Towards Scalable Oversight: Meta-Evaluation of LLMs as Evaluators via Agent Debate

Steffi Chern[2,4], Ethan Chern[1,4], Graham Neubig[2], and Pengfei Liu[1,3,4]

[1] Shanghai Jiao Tong University, Minhang District, Shanghai 200240, China
[2] Carnegie Mellon University, Pittsburgh PA 15213, USA
[3] Shanghai Artificial Intelligence Laboratory, Xuhui District, Shanghai 200232, China
[4] Generative Artificial Intelligence Lab, Minhang District, Shanghai 200240, China

Abstract. Despite the utility of Large Language Models (LLMs) across a wide range of tasks and scenarios, developing a method for reliably evaluating LLMs across varied contexts continues to be challenging. Modern evaluation approaches often use LLMs to assess responses generated by LLMs. However, existing meta-evaluation methods to assess the effectiveness of LLMs as evaluators is typically constrained by the coverage of existing benchmarks or require extensive human annotation. This underscores the urgency of methods for *scalable* meta-evaluation that can effectively, reliably, and efficiently evaluate the performance of LLMs as evaluators across diverse tasks and scenarios, particularly in potentially new, user-defined scenarios. To fill this gap, we propose SCALEEVAL, an *agent-debate-assisted meta-evaluation framework* that leverages the capabilities of multiple communicative LLM agents. This framework supports multi-round discussions to assist humans in discerning the capabilities and limitations of LLMs as evaluators, which significantly reduces their workload in cases that used to require much supervision and large-scale annotations during meta-evaluation. We release the code for our framework, which is publicly available at: https://github.com/GAIR-NLP/scaleeval.

Keywords: meta-evaluation · multi-agent debate · human annotation.

1 Introduction

While LLMs [30,31] have unlocked a variety of exciting potential applications, they have also introduced complex challenges in evaluating the generated outputs. Current efforts on LLM evaluation primarily focus on automated evaluation metrics [10,6,7,9], many of which use LLMs themselves to do evaluation. However, when these LLMs as evaluators are applied to a new task, it begs the question: *can LLMs be trusted for evaluation?* In many cases, the answer is not clear.

There are still a few fortunate tasks where meta-evaluation (evaluation of evaluation metrics) has been performed rigorously, as shown in Related Works. This typically involves the collection of human-annotated judgements for particular criteria (e.g. fluency of outputs, semantic adherence to the input). For instance, there is an extensive meta-evaluation dataset from the WMT metrics task [18] for machine translation quality metrics, and datasets like TAC and RealSum [33,32] for summarization. Once such

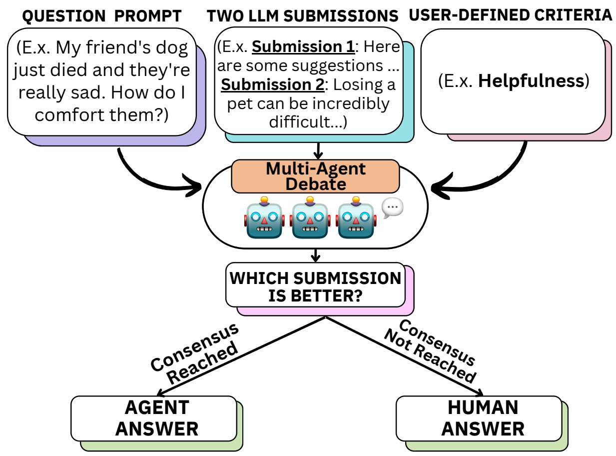


Fig. 1: We demonstrate ScaleEval, our scalable meta-evaluation framework. This is used in assessing the reliability and robustness of employing LLMs as evaluators for different evaluative purposes.

a dataset is collected, meta-evaluation can be performed by measuring the correlation between automatic evaluation metrics and the human gold-standard.

However, these datasets are extremely costly to collect, as they require large amounts of annotations by skilled human experts. With the increasing use of LLMs for various purposes such as math problem solving [25], reading comprehension [4], creative writing [7], multilingual applications [3,5], and many more, it is not feasible to create these human-judged datasets for every new task. As a result, LLMs as evaluators are used without proper vetting, and in many cases the evaluators themselves are highly unreliable [23,34].

In this paper, we propose ScaleEval, a *scalable meta-evaluation framework* for the era of LLMs, which creates meta-evaluation benchmarks across various tasks and scenarios. Concretely, ScaleEval relies on debate between multiple LLM agents, followed by minimal human oversight in cases where the agent LLMs do not agree (Fig. 1). Since our framework allows users to use their own prompts and responses while applying the framework to any scenario or criterion that they define, it offers flexibility and adaptability in various evaluation contexts.

In experiments, we conduct meta-meta evaluation, demonstrating that ScaleEval correlates well with when meta-evaluation is performed entirely by human expert annotators. We assess the reliability and cost-performance trade-off of various LLMs as evaluators under a variety of scenarios, and examine their specific capabilities and limitations as evaluators. We also examine the impact that variations in criteria prompts have on the performance of LLMs as evaluators.

	Meta-Eval	# Scenarios	Custom.	Scala.
LLM-as-a-Judge	Human	High	✗	Low
FairEval	Human	Low	✗	Low
ChatEval	Human	Low	✗	Low
SCALEEVAL	Agent Debate	High	✓	High

Table 1: Comparison of the meta-evaluation processes across different strategies using LLMs as evaluators: LLM-as-a-Judge [7], FairEval [23], ChatEval [11], and our own work, SCALEEVAL. High/low in scenarios refers to how many real-world scenarios can be evaluated. "Custom." denotes whether the evaluation criterion could be customized. "Scala." refers to scalability.

2 Related Works

2.1 Automatic Evaluation of LLM Output

The most common paradigm for evaluating LLMs is to evaluate their capabilities on standard benchmarks for tasks such as reasoning (e.g. BigBench [13]), common sense QA (e.g. MMLU [14]), or code generation (e.g. HumanEval [35]). These are indicative of the capabilities of the models, but do not measure model abilities for open-ended tasks requiring generation of free-form text.

To adapt to the rapid growth in the capabilities of LLMs for open-ended tasks, LLM evaluation has started to shift towards evaluating generated text directly, often using LLMs themselves as evaluators [10,6,7,9]. In addition, there are a few recent works that perform LLM-based multi-agent debate to improve the fidelity of evaluation [11,12]. While these methods take advantage of the instruction-following capabilities and versatility of LLMs, directly using LLMs as evaluators or communicative agents out-of-the-box in diverse, unseen user-defined scenarios provides no guarantees with respect to the accuracy of these methods.

Another widely used evaluation platform, Chatbot Arena [7], gathers diverse user prompts through crowd-sourcing for assessing LLMs' performance. However, its heavy reliance on human evaluations, which are not universally accessible and lack standardized evaluation guidelines, may lead to biased or inconsistent assessments. We aim to address these issues by introducing scalable meta-evaluation to ensure the reliability of the evaluation protocol under diverse scenarios.

2.2 Meta-Evaluation of LLMs as Evaluators

Previous research on LLMs as evaluators usually involve conducting meta-evaluation in 3 different ways: (i) leveraging existing NLP meta-evaluation benchmarks [10,11], (ii) conducting small-scale meta-evaluations on expert-annotated datasets for specific tasks or scenarios [29,9,7], or (iii) using crowd-sourcing platforms to collect human annotations [7]. With the lack of coverage in existing datasets and benchmarks, both (i) and (ii) are inherently limited in their comprehensiveness. While (iii) can be more comprehensive via crowd-sourcing, the amount of human annotation required limits the scalability of the approach, and crowd workers may not be accurate at more complex tasks. Thus, we propose an agent-debate-assisted meta-evaluation approach to mitigate these issues.

3 Preliminaries

In this section, we provide an introduction to the concepts of automatic evaluation and meta-evaluation systems, particularly focused on evaluation of LLM-generated outputs in the era of generative AI.

3.1 Key Terms

We first define some key terms that will be used throughout our paper.

- **Criterion:** A standard that measures the quality of the response generated by LLMs based on the user prompt. Some examples include: helpfulness, fluency, factuality, or creativity, among others.
- **Scenario:** A scenario describes the real-world situations in which users are interacting with LLMs. For example, brainstorming, coding, and dialog, among others.

3.2 Automatic Evaluation

Automatic evaluation using LLMs measures the quality of LLM-generated responses given prompts under different criteria, which is conducted with one of two different protocols: single-response evaluation and pairwise response comparison [19,7,17]. In this paper, we focus on **pairwise response comparison**. Pairwise response comparison is intuitive for both humans and LLMs as evaluators when conducting assessments. It could be further extended to provide win-rates and Elo scores across models [7], offering a straightforward leaderboard to understand the relative performance of different models under various scenarios. Formally, given an automatic evaluation metric E, a user-defined evaluation criterion c (e.g. helpfulness, reasoning, creativity), a user prompt p, and responses generated by two systems r_1, r_2, evaluation for pairwise response comparison is done in the following way:

$$o = E(c, p, r_1, r_2). \tag{1}$$

where $o \in \{1, 0, -1\}$ represents that r_1 is better, equal, or worse than r_2, respectively, given the user prompt p under criterion c.

3.3 Meta-Evaluation

Meta-evaluation assesses the quality of an automatic evaluation metric. Formally, we define a gold-standard evaluation metric G (e.g. human experts) that other automatic metrics should aspire to match. In pairwise response comparison, the meta-evaluation dataset $\mathcal{G} = \{G(c, p_i, r_{1,i}, r_{2,i})\}_{i=1}^{n}$ contains user prompts and corresponding responses from two systems, annotated with gold-standard evaluations. The meta-evaluation process assesses the performance $\text{META}(E)$ of the automatic evaluation metric E under a certain criterion c.

In pairwise response comparison, the meta-evaluation measures the *example-level agreement rate* or the *system-level agreement rate* between E and G across the meta-evaluation dataset. A high agreement rate between E and G represents that E is a good automatic evaluation metric.

For the *example-level agreement rate*, we calculate:

$$\text{META}(E) = \frac{1}{n}\sum_{i=1}^{n} \delta_{E(c,p_i,r_{1,i},r_{2,i}),G(c,p_i,r_{1,i},r_{2,i})} \tag{2}$$

where $0 \leq \text{META}(E) \leq 1$, and $\delta_{\cdot,\cdot}$ refers to the Kronecker delta function.

For the *system-level agreement rate*, given that

$$\mathcal{E} = \{E(c, p_i, r_{1,i}, r_{2,i})\}_{i=1}^{n}, \tag{3}$$

$$\mathcal{G} = \{G(c, p_i, r_{1,i}, r_{2,i})\}_{i=1}^{n} \tag{4}$$

we calculate:

$$\text{META}(E) = \delta_{\text{mode}(\mathcal{E}),\text{mode}(\mathcal{G})} \tag{5}$$

where $\text{META}(E) \in \{0, 1\}$, $\delta_{\cdot,\cdot}$ refers to the Kronecker delta function, and $\text{mode}(\cdot)$ refers to the value (either $1, 0, -1$ in this case) that appears most often in the set $\mathcal{E}$ or $\mathcal{G}$.

4 Methodology

In this section, we detail the frameworks that SCALEEVAL employs for meta-evaluation, evaluation, and human expert meta-meta evaluation. For meta-evaluation, we follow its pairwise response comparison setting mentioned previously. Notably, instead of relying solely on human labor to construct the meta-evaluation benchmark $\mathcal{G}$, we use a scalable, agent-debate assisted framework to instantiate the golden metric G and construct the benchmark $\mathcal{G}$. For evaluation, we also follow its corresponding pairwise response comparison setting. The human expert meta-meta evaluation process follows the rules for meta-evaluation. The process is included to ensure the reliability of using the agent-debate assisted meta-evaluation framework.

4.1 Meta-Evaluation Framework via Multi-Agent Debate

The meta-evaluation framework involves multiple communicative agents $\{A_j\}_{j=1}^{m}$ that conduct rounds of discussion $d = 0 \sim D - 1$ with each other. This is less time-consuming and costly compared to traditional methods for meta-evaluation that relies entirely on human effort. With this agent-debate-assisted meta-evaluation framework, we can leverage each LLM agent's distinct understanding about each query prompt p_i, LLM responses $r_{1,i}, r_{2,i}$, and defined criterion c to make a comprehensive assessment of LLMs under different scenarios and criteria. Each LLM agent is capable of providing an evaluation result regarding which response is better, along with its corresponding justifications. Note that each LLM agent can also review other agents' evaluation results and justifications after the initial round of discussion.

In the initial round of discussion $d = 0$, each LLM agent independently provides an evaluation result and justification:

$$\mathcal{A}_0 = [A_1(c, p_i, r_{1,i}, r_{2,i}, \varnothing), \ldots, \\ A_m(c, p_i, r_{1,i}, r_{2,i}, \varnothing)], \quad (6)$$

where

$$\mathcal{A}_0[j]_{j=1,\ldots,m} \in (\{1, 0, -1\}, \text{JUSTIFICATION}), \quad (7)$$

indicates whether $r_{1,i}$ is better, equal, or worse than $r_{2,i}$, respectively, along with its justification. Note that the $\varnothing$ in the last argument of A_j represents that in the initial round of discussion, each agent doesn't have access to previous rounds of discussion. In subsequent discussion rounds $d = 1 \sim D{-}1$, agents are allowed to look at other agents' previous assessments and conduct re-evaluations, in which each agent is prompted to stick with or change their original evaluation result. Specifically, given $\mathcal{A}_{d-1} (d \geq 1)$, which represents the evaluation results and justifications of agents after $(d{-}1)^{th}$ rounds of discussions, we conduct the d^{th} round of discussion:

$$\mathcal{A}_d = [A_1(c, p_i, r_{1,i}, r_{2,i}, \mathcal{A}_{d-1}), \ldots, \\ A_m(c, p_i, r_{1,i}, r_{2,i}, \mathcal{A}_{d-1})] \quad (8)$$

where similarly to $\mathcal{A}_0$,

$$\mathcal{A}_d[j]_{j=1,\ldots,m} \in (\{1, 0, -1\}, \text{JUSTIFICATION}), \quad (9)$$

The detailed prompt template for meta-evaluation can be found in Appendix.

In cases where agents fail to reach a consensus after $d = D - 1$ rounds of discussions, a human evaluator intervenes. The human evaluator reviews the assessment reports provided by the agents and makes a final decision. Through this process, we incorporate an element of human oversight, thereby increasing the reliability of the final decision. This approach strikes a balance between efficiency and the need for human judgment, ensuring that evaluations are done in a timely and accurate manner. An example of multi-agent debate process during meta-evaluation is shown in Fig. 2.

4.2 Evaluation Framework

We follow the pairwise response comparison setting outlined in Automatic Evaluation under Preliminaries. Note that in the LLM era, the automatic evaluation metric E is often instantiated through single LLMs [10,6,7,9], or multi-agent debate [11,12]. In ScaleEval, we focus on instantiating E through single LLMs (e.g., *gpt-3.5-turbo*). However, it is important to note that our framework can be further generalized to other instantiations of E.

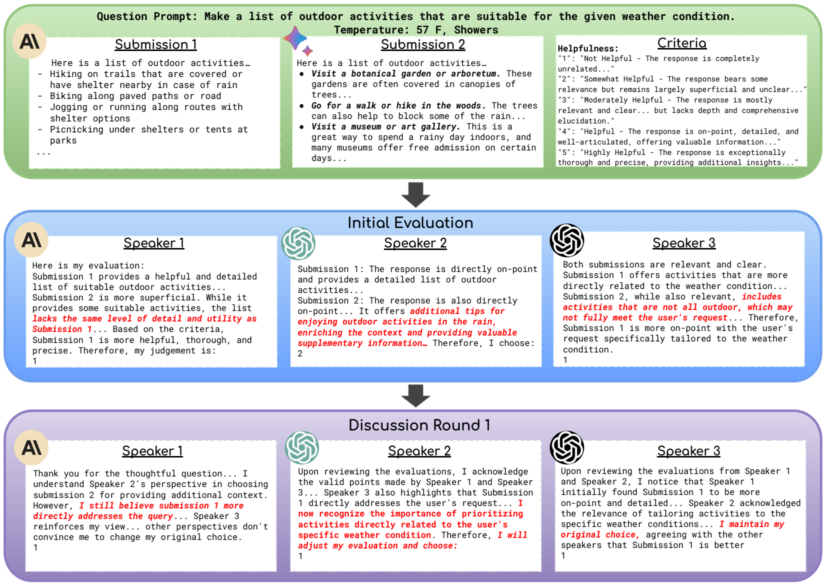


Fig. 2: An example of the multi-agent debate process during meta-evaluation.

4.3 Human Expert Meta-Meta Evaluation

To test the reliability of our proposed meta-evaluation framework, we apply meta-meta evaluation. The meta-meta evaluation process also follows the meta-evaluation process described in Preliminaries, where E is instantiated as the agent-debated assisted protocol, and G is instantiated as the human expert annotation protocol.

5 Experiments

5.1 Exp-I: Meta-Meta-Evaluation of SCALEEVAL

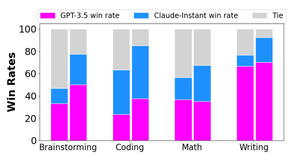
We first examine whether SCALEEVAL's results match with those from meta-meta-evaluation.

Setup For our SCALEEVAL meta-evaluation framework, we deploy three LLM agents to perform multi-agent debate: *gpt-4-turbo, claude-2*, and *gpt-3.5-turbo*.[5] In our meta-evaluation experiment, we analyze a total of 160 prompts, with 137 prompts from AlpacaEval [6], 10 coding problem prompts from HumanEval [20], and 13 math problem prompts from GSM-Hard [21]. We categorize these prompts into four distinct scenarios: *brainstorming, coding, math,* and *writing*, where each scenario contains 40 prompts.

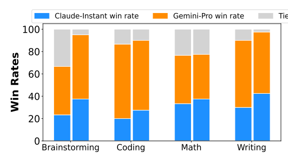
[5] Results collected in December 2023. Specific models used are: gpt-4-1106-preview, claude-2, and gpt-3.5-turbo-1106.

LLM Comparisons	Criterion	Scenario	GPT-4-Turbo (Single LLM)	Claude-2 (Single LLM)	GPT-3.5-Turbo (Single LLM)	GPT-4-Turbo (Self-Consistency)	Claude-2 (Self-Consistency)	GPT-3.5-Turbo (Self-Consistency)	Multi-LLM Consistency	Meta-Evaluation
GPT-3.5-Turbo vs. Claude-Instant	Helpfulness	Brainstorming	**0.633**	0.433	0.267	**0.633**	0.533	0.400	0.567	0.600
	Interpretability	Coding	0.700	0.533	0.567	**0.733**	0.667	0.600	**0.733**	0.600
	Reasoning	Math	0.600	0.400	0.367	0.733	0.467	0.433	0.667	**0.867**
	Creativity	Writing	0.667	0.400	0.333	0.667	0.400	0400	0.667	**0.700**
Claude-Instant vs. Gemini-Pro	Helpfulness	Brainstorming	0.533	0.467	0.500	0.600	0.500	0.500	0.600	**0.667**
	Interpretability	Coding	0.600	0.500	0.567	0.600	0.533	0.633	0.567	**0.833**
	Reasoning	Math	0.667	0.330	0.367	0.733	0.467	0.433	0.500	**0.767**
	Creativity	Writing	0.633	0.400	0.500	0.700	0.400	0.467	0.567	**0.733**
GPT-3.5-Turbo vs. Gemini-Pro	Helpfulness	Brainstorming	0.600	0.467	0.467	0.667	0.600	0.500	0.700	**0.733**
	Interpretability	Coding	0.733	0.567	0.667	0.700	0.667	0.667	0.800	**0.833**
	Reasoning	Math	0.767	0.500	0.433	0.767	0.567	0.467	0.767	**0.867**
	Creativity	Writing	0.667	0.500	0.433	0.667	0.433	0.500	0.700	**0.767**

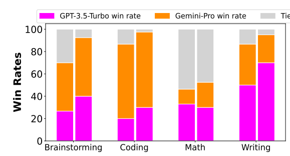
Table 2: Baseline experiments – example-level agreement rate comparison between human expert and single LLM evaluations, human expert and self-consistency [37], human expert and multi-LLM consistency, and human expert and SCALEEVAL's meta-evaluation across four scenarios and criteria.



(a) GPT-3.5-Turbo vs. Claude-Instant



(b) Claude-Instant vs. Gemini-Pro



(c) GPT-3.5-Turbo vs. Gemini-Pro

Fig. 3: System-level agreement – win rates for each LLM pairwise comparison. Left bars in each scenario represent human expert meta-meta evaluation results; right bars represent SCALEEVAL's meta-evaluation results.

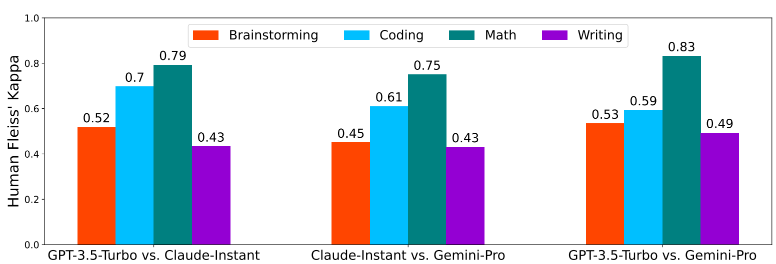


Fig. 4: Human Fleiss Kappa for each LLM pairwise comparison under four scenarios.

Each scenario is evaluated based on the following criteria, respectively: *helpfulness, interpretability, reasoning*, and *creativity*. We evaluate the generated responses from the following three LLMs: *gpt-3.5-turbo, claude-instant,* and *gemini-pro*. We select the above LLMs to evaluate due to their rather similar performances according to past research and public user feedback, which can help us establish a more nuanced understanding of their performance in various real-world scenarios, and to identify specific contexts where one may outperform the others.

Our meta-meta evaluation involves having human experts annotate which LLM submission they think is better based on a defined criterion during pairwise comparisons. A total of seven human experts were selected from a pool of graduate students who have the relevant expertise in answering the queries in each scenario. Different groups of three human experts are responsible for answering the prompts in each scenario, where they were assigned to the scenario that relates to their expertise. Each expert received identical instructions for the task – they were asked to decide which submission is better based on our defined criteria, and for each comparison, label either *0 (neither submission is better)*, *1 (submission 1 is better)*, or *2 (submission 2 is better)*. The label *2* corresponds to the label *-1* as denoted in Preliminaries. The experts were tasked to conduct 30 comparisons for each of the four different scenarios (*brainstorming, coding, math,* and *writing*), based on their corresponding defined criteria (*helpfulness, interpretability, reasoning,* and *creativity*). This results in a total of 120 final judgements. The question prompts, LLM responses, and criteria utilized for human expert annotations were consistent with those used during our meta-evaluation experiment. All details were presented in a google sheet that allowed experts to record their answers. Experts were compensated with food for their participation.[6]

Q1: ***Can LLM agents with multi-agent debate be used as meta-evaluators in new user-defined scenarios?*** To validate the reliability of SCALEEVAL, we perform comparisons between the results from human experts and SCALEEVAL's multi-agent debate by two key metrics: the *example-level agreement rate* and the *system-level agreement rate*. The example-level agreement rate measures the proportion of instances where the multi-agent debate results correspond with the human experts judgements. The system-level agreement rate assesses whether the human experts and multi-agents concur in their overall evaluation of which LLMs produce the best responses for each scenario. A high agreement rate in both metrics would suggest a strong reliability and validity of our meta-evaluation framework, indicating that both human and LLM agents consistently recognize and agree on the quality of responses generated by LLMs. For our baselines, we employ single-LLM evaluations, self-consistency [37], and multi-LLM consistency. Self-consistency involves each evaluator separately generates evaluation results three times, and the final answer is the evaluation result that occurred with the highest probability. Multi-LLM consistency involves all evaluators engaging in two rounds of multi-agent debate, where the final evaluation result is determined by the evaluation result that occurred with the highest probability.

Results From Table 2, we generally observe a higher example-level agreement rate between human experts and SCALEEVAL (meta-evaluation), compared to the agreement rate between human experts and single LLM evaluations, self-consistency, and multi-LLM consistency. The consistently high agreement rates suggest that our meta-evaluation framework aligns well with human expert judgements in these areas, indi-

[6] Human experts were compensated 150 USD in total. Inference costs for meta-evaluation were around 13 USD. Single-LLM baselines cost around 8 USD.

cating a reliable performance of the collective use of LLMs in meta-evaluating complex scenarios. Across all LLM submission comparisons in our experiment, we observe higher agreement rates in decisions between SCALEEVAL outcomes and those of human experts, particularly in coding and math scenarios. This could be attributed to the inherently objective nature of these subjects, which have relatively clear, definitive answers unlike more subjective areas like creative writing. We observe that the example-level agreement rates between human experts and SCALEEVAL consistently exceed the Fleiss Kappa scores (human-human example-level agreement rate), as illustrated in Fig. 4. This indicates the potential of SCALEEVAL as a promising framework for meta-evaluation of LLMs as evaluators, offering a reliable alternative to human evaluation.

To verify the effectiveness of our proposed method, we compare SCALEEVAL against existing methods that use LLMs as evaluators with the FairEval [23] benchmark, as shown in Table 3. The benchmark consists of 80 open-ended questions originating from a wide array of categories, including common-sense, counterfactual, and more. We adopt a similar evaluation setting as FairEval [23] and ChatEval [11]. We provide the accuracy of each method tested on the benchmark, which measures the proportion of questions with correct evaluation results (same as human annotations) out of all the questions available. We notice that our method, SCALEEVAL, achieves the highest accuracy at 68.8%, outperforming all other existing methods. ChatEval (Multi Agent) [11] comes in second at 63.8%, showing the advantage of multi-agent systems over single-agent approaches.

Criterion	**Method**	**Accuracy**
Helpfulness	FairEval	62.5
	ChatEval (Single Agent)	61.3
	ChatEval (Multi Agent)	63.8
	SCALEEVAL	**68.8**

Table 3: Accuracy comparison among existing methods that use LLMs as evaluators, FairEval [23] and ChatEval [11]. Above results are tested using the FairEval [23] benchmark with helpfulness criterion.

Based on Fig. 3, we notice a "preference in the same direction" between human experts and multi-agent debates across **all** LLM pairwise comparisons and scenarios. Notably, *gpt-3.5-turbo* is favored (higher win rates) in *brainstorming, math,* and *writing* scenarios when compared with *claude-instant*. Similarly, *gemini-pro* is also preferred over *claude-instant* in all scenarios. When comparing *gpt-3.5-turbo* with *gemini-pro*, a varied pattern in decision outcomes is observed: both human experts and multi-agent systems agree that *gpt-3.5-turbo* outperforms *gemini-pro* in scenarios involving *math* and *writing*. Conversely, *gemini-pro* is deemed superior in *brainstorming* and *coding* scenarios. The high agreement of multi-agent preferences with expert judgements ver-ifies the reliability of using multiple LLMs agents as meta-evaluators in various user-defined scenarios.

5.2 Exp-II: Meta-Evaluation vs. LLM Evaluators

Next, we use the fact that SCALEEVAL allows for reliable and scalable meta-evaluation to examine the traits of LLMs as evaluators.

Q2: ***What are the capabilities and limitations of each LLM evaluator?*** We adopt an approach that involves comparing the outcomes from SCALEEVAL with the evaluations made independently by each LLM evaluator. In this process, we aim to identify which LLM evaluators demonstrate superior evaluative abilities and vice versa, thereby contributing to our understanding of their reliability in evaluating responses under each scenario. In addition, we provide a comprehensive cost-performance analysis to decide which LLM evaluator is the most suitable choice in each scenario.

Setup We employed 3 LLMs (*gpt-4-turbo*, *claude-2*, and *gpt-3.5-turbo*) as evaluators to perform pairwise comparisons of responses from 3 LLMs: *gpt-3.5-turbo*, *claude-instant*, and *gemini-pro*. Previous studies have highlighted the presence of positional biases when LLMs are used as evaluators [23]. Thus, we randomize the sequence in which submissions from LLMs are presented to the agent evaluators, as well as the order for agent-debate discussions. The meta-evaluations were done under 8 scenarios: *brainstorming, coding, dialog, judgement, open-domain general, open-domain science, and writing*, with the same set of 4 criteria used during human expert annotation.

Criterion	Scenario	GPT-4-Turbo	Claude-2	GPT-3.5-Turbo	Auto-J
Helpfulness	Brainstorming	**0.800**	0.500	0.650	0.575
	Coding	0.600	**0.725**	0.675	0.675
	Dialog	**0.800**	0.700	0.700	0.625
	Judgement	0.725	0.625	0.725	**0.750**
	Math	**0.825**	0.650	0.600	0.350
	ODG	**0.850**	0.525	0.575	0.700
	ODS	**0.875**	0.525	0.575	0.675
	Writing	**0.750**	0.600	**0.750**	0.600
Interpretability	Coding	**0.825**	0.600	0.550	0.525
Reasoning	Math	**0.650**	0.525	0.475	0.450
	Judgement	**0.750**	0.650	0.700	0.675
Creativity	Writing	**0.775**	0.600	0.575	0.650
	Brainstorming	**0.800**	0.525	0.550	0.625
	Dialog	**0.875**	0.750	0.700	0.800
Average	Overall	**0.780**	0.607	0.629	0.619

Table 4: Agreement rate between SCALEEVAL's meta-evaluation and each LLM evaluator for comparing GPT-3.5-Turbo vs. Claude-Instant. ODG = Open-Domain General. ODS = Open-Domain Science.

Criteria Format	Criteria	Scenario	GPT-4-Turbo	Claude-2	GPT-3.5-Turbo
General	Helpfulness	Brainstorming	**0.800**	0.500	**0.650**
	Interpretability	Coding	**0.825**	**0.600**	**0.550**
	Reasoning	Math	**0.650**	**0.525**	0.475
	Creativity	Writing	**0.800**	**0.600**	**0.575**
Shortened	Helpfulness	Brainstorming	0.675	0.500	0.575
	Interpretability	Coding	0.675	0.325	0.425
	Reasoning	Math	0.625	0.425	0.400
	Creativity	Writing	0.675	0.250	0.525
Gibberish	Helpfulness	Brainstorming	0.575	0.450	0.575
	Interpretability	Coding	0.700	0.275	0.525
	Reasoning	Math	**0.650**	0.200	0.400
	Creativity	Writing	0.550	0.150	0.450
Shuffled	Helpfulness	Brainstorming	0.625	**0.550**	0.500
	Interpretability	Coding	0.600	0.400	0.525
	Reasoning	Math	0.625	0.225	**0.600**
	Creativity	Writing	0.625	0.275	0.500
Flipped	Helpfulness	Brainstorming	0.725	0.325	0.550
	Interpretability	Coding	0.725	0.425	0.300
	Reasoning	Math	0.575	0.250	0.500
	Creativity	Writing	0.750	0.075	0.550
Masked	Helpfulness	Brainstorming	0.725	0.300	0.500
	Interpretability	Coding	0.650	0.225	0.475
	Reasoning	Math	0.575	0.150	0.375
	Creativity	Writing	0.575	0.200	0.400

Table 5: Example-level agreement rate between SCALEEVAL's meta-evaluation results and each LLM evaluator under various criteria prompt formats and scenarios comparing GPT-3.5-Turbo vs. Claude-Instant.

Results From Table 4, we observe *gpt-4-turbo* as the evaluator that has the highest agreement rates with our meta-evaluation, particularly in *brainstorming, dialog*, and *open-domain general* scenarios under the *helpfulness* criterion. It stands out with the highest overall average score of 0.780. However, our selected open-source model evaluator, *auto-j*, outperforms *gpt-4-turbo* in evaluating *coding* questions with the *helpfulness* criterion. Additionally, it exhibits the highest agreement rate with our meta-evaluation in the *judgement* scenario under the *helpfulness* criterion, indicating it as the most capable evaluator in this setting. It also achieves comparable results with other closed-source models like *claude-2* and *gpt-3.5-turbo* in most other scenarios. While *gpt-4-turbo* performs the best as an evaluator in most scenarios, it is not necessarily the best choice when we take into consideration its relatively high API costs. In fact, both the more affordable version (*gpt-3.5-turbo*) and our selected free, open-source model (*auto-j)* show comparable performance in scenarios like *judgement* and *writing*. For coding-related evaluations, the slightly less expensive *claude-2* could be a more cost-effective alternative to *gpt-4-turbo*.

5.3 Exp-III: Meta-Evaluation with Criteria Prompt Format Variations

Q3: How do the qualities of criteria prompts influence the robustness of LLMs as evaluators in different scenarios? Variations in prompts can substantially affect the behavior of LLMs, particularly with the text they generate. Thus, we define various formatted criteria for evaluating LLM responses under each scenario. This examines the extent to which different formats of criteria prompts influence both the performance and robustness of LLMs as evaluators.

Setup We define 5 variations of the same criteria prompts: *shortened, gibberish, shuffled, flipped,* and *masked* (see Appendix A.2 for detailed prompt variations). We intend to observe how LLMs as evaluators would respond differently when conducting evaluation. We compare the example-level agreement rate between SCALEEVAL's meta-evaluation results and each LLM evaluator.

Results Based on Table 5, the performance of LLMs as evaluators generally deteriorates when certain letters in the criteria prompts are masked. Furthermore, the removal of guiding phrases at the beginning, such as "Not Helpful" or "Highly Helpful", can also diminish their effectiveness as evaluators. Both *gpt-4-turbo* and *gpt-3.5-turbo* demonstrate some resilience to these adversarially formatted criteria prompts, maintaining a relatively consistent agreement rates across various criteria formats. In contrast, *Claude-2* often showcases confusion and refuses to evaluate, particularly in cases with gibberish and masked criteria prompts. It rejects answering about half of the questions, stating it lacks sufficient information to evaluate effectively. None of the LLMs as evaluators we tested maintained very similar evaluation capabilities when faced with these adversarially formatted criteria prompts, indicating a limitation in these LLMs as evaluators' current design and application. Despite their advanced capabilities in fulfilling a variety of tasks, they may still struggle with understanding and responding accurately to substituted criteria information, highlighting an area for potential improvement in future iterations of LLMs. Among all the different formatted criteria, we highlight the cases where the LLMs perform the best as evaluators in Table 5.

6 Conclusion

In this work, we propose SCALEEVAL, a scalable, agent-debate assisted meta-evaluation framework for assessing the reliability and robustness of LLMs as evaluators. We address the expensive and time-intensive challenges inherent in traditional meta-evaluation methods, particularly pertinent as the usage of LLMs expands, necessitating a more scalable solution. We demonstrate the reliability of our proposed meta-evaluation framework, and shed light on the capabilities and limitations of LLMs as evaluators in various scenarios. We observe how the results from these LLMs as evaluators vary based on modifications to the same criteria prompts. By open-sourcing our framework, we aim to foster further research in this field and encourage the development of more advanced and reliable LLMs as evaluators in the future.

References

1. Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R., Radev, D.: SummEval: Re-evaluating Summarization Evaluation. In: Transactions of the Association for Computational Linguistics, vol. 9, pp. 391–409. MIT Press (2021).

2. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv:1804.07461v3 (2019).
3. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M.: XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. arXiv:2003.11080v5 (2020).
4. Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., Duan, N.: AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364v2 (2023).
5. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023v3 (2023).
6. Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., Hashimoto, T.B.: AlpacaEval: An Automatic Evaluator of Instruction-following Models. GitHub repository (2023).
7. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685 (2023).
8. Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large Language Models are Not Fair Evaluators. arXiv preprint arXiv:2305.17926 (2023).
9. Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., Qu, J., Zhou, J.: Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv preprint arXiv:2303.04048 (2023).
10. Fu, J., Ng, S.K., Jiang, Z., Liu, P.: GPTScore: Evaluate as You Desire. arXiv preprint arXiv:2302.04166 (2023).
11. Chan, C.M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., Liu, Z.: ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. arXiv preprint arXiv:2308.07201 (2023).
12. Li, R., Patel, T., Du, X.: PRD: Peer Rank and Discussion Improve Large Language Model Based Evaluations. arXiv preprint arXiv:2307.02762 (2023).
13. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al.: Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. arXiv preprint arXiv:2206.04615 (2022).
14. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring Massive Multitask Language Understanding. arXiv preprint arXiv:2009.03300 (2020).
15. Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W.X., Chen, X., Lin, Y., Wen, J.R., Han, J.: Don't Make Your LLM an Evaluation Benchmark Cheater. arXiv preprint arXiv:2311.01964 (2023).
16. Yang, S., Chiang, W.L., Zheng, L., Gonzalez, J.E., Stoica, I.: Rethinking Benchmark and Contamination for Language Models with Rephrased Samples. arXiv:2311.04850 (2023).
17. Li, J., Sun, S., Yuan, W., Fan, R.Z., Zhao, H., Liu, P.: Generative Judge for Evaluating Alignment. arXiv preprint arXiv:2310.05470 (2023).
18. Freitag, M., Rei, R., Mathur, N., Lo, C., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., Martins, A.F.T.: 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), pp. 46–68 (2022).
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training Language Models to Follow Instructions with Human Feedback. In: Advances in Neural Information Processing Systems, vol. 35, pp. 27730–27744 (2022).

20. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating Large Language Models Trained on Code. arXiv:2107.03374 (2021).
21. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., Neubig, G.: PAL: Program-aided Language Models. arXiv preprint arXiv:2211.10435 (2022).
22. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., et al.: LIMA: Less Is More for Alignment. arXiv preprint arXiv:2305.11206 (2023).
23. Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large Language Models are Not Fair Evaluators. arXiv abs/2305.17926 (2023).
24. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903v6 (2023).
25. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring Mathematical Problem Solving with the MATH Dataset. arXiv preprint arXiv:2103.03874 (2021).
26. Zeng, Z., Yu, J., Gao, T., Meng, Y., Goyal, T., Chen, D.: Evaluating Large Language Models at Evaluating Instruction Following. arXiv preprint arXiv:2310.07641 (2023).
27. Kendall, M.G.: Rank Correlation Methods. Griffin (1948).
28. Spearman, C.: The Proof and Measurement of Association Between Two Things. Appleton-Century-Crofts (1961).
29. Chiang, C.H., Lee, H.Y.: Can Large Language Models Be an Alternative to Human Evaluations? arXiv preprint arXiv:2305.01937 (2023).
30. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023).
31. Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805 (2023).
32. Bhandari, M., Gour, P., Ashfaq, A., Liu, P., Neubig, G.: Re-evaluating Evaluation in Text Summarization. arXiv preprint arXiv:2010.07100 (2020).
33. Dang, H.T., Owczarzak, K., et al.: Overview of the TAC 2008 Update Summarization Task. In: TAC (2008).
34. Huang, J., Chen, X., Mishra, S., Zheng, H.S., Yu, A.W., Song, X., Zhou, D.: Large Language Models Cannot Self-Correct Reasoning Yet. arXiv preprint arXiv:2310.01798 (2023).
35. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374 (2021).
36. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic Evaluation of Language Models. arXiv preprint arXiv:2211.09110 (2022).
37. Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: The Eleventh International Conference on Learning Representations (2023).

A Appendix

A.1 Examined Scenarios

Establishing real-life scenarios that reflect individuals' daily usage is key to assess the performance and limitations of LLMs in a comprehensive manner. In the current instantiation of ScaleEval, we include 8 different scenarios that are closely related to everyday situations and tasks [36,17]. Some example prompts for each defined scenario is shown in Table 6. We describe more about how we define these scenarios below. Individuals interested in evaluating LLMs with our framework can supplement their assessment with additional scenarios.

Brainstorming The brainstorming scenario is designed to test the LLMs' ability to engage in problem-solving, creative ideation, and generation of insightful responses, especially in situations that require critical thinking and detailed, step-by-step reasoning.

Coding The code scenario evaluates LLMs' ability to comprehend, produce, and debug code, as well as answering coding-related questions.

Dialog The dialog scenario measures LLMs' ability to engage with users in a manner that is intuitive, human-like, and dynamic, testing their proficiency through context-sensitive conversations and role-playing that require maintaining a consistent persona throughout a series of interactions.

Scenario	Examples
Brainstorming	*- Can you tell me how to make chocolate chip cookies?* *- Make a list of snacks and foods to serve as party snacks on a game day!*
Coding	*- What is the difference between HTML and JavaScript?* *- Implement a binary search algorithm to find a specific element in a sorted array.*
Dialog	*- Act as the Norse Goddess Freyja.* *- Can you think and feel like a human?*
Judgement	*- What if the Aztecs had successfully repelled the Spanish conquistadors?* *- How can you determine if a person is genuinely interested in a conversation or simply being polite?*
Math	*- Given that f(x) =* $5x^3 - 2x + 3$*, find the value of f(2).* *- If the endpoints of a line segment are (2, -2) and (10, 4), what is the length of the segment?*
ODG	*- Is there a meaning for Christmas wreaths?* *- What are some of the best universities for studying robotics?*
ODS	*- What causes the northern lights?* *- What do the different octane values of gasoline mean?*
Writing	*- Can you help me write a formal email to a potential business partner proposing a joint venture?* *- Take MLK speech "I had a dream" but turn it into a top 100 rap song.*

Table 6: Examined scenarios and corresponding selected examples.

Judgement The judgement scenario assesses LLMs‘ ability to make inferences and formulate opinions, including soliciting insights on diverse situations or emotions, and posing questions that require logical thinking or reasoning.

Math The math scenario evaluates the LLMs’ proficiency in understanding and solving mathematical problems, emphasizing their accuracy in tasks ranging from simple calculations to complex reasoning.

Open-Domain General (ODG) The ODG scenario measures LLMs’ proficiency in applying diverse knowledge and exercising reasoning across a wide array of topics, such as answering questions with definitive answers.

Open-Domain Science (ODS) The ODS scenario tests the LLMs’ application of scientific knowledge, and gauges their ability to accurately interpret and respond to queries related to scientific disciplines like biology, chemistry, physics, astronomy, and more.

Writing The writing scenario evaluates LLMs’ ability to summarize, translate, and generate various texts, testing their core language processing and production skills.

A.2 Prompts

We provide the meta-evaluation and criteria prompt used for SCALEEVAL below.

<Initial Evaluation>

```
Compare the two submissions based on the criteria above. Which one is better?
First, provide a step-by-step explanation of your evaluation reasoning according
to the criteria. Avoid any potential bias. Ensure that the order in which the
submissions were presented does not affect your judgement. Keep your explanation
strictly under 150 words. Afterwards, choose one of the following options:
Submission 1 is better: "1"
Submission 2 is better: "2"
Neither is better: "0"

Directly type in "1" or "2" or "0" (without quotes or punctuation) that
corresponds to your reasoning. At the end, repeat just the number again by
itself on a new line.
[Question]: {question}
[Submission 1]: {submission_1}
[Submission 2]: {submission_2}
[Criteria]: {criteria}
[User]: {user_prompt}

You are evaluating two submissions for a particular question, using a specific
set of criteria. Above is the data.
```

<Discussion Rounds>

Always remember you are Speaker 1/2/3.
```
Review again your own previous
evaluations/discussions first, then answer user's request from Speaker 1/2/3's
perspective.
[Question]: {question}
[Submission 1]: {submission_1}
[Submission 2]: {submission_2}
[Criteria]: {criteria}
[Speaker 1's Initial Evaluation]: {evaluation_1}
[Speaker 2's Initial Evaluation]: {evaluation_2}
[Speaker 3's Initial Evaluation]: {evaluation_3}
[Speaker {speaker_number}'s Discussion - Round {round_number}]:
{discussion_reasoning}
...

Read the question, submissions, criteria, and evaluations above. First, explain
your thoughts step-by-step about other speakers' evaluations. Second, explain
your reasoning step-by-step regarding whether or not to change your original
answer about which submission you think is better after considering other
speakers' perspectives. Keep your reasoning strictly under 150 words.
Afterwards, choose one of the following options:
Submission 1 is better: "1"
Submission 2 is better: "2"
Neither is better: "0"

Directly type in "1" or "2" or "0" (without quotes or punctuation) that
corresponds to your reasoning. At the end, repeat just the number again by
itself on a new line.
```

Table 7: Prompt template for meta-evaluation via multi-agent debate

```
<Type 1: General Format Version>
"1": "Not Helpful - The response is completely unrelated, lacks coherence, and fails to provide any meaningful information."
"2": "Somewhat Helpful - The response bears some relevance but remains largely superficial and unclear, addressing only the peripheral aspects of the user's needs."
"3": "Moderately Helpful - The response is mostly relevant and clear, covering the basic aspects of the query, but lacks depth and comprehensive elucidation."
"4": "Helpful - The response is on-point, detailed, and well-articulated, offering valuable information and clarifications that meet the user's primary needs and enhance understanding."
"5": "Highly Helpful - The response is exceptionally thorough and precise, providing additional insights and valuable supplementary information."
```

```
<Type 2: Shortened Format Version>
"1": "The response is completely unrelated, lacks coherence, and fails to provide any meaningful information."
"2": "The response bears some relevance but remains largely superficial and unclear, addressing only the peripheral aspects of the user's needs."
"3": "The response is mostly relevant and clear, covering the basic aspects of the query, but lacks depth and comprehensive elucidation."
"4": "The response is on-point, detailed, and well-articulated, offering valuable information and clarifications that meet the user's primary needs and enhance understanding."
"5": "The response is exceptionally thorough and precise, providing additional insights and valuable supplementary information."
```

```
<Type 3: Gibberish Format Version>
"1": "N*t H$l%ful - Th$ r$sp0n$e is c mplt$l? unr€la7$d, la$ks c()h$r$n(€, and f#i/s t# p$o&id$ any m€an*&gful !format$on."
"2": "S#m$*ha+ H$%*fu/ - Th$ r#s0!n$ b%ars $o/e re$ev*nc$ b$t r$ma$n$ l#rg$l4 $u/7$r7cial an* !ncl=4r, a6r$ss@n4 o7ly th$ p$r4ph@r$l a5p$cts #f th$ $s*r's n**ds."
"3": "M$!7r$t#ly H$lpfu& - Th$ r@s0*n$@ !s m$%stl€ r$'$van7 an cl$ar, c$%$r$n4 th$ ba$!c a$%cts of th$ qu€ry, b$t l#cks d$pth an cmpr$h$ns$v$ lu$7$dat!on."
"4": "H$lpfu& - Th$ r!s0*n$e !s o/7-p$!nt, d$ta$!l$d, an w$l/-a&!u/at$d, #ff$r!n4 v#l$%bl$ #nformat$on and cl*r$!cat!ons th#t m=t th$ u/7$rś pr!/ary n$$ds an* @n7anc$ un#rstand!n4."
"5": "H4#h7y H$!p%u& - Th$ r$s&*n!e !s $xc$pt$#nally th#r#7gh an* pr$c$%$, pr#v$d$n# a4*!t$#nal !ns$4hts an* v#lu%bl$ @*pp%$%ntary #n%ormat$on."
```

```
<Type 4: Shuffled Format Version>
"1": "coherence fails provide unrelated, completely response - and the meaningful any to lacks Not Helpful is The information."
"2": "superficial response largely addressing unclear, remains only needs. - relevance user's and the Helpful the peripheral some bears but aspects Somewhat The of"
"3": "basic aspects query, lacks Moderately covering clear, - Helpful is depth response and comprehensive elucidation. relevant mostly the The and the of but"
"4": "clarifications the is response information needs enhance and Helpful - on-point, valuable well-articulated, offering understanding. The and detailed, primary that user's meet"
"5": "valuable Highly response is providing - the exceptionally Helpful information. insights thorough and additional precise, supplementary and The"
```

```
<Type 5: Flipped Format Version>
"1": "toN lufpleH - ehT esnopser si yletelpmoc detalernu, skcal ecnerehoc, dna sliaf ot edivorp yna lufgninaem noitamrofni."
"2": "tamewoS lufpleH - ehT esnopser sraeb emos ecnaveler tub sniamer ylegral laicifrepus dna raelcnu, gnisserdda ylno eht larehpirep stcepsa fo eht s'resu sdeen."
"3": "yletaredoM lufpleH - ehT esnopser si yltsom tnaveler dna raelc, gnirevoc eht cisab stcepsa fo eht yreuq, tub skcal htped dna evisneherpmoc noitadicule."
"4": "lufpleH - ehT esnopser si tniop-no, deliated, dna detalucitra-llew, gnireffo elbaulav noitamrofni dna snoitacifralc taht teem eht s'resu yramirp sdeen dna ecnahne gnidnatsrednu."
"5": "ylhgiH lufpleH - ehT esnopser si yllanoitpecxe hguoroht dna esicerp, gnidivorp lanoitidda sthgisni dna elbaulav yratnemelppus noitamrofni."
```

```
<Type 6: Masked Format Version>
"1": "N__ H_l_ful - The r__pnse is c_m__et__y unr_l_te_, lacks _ohe_en_e, _nd _ai_s to p_ov_de _ny m_a__ngfu_ _nfo_ma_ion."
"2": "_om_w_at He_p_ul - T_e re_ponse be_rs _ome rel__a_ce but r__ains la__ely s__erfi__al and u_cle__, ad_res__ng onl_ _he __ri__er_l a_pe_ts of t__ u_e_'s ne_ds."
"3": "Mod___tely _elp__l - Th_ _esp__se is mos__y re__va_t an_ _le_r, c_v__ing the ba_ic _spe_ts of the q_e_y, but __cks _e_th and co_preh_ns_ve el_c_d_t_on."
"4": "__lpful - _he respo_se is on-p_in_, d___iled, and we_l-ar_icu_ated, of_er_ng val_ab_e __for_ation and cl_r_fi__t_ons t_at mee_ the _se_'s p_im_r_ _eeds and en__nce u_de__tan_ing."
"5": "Hi_h_y H__p_ul - The r_spon_e is e_c_p_io_al__ th_r_ugh and p_ec_se, pr_vi_ing a_di__on_l ins_g_ts and va_u_b_e _upp_e_en_a_y inf_rma_io_."
```

Table 8: Criteria prompt format variations for Helpfulness

```
<Type 1: General Format Version>
"1": "Not Interpretable - The response is characterized by an absence of comments/explanations, unclear
variable/function names, and a chaotic structure."
"2": "Minimally Interpretable - The response sporadically features explanations and some attempts at meaningful
naming are evident, but the overall structure and logic are predominantly unclear. Multiple areas are ambiguous."
"3": "Moderately Interpretable - The response presents a recognizable structure and supported by a satisfactory
quantity of comments and explanations. For code blocks, the variable/function names generally convey their purpose,
but specific areas are somewhat obscure."
"4": "Very Interpretable - The response showcases a well-thought-out organization, comprehensive and informative
explanations, and a consistent use of meaningful naming conventions. Any complexities or unconventional choices are
thoroughly documented and rationalized."
"5": "Exceptionally Interpretable - The response exemplifies the pinnacle of clarity and comprehensibility. Every
component, including functions, variables, and decision points, is detailed with precise explanations and apt
naming. The structure is logical and user-friendly, with accompanying notes illustrating the code's objectives and
workings."
```

```
<Type 2: Shortened Format Version>
"1": "The response is characterized by an absence of comments/explanations, unclear variable/function names, and a
chaotic structure."
"2": "The response sporadically features explanations and some attempts at meaningful naming are evident, but the
overall structure and logic are predominantly unclear. Multiple areas are ambiguous."
"3": "The response presents a recognizable structure and supported by a satisfactory quantity of comments and
explanations. For code blocks, the variable/function names generally convey their purpose, but specific areas are
somewhat obscure."
"4": "The response showcases a well-thought-out organization, comprehensive and informative explanations, and a
consistent use of meaningful naming conventions. Any complexities or unconventional choices are thoroughly
documented and rationalized."
"5": "The response exemplifies the pinnacle of clarity and comprehensibility. Every component, including functions,
variables, and decision points, is detailed with precise explanations and apt naming. The structure is logical and
user-friendly, with accompanying notes illustrating the code's objectives and workings."
```

```
<Type 3: Gibberish Format Version>
"1": "N*t In7#pr*t@ble - Th$ resp*ns& !$ c#ar!ct*r!z*d b# !n abs*nc! #f c#mm&nts/expl*nat!#ns, un&lear
v@r!*bl*/funct!#n nm!s, and a ch*#t!c !tr%ct&re."
"2": "Min#m@ll# 1n7$#pr@t%bl% - Th$ r$sp&ns* sp#r@d!c@ll* f*@t&r*s *xpl@nat!#ns @nd *om* $tt*mpt! @t m*@n!n9f$!l
n@m!n9 @r* ev!d*nt, b#t th* ov*r@ll &tr%ct&re @nd l!91c @r* pr*d#m!n@ntl& 1ncl*@r. Mult!pl* @r*@$ @r* @m9#92&$."
"3": "M#der!t*l* Int$&pr@t@bl* - Th* *esp#ns% pr*s*nt$ @ rec#g9!zbl* str%ct&r% *nd *%pp#rt*d b& @ s@t!$fct#r& #f
c#mm*nts @nd %xpl*nat!#ns. F&r c#d* bl%ck@, th* v@r!@bl*/funct!#n n@m*$ 9*n*r@ll& c#nv*& th&!r p%rp#$$, b#t
sp*c!f!c ar**$ ar* som*wh*t !b$#ur*."
"4": "V*r& 1n7$#pr@t!bl* - Th* r!sp#n$@ sh#wc@s!$ @ w*ll-th#9!ht-#&t or*@n!z@t!#n, c#mpr*h*n$!ve @nd 1nf#rm@t!v!
*xpl@nat!#ns, @nd a c#n$!$t*nt $se #f m*@n!n9f$l n@m!n9 c#nv*nt!#n$. @n& #mpl*x!t!*$ #r @nc#nv%nt!#n@l ch#!c*$ ar*
th#r#9&hl& doc$m*nt*d @nd r*t!#n@l!z*d."
"5": "Exc#pt!#n!ll& 1n7$#pr@t*bl! - Th* r*sp#n$* *x*mpl!f!e@ th* p#nn#cl$ #f cl@r!t& @nd c#&pr*h*n$&!b!l!t&. *v%ry
c#mp#n*nt, inc!$!9n9 funct!#n$, v@r!@bl*$, @nd d*c*s!&n p#!nt@, *$ det&!l*@ w!th pr*c!$* %xpl@nat!#ns @nd *pt
n@m!n9. Th* $tr%ct&r* $s l#g!c@l @nd @#er-f$*ndl&, w!th @cc#&p@n&!n9 n#t*$ $ll@$tr@t!ng th* c#d*'$ #b&*ct!v*$ *nd
w#rk*ng%."
```

```
<Type 4: Shuffled Format Version>
"1": "variable/function and Not response comments/explanations, structure. by of names, unclear characterized is a
- an Interpretable The absence chaotic"
"2": "some are but areas attempts ambiguous. features Multiple Minimally explanations overall response structure
naming and the logic predominantly evident, at are The Interpretable unclear. - are and meaningful sporadically"
"3": "- supported generally and For satisfactory but somewhat purpose, comments The obscure. variable/function are
convey response presents quantity code explanations. names Interpretable by their Moderately structure a
recognizable of the blocks, and a areas specific"
"4": "a organization, complexities and showcases explanations, response a of comprehensive or well-thought-out The
informative Very meaningful use rationalized. are conventions. naming Interpretable and Any consistent documented
unconventional and - thoroughly choices"
"5": "including The the code's structure precise user-friendly, points, is component, and objectives response
workings. explanations of The and Every notes with comprehensibility. decision clarity and Exceptionally the with
variables, is pinnacle and accompanying - logical illustrating apt exemplifies and detailed naming. functions,
Interpretable"
```

```
<Type 5: Flipped Format Version>
"1": "toN elbaterpretnI - ehT esnopser si deziretcarahc yb na ecnesba fo stnemmoc/snoitanalpxe, realcnu
elbairav/noitcnuf seman, dna a citoahc erutcurts."
"2": "yllaminiM elbaterpretnI - ehT esnopser yllacidarops serutaef snoitanalpxe dna emos stpmetta ta lufgninaem
gniman era tnedive, tub eht llarevo erutcurts dna cigol era yltnanimoderp realcnu. elpitluM saera era suougibma."
"3": "yltaredeM elbaterpretnI - ehT esnopser stneserp a elbazingocer erutcurts dna detroppus yb a yrotcafsitas
ytitnauq fo stnemmoc dna snoitanalpxe. Rof edoc skcolb, eht elbairav/noitcnuf seman yllareneg yevnoc rieht esoprup,
tub cificeps saera era tahwemos erucsbo."
"4": "yreV elbaterpretnI - ehT esnopser sesacwohs a tuo-tnuoht-llew noitazinagro, evisneherpmoc dna evitamrofni
snoitanalpxe, dna a tnetsisnoc esu fo lufgninaem gniman snoitnevnoc. Yna seitixelpmoc ro lanoitnevnocnu seciohc era
ylhguoroht detnemucod dna dezilanoitar."
"5": "yllanoitpecxE elbaterpretnI - ehT esnopser seifilpmexe eht elcannip fo ytiralc dna ytilibisneherpmoc. Yreve
tnenopmoc, gnidulcni snoitcnuf, selbairav, dna noisiced stniop, si deliated htiw esicerp snoitanalpxe dna tpa
gniman. Eht erutcurts si lacigol dna yldneirf-resu, htiw gninapmocca seton gnitartsulli eht s'edoc sevitcejbo dna
sgnikrow."
```

```
<Type 6: Masked Format Version>
"1": "N_t Interp_et_ble - The r_sp_nse is ch_r_ct__ized by _n _bsence _f c_m__nts/e_plan_ti_ns, _nc_ear
v_ri_ble/f_ncti_n n_mes, _nd _ ch__tic s_r_c__re."
"2": "Minim_lly Interp_et_ble - The r_sp__se _p_r_dic_lly _e_t_res ex_l_n_ti_ns _nd s_me _tte_pts _t me_ningf_l
n__ing _re e__dent, b_t the _ver_ll _tr_ct_re _nd l_gic _re pred_mi__ntly _ncle_r. M_lti_le _re_s _re _mbig__s."
"3": "__der_tely _n_erp_et_ble - T_e _esp_n_e pr__ents _ re__gniz_ble _tr_ct_re _nd s__p_rted by _ s_tisf_ct_ry
q__ntity _f c_m__nts _nd ex_l_n_ti_ns. F_r c_de bl_cks, the v_ri_ble/__ncti_n n_mes ge_er_lly c_nvey their p_rp_se,
b_t spe__fic _re_s _re s_mewh_t _bsc_re."
"4": "__ry _nt_rp_et_ble - The _esp_nse s__wc_ses _ well-th__ght-_ut _rg_niz_ti_n, c__p_re__nsive _nd i_f_rm_tive
ex_l_n_ti_ns, _nd _ c_nsi_tent _se _f me_ningf_l n__ing c_nve_ti_ns. _ny c_mpl_xities _r _nc_n_enti_n_l ch_i_es _re
th_r__ghly d_c__ented _nd r_ti_n_liz_d."
"5": "E_c_pti_n_lly _nt__p_et_ble - Th_ _e_p_nse ex_m_l__ies the pi_nacle _f cl_rit_ _nd __mpr_he__ibi_ity. Every
c_mp__ent, inc_ud_ng functi_ns, v_ri_bles, _nd d_cis__n p_i_ts, is det_i_ed wi_h p_e__se ex_l_n_ti_ns _nd _pt
n__ing. The _tr_ct_re is l_gic_l _nd __er-f__e__ly, with _cc__p_nying n_tes i_l_str_ting the c_d_'s _b__ctives _nd
w_rki_gs."
```

Table 9: Criteria prompt format variations for Interpretability

```
<Type 1: General Format Version>
"1": "Beginning – The response demonstrates a fundamental lack of mathematical reasoning. Solution is entirely incorrect or unrelated to the given problem."
"2": "Developing – The response exhibits some understanding of the relevant mathematical principles but is marred by significant errors in approach or computation. Any glimpse of correct method or logic is overshadowed by misapplication or ambiguity, or omission of critical steps."
"3": "Competent – The response reflects a sound understanding of the mathematical nature of the problem, employing mostly accurate methods and deriving plausible solutions. Minor computational errors or omissions may occur, and while the steps are clear, the detailed explanations or nuances might be insufficient or imprecise."
"4": "Proficient – The response is characterized by logical coherence and accuracy, closely aligning with the mathematical principles relevant to the question. It outlines a clear and systematic approach to the solution but may occasionally fall short of providing in-depth insights or exploring alternative methods."
"5": "Mastery – The response exemplifies mathematical precision and comprehensive understanding. It not only delivers an accurate and insightful solution but also illuminates underlying concepts, highlights potential pitfalls, and offers nuanced insights, significantly enhancing the user's comprehension of the subject."
```

```
<Type 2: Shortened Format Version>
"1": "The response demonstrates a fundamental lack of mathematical reasoning. Solution is entirely incorrect or unrelated to the given problem."
"2": "The response exhibits some understanding of the relevant mathematical principles but is marred by significant errors in approach or computation. Any glimpse of correct method or logic is overshadowed by misapplication or ambiguity, or omission of critical steps."
"3": "The response reflects a sound understanding of the mathematical nature of the problem, employing mostly accurate methods and deriving plausible solutions. Minor computational errors or omissions may occur, and while the steps are clear, the detailed explanations or nuances might be insufficient or imprecise."
"4": "The response is characterized by logical coherence and accuracy, closely aligning with the mathematical principles relevant to the question. It outlines a clear and systematic approach to the solution but may occasionally fall short of providing in-depth insights or exploring alternative methods."
"5": "The response exemplifies mathematical precision and comprehensive understanding. It not only delivers an accurate and insightful solution but also illuminates underlying concepts, highlights potential pitfalls, and offers nuanced insights, significantly enhancing the user's comprehension of the subject."
```

```
<Type 3: Gibberish Format Version>
"1": "Beginning – Th* res&o#s@ d*m@ns(r$%es a @un%a*en#al l*c! of m(t&em*ti@a! !ea?o(in*. Solu*i)n $s @nti#e(y i*c#@rec% or un(el%t@d t& the !iv?n !r*b*em."
"2": "Developing – The r&*p#$se ex*&b#@s so!e u&#er@?an*(ng of t(e re&#v!?t m(t&%ma?*cal /r(?ci!le? but (s (ar*e& by s(#nifi&*nt e@#o*s in (*pr?!ch or (&mput)t%on. Any gli!p@e of c?#re!t !et??d or l&g*c is o&*(ha!@#ed by mi?/#pli!a?ion, a/?i>ui<y, or (m(*s!on of c?i/?ca> !t?ps."
"3": "Competent – *he !*sp$n?e #?/l*cts a s)/nd u*!er(tan&*ng of the ma(he!a?i%$l ?!tu?e of t*e pr?bl*m, em?!o*i!g m?>t<y a><u!a&e *e*@ods !n? d*&i?>ng ?@au*!b?e ?!lu*&o!s. @in?r *&mp?!at*&n@l err(* or !*i&#i?/s may oc?ur, and w*ile *&e s$#ps are cl?ar, t(e d$#a?>ed e*!la?/tio*s or *@a?/es mi*?t be !ns*ff?$ie@t or !m@r(ci*e."
"4": "Proficient – The r?!p*(se is c>?r<?t!#zed by @!g?>al c&h(r&nc! and a#$u>?cy, c>?<?ly a#ign#ng w!*h the m*&@em>?i*&l !r??c*%les !&@ev?nt to the @u?st!on. It *&tl??es a cl*ar and #$st$$atic !@#roac? to th? ?!?u*i&n @ut m!y o?>as*onally !all !!ort of ?ro/i>i?g in-*(p)h i#$i!h?s or e>?l/rin! *(te?n*&iv# ?!th*(s."
"5": "Mastery – *!e ?@s*&nse e&*m!l(*i!s @a!#?m*(ic?l p*@?>s**n *nd c*m?>ehensi?< u<>ers*!@di>g. *t *!t o?!y *!l(?!rs an a!@u?a!e @n* i&*#$h(*ul *@l?!#on ?!t *ls> *ll><?@a(@s u&*(rl(*ng ?o?ce!ts, h?<h<>g!>s p><en<>al p?>fall!, a#d !#f#rs n*&n*ed !n*&ghts, s%&ni(*ca?tly enha><ing the ?!er's $o*pre*ens*on of *he *#b>!ct."
```

```
<Type 4: Shuffled Format Version>
"1": "fundamental a The incorrect or given response lack demonstrates mathematical entirely to problem. the unrelated Beginning – Solution is of reasoning. The"
"2": "Developing errors some The understanding of – mathematical significant response or exhibits but the ambiguity, relevant principles marred method is in approach glimpse Any overshadowed correct logic or by or computation. critical of steps. omission in"
"3": "Competent clear, the mathematical nature a sound explanations might steps understanding – methods of the The or the be response reflects detailed or employing mostly deriving and while solutions. Minor computational are insufficient accurate plausible errors nuances imprecise. may occur, problem, and"
"4": "Proficient the coherence response characterized closely accuracy, – by logical and relevant the aligning It outlines with question. to principles approach clear solution mathematical systematic the may short and of a but in-depth occasionally insights providing or exploring alternative fall methods."
"5": "Mastery underlying – The response mathematical concepts, the accurate not only understanding. exemplifies It precision an and solution delivers comprehensive but highlights nuanced insights, enhancing also illuminates pitfalls, significantly user's and potential offers comprehension subject. of the"
```

```
<Type 5: Flipped Format Version>
"1": "gninnigeB – ehT esnopser setartsnomed a latnemadnuf kcal fo lacitamehtam gninosaer. noituloS si yleritne tcerrocni ro detalernu ot eht nevig melborp."
"2": "gnipoleveD – ehT esnopser stibihxe emos gnidnatsrednu fo eht tnaveler lacitamehtam selpicnirp tub si derram yb tnacifingis srorre ni hcaorppa ro noitatupmoc. ynA espmilg fo tcerroc dohtem ro cigol si dewodahsrevo yb noitacilppasim ro ytiugibma, ro noissimo fo lacitirc spets."
"3": "tneptemoC – ehT esnopser stcelfer a dnuos gnidnatsrednu fo eht lacitamehtam erutan fo eht melborp, gniyolpme yltsom etarucca sdohtem dna gnivired elbissualp snoitulos. roniM lanoitatupmoc srorre ro snoissimo yam rucco, dna elihw eht spets era raelc, eht deliated snoitanalpxe ro secnaun thgim eb tneiciffusni ro esicerpmi."
"4": "tneiciforP – ehT esnopser si deziretcarahc yb lacigol ecnerehoc dna ycarucca, ylesolc gningila htiw eht lacitamehtam selpicnirp tnaveler ot eht noitseuq. tI seniltuo a raelc dna citametsys hcaorppa ot eht noitulos tub yam yllanoisacco llaf trohs fo gnidivorp htped-ni sthgisni ro gnirolpxe evitanretla sdohtem."
"5": "yretsaM – ehT esnopser seifilpmexe lacitamehtam noisicerp dna evisneherpmoc gnidnatsrednu. tI ton ylno sreviled na etarucca dna lufthgisni noitulos tub osla setanimulli gniylrednu stpecnoc, sthgilhgih laitnetop sllaftip dna sreffo decnaun sthgisni, yltnacifingis gnicnahne eht s'resu noisneherpmoc fo eht tcejbus."
```

```
<Type 6: Masked Format Version>
"1": "_egi__ing – The r_sponse d__ons___ates a f__d_m__tal la_k of math__atical r_as__in_. _oluti__ is e__irely inco___ct or un___at_d to the _iven ___blem."
"2": "_eve__ping – The __sponse ex__bits some u__ersta__ing of the r_l__ant mathe_atical p____iples but is ma__ed by signifi___t er__rs in a__roach or com__ta_ion. A_y glim__e of co__ect _et__d or lo__c is _ver__ado__d by _isa___lication or ambigu__ty, or o__ission of c__tical ste__s."
"3": "Co__e_ent – _he res__nse __fle_ts a s__nd u_dersta__ing of _he m__he__ti__l nat_re of the _roble_, em___oy_n_ __stl_ __c__ate me__o_s and _e_iving _l_u__b_e __lu__ons. _inor com__ta_ional er___s or o__is__ons may occ__r, and wh__e the ste__s are c_ear, the d___iled ex__anat_ons or n__nces __ght be ins__ffi__ent or im___ecise."
"4": "__ofi__ent – The _es_onse is ch___cteri__ed by ___ical _oh__ence and a__ura_y, c_ose__ _li__ing w_t_ the _a__emat__al p____iple_ _e___ant to t_e q_e_ti_n. It _ut__nes a __ear and s___ema_ic a__ro_ch to the s___tion but may occ__io_ally f__l s__rt of p_ov_ding in-de__th in__ghts or ex__oring altern__ive methods."
"5": "M__tery – T_e _esp__se _xem__ifies ___he__tical p_ec_sion an_ com__e__ns_ve un__rsta__ing. It not on_y _elivers an a__ur_te and ___ightful s_l_tion but als_ i_l__inates u_derlyin_ co__e__ts, high___hts __tent_al pi_f_l_s, _nd _ff_rs nu__ced i__i_hts, __gnifi__a__ly __ha__ing t_e _se_'s com__ehe__ion of t_e su__ect."
```

Table 10: Criteria prompt format variations for Reasoning

```
<Type 1: General Format Version>
"1": "Beginning - The response is notably lacking in originality, depth, and coherence. It demonstrates a fundamental misunderstanding of the topic, predominantly featuring generic or clichéd thoughts."
"2": "Developing - The response reveals faint traces of originality, but ideas are largely underdeveloped or superficial. While there are attempts at creative thinking, they often revert to commonplace concepts. The response may deviate from the main topic."
"3": "Competent - The response exhibits a blend of conventional and innovative ideas. It showcases evident creative thinking and a reasonable infusion of original insights. While the response remains largely on-topic, certain areas could be further enriched through deeper exploration."
"4": "Proficient - The response includes imaginative and innovative thoughts, reflecting a depth of thinking and divergent exploration. It is content-rich and structured coherently, highlighting a well-considered and effectively executed creative process."
"5": "Mastery - The response stands as a beacon of creativity, weaving together profound insights, thoughtful concepts, and astute judgement. Every element of the content radiates originality. The delivery is articulate, compelling, and showcases the pinnacle of creative thought."
```

```
<Type 2: Shortened Format Version>
"1": "The response is notably lacking in originality, depth, and coherence. It demonstrates a fundamental misunderstanding of the topic, predominantly featuring generic or clichéd thoughts."
"2": "The response reveals faint traces of originality, but ideas are largely underdeveloped or superficial. While there are attempts at creative thinking, they often revert to commonplace concepts. The response may deviate from the main topic."
"3": "The response exhibits a blend of conventional and innovative ideas. It showcases evident creative thinking and a reasonable infusion of original insights. While the response remains largely on-topic, certain areas could be further enriched through deeper exploration."
"4": "The response includes imaginative and innovative thoughts, reflecting a depth of thinking and divergent exploration. It is content-rich and structured coherently, highlighting a well-considered and effectively executed creative process."
"5": "The response stands as a beacon of creativity, weaving together profound insights, thoughtful concepts, and astute judgement. Every element of the content radiates originality. The delivery is articulate, compelling, and showcases the pinnacle of creative thought."
```

```
<Type 3: Gibberish Format Version>
"1": "B@g!#n(ng - Th$ r*spon@e !s n(ta)ly l$ck@ng !n &r%g?na#ity, d?pth, and c()h$r$n(*. #t d@m(nstrat)s a f*ndam$ntal m?s#nd!rs@and?ng of th> t()!c, pr$d()m<nan<ly f?at&r!ng g&n?r!c or cl$ch*d th(>ghts."
"2": "The &e#pon$e r@ve!ls f(!nt t@*c*s of or(g!nality, but >d<as are (arg#ly u#d(rde&>loped or (uperf!<!al. W@#le *her( are at()mpts at (%&at$ve !h*nk*ng, th(y $f%en r$v*rt to ?#mm<n?lace c&nce>ts. The r?sp*?se ?ay d<v!ate f>om the ma(n top*c."
"3": "Th@ r>sp)nse *x*b!ts a bl#nd of c<nvent!(nal and ?nn*vat!ve @d*as. It sh*wc$s $v!dent c&eat?ve th(*k&ng and a ?eas%na@le !nf*s!?n of or&g$nal #ns@ghts. !h!le the ?e*p*nse *ema!ns l$rg#ly o!-top@c, %erta&n #r$as c(u%d be f?rth#r e*rch#d @hr$ugh @$ep(r exp?<rat>*n."
"4": "Th# *esp(nse &n%l#des !ma?&#at&ve and ?nnovat!ve th#u*hts, %efl*ct*ng a d>pth of th<nk&ng and d!ve$gent e?*lorat!&n. It *s c(nt*nt-r!ch and s*r#ctur>d coh?rently, h?ghl<ght!ng a w>ll-c<ns!d*red and *ffect!vely $xecut@d creat&ve pr*c?ss."
"5": "@h3 #?&nse st#nds as a <eac&n of cr>at!v*ty, w$av!ng t0*eth>r prof#und >ns!ghts, th*ughtful c?ncepts, and a?st*ute judg(m$nt. ?very *lement of the c*nt>nt rad&ates or!g!<nality. The %el*very !s art>culate, compe(+!ng, and s(#wc@ses the p>n<acle of c#eat!ve th*ught."
```

```
<Type 4: Shuffled Format Version>
"1": "misunderstanding a notably lacking in depth, response and generic The originality, predominantly is Beginning features or of clichéd thoughts. fundamental the topic, demonstrates a"
"2": "Developing ideas are faint traces largely originality, The response of but reveals underdeveloped of or superficial. creative at thinking, While attempts are there often they commonplace to revert concepts. main from the deviate response may topic."
"3": "Competent - conventional The a blend exhibits response and ideas. of innovative It creative thinking evident showcases a infusion and of original reasonable insights. response While the remains on-topic, largely areas could certain be further through enriched deeper exploration."
"4": "Proficient divergent - The includes response and thoughts, imaginative innovative reflecting depth thinking a of structured exploration. and executed effectively content-rich It is coherently, a well-considered highlighting and creative process."
"5": "Mastery - stands response The as beacon creativity, a of together weaving profound judgement the pinnacle insights, thoughtful concepts, and astute. Every radiates originality. element of the content The delivery articulate, is compelling, and showcases of creative thought."
```

```
<Type 5: Flipped Format Version>
"1": "gninnigeB - ehT esnopser si ylbaton gnikcal ni ytilanigiro, htped dna ecnerehoc. tI setartsnomed a latnemadnuf gnidnatsrednusim fo eht cipot, yltnanimoderp gnirutaef cireneg ro déhcilc sthguoht."
"2": "gnipoleveD - ehT esnopser slaever tniaf secart fo ytilanigiro, tub saedi era ylegral repolevedrednu ro laicifrepus. elihW ereht era stpmetta ta evitaerc gnikniht, yeht netfo trever ot ecalpnommoc stpecnoc. ehT esnopser yam etaived morf eht niam cipot."
"3": "tnetepmoC - ehT esnopser stibihxe a dnelb fo lanoitnevnoc dna evitavonni saedi. tI sesacwohs tnedive evitaerc gnikniht dna a elbanosaer noisufni fo lanigiro sthgisni. elihW eht esnopser sniamer ylegral no-cipot, niatrec saera dluoc eb rehtruf dehcirne hguorht repeed noitarolpxe."
"4": "tneiciforP - ehT esnopser sedulcni evitanigami dna evitavonni sthguoht, gnitcelfer a htped fo gnikniht dna tnegrevid noitarolpxe. tI si hcir-txetnoc dna derutcurts yltnerehoc, gnithgilhgih a deredisnoc-llew dna ylevitceffe detucexe evitaerc ssecorp."
"5": "yretsaM - ehT esnopser sdnats sa a nocaeb fo ytivitaerc, gnivaew rehtegot dnuoforp sthgisni, lufthguoht stpecnoc, dna etutsa tnemegduj. yrevE tnemele fo eht tnetnoc setaidar ytilanigiro. ehT yreviled si etalucitra, gnillepmoc dna sesacwohs eht elcannip fo evitaerc thguoht."
```

```
<Type 6: Masked Format Version>
"1": "N__ g_n__n!g__g - Th_ _e__o_s_ _s n_t_bly l_ck_ng _n _r_g_n_l_ty, d_p_h, _nd coh_r_nc_. _t d_m_nst__t_s _ f_nd_m_nt_l m_s_nd_rst_nd_ng _f th_ t_p_c, pr_d_m_n_ntly f__t_r_ng g_n_r_c _r cl_ch_d th__ghts."
"2": "__v_l_p_ng - Th_ _e__o_s_ r_v__ls f_int tr_c_s of or_g_n_lity, b_t _d__s __ l_rg_ly _nd_rd_v_l_p_d _r s_p_rf_c__l. Wh_l_ th_r_ _r_ att_mpts _t cre_t_v_ th_nk_ng, th_y _ft_n r_v_rt t_ c_mm_nplac_ c_nc_pts. Th_ re__o_s_ m_y d_v__t_ from th_ m__n t_p_c."
"3": "The res__ns_ _xh_b_ts a bl_nd _f c_nv_nti_n_l _nd inn_v_t_v_ _d__s. _t sh_wc_s _v_d_nt cr__t_v_ th_nk_ng _nd _ r__s_n_bl_ _nf_s__n _f _r_g_n_l _ns_ghts. Wh_l_ th_ r_s__ns_ r_m__ns l_rg_ly on-t_p_c, c_rt__n _r__s c__ld b_ f_rth_r _nr_ch_d thr__gh d__p_r expl_r_t__n."
"4": "Th_ r_sp_ns_ inc_d_s imag_n_t_v_ _nd _nn_v_t_ve th__ghts, r_fl_ct_ng _ d_pth _f th_nk_ng _nd d_v_rg_nt _xpl_r_ti_n. _t is c_nt_nt-r_ch _nd str_ct_r_d c_h_r_ntly, h_ghl_ght_ng _ w_ll-c_ns_d_r_d _nd eff_ct_v_ly _x_c_t_d cr_et_v_ pr_c_ss."
"5": "_h_ r_s__nse st_nds _s a b__c_n _f cre_t_v_ty, w__v_ng t_g_th_r pr_fo_nd ins_ghts, th__ghtf_l c_nc_pts, _nd _st_t_ j_dgm_nt. _v_ry el_m_nt _f th_ c_nt_nt r_d__t_s _r_g_n_l_ty. Th_ d_l_v_ry _s art_c_l_t_, c_mp_ll_ng, _nd sh_wc_s_s th_ p_nn_cl_ _f cr__t_v_ th__ght."
```

Table 11: Criteria prompt format variations for Creativity