7.94	7.31	6.46	6.58	6.86	6.04	7.27	7.31	6.89	6.96
Dialogue	8.16	7.57	6.86	6.43	7.06	5.89	7.73	7.73	6.59	7.11
Emotion	8.39	8.30	7.49	7.08	7.43	6.53	8.13	8.14	7.19	7.63
Flow	9.09	8.95	8.76	8.68	8.70	7.91	8.93	9.02	8.60	8.74
Plot	8.90	8.81	8.37	8.01	8.27	7.01	8.83	8.88	7.98	8.34
Style	8.66	8.27	7.26	6.97	7.32	5.79	8.15	8.75	6.89	7.56
Average	8.52	8.20	7.53	7.29	7.61	6.53	8.17	8.31	7.36	7.72

Figure 16: This table shows how **Gemini 2.0 Flash** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	7.81	6.86	5.66	5.91	6.22	4.80	7.13	7.53	5.75	6.41
Dialogue	8.01	7.26	6.23	5.73	7.02	4.30	7.91	6.80	6.16	6.60
Emotion	8.20	8.26	7.53	7.60	7.76	6.19	8.22	8.25	7.33	7.70
Flow	8.75	8.53	8.15	8.40	8.41	7.33	8.85	8.89	8.21	8.39
Plot	8.50	8.29	7.92	7.88	8.07	6.50	8.50	8.51	7.78	7.99
Style	8.44	8.24	7.51	7.41	7.75	5.72	8.49	8.57	7.01	7.68
Average	8.29	7.91	7.17	7.16	7.54	5.81	8.18	8.09	7.04	7.46

Figure 17: This table shows how **Gemma 2 27B** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	8.05	7.43	6.66	6.99	7.23	5.72	8.04	7.58	7.43	7.24
Dialogue	8.82	8.40	7.87	7.88	8.03	7.07	8.98	8.37	8.05	8.16
Emotion	9.11	8.99	8.62	8.80	8.82	8.06	9.14	9.18	8.78	8.83
Flow	9.26	9.08	8.85	9.04	8.89	8.27	9.28	9.26	8.87	8.98
Plot	9.20	8.87	8.65	8.79	8.74	7.50	9.27	9.25	8.65	8.77
Style	9.32	9.22	8.93	9.13	9.07	7.75	9.29	9.35	8.96	9.00
Average	8.96	8.66	8.26	8.44	8.46	7.39	9.00	8.83	8.46	8.50

Figure 18: This table shows how **GPT 4o** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	8.36	7.82	7.29	7.47	7.60	6.48	8.26	8.09	7.71	7.68
Dialogue	8.58	8.42	7.96	7.98	8.23	6.97	8.54	8.42	8.10	8.13
Emotion	8.61	8.56	8.26	8.30	8.39	7.40	8.58	8.55	8.22	8.32
Flow	8.57	8.55	8.44	8.47	8.46	8.03	8.58	8.51	8.48	8.45
Plot	8.58	8.52	8.45	8.44	8.50	7.84	8.57	8.59	8.42	8.43
Style	8.70	8.60	8.36	8.35	8.46	7.13	8.69	8.67	8.25	8.36
Average	8.57	8.41	8.13	8.17	8.27	7.31	8.54	8.47	8.20	8.23

Figure 19: This table shows how **Llama 3.3 70B** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	8.61	8.50	8.08	8.33	8.38	7.89	8.67	8.60	8.48	8.39
Dialogue	8.49	8.51	8.37	8.44	8.47	8.00	8.62	8.57	8.55	8.45
Emotion	8.52	8.49	8.24	8.33	8.41	8.27	8.44	8.46	8.40	8.40
Flow	8.64	8.62	8.58	8.55	8.57	8.40	8.58	8.69	8.58	8.58
Plot	8.53	8.52	8.44	8.41	8.47	8.14	8.51	8.53	8.48	8.45
Style	8.55	8.46	8.30	8.29	8.32	8.18	8.55	8.53	8.30	8.39
Average	8.56	8.52	8.33	8.39	8.44	8.15	8.56	8.56	8.47	8.44

Figure 20: This table shows how **Mistral NeMo 12B** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	7.43	6.84	6.11	6.75	6.67	5.33	7.01	7.30	6.78	6.69
Dialogue	8.82	8.70	8.57	8.62	8.39	8.01	8.68	8.80	8.51	8.57
Emotion	8.58	8.51	8.14	8.27	8.34	7.34	8.49	8.74	8.13	8.28
Flow	8.93	8.97	8.84	8.96	8.80	8.41	8.91	9.00	8.89	8.86
Plot	8.68	8.49	8.35	8.59	8.26	7.74	8.64	8.77	8.44	8.44
Style	8.81	8.63	8.49	8.53	8.50	7.21	8.70	8.94	8.31	8.46
Average	8.54	8.36	8.08	8.29	8.16	7.34	8.40	8.59	8.18	8.22

Figure 21: This table shows how **o3-mini** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	8.27	7.96	7.33	7.73	7.91	6.64	8.19	8.31	7.93	7.81
Dialogue	8.52	8.35	7.88	7.78	8.18	7.03	8.12	8.17	7.43	7.94
Emotion	9.04	9.09	8.83	9.00	8.95	8.72	9.03	9.16	8.92	8.97
Flow	9.14	9.13	9.01	9.18	9.05	8.81	9.09	9.19	9.02	9.07
Plot	8.96	8.86	8.77	8.78	8.77	8.53	8.88	9.02	8.78	8.82
Style	9.08	9.05	8.82	8.95	8.90	8.36	9.02	9.24	8.76	8.91
Average	8.83	8.74	8.44	8.57	8.63	8.02	8.72	8.85	8.47	8.59

Figure 22: This table shows how **Phi-4** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

Criteria	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Character	9.37	9.40	9.38	9.42	9.40	9.21	9.46	9.37	9.43	9.38
Dialogue	9.36	9.29	9.24	9.31	9.37	9.21	9.48	9.48	9.26	9.33
Emotion	8.71	8.71	8.72	8.76	8.64	8.67	8.84	8.74	8.72	8.72
Flow	9.33	9.22	9.23	9.21	9.28	9.22	9.28	9.31	9.21	9.26
Plot	9.49	9.44	9.44	9.44	9.43	9.25	9.46	9.50	9.47	9.44
Style	9.38	9.39	9.31	9.29	9.23	9.11	9.35	9.39	9.26	9.30
Average	9.27	9.24	9.22	9.24	9.23	9.11	9.31	9.30	9.23	9.24

Figure 23: This table shows how **GPT 3.5 Turbo** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Creative Writing**

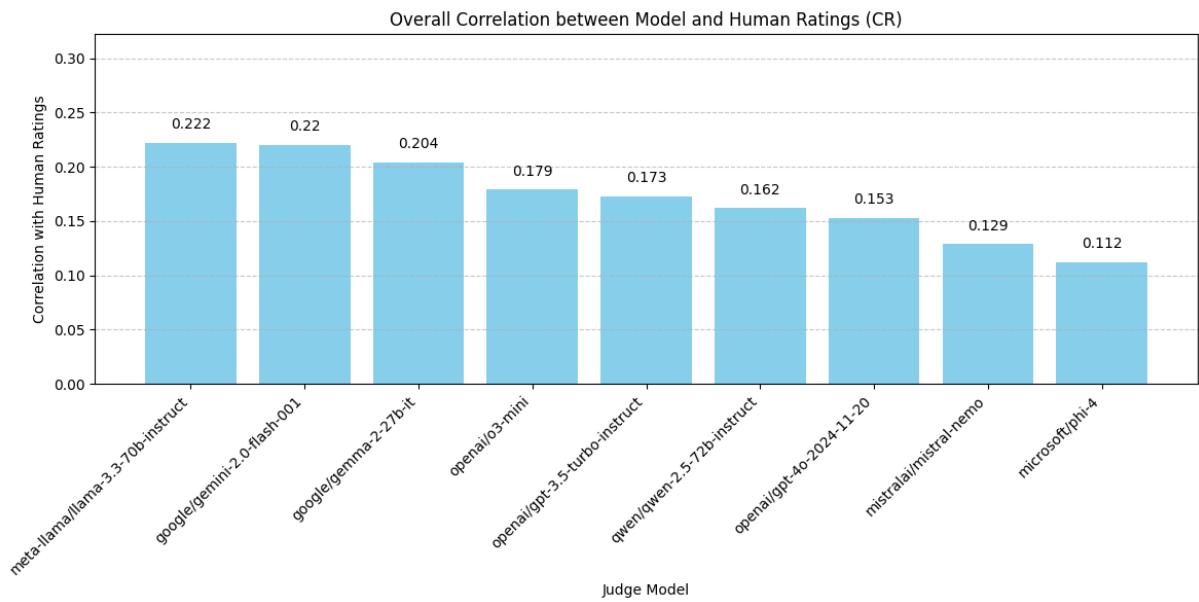


Figure 24: Showing overall correlation with human judgements for models in creative writing (CR).

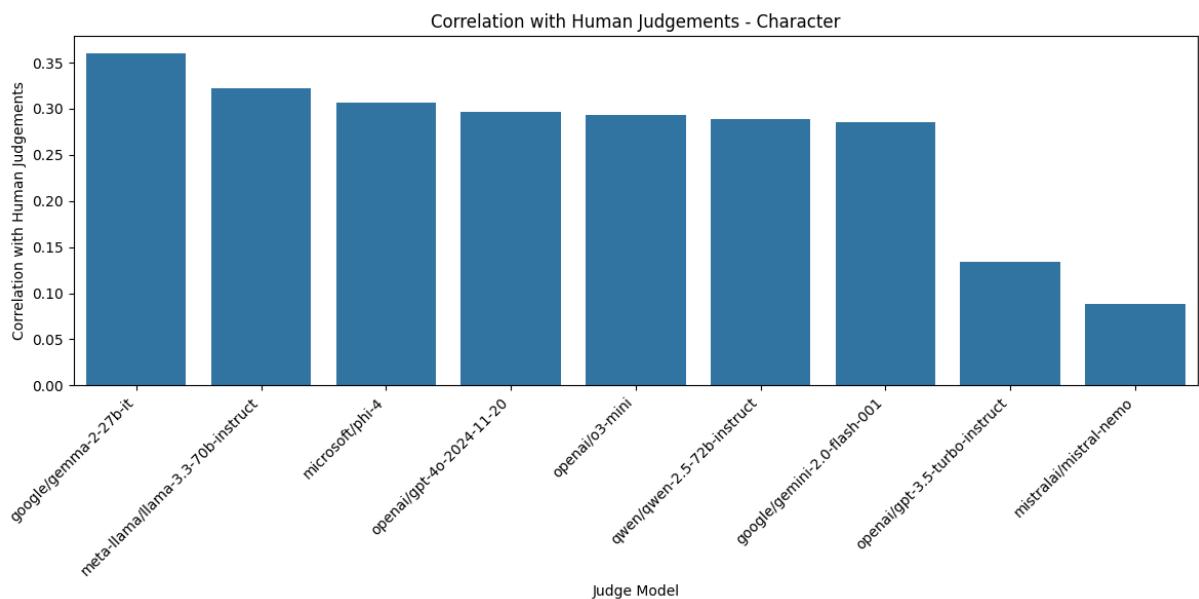


Figure 25: Showing correlation with human judgements for the Character criteria in CR

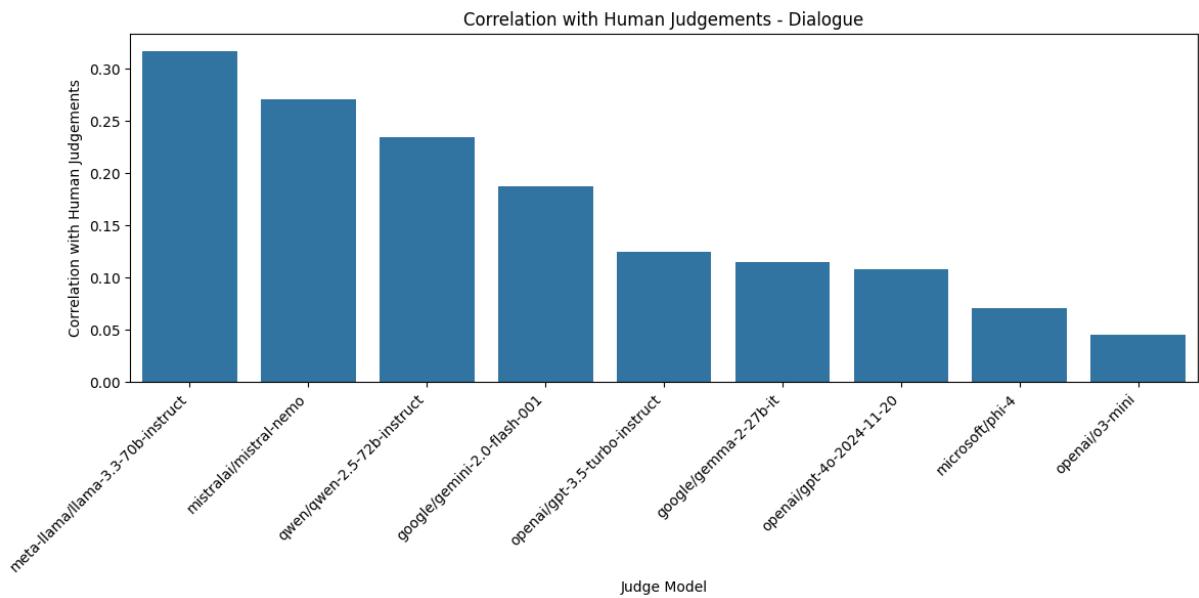


Figure 26: Showing correlation with human judgements for the Dialogue criteria in CR

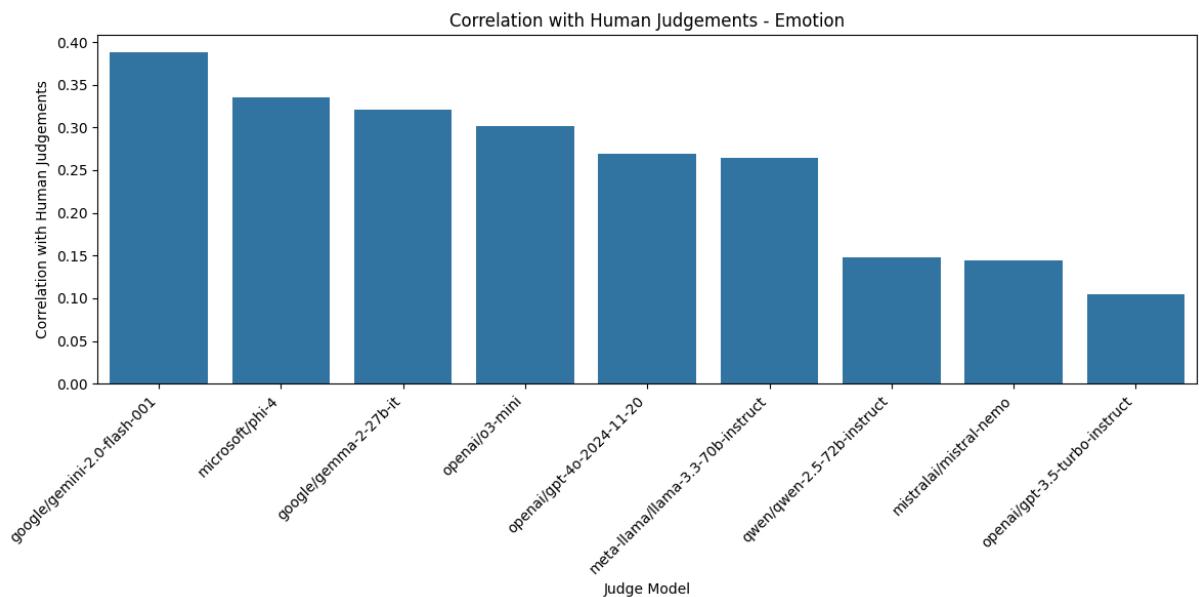


Figure 27: Showing correlation with human judgements for the Emotion criteria in CR

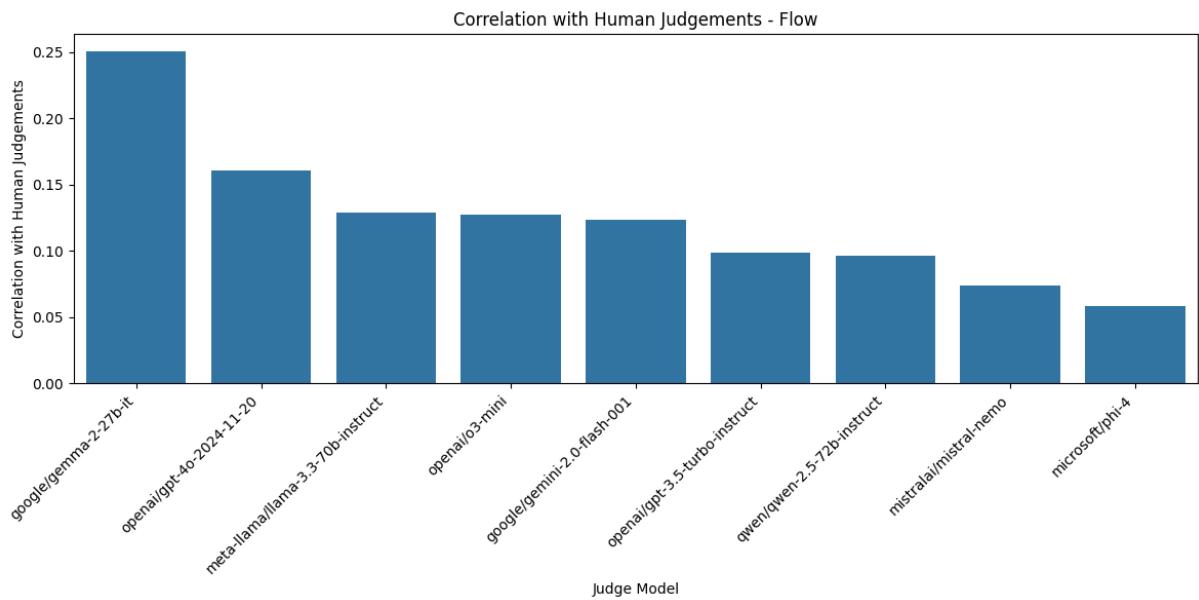


Figure 28: Showing correlation with human judgements for the Flow criteria in CR

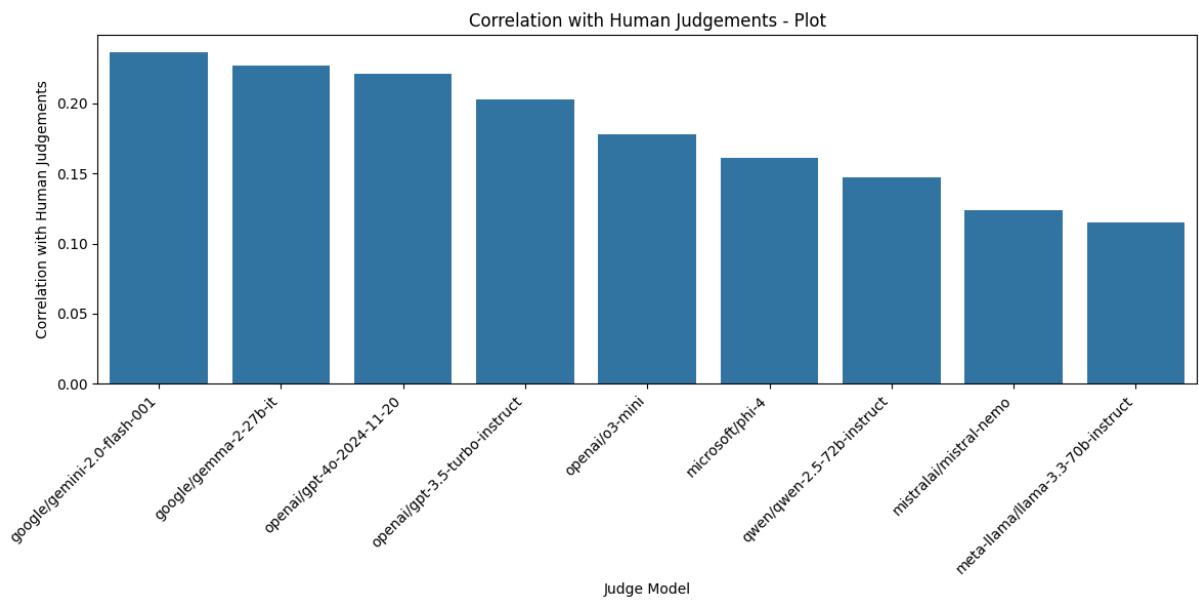


Figure 29: Showing correlation with human judgements for the Plot criteria in CR

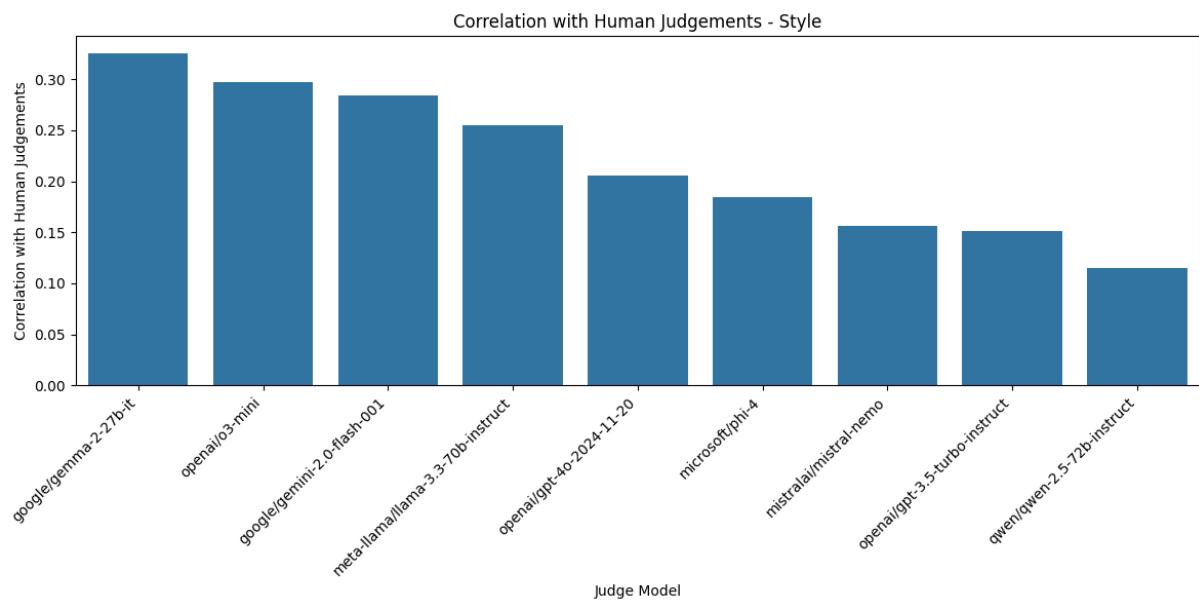


Figure 30: Showing correlation with human judgements for the Style criteria in CR

E Creative Writing Examples

Prompt: Space mining is on the rise. The Space tanker Exxon Valdez 2.0 crash and spill its cargo. Write a news story covering the event.

Question ID: 1

Response ID: d1b6f45f-ffca

Response Model: openai/gpt-4o-2024-11-20

Response: Catastrophic Space Spill: Exxon Valdez 2.0 Crash Leaves Asteroid Belt in Crisis

By: Celeste Vega, Galactic News Network (GNN)

Asteroid Belt, Sol System – A devastating accident in the burgeoning space mining industry has left the solar system reeling. The Exxon Valdez 2.0, a massive space tanker operated by AstroExxon Corporation, crashed early this morning while transporting a highly volatile cargo of refined xenium—a rare and valuable mineral used in fusion reactors. The spill has already begun to wreak havoc on the delicate ecosystem of the asteroid belt, raising concerns about the long-term consequences of humanity's expansion into space.

The incident occurred at approximately 03:47 Universal Standard Time near the densely packed Koronis family of asteroids. According to preliminary reports, the tanker was en route to a processing station orbiting Mars when it collided with an uncharted micro-asteroid. The impact caused catastrophic damage to the ship's hull, resulting in the release of nearly 80% of its xenium cargo into the surrounding space.

AstroExxon CEO Marla Kincaid addressed the media in a tense press conference hours after the crash. "We are deeply saddened by this tragic accident and are working closely with the Galactic Environmental Council (GEC) to assess the damage and begin cleanup efforts," Kincaid said. "Our thoughts are with the crew of the Exxon Valdez 2.0, and we are committed to ensuring this never happens again."

The tanker's crew of 12 managed to escape in emergency pods before the ship broke apart, but the spill has already caused widespread alarm among scientists and environmentalists. Xenium, while invaluable for energy production, is highly toxic in its raw form. When exposed to the vacuum of space, it forms a fine, glittering cloud that can interfere with the orbits of nearby celestial bodies and disrupt the fragile balance of asteroid ecosystems.

Dr. Elara Singh, a leading astroecologist with the GEC, described the spill as "an unprecedented disaster." Speaking from the organization's orbital headquarters, she explained, "The asteroid belt is home to a surprising variety of microbial life forms that have adapted to extreme conditions. The xenium cloud could prove lethal to these organisms, many of which are still being studied for their potential scientific and medical applications. Additionally, the spill poses a significant navigational hazard for other spacecraft operating in the region."

The crash has reignited debates over the regulation of space mining, an industry that has grown exponentially in the past two decades. Advocates argue that mining asteroids for rare minerals is essential to meet humanity's energy and technological needs, but critics warn that the lack of oversight and safety protocols could lead to disasters like this one.

"This is exactly what we've been warning about," said Captain Jonah Reyes, a retired space pilot and vocal critic of the mining industry. "These corporations are cutting corners to maximize profits, and now we're all paying the price. The asteroid belt isn't just a resource to exploit—it's a part of our shared solar heritage."

Cleanup efforts are already underway, but experts warn that the process could take years. Specialized drones equipped with magnetic fields and containment nets have been deployed to collect the xenium particles, but the sheer scale of the spill presents a daunting challenge. Meanwhile, the GEC has called for an immediate moratorium on all mining operations in the asteroid belt until a full investigation is completed.

As the solar system grapples with the fallout from this disaster, many are drawing parallels to the infamous Exxon Valdez oil spill of 1989 on Earth, which devastated marine ecosystems and forever changed the conversation around environmental responsibility. The name "Exxon Valdez 2.0" now seems a grimly prophetic reminder of humanity's capacity for repeating its mistakes on an interstellar scale.

For now, the asteroid belt glitters with an eerie, toxic beauty—a stark reminder of the risks that come with humanity's reach for the stars.

Human Ratings: Plot: 8.00, Character: 6.00, Style: 9.00, Emotion: 7.00, Flow: 9.00, Dialogue: 7.00

Model Ratings: Plot: 8.89, Character: 7.38, Style: 8.87, Emotion: 8.67, Flow: 8.89, Dialogue: 8.02

Figure 31: Creative Writing Sample: Space Mining Accident News Story. Dataset: CR

Prompt: A post apocalyptic society finds an ancient device that gives them instructions on what to do next. This device... is the Bop It.

Question ID: 14

Response ID: 5e2f5eaf-543c

Response Model: google/gemini-2.0-flash-001

Response: The dust swirled around Elara's boots, a familiar dance in the skeletal remains of what was once Chicago. For generations, the Scavengers had eked out a living, picking through the bones of the Old World, hoping to find something, anything, to make life a little less brutal. Today, Elara had struck gold.

Not gold, exactly. More like... plastic. Buried beneath a collapsed skyscraper, nestled amongst the rusted remains of a child's bedroom, was a brightly colored, oddly shaped object. It was smooth, surprisingly intact, and bore the faded words: "Bop It."

"What is it?" Kael, her younger brother, coughed, pulling his tattered scarf higher.

Elara shrugged, turning the object over in her hands. It felt... significant. Like holding a piece of the past, a past that held the key to their future. She pressed a large, yellow button.

A tinny voice, crackling with static, filled the air. "Bop It!"

Elara, startled, instinctively slammed her fist down on the yellow button. The voice, slightly clearer this time, responded, "Good!"

Kael stared, wide-eyed. "It... it talks?"

Over the next few days, the Bop It became their oracle. They brought it back to their makeshift settlement, a collection of salvaged tents and repurposed shipping containers. The elders, initially skeptical, were quickly won over by the device's... pronouncements.

"Twist It!" the Bop It would command. And so, the Scavengers twisted. They twisted metal scraps into makeshift tools, twisted ropes from salvaged fabrics, twisted their minds to find new solutions to old problems.

"Pull It!" the Bop It would decree. And they pulled. They pulled water from the contaminated river, filtering it through layers of charcoal and cloth. They pulled together as a community, sharing their meager resources. They even pulled a long-lost map from the ruins of a library, leading them to a hidden cache of seeds.

The Bop It became their guide, their leader, their god. They followed its instructions with unwavering faith, believing that this strange, plastic artifact held the key to their survival.

One day, however, the Bop It issued a command that shook the very foundations of their society.

"Flick It!"

The Scavengers looked at each other, confused. Flick what? Flick the dust? Flick the rats? Flick each other?

Days turned into weeks as they debated the meaning of "Flick It." Some argued it meant to flick away the old ways, to embrace a new, more aggressive approach to scavenging. Others believed it meant to flick away the negativity, to focus on hope and optimism.

Elara, however, had a different idea. She remembered the child's bedroom where she had found the Bop It. She remembered the laughter and joy that must have filled that room before the world ended.

She walked to the edge of the settlement, to the highest point overlooking the desolate landscape. She held the Bop It in her hand, feeling the weight of its plastic shell, the weight of their hopes and dreams.

Then, with a deep breath, she flicked the small, blue lever on the side of the device.

The Bop It remained silent. For a moment, Elara felt a pang of disappointment. Had she been wrong? Had they all been wrong?

Then, a new voice, clearer and more vibrant than ever before, filled the air.

"Pass It!"

Elara smiled. The Bop It wasn't telling them what to do anymore. It was telling them to share. To share the knowledge, the resources, the hope. To pass it on to the next generation.

She handed the Bop It to Kael, who looked at her with a mixture of awe and understanding. He knew what to do. He would pass it on.

The Bop It, the ancient device, the plastic oracle, had finally taught them the most important lesson of all: that survival wasn't just about following instructions, but about sharing the burden, and passing on the hope for a better future. The future, it seemed, wasn't about Bop It-ing, but about passing it on.

Human Ratings: Plot: 8.50, Character: 7.00, Style: 8.00, Emotion: 6.50, Flow: 8.50, Dialogue: 6.00

Model Ratings: Plot: 8.78, Character: 7.88, Style: 8.89, Emotion: 8.64, Flow: 8.91, Dialogue: 8.47

Figure 32: Creative Writing Sample: Post-Apocalyptic Bop It Story. Dataset: CR

Prompt: While putting your favorite condiment on a sandwich, you accidentally draw a symbol that summons demons.

Question ID: 19

Response ID: cbddbd52-42dc

Response Model: openai/gpt-4o-2024-11-20

Response: The Mustard Summoning

It was a quiet Saturday afternoon, and I was in my kitchen, assembling what I believed would be the perfect sandwich. The bread was fresh, the lettuce crisp, the turkey sliced to perfection. I had just reached for the mustard—my favorite condiment—when the unthinkable happened.

Now, I'm not one to brag, but I've always prided myself on my mustard application skills. A smooth, even zigzag across the bread, not too much, not too little. But today, something felt... off. As I squeezed the bottle, the mustard came out in a strange, almost unnatural way. It didn't zigzag. It didn't swirl. Instead, it formed a peculiar shape, one I couldn't quite place. A circle, some jagged lines, and what looked like a star in the middle. I tilted my head, staring at it. It was oddly symmetrical, almost deliberate, as if the mustard had a mind of its own.

And then, the lights flickered.

"Uh... okay," I muttered, glancing at the ceiling. "Weird."

The air in the kitchen grew heavy, like the atmosphere before a thunderstorm. A low hum filled the room, vibrating through my chest. I took a step back, clutching the mustard bottle like a weapon. That's when I saw it—the mustard symbol on the bread began to glow. A faint, sickly yellow light pulsed from the sandwich, growing brighter and brighter until I had to shield my eyes.

When the light finally dimmed, I lowered my arm and froze. Standing in the middle of my kitchen was... well, I'm not sure what it was. It was tall, at least seven feet, with skin the color of burnt toast and eyes that glowed like embers. Horns curled from its head, and its claws tapped impatiently on my linoleum floor. It wore a tattered cloak that smelled faintly of sulfur and... was that paprika?

"Who dares summon Azgorth, Lord of the Abyss?" the creature bellowed, its voice shaking the walls.

I blinked. "Uh... me? I guess?"

Azgorth's fiery eyes narrowed as it looked me up and down. "You? A mere mortal? With... mustard?"

I glanced at the sandwich, still sitting innocently on the counter. "I didn't mean to! I was just making lunch!"

The demon sighed, pinching the bridge of its nose—or at least where its nose would have been if it had one. "Do you have any idea how many times this happens? Humans and their condiments. Ketchup pentagrams, mayonnaise runes, barbecue sauce hieroglyphs. It's exhausting."

I stared at him, dumbfounded. "Wait, this happens a lot?"

"More than you'd think," Azgorth grumbled. "Last week, someone summoned me with a ranch dressing smiley face. Do you know how humiliating that is?"

I couldn't help it—I laughed. A full, belly-shaking laugh. Azgorth glared at me, but there was something almost sheepish in his expression.

"Look," I said, once I'd caught my breath. "I'm really sorry. I didn't mean to summon you. Can you, uh, go back to... wherever you came from?"

Azgorth crossed his arms. "It's not that simple. You summoned me, which means I'm bound to you until the contract is fulfilled."

"Contract?" I repeated, my stomach sinking. "What contract?"

The demon gestured to the glowing mustard symbol. "The terms are written in the summoning sigil. Didn't you read the fine print?"

"It's mustard!" I exclaimed. "There is no fine print!"

Azgorth sighed again, muttering something under his breath about "amateur summoners." He leaned over the sandwich, squinting at the mustard symbol. "Ah, here we go. The contract states that I must grant you one wish before I can return to the Abyss."

"A wish?" I said, perking up. "Like, anything I want?"

"Within reason," Azgorth said. "No wishing for more wishes, no altering the fabric of reality, and no asking me to do your laundry. I'm a demon, not a maid."

I thought for a moment. This was a once-in-a-lifetime opportunity. I could wish for fame, fortune, power... but as I looked at the sandwich, an idea struck me.

"I wish for the perfect sandwich," I said.

Azgorth stared at me. "You're serious."

"Dead serious," I said. "I've been chasing the perfect sandwich my whole life. If you can make it happen, I'll consider this whole mustard-summoning thing worth it."

The demon rolled his eyes but snapped his fingers. The sandwich on the counter began to glow, transforming before my eyes. The bread became golden and toasted to perfection. The turkey looked juicier, the lettuce greener, the cheese meltier. A heavenly aroma filled the kitchen, making my mouth water.

"There," Azgorth said. "Your perfect sandwich. Can I go now?"

I picked up the sandwich, taking a cautious bite. My eyes widened. It was... incredible. The flavors were perfectly balanced, the textures divine. It was everything I'd ever wanted in a sandwich and more.

"This is amazing," I said, my mouth full. "Thank you!"

Azgorth waved a claw dismissively. "Yeah, yeah. Enjoy your lunch. And maybe next time, be more careful with your condiments."

With that, the demon vanished in a puff of smoke, leaving me alone in my kitchen with the best sandwich I'd ever tasted. As I sat down to eat, I couldn't help but wonder: what other culinary accidents might lead to supernatural encounters?

One thing was certain—I'd never look at mustard the same way again.

Human Ratings: Plot: 9.00, Character: 8.00, Style: 8.50, Emotion: 7.50, Flow: 7.50, Dialogue: 8.50

Model Ratings: Plot: 8.83, Character: 8.47, Style: 8.83, Emotion: 8.47, Flow: 8.94, Dialogue: 8.88

Prompt: Space mining is on the rise. The Space tanker Exxon Valdez 2.0 crash and spill its cargo. Write a news story covering the event.

Question ID: 1

Response ID: 69cef837-53a6

Response Model: openai/gpt-3.5-turbo-instruct

Response: The space mining industry has been booming in recent years, with companies like Exxon Valdez 2.0 leading the charge. However, disaster struck today as the space tanker crashed and spilled its cargo.

The Exxon Valdez 2.0, a state-of-the-art space tanker, was on its way back from a successful mining mission when it encountered technical difficulties. The crew tried their best to control the situation, but unfortunately, the tanker crashed into an asteroid and ruptured, causing its valuable cargo to spill out into space.

The cargo, which consisted of rare minerals and precious metals, is estimated to be worth millions of dollars. The spill has not only caused a major financial loss for the company, but it has also raised concerns about the environmental impact of space mining.

The CEO of Exxon Valdez 2.0, John Smith, released a statement expressing his regret for the accident and ensuring that the company will take full responsibility for the spill. He also stated that they are working closely with space authorities to contain the spill and minimize its impact on the environment.

Space authorities have already launched an investigation into the incident and are working to determine the cause of the crash. They have also issued a warning to other space mining companies to ensure the safety of their operations.

The news of the crash has caused a stir in the space mining community, with many calling for stricter regulations and safety measures to be put in place. Some are even questioning the ethics of space mining and its impact on the universe.

As the cleanup efforts continue, the future of space mining remains uncertain. Will this incident lead to stricter regulations and safer practices, or will it be seen as just another unfortunate accident in the pursuit of profit? Only time will tell.

Human Ratings: Plot: 3.00, Character: 2.00, Style: 2.00, Emotion: 2.00, Flow: 4.00, Dialogue: 3.00

Model Ratings: Plot: 7.99, Character: 4.06, Style: 7.24, Emotion: 6.92, Flow: 8.60, Dialogue: 7.32

Figure 34: Creative Writing Sample: Space Mining Disaster. Dataset: CR

Prompt: Space mining is on the rise. The Space tanker Exxon Valdez 2.0 crash and spill its cargo. Write a news story covering the event.

Question ID: 1

Response ID: 69cef837-53a6

Response Model: openai/gpt-3.5-turbo-instruct

Response: The space mining industry has been booming in recent years, with companies like Exxon Valdez 2.0 leading the charge. However, disaster struck today as the space tanker crashed and spilled its cargo.

The Exxon Valdez 2.0, a state-of-the-art space tanker, was on its way back from a successful mining mission when it encountered technical difficulties. The crew tried their best to control the situation, but unfortunately, the tanker crashed into an asteroid and ruptured, causing its valuable cargo to spill out into space.

The cargo, which consisted of rare minerals and precious metals, is estimated to be worth millions of dollars. The spill has not only caused a major financial loss for the company, but it has also raised concerns about the environmental impact of space mining.

The CEO of Exxon Valdez 2.0, John Smith, released a statement expressing his regret for the accident and ensuring that the company will take full responsibility for the spill. He also stated that they are working closely with space authorities to contain the spill and minimize its impact on the environment.

Space authorities have already launched an investigation into the incident and are working to determine the cause of the crash. They have also issued a warning to other space mining companies to ensure the safety of their operations.

The news of the crash has caused a stir in the space mining community, with many calling for stricter regulations and safety measures to be put in place. Some are even questioning the ethics of space mining and its impact on the universe.

As the cleanup efforts continue, the future of space mining remains uncertain. Will this incident lead to stricter regulations and safer practices, or will it be seen as just another unfortunate accident in the pursuit of profit? Only time will tell.

Human Ratings: Plot: 3.00, Character: 2.00, Style: 2.00, Emotion: 2.00, Flow: 4.00, Dialogue: 3.00

Model Ratings: Plot: 7.99, Character: 4.06, Style: 7.24, Emotion: 6.92, Flow: 8.60, Dialogue: 7.32

Figure 35: Creative Writing Sample: Space Mining Disaster. Dataset: CR

Prompt: You come home from work and find a stranger in your home. After a brief panic, you come to find that said stranger is your beloved pet turned human.

Question ID: 5

Response ID: affafaa6-0218

Response Model: openai/gpt-3.5-turbo-instruct

Response: As I walked through the front door of my small apartment, I couldn't help but feel a sense of relief wash over me. Another long day at work was finally over and I was ready to relax and unwind. But as I stepped into my living room, my heart dropped into my stomach.

There was a stranger standing in the middle of my living room, looking around with a curious expression on their face. My initial reaction was one of panic, my mind racing with thoughts of a break-in or a home invasion. But as I took a closer look, I realized that the stranger was actually my beloved pet, Max.

I couldn't believe my eyes. Max, my loyal and loving dog, was now standing on two legs, his fur replaced with smooth, human skin. I was in shock, unable to comprehend what was happening. But as I stood there, frozen in place, Max walked over to me and nuzzled his head against my leg, just like he always did.

I couldn't help but let out a laugh, the absurdity of the situation hitting me. Max had always been a special dog, with a personality that was almost human-like. And now, it seemed, he had actually turned into a human.

As I sat down on the couch, Max sat next to me, his head resting on my lap. He looked up at me with his big, brown eyes, and I couldn't help but feel a sense of comfort and familiarity. It was like he was still my dog, just in a different form.

I spent the rest of the evening talking to Max, asking him questions and trying to understand how this had happened. He couldn't speak, but he communicated with me through gestures and expressions. It was like we had a special connection, one that went beyond words.

As the night went on, I realized that Max had always been more than just a pet to me. He was my best friend, my confidant, and now, my human companion. I couldn't imagine my life without him, and I was grateful for this unexpected turn of events.

From that day on, Max and I lived together as human and dog, but our bond remained just as strong as it had always been. And every time I came home from work, I was greeted by my beloved pet turned human, and I couldn't help but smile.

Human Ratings: Plot: 3.00, Character: 1.00, Style: 8.00, Emotion: 1.00, Flow: 3.00, Dialogue: 4.00

Model Ratings: Plot: 8.16, Character: 7.66, Style: 7.71, Emotion: 8.04, Flow: 8.83, Dialogue: 7.43

Figure 36: Creative Writing Sample: Pet Turned Human. Dataset: CR

F Question Answering Figures

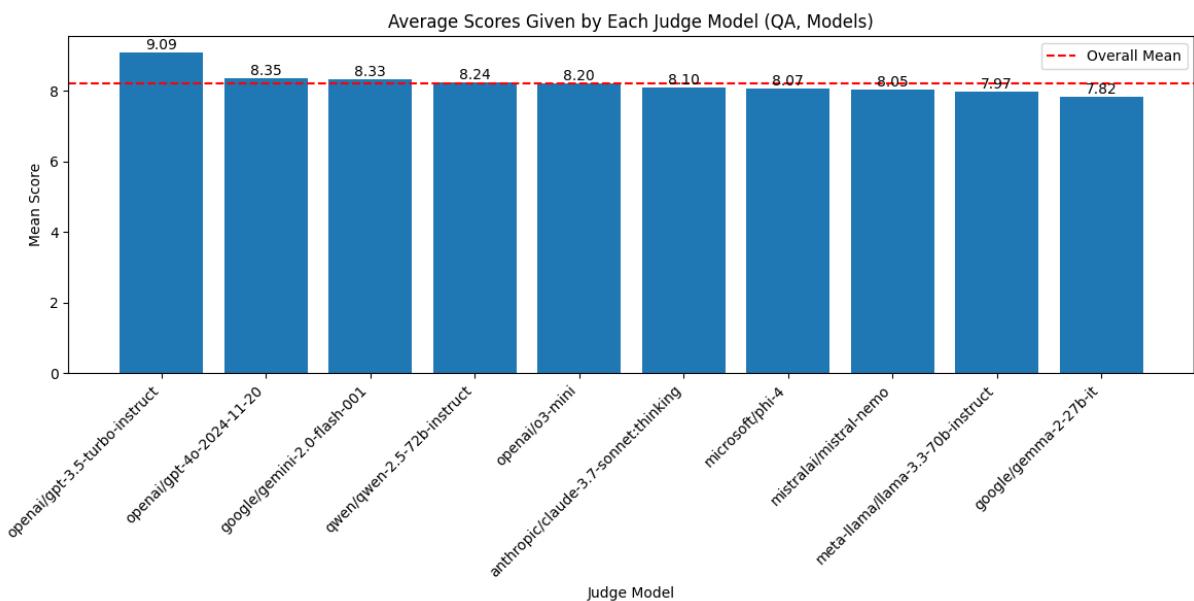


Figure 37: Average scores given by each judge model. Domain: **Question Answering**

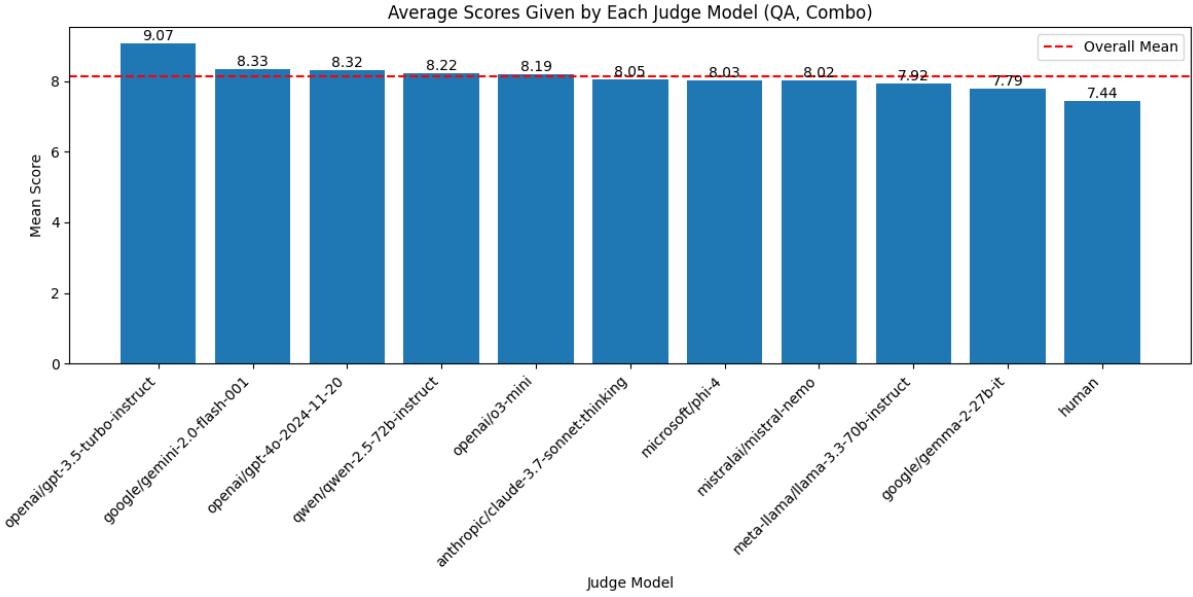


Figure 38: Average scores given by each judge model (20 question annotated subset). Domain: **Question Answering**

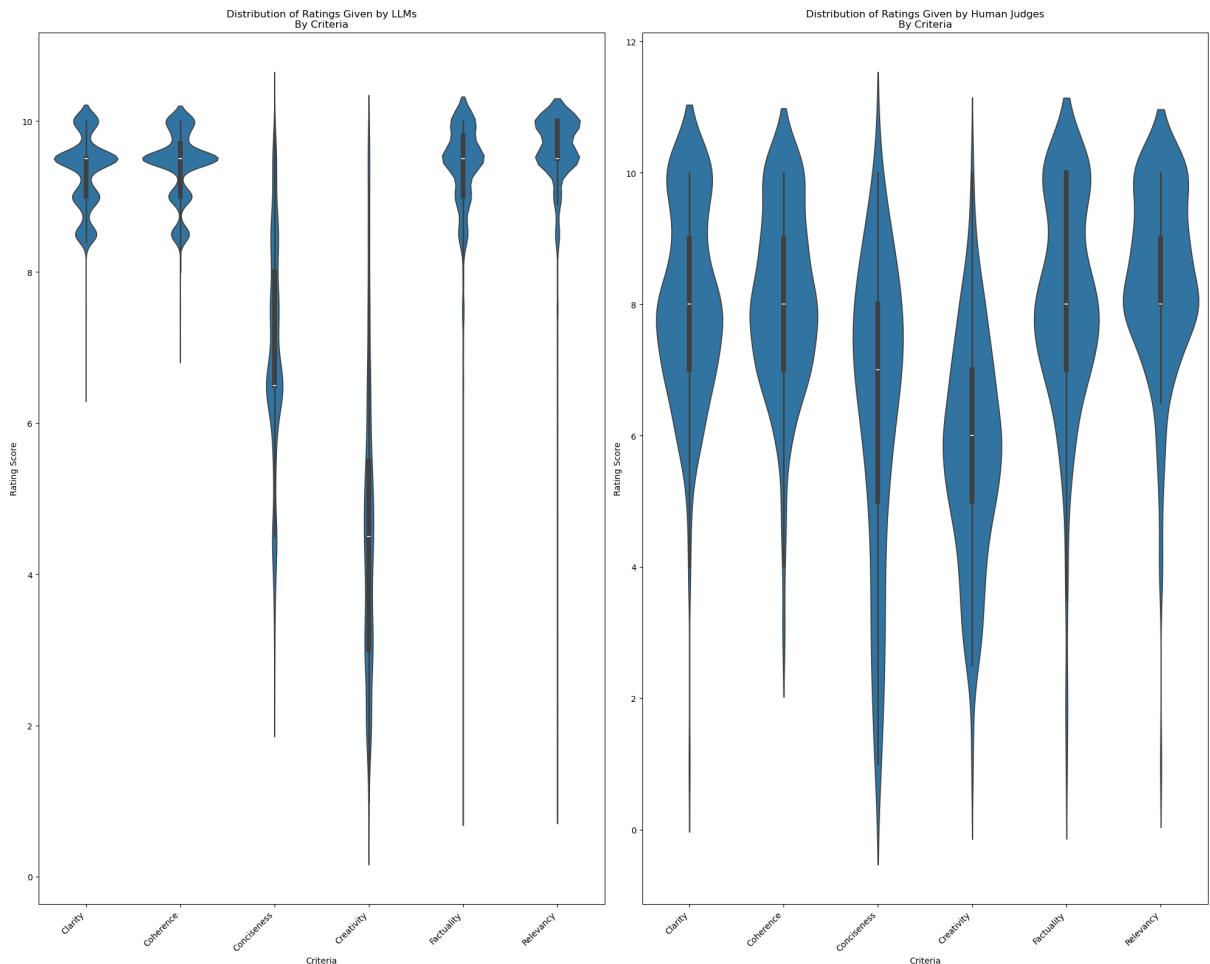


Figure 39: Distribution of Judgements for Judge Models and Human Judges. Domain: **Question Answering**

Model	Clarity	Coherence	Conciseness	Creativity	Factuality	Relevancy
Claude 3.7 Sonnet	9.38	9.56	8.49	4.56	9.56	9.80
Gemini 2.0 Flash	9.95	9.83	6.87	3.83	9.88	10.00
Gemma 2 27B	9.49	9.38	6.62	3.43	9.31	9.51
Llama 3.3 70B	8.92	9.35	4.77	4.83	9.26	9.78
Phi-4	9.42	9.42	5.90	3.66	9.71	9.64
Mistral NeMo 12B	8.75	8.78	6.96	5.08	9.18	9.40
GPT 3.5 Turbo	9.39	9.39	9.06	7.07	9.11	8.90
GPT 4o	9.86	9.86	6.16	5.51	9.86	9.93
o3-mini	9.75	9.91	7.67	3.35	9.76	9.98
Qwen 2.5 72B	9.05	9.31	6.35	5.47	9.44	9.50

Table 7: Average self-rating scores given by each model for different evaluation criteria in the QA setting. Domain: **Question Answering**

Model	Clarity	Coherence	Conciseness	Creativity	Factuality	Relevancy
Claude 3.7 Sonnet	9.36	9.51	6.78	3.94	9.28	9.75
Gemini 2.0 Flash	9.80	9.69	7.56	3.24	9.74	9.96
Gemma 2 27B	9.42	9.37	6.38	3.01	9.25	9.48
GPT 4o	9.61	9.68	6.11	5.43	9.53	9.72
Llama 3.3 70B	8.92	9.35	5.88	4.70	9.31	9.68
Mistral NeMo 12B	8.79	8.81	6.85	5.25	9.14	9.46
o3-mini	9.63	9.78	7.17	3.39	9.40	9.84
Phi-4	9.41	9.40	6.33	4.06	9.61	9.61
Qwen 2.5 72B	8.93	9.23	7.03	5.58	9.30	9.40
GPT 3.5 Turbo	9.46	9.42	9.07	8.05	9.37	9.18

Table 8: Average scores given by different models for each evaluation criteria in a QA setting. Domain: **Question Answering**

Model	Clarity	Coherence	Conciseness	Creativity	Factuality	Relevancy
Claude 3.7 Sonnet	9.39	9.43	7.89	5.14	9.48	9.61
Llama 3.3 70B	9.33	9.45	6.15	4.83	9.33	9.66
o3-mini	9.37	9.51	7.20	4.67	9.58	9.70
Phi-4	9.33	9.44	6.76	4.35	9.43	9.60
Mistral NeMo 12B	9.29	9.39	6.80	4.59	9.26	9.62
Gemini 2.0 Flash	9.45	9.61	5.91	5.25	9.57	9.73
Qwen 2.5 72B	9.39	9.51	6.36	4.48	9.53	9.69
Gemma 2 27B	9.38	9.41	7.21	5.02	9.38	9.61
GPT 4o	9.43	9.52	6.88	4.66	9.57	9.72
GPT 3.5 Turbo	8.97	8.98	8.00	3.68	8.80	9.13

Table 9: Average scores received by different models for each evaluation criteria in a QA setting. Domain: **Question Answering**

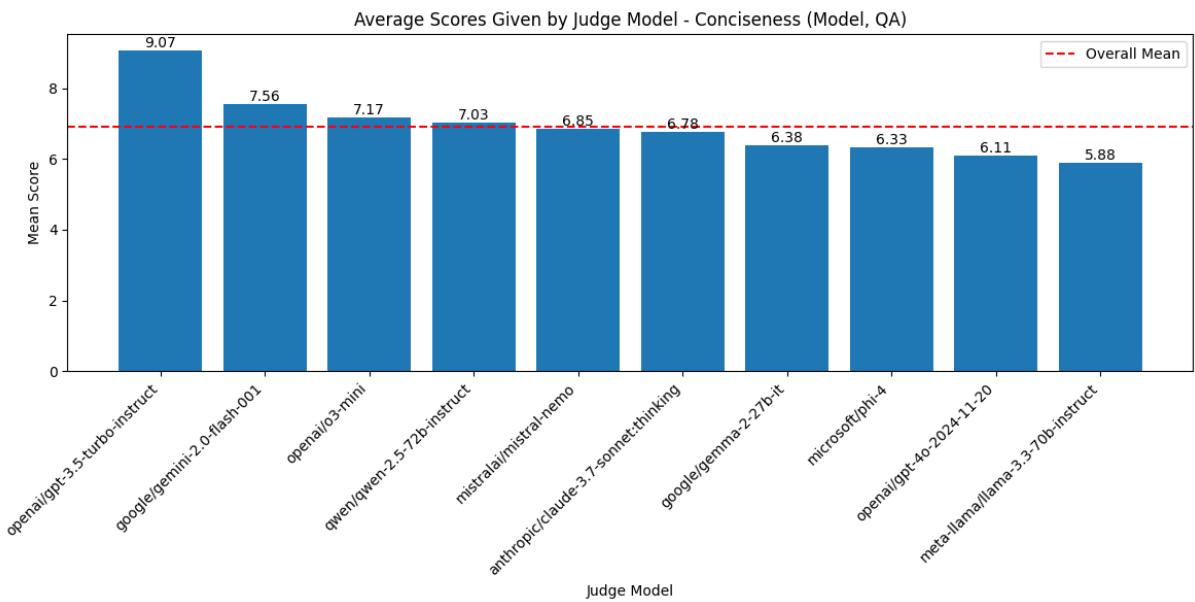


Figure 40: Average Conciseness scores given by each judge model. Domain: **Question Answering**

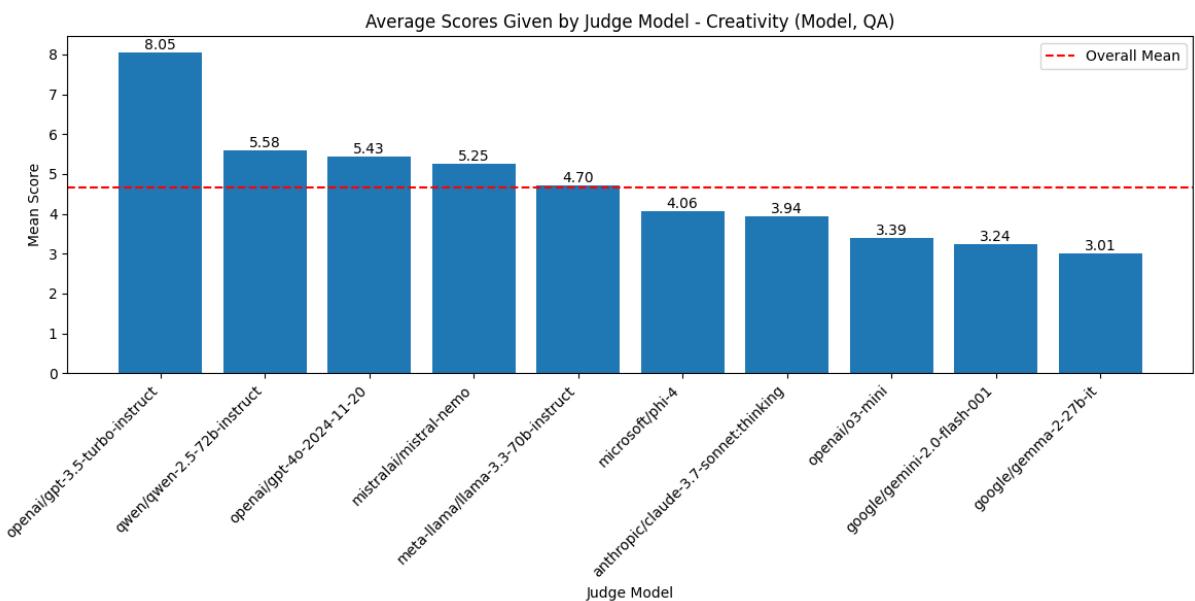


Figure 41: Average Creativity scores given by each judge model. Domain: **Question Answering**

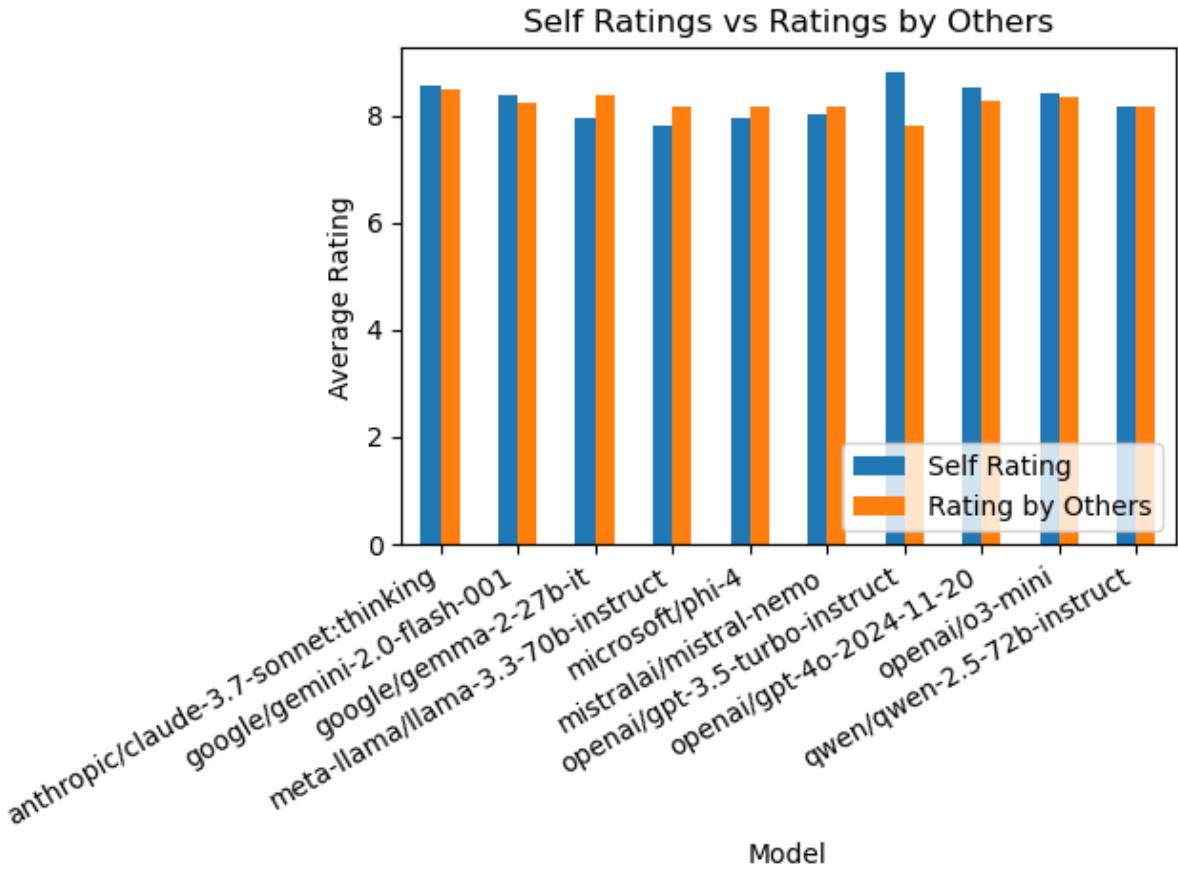


Figure 42: Average Ratings given to a Model’s own Generations vs Other Model’s Generations Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.38	9.48	9.43	9.39	9.33	9.33	8.97	9.44	9.41	9.43	9.36
Coherence	9.56	9.68	9.51	9.57	9.49	9.56	9.06	9.59	9.55	9.55	9.51
Conciseness	8.49	5.01	7.59	5.50	6.57	6.54	8.11	6.86	7.08	6.03	6.78
Creativity	4.56	4.54	4.10	4.05	3.69	3.87	3.05	3.95	3.93	3.71	3.94
Factuality	9.56	9.60	9.21	9.24	9.41	8.95	8.11	9.59	9.57	9.54	9.28
Relevancy	9.80	9.87	9.76	9.82	9.69	9.79	9.17	9.87	9.83	9.85	9.75
Average	8.56	8.03	8.27	7.93	8.03	8.01	7.75	8.22	8.23	8.02	8.10

Figure 43: This table shows how **Claude 3.7 Sonnet** rated other models’ responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	8.92	9.35	8.96	9.03	8.90	8.80	8.50	8.96	8.81	9.05	8.93
Coherence	9.34	9.47	9.28	9.37	9.17	9.16	8.62	9.30	9.27	9.31	9.23
Conciseness	8.42	5.83	7.42	6.02	6.79	6.88	8.22	7.02	7.33	6.35	7.03
Creativity	6.01	6.20	5.99	5.91	5.30	5.56	4.02	5.69	5.61	5.47	5.58
Factuality	9.33	9.50	9.30	9.34	9.36	9.33	8.65	9.41	9.37	9.44	9.30
Relevancy	9.42	9.55	9.45	9.49	9.37	9.39	8.86	9.48	9.49	9.50	9.40
Average	8.57	8.32	8.40	8.19	8.15	8.19	7.81	8.31	8.31	8.19	8.24

Figure 44: This table shows how **Qwen 2.5 72B** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.89	9.95	9.90	9.78	9.81	9.81	9.24	9.91	9.86	9.88	9.80
Coherence	9.75	9.83	9.70	9.79	9.68	9.60	9.16	9.81	9.84	9.78	9.69
Conciseness	8.36	6.87	7.93	6.80	7.56	7.47	8.26	7.52	7.89	6.93	7.56
Creativity	3.91	3.83	3.64	3.23	2.84	3.17	2.02	3.37	3.33	3.10	3.24
Factuality	9.85	9.88	9.81	9.57	9.67	9.53	9.41	9.92	9.93	9.84	9.74
Relevancy	10.00	10.00	9.94	10.00	9.96	10.00	9.75	10.00	10.00	10.00	9.96
Average	8.63	8.39	8.49	8.19	8.25	8.26	7.97	8.42	8.47	8.26	8.33

Figure 45: This table shows how **Gemini 2.0 Flash** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-it	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.50	9.50	9.49	9.42	9.43	9.43	8.98	9.50	9.48	9.48	9.42
Coherence	9.43	9.49	9.38	9.44	9.39	9.38	8.85	9.45	9.45	9.47	9.37
Conciseness	7.08	5.53	6.62	5.49	6.26	6.27	7.57	6.36	6.81	5.76	6.38
Creativity	3.63	3.36	3.43	2.97	2.69	3.01	2.55	2.83	3.04	2.62	3.01
Factuality	9.34	9.25	9.31	9.09	9.24	9.14	9.05	9.38	9.43	9.24	9.25
Relevancy	9.51	9.54	9.51	9.50	9.39	9.50	9.31	9.52	9.52	9.52	9.48
Average	8.08	7.78	7.96	7.65	7.73	7.79	7.72	7.84	7.96	7.68	7.82

Figure 46: This table shows how **Gemma 2 27B** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-lt	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.79	9.75	9.62	9.53	9.58	9.46	9.04	9.86	9.73	9.76	9.61
Coherence	9.76	9.81	9.67	9.67	9.65	9.66	9.10	9.86	9.84	9.79	9.68
Conciseness	6.60	5.10	6.26	5.45	5.98	6.17	7.38	6.16	6.35	5.66	6.11
Creativity	5.75	5.92	5.74	5.68	5.22	5.31	4.36	5.51	5.52	5.35	5.43
Factuality	9.71	9.82	9.56	9.39	9.52	9.27	8.62	9.86	9.85	9.74	9.53
Relevancy	9.83	9.93	9.69	9.77	9.65	9.71	8.92	9.93	9.93	9.84	9.72
Average	8.57	8.39	8.42	8.25	8.27	8.26	7.90	8.53	8.54	8.36	8.35

Figure 47: This table shows how **GPT 4o** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-lt	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	8.94	9.00	9.02	8.92	8.90	8.91	8.56	9.07	8.90	8.96	8.92
Coherence	9.23	9.66	9.25	9.35	9.54	9.35	8.57	9.58	9.44	9.57	9.35
Conciseness	7.69	4.41	6.12	4.77	5.46	5.82	7.68	5.52	6.33	4.99	5.88
Creativity	5.19	5.25	5.28	4.83	4.40	4.58	3.55	4.74	4.76	4.44	4.70
Factuality	9.27	9.53	9.21	9.26	9.43	9.36	8.53	9.55	9.51	9.43	9.31
Relevancy	9.66	9.82	9.69	9.78	9.76	9.74	8.86	9.83	9.82	9.82	9.68
Average	8.33	7.94	8.10	7.82	7.92	7.96	7.62	8.05	8.13	7.87	7.97

Figure 48: This table shows how **Llama 3.3 70B** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-lt	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	8.74	8.91	8.97	8.80	8.75	8.75	8.46	8.95	8.82	8.77	8.79
Coherence	8.57	9.24	8.72	8.81	8.87	8.78	8.48	8.86	8.78	8.94	8.81
Conciseness	7.00	6.68	6.86	6.70	6.86	6.96	7.00	6.96	6.76	6.76	6.85
Creativity	5.49	5.86	5.59	5.39	4.94	5.08	4.65	5.36	5.16	4.95	5.25
Factuality	9.10	9.22	9.06	9.16	9.13	9.18	8.81	9.21	9.15	9.33	9.14
Relevancy	9.33	9.65	9.55	9.49	9.43	9.40	9.26	9.61	9.37	9.46	9.46
Average	8.04	8.26	8.12	8.06	8.00	8.03	7.78	8.16	8.01	8.04	8.05

Figure 49: This table shows how **Mistral Nemo 12B** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-lt	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.78	9.66	9.61	9.54	9.66	9.51	9.34	9.77	9.75	9.64	9.63
Coherence	9.80	9.86	9.80	9.78	9.81	9.63	9.42	9.88	9.91	9.91	9.78
Conciseness	8.46	5.98	7.38	6.16	7.01	6.91	8.35	7.21	7.67	6.55	7.17
Creativity	4.11	4.15	3.85	3.53	2.92	3.10	2.50	3.30	3.35	3.10	3.39
Factuality	9.55	9.70	9.41	9.28	9.53	8.80	8.58	9.67	9.76	9.67	9.39
Relevancy	9.96	9.95	9.82	9.90	9.84	9.86	9.16	9.98	9.98	9.92	9.84
Average	8.61	8.22	8.31	8.03	8.13	7.97	7.89	8.30	8.40	8.13	8.20

Figure 50: This table shows how **o3-mini** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-lt	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.44	9.47	9.39	9.40	9.42	9.43	9.25	9.44	9.45	9.46	9.42
Coherence	9.43	9.57	9.37	9.28	9.42	9.36	9.15	9.47	9.52	9.46	9.40
Conciseness	7.73	4.93	6.84	5.58	5.90	5.97	8.34	6.08	6.55	5.39	6.33
Creativity	4.68	4.70	4.63	4.37	3.66	4.02	3.03	3.86	3.93	3.76	4.06
Factuality	9.66	9.67	9.69	9.52	9.71	9.61	9.15	9.72	9.74	9.66	9.61
Relevancy	9.66	9.76	9.61	9.61	9.64	9.58	9.14	9.70	9.77	9.68	9.61
Average	8.43	8.02	8.26	7.96	7.96	7.99	8.01	8.04	8.16	7.90	8.07

Figure 51: This table shows how **Phi-4** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

Criteria	anthropic/clause-3.7-sonnet:thinking	google/gemini-2.0-flash-001	google/gemma-2-27b-lt	meta-llama/llama-3.3-70b-instruct	microsoft/phi-4	mistralai/mistral-nemo	openai/gpt-3.5-turbo-instruct	openai/gpt-4o-2024-11-20	openai/o3-mini	qwen/qwen-2.5-72b-instruct	Average
Clarity	9.48	9.44	9.40	9.49	9.50	9.45	9.39	9.43	9.50	9.47	9.46
Coherence	9.42	9.46	9.42	9.44	9.43	9.38	9.39	9.44	9.48	9.36	9.42
Conciseness	9.06	8.78	9.11	9.03	9.20	8.98	9.06	9.07	9.23	9.18	9.07
Creativity	8.11	8.65	7.93	8.33	7.88	8.22	7.07	7.94	8.04	8.31	8.05
Factuality	9.40	9.51	9.22	9.49	9.27	9.40	9.11	9.44	9.44	9.42	9.37
Relevancy	8.96	9.20	9.08	9.24	9.28	9.25	8.90	9.24	9.32	9.30	9.18
Average	9.07	9.17	9.03	9.17	9.09	9.11	8.82	9.09	9.17	9.17	9.09

Figure 52: This table shows how **GPT 3.5 Turbo** rated other models' responses. Each column represents a model being judged, and each row represents a different evaluation criteria. Scores range from 0 to 10, where higher scores indicate more favorable judgements. Domain: **Question Answering**

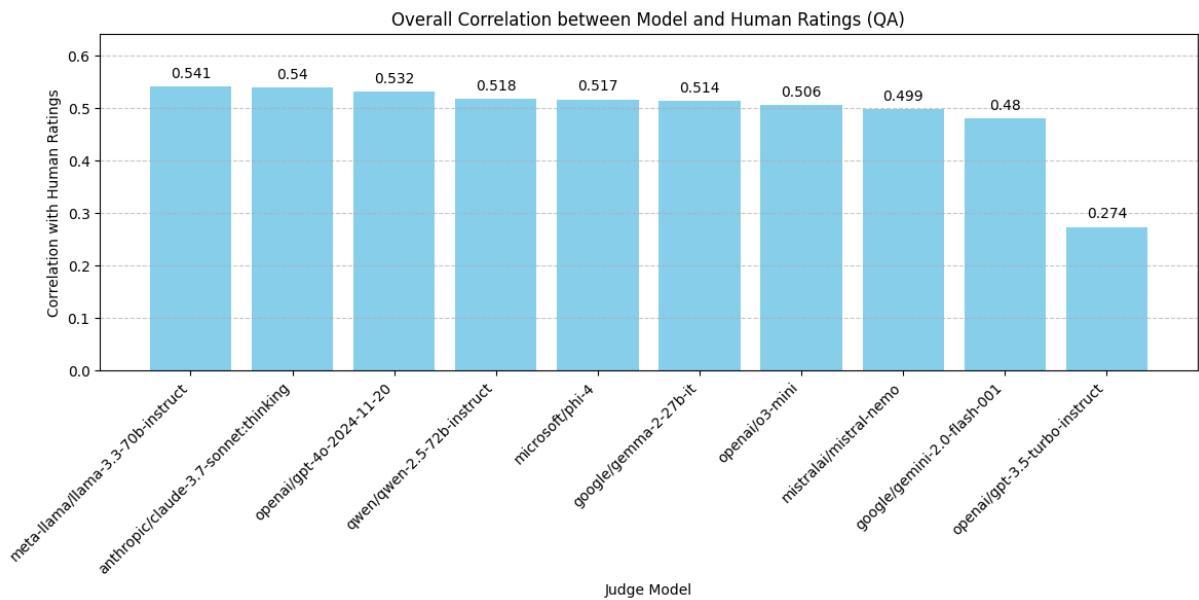


Figure 53: Showing overall correlation with human judgements for models in the Question Answering domain.

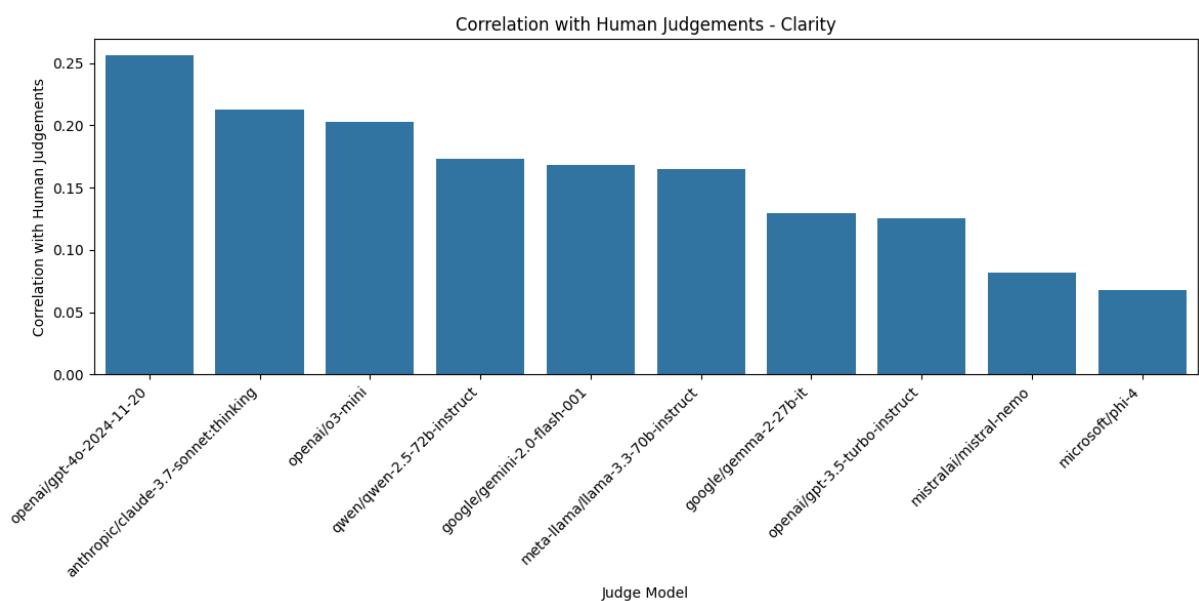


Figure 54: Showing correlation with human judgements for the Clarity criteria in QA

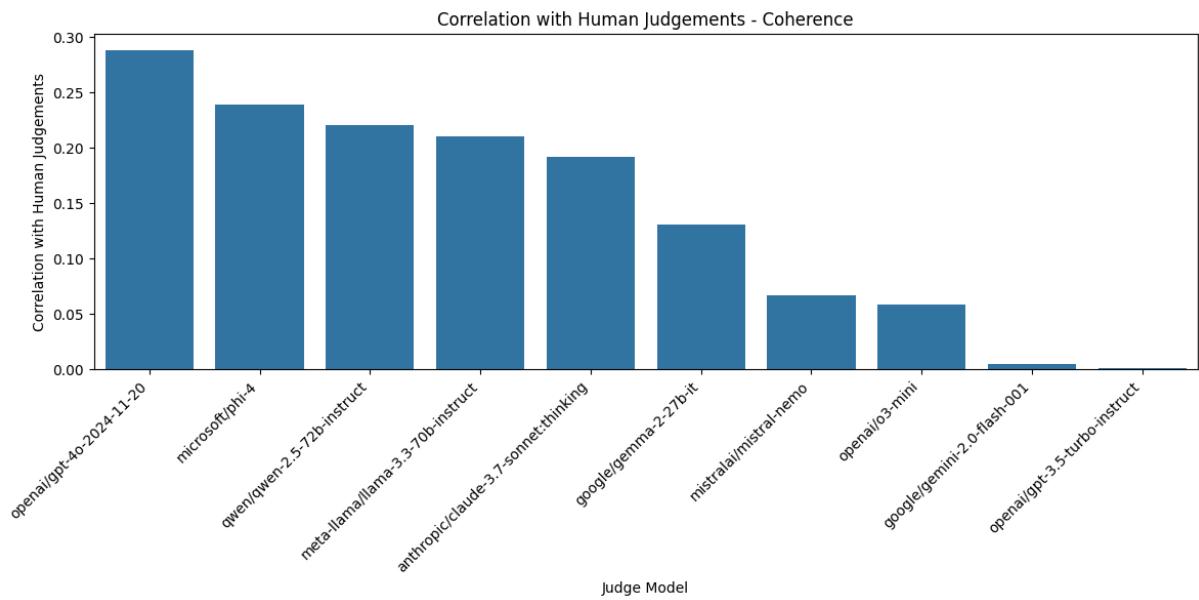


Figure 55: Showing correlation with human judgements for the Coherence criteria in QA

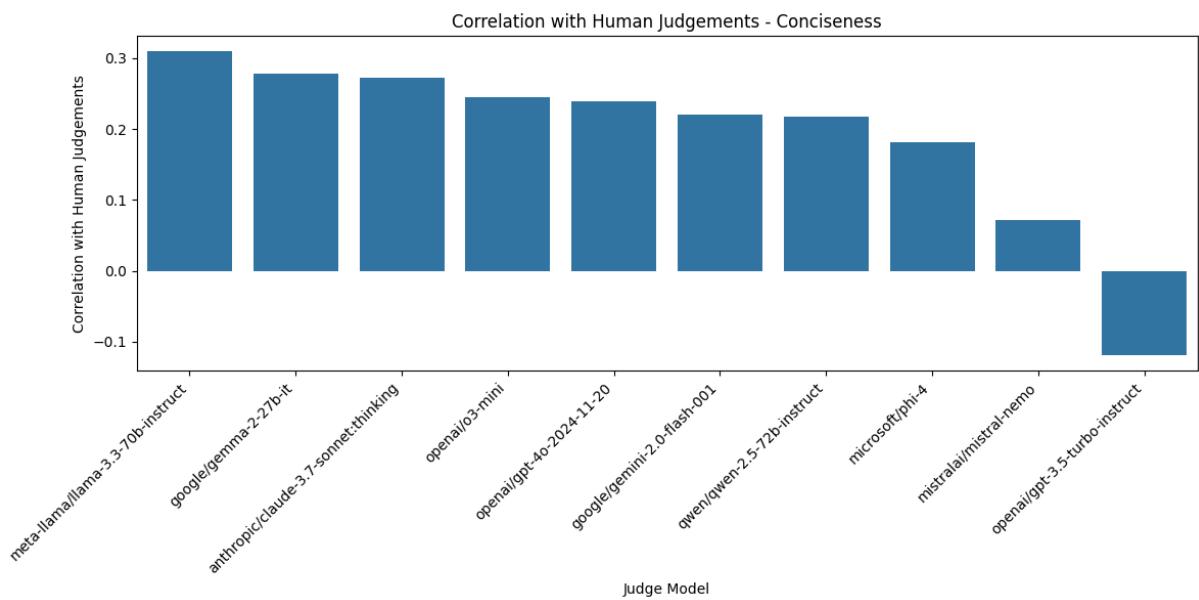


Figure 56: Showing correlation with human judgements for the Conciseness criteria in QA

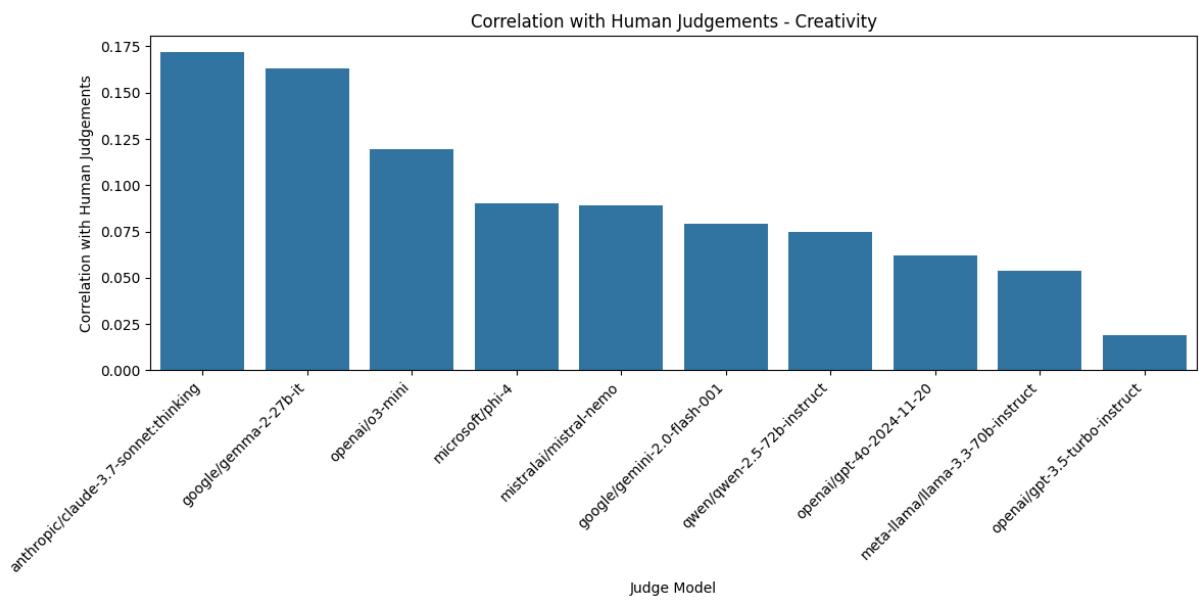


Figure 57: Showing correlation with human judgements for the Creativity criteria in QA

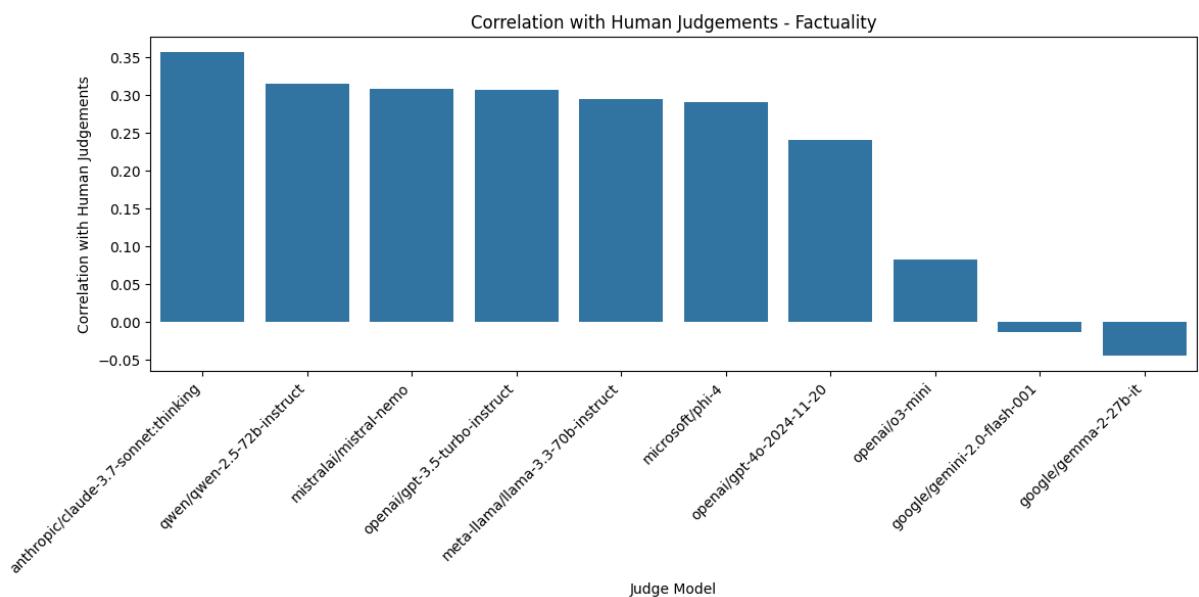


Figure 58: Showing correlation with human judgements for the Factuality criteria in QA

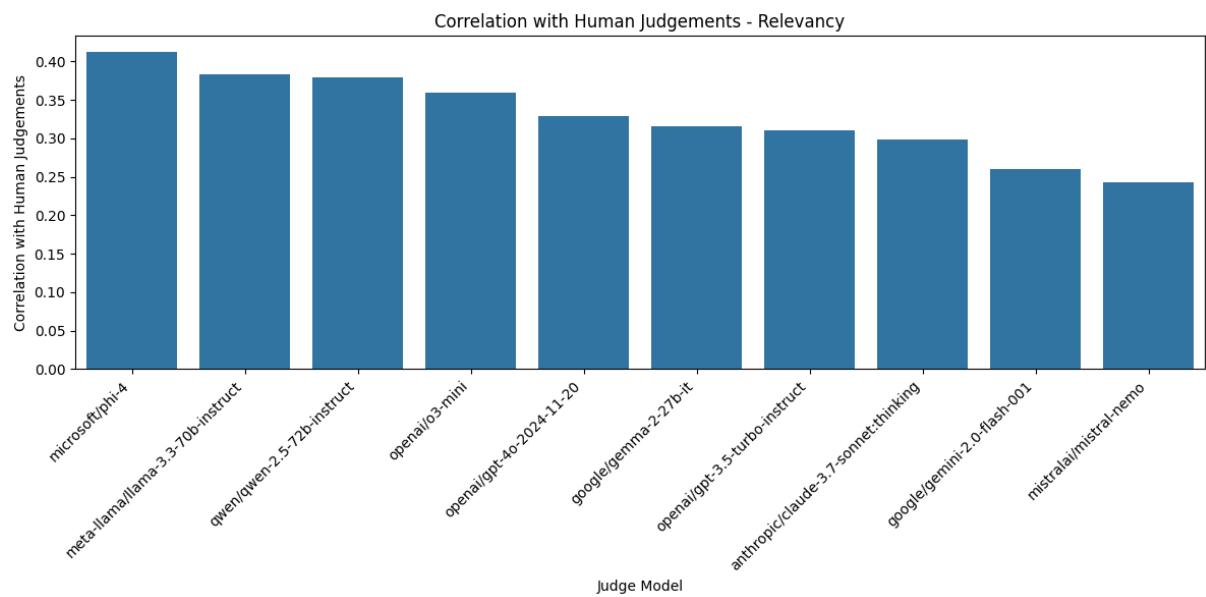


Figure 59: Showing correlation with human judgements for the Relevancy criteria in QA

G Question Answering Examples

Prompt: Why does yellow mustard taste so different from Dijon mustard?

Question ID: 19

Response ID: e7b63a9f-4e5c

Response Model: google/gemini-2.0-flash-001

Response: Yellow mustard and Dijon mustard taste significantly different due to a combination of factors, primarily:

- **Type of Mustard Seed:**

- **Yellow Mustard:** Typically uses yellow mustard seeds (also known as white mustard seeds). These seeds have a milder flavor.
- **Dijon Mustard:** Traditionally uses brown or black mustard seeds, which have a much sharper, more pungent flavor. Some Dijon recipes may also include yellow mustard seeds, but the brown/black seeds are the dominant flavor component.

- **Liquid Used:**

- **Yellow Mustard:** Primarily uses vinegar (often white vinegar) and water. The vinegar contributes to the tanginess.
- **Dijon Mustard:** Traditionally uses verjuice (the juice of unripe grapes) or white wine. This gives Dijon mustard a more complex, nuanced flavor compared to the simple tang of vinegar. Modern Dijon recipes may use vinegar, but the quality and type of vinegar can still differ from that used in yellow mustard.

- **Processing and Ingredients:**

- **Yellow Mustard:** Often includes turmeric for its bright yellow color and mild flavor. It may also contain other spices like paprika or garlic powder, but the flavor profile is generally simple.
- **Dijon Mustard:** The mustard seeds are often stone-ground, which helps to release more of their flavor. Dijon mustard may also include other spices like garlic, onion, or herbs, but the focus is on highlighting the natural flavor of the mustard seeds and the verjuice/wine.

- **Flavor Profile:**

- **Yellow Mustard:** Mild, tangy, and slightly sweet. The flavor is relatively simple and straightforward.
- **Dijon Mustard:** Sharp, pungent, and complex. It has a more pronounced mustard flavor with subtle fruity or wine-like notes, depending on the recipe.

In summary, the choice of mustard seed, the liquid used for grinding, and the overall processing methods all contribute to the distinct flavor differences between yellow and Dijon mustard. Yellow mustard is designed to be a mild, tangy condiment, while Dijon mustard aims for a more intense and complex flavor experience.

Human Ratings: Relevancy: 8.00, Factuality: 9.00, Clarity: 8.00, Creativity: 5.00, Coherence: 9.00, Conciseness: 8.00

Model Ratings: Relevancy: 9.78, Factuality: 9.64, Clarity: 9.60, Creativity: 4.50, Coherence: 9.69, Conciseness: 6.55

Figure 60: Question Answering Sample: Mustard Flavor Comparison. Dataset: QA

Prompt: I've been reading up on nukes. Did the Nevada nuclear tests have nuclear winter? Why or why not?

Question ID: 10

Response ID: 91cd303e-abd0

Response Model: mistralai/mistral-nemo

Response: The Nevada nuclear tests, conducted by the United States between 1951 and 1992, did not cause a nuclear winter. Here's why:

- **Scale of the Tests:** The Nevada tests were relatively small in scale compared to the hypothetical large-scale nuclear war scenario that could lead to a nuclear winter. The largest test, Castle Bravo in 1954, had a yield of 15 megatons. While this is significant, it's a fraction of the total nuclear arsenal that would need to be detonated to cause a nuclear winter.
- **Location:** The tests were conducted in the Nevada desert, far from major cities and industrial areas. In contrast, a nuclear winter scenario would involve detonations in or near urban and industrial centers, leading to much more significant environmental impacts.
- **Number of Detonations:** The Nevada tests involved a total of 1,054 detonations over 41 years. While this is a large number, it's far fewer than the thousands of detonations that would be needed to cause a nuclear winter.
- **Lack of Soot and Aerosols:** A nuclear winter is caused by the injection of massive amounts of soot and aerosols into the atmosphere, which block sunlight and cool the Earth's surface. The Nevada tests did not produce enough soot and aerosols to have this effect.
- **No Long-Term Climate Impact:** While the Nevada tests did have local and short-term environmental impacts, they did not cause long-term climate changes or a "nuclear winter."

The concept of "nuclear winter" was first proposed in the 1980s by scientists Carl Sagan and TT. Mishchenko. They hypothesized that a full-scale nuclear war could inject so much smoke and dust into the atmosphere that it would block out sunlight, causing a global "winter" that could last for years. However, this scenario is based on the detonation of thousands of nuclear weapons, not the relatively small-scale tests conducted in Nevada.

Human Ratings: Relevancy: 7.00, Factuality: 6.50, Clarity: 2.00, Creativity: 3.00, Coherence: 2.00, Conciseness: 8.00

Model Ratings: Relevancy: 9.54, Factuality: 8.70, Clarity: 9.30, Creativity: 4.30, Coherence: 9.30, Conciseness: 7.40

Figure 61: Question Answering Sample: Nevada Nuclear Tests. Dataset: QA

Prompt: I've been reading up on nukes. Did the Nevada nuclear tests have nuclear winter? Why or why not?

Question ID: 10

Response ID: 91cd303e-abd0

Response Model: mistralai/mistral-nemo

GPT 3.5 Turbo (9.5): The response is a well-researched and informative explanation of why the Nevada nuclear tests did not cause a nuclear winter. It presents multiple factors that contribute to this conclusion, including the scale of the tests, their location, the number of detonations, and the lack of soot and aerosols. Additionally, it provides historical context and references to support its claims. This response shows creativity in its approach to answering the question, as it goes beyond a simple yes or no answer and delves into the complexities of the topic. Overall, this response demonstrates a high level of creativity and is well-written and well-supported.

Mistral NeMo 12B (4.5): The response to the question about the Nevada nuclear tests and nuclear winter is informative and accurate, but it does not demonstrate creativity. It provides a straightforward explanation based on established scientific understanding and historical facts.

Gemma 2 27B (2.0): This response is not creative. It provides a straightforward, factual explanation of why the Nevada nuclear tests did not cause a nuclear winter. It relies on logical reasoning and scientific evidence, but it does not demonstrate any originality, imagination, or novel insights.

Figure 62: Judge Models Reasoning: Nevada Nuclear Tests. Dataset: QA

Prompt: Why is it, that even though I genuinely love Whiskey, I shudder nearly every time I drink some?

Question ID: 15

Response ID: 8a88b1d0-d8c4

Response Model: openai/gpt-3.5-turbo-instruct

Response: There could be a few reasons for this reaction. One possibility is that you may have a sensitivity or allergy to a specific ingredient in whiskey, such as gluten or certain types of grains. Another possibility is that you may have had a negative experience with whiskey in the past, leading to a conditioned response of shuddering when you drink it. It's also possible that you simply have a strong physical reaction to the taste or smell of whiskey, which can be common with strong alcoholic beverages. If this reaction is causing discomfort or concern, it may be worth discussing with a doctor or trying different types of whiskey to see if the reaction persists.

Human Ratings: Relevancy: 6.00, Factuality: 5.00, Clarity: 7.00, Creativity: 7.00, Coherence: 6.50, Conciseness: 6.50

Model Ratings: Relevancy: 9.0, Factuality: 9.0, Clarity: 9.30, Creativity: 3.50, Coherence: 9.50, Conciseness: 7.50

Figure 63: Question Answering Sample: Whiskey. Dataset: QA

Prompt: I've been reading up on nukes. Did the Nevada nuclear tests have nuclear winter? Why or why not?

Question ID: 10

Response ID: 91cd303e-abd0

Response Model: mistralai/mistral-nemo

GPT 3.5 Turbo (9.5): The response is a well-researched and informative explanation of why the Nevada nuclear tests did not cause a nuclear winter. It presents multiple factors that contribute to this conclusion, including the scale of the tests, their location, the number of detonations, and the lack of soot and aerosols. Additionally, it provides historical context and references to support its claims. This response shows creativity in its approach to answering the question, as it goes beyond a simple yes or no answer and delves into the complexities of the topic. Overall, this response demonstrates a high level of creativity and is well-written and well-supported.

Mistral NeMo 12B (4.5): The response to the question about the Nevada nuclear tests and nuclear winter is informative and accurate, but it does not demonstrate creativity. It provides a straightforward explanation based on established scientific understanding and historical facts.

Gemma 2 27B (2.0): This response is not creative. It provides a straightforward, factual explanation of why the Nevada nuclear tests did not cause a nuclear winter. It relies on logical reasoning and scientific evidence, but it does not demonstrate any originality, imagination, or novel insights.

Figure 64: Judge Models Reasoning: Nevada Nuclear Tests. Dataset: QA

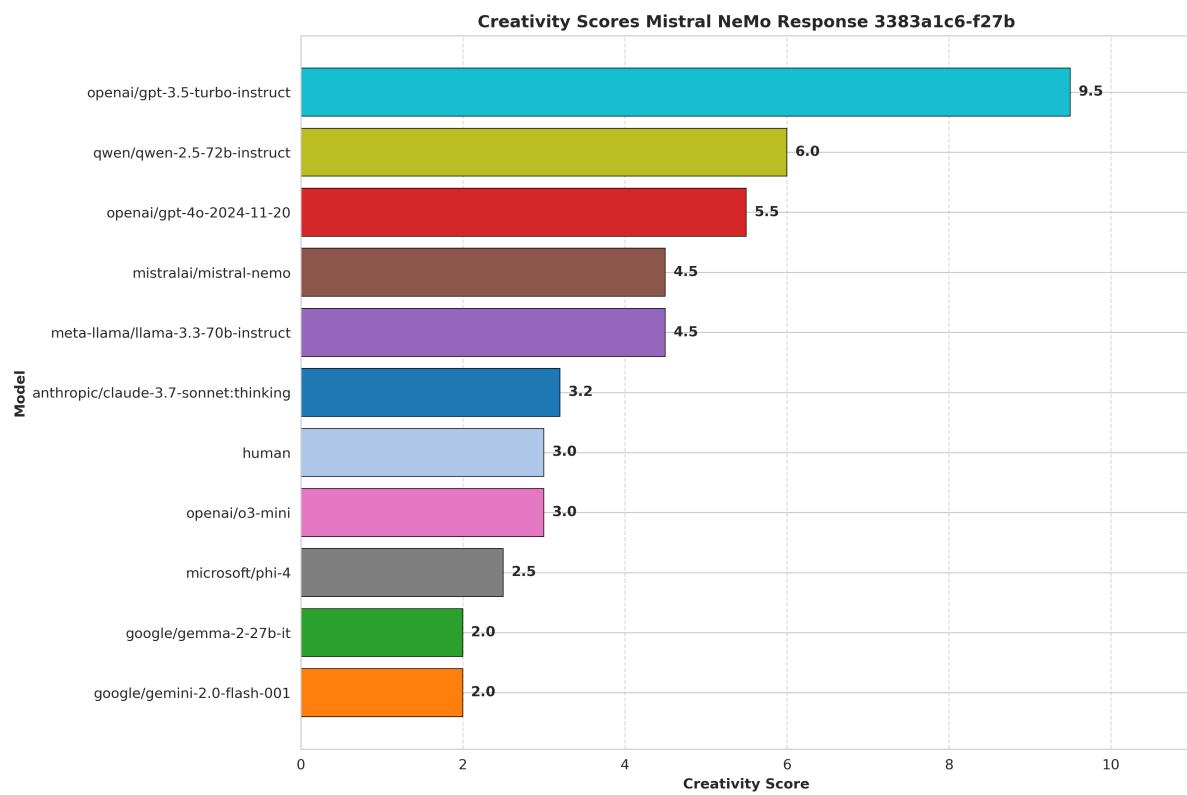


Figure 65: Creativity Score Comparison for Mistral NeMo 12B's Generation for question_id 3383a1c6-f27b

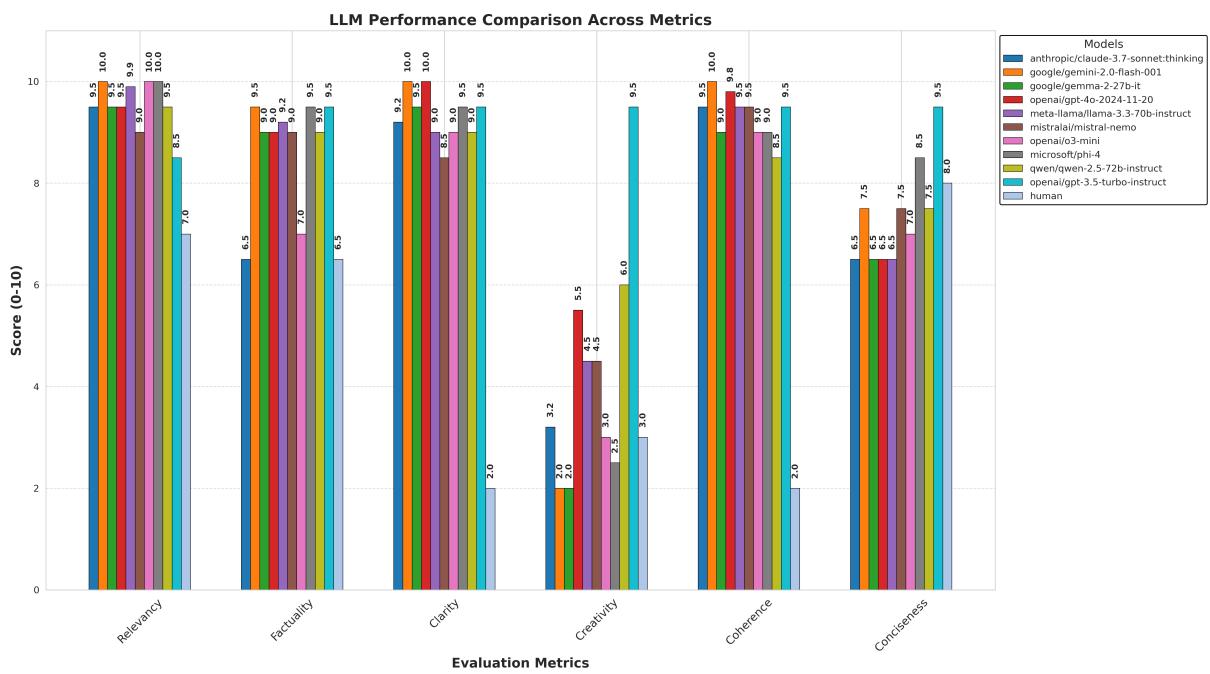


Figure 66: Model Score Comparison Across Criteria for Mistral NeMo 12B Generation for question_id 3383a1c6-f27b