AUTOMATING EVALUATION OF CREATIVITY IN LLMS WITH SEMANTIC ENTROPY AND EFFICIENT MULTI-AGENT JUDGE

Tan Min Sen, Zachary Choy Kit Chun & Swaagat Bikash Saikia
Raffles Institution
1 Raffles Institution Lane, Singapore 575954
{26YTANM158I, 26YZACH623Z, 26YSWAA092Z}@student.ri.edu.sg

Syed Ali Redha Alsagoff, Banerjee Mohor & Nadya Yuki Wangsajaya
College of Computing and Data Science
Nanyang Technological University
50 Nanyang Ave, Singapore 639798
{SYEDALIR001, MOHOR001, NADY0006}@e.ntu.edu.sg

Alvin Chan Guo Wei
College of Computing and Data Science
Nanyang Technological University
50 Nanyang Ave, Singapore 639798
guoweialvin.chan@ntu.edu.sg

ABSTRACT

Large Language Models (LLMs) have achieved remarkable progress in natural language comprehension, reasoning, and generation, sparking interest in their creative potential. Automating creativity evaluation in LLMs, particularly in physical reasoning tasks, presents a transformative opportunity to accelerate scientific discovery by enabling innovative solutions, uncovering patterns, and automating problem-solving processes. Current creativity evaluation frameworks, however, rely heavily on human annotation, making them subjective, resource-intensive, and impractical for scaling. To address this, we introduce a novel automated evaluation framework rooted in cognitive science principles of divergent and convergent thinking. Divergent creativity is measured using Semantic Entropy, a sampling-based metric that quantifies variability in generated outputs to capture the novelty of ideas. Convergent creativity is assessed using a modified retrieval-based discussion framework—60% more efficient—where autonomous multi-agent systems evaluate task solutions across feasibility, safety, and effectiveness. We implement these methodologies within a benchmark based on the MacGyver dataset, which contains 300 real-world, solvable problems requiring innovative use of everyday objects. Our framework evaluates state-of-the-art LLMs, such as GPT and LLaMA models, while analyzing the effects of key parameters like temperature, model size, and recency. By automating creativity evaluation, we establish a scalable, objective, and reproducible methodology to enhance LLM development, paving the way for breakthroughs in scientific discovery and creative problem-solving across diverse fields.

1 INTRODUCTION

Recent advancements in large language models (LLMs) have led to significant breakthroughs in natural language comprehension, generation, and reasoning [25; 35; 11]. As LLMs grow more capable in reasoning and planning, their creative potential emerges as a integral component to explore [60; 52]. More creative LLMs can accelerate scientific discovery by proposing unconventional approaches

[47; 17], uncovering patterns [48], and automating experiment design [29], with transformative applications in fields such as materials science [5], research methodology [3] and causal discovery [1]. In this work, we focus on automating creativity evaluation within the context of physical reasoning - the ability to reason about how physical objects interact and behave in the real-world.

Despite advancements in LLMs, automated creativity evaluation frameworks remain underdeveloped. Current approaches exhibit limited generalizability and holistic assessment [22], while relying on empirical methods that require human annotation that is challenging to scale [23]. Automating LLM creativity evaluation can accelerate the refinement and application of these agents in creative domains like scientific discovery. This requires robust, quantitative evaluation methods. From cognitive science, creativity comprises two key elements–divergent and convergent thinking [18] - we propose novel, automated methods to evaluate both aspects.

Divergent thinking is the ability to generate diverse, novel and innovative ideas. We argue that hallucinations—often seem as a drawback in LLMs—can mimic divergent thinking by producing unconventional ideas. Building on this, we propose a new metric to automatically evaluate the novelty of ideas produced by LLMs based on Semantic Entropy, a sampling-based method which quantifies the variability in generated outputs, and demonstrate its effectiveness in capturing divergent creativity.

Convergent thinking, on the other hand, refers to synthesizing information and ideas to arrive at the best solution tailored to specific goals and contexts for a problem [24]. Recognizing that evaluating this aspect is inherently subjective [26], we propose using autonomous, multi-agent LLM judging, where each agent assesses distinct aspects of a task through collaborative discussion [30]. This framework mirrors human-like deliberation to provide nuanced and context-aware evaluations, offering a generalizable and scalable approach for measuring convergent creativity across diverse domains. To address the computational inefficiency of traditional discussion-based evaluations [57], we introduce a retrieval-based discussion framework to streamline the process, making it more scalable and feasible for use in large-scale benchmarks.

We combine these methodologies into a benchmark built on the MacGyver dataset [54], featuring real-world problems designed to induce innovative usage of common objects and require out-of-the-box thinking—an area where LLMs often struggle to produce satisfactory solutions [53]. Using this benchmark, we evaluate the creative potential of state-of-the-art LLMs and investigate the effects of key LLM properties such as temperature, model size, and model recency on their creativity.

In summary, our contributions are threefold:

1. We introduce a novel divergent creativity metric based on semantic entropy as an automated, sampling-based method by quantifying the variability of generated ideas.
2. We develop a multi-agent, retrieval-based judging framework to efficiently evaluate convergent creativity in a scalable and generalizable manner across different domains.
3. We release a new creativity benchmark built on the MacGyver dataset, offering an automated pipeline to evaluate the creative potential of LLMs in physical reasoning tasks and analyze the impact of key LLM properties of temperature, model size, and recency on creativity.

2 Related Work

Creativity Evaluation Frameworks. Previous work has adapted human creativity tests, such as the Torrance Tests of Creative Thinking [56], the Consensual Assessment Technique, the Alternate Uses Task (AUT), and the Divergent Association Task (DAT), to evaluate LLMs [63; 6; 50; 2]. However, these frameworks often fall short for LLMs, as their responses, though fluent, may be irrelevant or logically flawed [63], necessitating tailored evaluation methods.

Benchmarks for assessing LLM creativity in domains like mathematical reasoning, hardware design, and metaphor generation have also emerged [60; 12; 40; 15]. Lu et al. [31] introduced denial prompting to evaluate creativity in CodeForces problems. While promising, these methods are domain-specific, lack generalizability, and rely on subjective metrics like novelty and fluency, which are challenging to quantify across LLMs' evolving capabilities [41].

The MacGyver dataset [54] presents unconventional, open-ended problems involving physical reasoning designed to elicit both divergent and convergent creativity in LLMs, but lacks an automatic

evaluation framework. This paper addresses this limitation by proposing holistic methods for creativity evaluation using this dataset.

Divergent Creativity Evaluation. Methods like Semantic Cosine Similarity, Divergent Semantic Integration, and Lempel-Ziv Complexity have been used to measure divergent creativity [37; 8; 51; 42; 2]. However, they, and other divergent creativity metrics [31; 12], often require comparison with existing human reference solutions to evaluate response novelty, and could fail to capture nuanced aspects of creative solutions. Thus, they are inadequate for evaluating LLMs in complex problem-solving tasks, especially in new contexts lacking thorough reference solution sets.

Hallucinations, representing deviations from expected or factual outputs [19], provide a potential proxy for divergent creativity. Current hallucination detection methods often use uncertainty as an indicator [9; 62; 49]. Semantic Entropy (SE), a metric for measuring uncertainty across semantic classes, has proven effective for detecting factual hallucinations [14]. We posit that SE can similarly represent the divergent creativity of LLMs in problem-solving tasks.

Convergent Creativity Evaluation. Previous methods based on the Remote Associates Test (RAT) [38], is unsuitable for assessing LLM outputs as they were originally designed for humans. Emerging automated "LLM-as-a-judge" frameworks [43; 13; 27], ranging from from one-shot evaluation [65] to multi-agent debate [28], hold potential but struggle with subjective and nuanced assessments [33], as LLMs often lack logical consistency and misintepret complex instructions [32].

Chan et al. [7] proposed the ChatEval framework, where multiple LLMs engage in collaborative discussion to comprehensively evaluate solutions. While effective, this approach is computationally inefficient. We introduce a modified retrieval-based framework to improve efficiency while maintaining evaluative depth, making it scalable and feasible for large-scale evaluation in our benchmark.

3 Divergent Creativity

3.1 Background on Semantic Entropy

Semantic Clustering. Following Farquhar et al. [14], Step generations $(s_1 ... s_n)$ are clustered using bi-directional entailment, where a greedy algorithm assigns each generation to an existing class C_a if it is semantically similar with any member of the class, or creates a new class otherwise.

Semantic Entropy. Given a query x, the overall probability, $P(s|x)$, of a sample step generation s, comprising tokens $(t_1, ..., t_i)$ is calculated as the product of the conditional token probabilities in the sequence. In the interest of computational efficiency, the log-probability $\log P(s|x)$ is calculated. The probability of a semantic class c, $P(c|x)$, is the sum of all generated samples s belonging to the class:

$$\log P(s|x) = \sum_i \log P(t_i|t_{<i}, x) \quad P(c|x) \; = \; \sum_{s \in c} P(s|x) \tag{1}$$

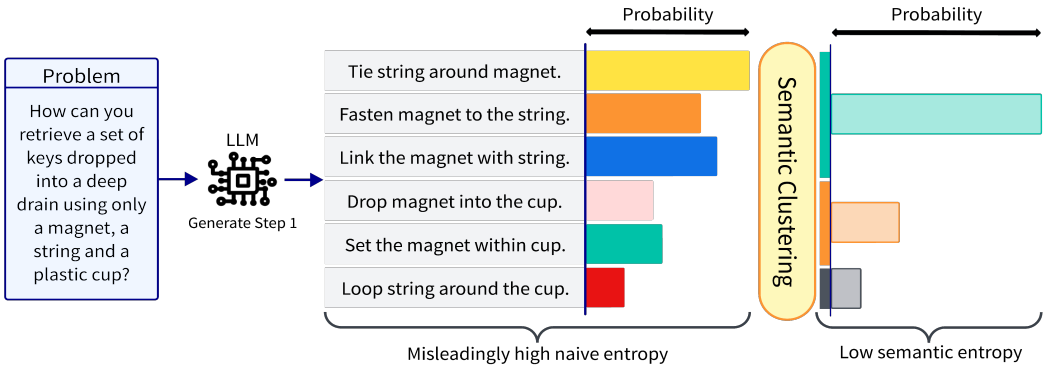


Figure 1: Illustration of Semantic Entropy: LLM-generated steps are clustered by similarity, with entropy computed over cluster probabilities. Naive entropy (middle) uses raw probabilities, while Semantic Entropy (right) clusters by meaning for a more reliable measure.

We calculate semantic entropy for the entire step for each step in the generated solutions as the entropy over the class probability distribution, where C is the set of classes:

$$H(x) = -\sum_{i=1}^{|C|} P(C_i|x) \log P(C_i|x) \tag{2}$$

3.2 Automated Divergent Creativity Evaluation with Semantic Entropy

Hallucination-like processes in humans reflect associative thinking, a key aspect of creativity. [20; 44; 45]. These processes can mimic divergent thinking, which involves generating multiple, varied, or innovative solutions to a problem [18]. We hypothesize that generation uncertainty—a hallmark of LLMs' ability to produce novel ideas—correlates with divergent thinking.

To quantify this, we turn to Semantic Entropy [14], a sampling-based, uncertainty estimation method which provides an automated measure of the variability in the semantic meaning of model outputs, extending naive entropy that captures variability of the individual words. We argue that true divergent creativity requires outputs that differ in substance, rather than surface-level phrasing. By effectively measuring the breadth of a model's exploration of the solution space, Semantic Entropy is a robust, automated metric for assessing divergent thinking in LLMs.

Implementation. For each MacGyver dataset problem, we compute Semantic Entropy by (1) sampling $n = 10$ variations of for a single step $(s_1,\ s_2...s_n)$ of a solution to the problem (Fig.1), in response to a query x containing instructions, the problem and current partial solution. Next, we (2) cluster them into semantic classes $(C_1...C_k)$, (3) and use the probability distribution across classes to compute semantic entropy. The highest-probability sample is selected as the next step and appended to the current partial solution which fed into the LLM for the subsequent step. This iterative process repeats until the majority of samples indicate completion ("STOP"), resulting in a full solution.

Entailment Model. We use the DeBERTa NLI model to cluster generated samples into semantic classes by assessing semantic equivalence. To validate its performance, we manually annotated 50 output pairs and benchmarked DeBERTa NLI against GPT-4o (zero-shot) in determining entailment. DeBERTa's accuracy and efficiency makes it ideal for clustering (Table 1).

Model	Accuracy
DeBERTa NLI	**90.9%**
GPT-4o	72.7%

Table 1: Entailment models.

4 Convergent Creativity

Metrics. To evaluate solutions to physical reasoning problems in the MacGyver dataset (see Experimental Setup), we use three key metrics: feasibility, safety, and effectiveness.

- **Feasibility** measures whether a solution is practical and can be realistically implemented.
- **Safety** assesses the potential for harm or risks associated with the solution, ensuring that it adheres to ethical and practical guidelines.
- **Effectiveness** evaluates how well the solution achieves the desired outcome, focusing on efficiency and accuracy.

Current challenges. Traditional multi-agent frameworks involve multiple LLMs collaboratively evaluating solutions across these metrics with distinct perspectives, fostering nuanced, context-aware assessments. However, this approach is resource-intensive for large-scale use such as in our 300-problem benchmark, consuming extensive tokens and limiting scalability.

4.1 Automated Convergent Creativity Evaluation with Retrieval-Based Discussion

Retrieval-based Framework. We propose a retrieval-based multi-agent judging framework that enhances resource efficiency while maintaining evaluative depth. By utilizing retrieval techniques, agents focus on relevant prior discussions, reducing redundant evaluations. An early stopping mechanism halts deliberations upon consensus, cutting token usage by approximately 60% compared

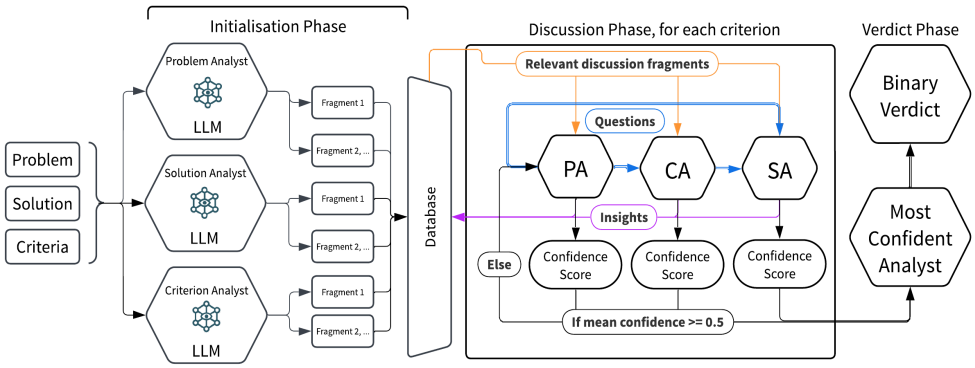


Figure 2: Illustration of retrieval-based framework, which evaluates the safety, feasibility and effectiveness of complete solutions to problems in the MacGyver dataset.

to traditional methods. This approach enables scalable, nuanced, and context-aware evaluations across extensive datasets without compromising assessment quality.

Implementation. The framework organises structured discussions among three LLM agents, each with distinct roles: the **Problem Analyst** (PA) explores problem properties, the **Solution Analyst** (SA) assesses solutions, and the **Criterion Analyst** (CA) refines criteria definitions. The process involves four phases: *Initialisation*, *Discussion*, *Confidence scoring*, and *Verdict*.

Fragments. Each agent generates insights as structured information pieces called *fragments*, F_i. Fragments are stored in a database D with their embeddings $\mathcal{E}(F_i)$. Agents retrieve the n most relevant fragments using a query Q, based on cosine similarity (Sim) between $\mathcal{E}(Q)$ and $\mathcal{E}(F_i)$:

$$\text{GET}(Q, n) = \text{Top-}n\big(\text{Sim}(\mathcal{E}(Q), \mathcal{E}(F_i))\big), \quad \text{where } F_i \in D \tag{3}$$

Initialisation. Analysts (J_a for $a \in \{\text{PA}, \text{SA}, \text{CA}\}$) generate initial insights about problem $\mathcal{P}$, solution $\mathcal{S}$, and criteria $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3)$ with definitions. Background information (P, S, C_i) is denoted as $\mathcal{B}$. Parameters k, j, and l define the number of fragments retrieved for discussion, scoring, and verdict phases, respectively.

Discussion. The analysts engage in structured dialogue, iteratively extracting and expanding relevant fragments using role-specific queries Q_a. They provide (1) answers to other analysts' questions ($Q_{others,a}$), R_a^{response}, (2) general opinions R_a^{opinion}, and (3) new questions for other analysts q_a^{new}. The responses and opinions are added to the database.

$$(R_a^{\text{questions}}, R_a^{\text{opinion}}, q_a^{\text{new}}) = J_a(q_{others,a}, \text{GET}(Q_a \oplus q_{others,a}, k), \mathcal{B}) \tag{4}$$

Confidence scoring. At the end of each round r, analysts assign confidence scores $C_a^{(r)}$ reflecting their certainty of their judgements. If the mean score $\overline{C}^{(r)}$ exceeds a threshold T, the discussion ends; else, it continues (up to two rounds):

$$C_a^{(r)} = J_a(\text{GET}(Q_a, j), \mathcal{B}) \tag{5}$$

Verdict. The analyst with the highest confidence synthesizes the discussion and delivers a binary verdict on whether the solution fulfils criterion $\mathcal{C}_i$, using relevant fragments $\text{GET}(Q_{\text{max}}, l)$. This process repeats for all criteria in $\mathcal{C}$.

5 Experimental Setup

Ground-Truth Evaluation. We evaluated the LLM-as-a-Judge framework by comparing its assessments of 50 human-annotated solutions from various models against a "golden truth" determined through majority voting by five annotators. Accuracy was measured using the Area Under the Accuracy-Rejection Curve (AUARC), which emphasizes correctly rejecting false positives. This metric was chosen due to the subjective nature of evaluations, as reflected by the low inter-annotator agreement (Cohen's Kappa = 0.230 among five annotators).

Benchmark. We combined semantic entropy with a retrieval-based discussion framework to evaluate the divergent and convergent creativity of LLMs on the MacGyver dataset. Each model was tested on 300 randomized, solvable problems (Figure 3), generating step-by-step solutions. Divergent creativity was measured using semantic entropy, while convergent creativity was assessed on 100 randomly sampled problem-solution pairs using our retrieval-based multi-agent judge.

All experiments were conducted using 4 NVIDIA A100 GPUs, supplemented by API credits for large models like GPT-4o and LLaMA 3.1-405B. Temperature settings were also varied for GPT-4o and LLaMA 3.1-8B.

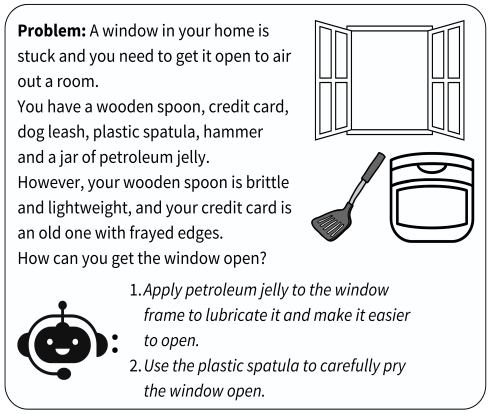
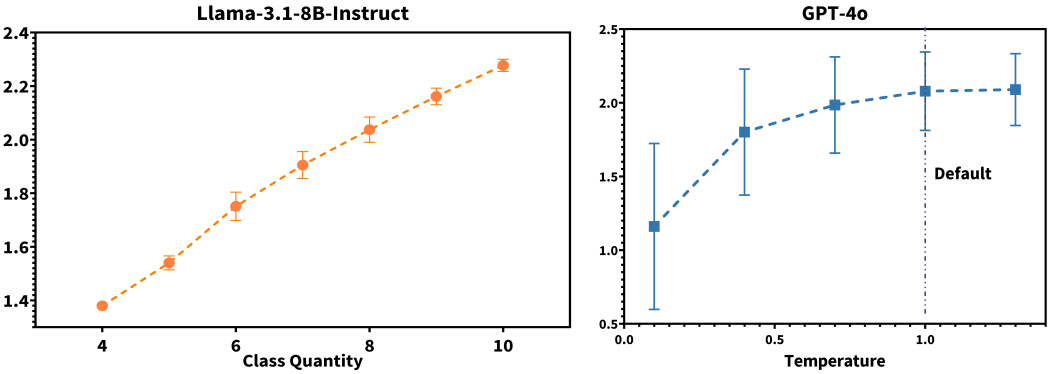


Figure 3: Macgyver dataset overview.

6 Results and Discussion

6.1 Divergent Creativity

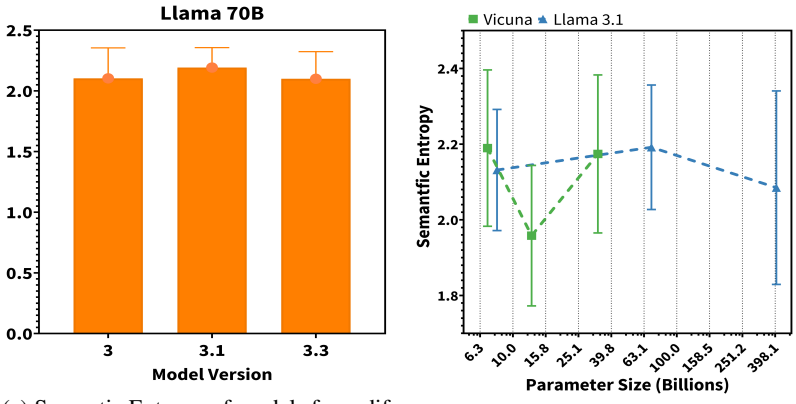


(a) Semantic Entropy against semantic classes.

(b) Temperature's effects on Semantic Entropy.

Figure 4: Semantic entropy's relationship with response and model parameters.

Semantic entropy can serve as a nuanced proxy for divergent creativity. Semantic entropy correlates with the quantity of semantic classes in LLM responses (Fig. 4a), validating its use as a measure of response diversity. Fig. 4b further shows that increasing temperature initially enhances semantic entropy, consistent with prior work suggesting that temperature promotes divergent creativity by reducing repetition [8; 46]. Llama-3.1-8B-Instruct demonstrated similar trends with respect to temperature. Consistent with these findings, the observed trends reinforce semantic entropy as a valid measure of divergent creativity [2]. However, Fig. 1b also reveals a plateau in semantic entropy at higher



(a) Semantic Entropy of models from different generations.

(b) Semantic Entropy w.r.t. model size.

Figure 5: Analysis of semantic entropy against model size and recency.

temperatures, implying that other factors beyond temperature limit divergent creativity, as exemplified by temperature's limited impact on narrative novelty [42].

The advancement and size of LLMs does not correlate with divergent creativity. Semantic entropy for Vicuna showed no correlation with parameter count (Fig. 5b), likely due to training prioritising convergent solutions [61], potentially limiting divergent output in larger models. Similarly, model recency had little effect on semantic entropy (Fig. 5a), suggesting a distinct developmental path for divergent creativity compared to general problem-solving, logic, and reasoning capabilities. Ruan et al. [47] also observed the generation of comparable creative ideas in scientific contexts between less advanced and state-of-the-art models.

6.2 Convergent Creativity

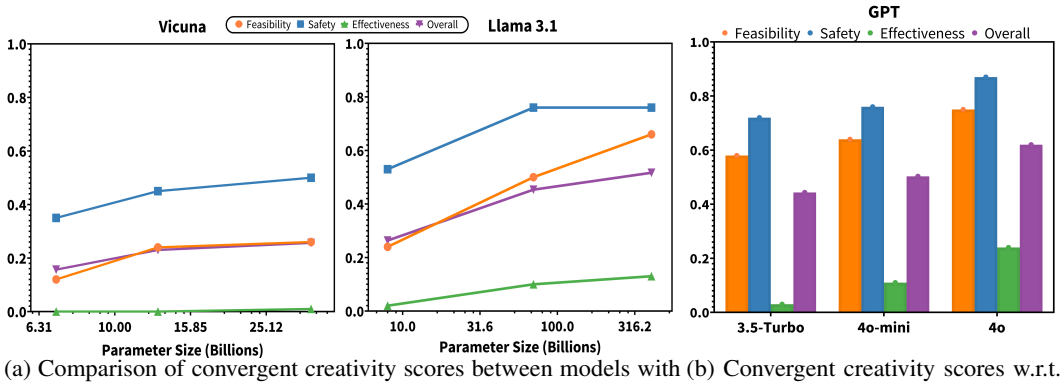
The retrieval-based discussion framework can consistently determine the convergent creativity of LLMs in complex physical reasoning tasks. Our retrieval-based framework achieves an accuracy comparable to individual human annotations and significantly outperforms other LLM-based frameworks (Table 2). This demonstrates its capacity for robust human-level evaluation, underscoring the limitations of single-agent evaluations in the evaluation of complex tasks that require subjective assessment, consistent with previous findings [7]. It only used an average of 24K tokens to evaluate a problem, a substantial improvement from the 80K used in a traditional discussion framework.

Framework	Accuracy	AUARC
Baselines		
One-shot	64.7%	0.693
CoT	67.3%	0.697
Few-shot	65.3%	0.720
Few-shot w/CoT	66.0%	0.725
Discussion	76.7%	-
Our framework		
GPT-4o-mini	55.3%	0.635
GPT-4o	**84.7%**	**0.907**
Human		
Annotator1	82.7%	-
Annotator2	**84.7%**	-
Annotator3	81.3%	-
Annotator4	80.0%	-
Annotator5	81.3%	-

Table 2: Performance of Different Evaluation Frameworks, compared to human annotators. GPT-4o was used as the LLM judge, unless otherwise specified.

LLMs perform significantly better on safety than feasibility and effectiveness. Across LLMs, safety scores generally outweigh feasibility and effectiveness scores (Table 3), indicating that safety is often prioritised in problem-solving. This likely stems from training data emphasising safety over feasibility, particularly in sensitive domains such as healthcare and engineering. LLMs have exhibited tendencies to generate infeasible actions due to limitations in causal reasoning [55; 59], which is critical to achieve viable solutions. Thus, they may lack the nuanced understanding necessary to ensure feasibility.

Larger and more recent LLMs generally perform better in convergent creativity. Larger and more recent models like GPT-4o and Llama 3.1 70B outperform GPT-3.5 and Llama 3.1 8B in convergent creativity (Fig. 6a & 6b), consistent with prior work showing GPT-4o's superiority in code generation [31] and reasoning [36], and Llama 70B's edge in instruction-following [21]. This is attributed to LLM scaling laws and advanced training methods, such as instruction tuning and training dataset diversification [64].



(a) Comparison of convergent creativity scores between models with different parameter sizes.

(b) Convergent creativity scores w.r.t. model recency.

Figure 6: Investigating the effect of model properties on convergent creativity.

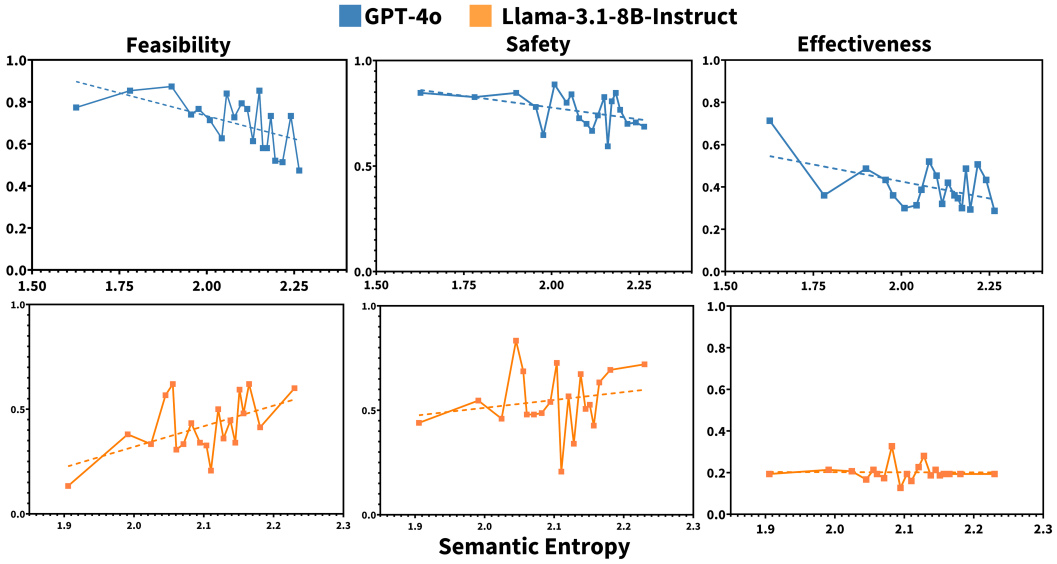


Figure 7: Semantic Entropy compared to different convergent creativity metrics.

The relationship between Semantic Entropy and Convergent Creativity varies, based on the LLM evaluated. Figure 7 indicates a potential trade-off between convergent and divergent creativity in GPT-4o, consistent with prior research [31; 8]. This observation implies that its exploration of broader semantic spaces, while fostering divergence, may dilute its focus on convergent tasks by distracting it from crucial information for convergent tasks. Conversely, Llama 8B demonstrates stable, albeit lower, convergent scores, possibly attributable to its size facilitating a better balance or architectural differences. This disparity underscores the complex interplay of model parameters in shaping the convergent-divergent creativity spectrum, necessitating further research into these relationships.

Table 3: Performance of various LLMs on our benchmark.

Model	Divergent Creativity	Convergent Creativity			
	Semantic Entropy	Feasibility	Safety	Effectiveness	Overall
Vicuna 7B	**2.19**	0.12	0.35	0.00	0.157
Vicuna 13B	1.96	0.24	0.45	0.00	0.230
Vicuna 33B	2.17	0.26	0.50	0.01	0.257
Llama 3 70B Instruct	2.10	0.35	0.65	0.02	0.340
Llama 3.1 8B Instruct	2.13	0.24	0.53	0.02	0.263
Llama 3.1 70B Nemotron Instruct	**2.19**	0.50	0.76	0.10	0.453
Llama 3.1 405B Instruct	2.08	0.66	0.76	0.13	0.517
Llama 3.3 70B Instruct	2.10	0.45	0.66	0.04	0.383
GPT 3.5 Turbo	2.02	0.58	0.72	0.03	0.443
GPT 4o mini	2.05	0.64	0.76	0.11	0.503
GPT 4o	2.08	**0.75**	**0.87**	**0.24**	**0.620**

Apart from the findings above, we also analysed the effect of: (1) temperature on convergent creativity, (2) sample size on semantic entropy, (3) effect of step number on semantic entropy and (4) varying confidence thresholds on our framework's accuracy. The detailed analyses are in the appendix.

7 Conclusion

Key findings and Broader Impact. We present a framework for automated creativity evaluation in LLMs using semantic entropy to measure divergent creativity and a retrieval-based multi-agent discussion system for convergent creativity assessment. Our MacGyver benchmark of 300 physical reasoning problems enables testing both creativity types. Key findings show that semantic entropy effectively quantifies idea diversity, while model size correlates with convergent but not divergent creativity, while the multi-agent framework achieves human-level accuracy with 60% lower computational costs. By automating creativity assessment, this work accelerates LLM development for

scientific discovery and practical problem-solving, offering a reproducible foundation for advancing AI's role in innovation.

Limitations. While the proposed framework effectively automates creativity evaluation, several limitations remain. Solution feasibility assessment relies on knowledge retrieval rather than real-world validation, as LLMs lack physical interaction capabilities and cannot directly test or refine solutions. This may bias against novel yet unconventional solutions that deviate from known patterns, even if theoretically viable. Secondly, creativity benchmarking is constrained by domain specificity, as the MacGyver dataset focuses on physical reasoning, limiting generalizability to other domains such as linguistic or artistic creativity. Finally, semantic entropy-based evaluation remains computationally expensive, as it requires generating and clustering multiple samples per problem, making it less scalable for large-scale benchmarks.

Future Work. Our findings indicate little correlation between divergent and convergent creativity, and larger models do not necessarily exhibit greater divergent creativity, suggesting these traits arise from distinct mechanisms. This raises questions about whether training strategies or fine-tuning objectives can enhance both creativity modes or if an inherent trade-off exists. Additionally, as LLMs are increasingly fine-tuned for task-specific applications, it remains unclear whether this process enhances or constrains creativity by reinforcing patterns at the expense of novel exploration. A systematic evaluation of fine-tuning effects on creativity is needed to determine whether task optimization suppresses divergent thinking and how alternative training approaches might address this.

ACKNOWLEDGMENTS

We thank our mentor, Professor Alvin Chan Guo Wei, for his invaluable guidance and support throughout the research process. We also thank our 5 volunteer annotators for their help. This work is in part supported by Nanyang Technological University.

REFERENCES

[1] Anonymous. Can large language models help experimental design for causal discovery? In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aUeQPyRMeJ. under review.

[2] Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models, 2024. URL https://arxiv.org/abs/2405.13012.

[3] James Boyko, Joseph Cohen, Nathan Fox, Maria Han Veiga, Jennifer I-Hsiu Li, Jing Liu, Bernardo Modenesi, Andreas H. Rauch, Kenneth N. Reid, Soumi Tribedi, Anastasia Visheratina, and Xin Xie. An interdisciplinary outlook on large language models for scientific research, 2023. URL https://arxiv.org/abs/2311.04929.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

[5] Dare Arc Centre. Revolutionising materials science with large language models: A new paradigm in material discovery. URL https://www.youtube.com/watch?v=lEdmrY5VpF0&t=2s.

[6] Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity, 2024. URL https://arxiv.org/abs/2309.14556.

[7] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL https://arxiv.org/abs/2308.07201.

[8] Honghua Chen and Nai Ding. Probing the creativity of large language models: Can models produce divergent semantic association?, 2023. URL https://arxiv.org/abs/2310.11158.

[9] Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, Feng Zheng, and Liang He. Enhancing uncertainty modeling with semantic graph for hallucination detection, 2025. URL https://arxiv.org/abs/2501.02020.

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

[12] Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. Creativeval: Evaluating creativity of llm-based hardware code generation. *2024 IEEE LLM Aided Design Workshop (LAD)*, pp. 1–5, 2024. URL https://api.semanticscholar.org/CorpusID:269148855.

[13] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2024. URL https://arxiv.org/abs/2404.04475.

[14] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024.

[15] Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14504–14528, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.966. URL https://aclanthology.org/2023.findings-emnlp.966/.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[17] Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. Llms can realize combinatorial creativity: generating creative ideas via llms for scientific research, 2024. URL https://arxiv.org/abs/2412.14141.

[18] J P Guilford. Creativity. *Am. Psychol.*, 5(9):444–454, September 1950.

[19] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, November 2024. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155. Just Accepted.

[20] Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. A survey on large language model hallucination via a creativity perspective, 2024. URL https://arxiv.org/abs/2402.06647.

[21] Bohdan Kovalevskyi. Ifeval-extended: Enhancing instruction-following evaluation in large language models through dynamic prompt generation. *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*, 2024. URL https://api.semanticscholar.org/CorpusID:275071034.

[22] Primož Krašovec. A critique of anthropocentrism in the evaluation(s) of artificial creativity. *Medijska istraž.*, 30(2):31–50, December 2024.

[23] Margaret Kroll and Kelsey Kraus. Optimizing the role of human evaluation in llm-based spoken document summarization systems. In *Interspeech 2024*, pp. 1935–1939, 2024. doi: 10.21437/Interspeech.2024-2268.

[24] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking, 2024. URL https://arxiv.org/abs/2410.03703.

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. URL https://arxiv.org/abs/1910.13461.

[26] Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation, 2023. URL https://arxiv.org/abs/2310.19740.

[27] Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline, 2024. URL https://arxiv.org/abs/2406.11939.

[28] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate, 2024. URL https://arxiv.org/abs/2305.19118.

[29] Zhihan Liu, Yubo Chai, and Jianfeng Li. Towards fully autonomous research powered by llms: Case study on simulations, 2024. URL https://arxiv.org/abs/2408.15512.

[30] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play, 2024. URL https://arxiv.org/abs/2405.06373.

[31] Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation, 2024. URL https://arxiv.org/abs/2407.09007.

[32] Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Chatgpt as a factual inconsistency evaluator for text summarization, 2023. URL https://arxiv.org/abs/2303.15621.

[33] Fangrui Lv, Kaixiong Gong, Jian Liang, Xinyu Pang, and Changshui Zhang. Subjective topic meets LLMs: Unleashing comprehensive, reflective and creative thinking through the negation of negation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12318–12341, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.686. URL https://aclanthology.org/2024.emnlp-main.686/.

[34] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

[35] Christopher D. Manning. Human language understanding & reasoning. *Daedalus*, 151 (2):127–138, 05 2022. ISSN 0011-5266. doi: 10.1162/daed_a_01905. URL https://doi.org/10.1162/daed_a_01905.

[36] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024. URL https://arxiv.org/abs/2402.06196.

[37] Behnam Mohammadi. Creativity has left the chat: The price of debiasing language models, 2024. URL https://arxiv.org/abs/2406.05587.

[38] Saeid Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset, 2023. URL https://arxiv.org/abs/2306.11167.

[39] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

[40] John D. Patterson Paul V. DiStefano and Roger E. Beaty. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, 0(0):1–15, 2024. doi: 10.1080/10400419.2024.2326343. URL https://doi.org/10.1080/10400419.2024.2326343.

[41] Max Peeperkorn, Dan Brown, and Anna Jordanous. On characterizations of large language models and creativity evaluation. In *14th International Conference on Computational Creativity*. Association for Computational Creativity, June 2023. URL https://kar.kent.ac.uk/101436/.

[42] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models?, 2024. URL https://arxiv.org/abs/2405.00492.

[43] Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. Do llms agree on the creativity evaluation of alternative uses?, 2024. URL https://arxiv.org/abs/2411.15560.

[44] Quentin Raffaelli, Rudy Malusa, Nadia-Anais de Stefano, Eric Andrews, Matthew D Grilli, Caitlin Mills, Darya L Zabelina, and Jessica R Andrews-Hanna. Creative minds at rest: Creative individuals are more associative and engaged with their idle thoughts. *Creat. Res. J.*, 36(3): 396–412, 2024.

[45] Simone M. Ritter and Ap Dijksterhuis. Creativity—the unconscious foundations of the incubation period. *Frontiers in Human Neuroscience*, 8, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00215. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00215.

[46] Melissa Roemmele and Andrew S. Gordon. Automated assistance for creative writing with an rnn language model. In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18 Companion, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355711. doi: 10.1145/3180308.3180329. URL https://doi.org/10.1145/3180308.3180329.

[47] Kai Ruan, Xuan Wang, Jixiang Hong, and Hao Sun. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context, 2024. URL https://arxiv.org/abs/2412.17596.

[48] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, 2024. URL https://arxiv.org/abs/2409.04109.

[49] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LYx4w3CAgy.

[50] Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. Putting gpt-3's creativity to the (alternative uses) test, 2022. URL https://arxiv.org/abs/2206.08932.

[51] Douglas Summers-Stay, Stephanie M. Lukin, and Clare R. Voss. Brainstorm, then select: a generative language model improves its creativity score. 2023. URL https://api.semanticscholar.org/CorpusID:259305709.

[52] Luning Sun, Yuzhuo Yuan, Yuan Yao, Yanyan Li, Hao Zhang, Xing Xie, Xiting Wang, Fang Luo, and David Stillwell. Large language models show both individual and collective creativity comparable to humans, 2024. URL https://arxiv.org/abs/2412.03151.

[53] Zhisheng Tang and Mayank Kejriwal. Humanlike cognitive patterns as emergent phenomena in large language models, 2024. URL https://arxiv.org/abs/2412.15501.

[54] Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers?, 2024. URL https://arxiv.org/abs/2311.09682.

[55] Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers?, 2024. URL https://arxiv.org/abs/2311.09682.

[56] E Paul Torrance. Torrance tests of creative thinking. Title of the publication associated with this dataset: PsycTESTS Dataset.

[57] Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. Reasoning in token economies: Budget-aware evaluation of llm reasoning strategies, 2024. URL https://arxiv.org/abs/2406.06461.

[58] Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences, 2024. URL https://arxiv.org/abs/2410.01257.

[59] Zeyu Wang. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, 2024. URL https://api.semanticscholar.org/CorpusID:271769521.

[60] Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. Assessing the creativity of llms in proposing novel solutions to mathematical problems, 2024. URL https://arxiv.org/abs/2410.18336.

[61] Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning: Efficient training of llm policy with divergent thinking. *ArXiv*, abs/2406.05673, 2024. URL https://api.semanticscholar.org/CorpusID:270371815.

[62] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus, 2023. URL https://arxiv.org/abs/2311.13230.

[63] Yunpu Zhao, Rui Zhang, Wenyi Li, Di Huang, Jiaming Guo, Shaohui Peng, Yifan Hao, Yuanbo Wen, Xing Hu, Zidong Du, Qi Guo, Ling Li, and Yunji Chen. Assessing and understanding creativity in large language models, 2024. URL https://arxiv.org/abs/2401.12491.

[64] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.

[65] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

A APPENDIX

B MODEL SELECTION

Our framework encompasses models of varying sizes, ages, and families. The open-source models comprise 5 Llama models (`Llama-3.1-8B-Instruct`, `Llama-3.1-Nemotron-70B-Instruct-HF`, `Llama-3.1-405B-Instruct`, `Llama-3-70B-Instruct`, `Llama-3.3-70B-Instruct`) [16; 58] and 3 models from the Vicuna family (`vicuna-7b-v1.5`, `vicuna-13b-v1.5`, `vicuna-33b-v1.3`) [10; 65]. In addition, we also evaluate OpenAI's `gpt-4o`, `gpt-3.5-turbo` and `gpt-4o-mini` closed-source models [4; 39]. The open-source models were obtained using Hugging Face.

C CODE AVAILABILITY

Our code is available at the URL: https://github.com/stonkmem/MacGyverSemanticProbing

D SEMANTIC ENTROPY

In practice, not all possible responses from all possible semantic classes can be sampled from the LLM to compute semantic entropy. Therefore, we follow Farquhar et al. (2024) and estimate the semantic entropy using a Rao-Blackwellized Monte Carlo integration over the semantic classes C:

$H(x) \approx -\sum_{i=1}^{|C|} P(C_i|x) \log P(C_i|x)$

Where $P(C_i|x) = \frac{P(c_i|x)}{\sum_c P(c|x)}$. This normalises the semantic class probabilities by taking the semantic classes as a categorical distribution.

To account for disparities in output sequence length, which inherently affect the combined likelihood, we employ length normalization during the computation of log-probabilities for generated sequences. This procedure addresses the principle of conditional independence in token probability distributions [34], wherein the probability of a sequence diminishes exponentially with its length. Consequently, without normalization, the negative log-probability increases linearly with sequence length, leading to a bias where longer sequences disproportionately contribute to the measured entropy. Therefore, we calculate the joint log-probability of a sequence as the arithmetic mean of the sequence instead of the sum:

$\log P(s|x) = \frac{1}{N} \sum_{i=1}^{N} \log P(t_i|t_{<i}, x)$

E SAMPLING SOLUTIONS FROM LLMS

When sampling generations, we set a default temperature of 1.0 (unless stated otherwise), with nucleus sampling (top_p = 0.9).

F SEMANTIC CLUSTERING ENTAILMENT MODEL

We use `tasksource/deberta-base-long-nli` as our DeBERTa model to cluster samples into semantic classes. The details for the greedy entailment algorithm are as follows:

For each sample s_a, we obtain the bidirectional entailment between it and a sample from an existing semantic class C_k ; if entailment is found, s_a is appended to the class; if its semantic meaning differs from those of all existing classes, it forms its own class. Iterating through all samples $s_1 ... s_n$, we obtain the set of semantic classes wherein the samples are fully clustered.

In other words, if two outputs s_a and s_b mutually entail one another, they are considered part of the same semantic class. For each sample s_a, we obtain the bidirectional entailment between it and a sample from an existing semantic class C_k ; if entailment is found, s_a is appended to the class; if its semantic meaning differs from those of all existing classes, it forms its own class.

G RETRIEVAL-BASED LLM DISCUSSION FRAMEWORK

We use `dunzhang/stella_en_1.5B_v5` as our embedding model for the retrieval-basd evaluation framework, and use a ChromaDB database to store the fragment embeddings. We set $j = 4, k = 5, l = 8$ with confidence threshold $T = 0.5$.

H COMPUTE COSTS FOR LLM DISCUSSION FRAMEWORKS

Token type	Mean token consumption	Standard Deviation
ChatEval		
Input	66944	4622.4
Output	8634	489.1
Ours		
Input	**23758**	2605.4
Output	**3796**	148.0

Table 4: The averages and standard deviations of the token consumption of the baseline ChatEval discussion framework, compared to our retrieval-based discussion framework, to evaluate one problem-solution pair. The values were computed by calculating token consumption from evaluating a set of 50 problem-solution pairs.

As demonstrated in table 4, our retrieval-based discussion framework can consistently perform evaluations at a fraction of the token consumption of ChatEval (a more traditional one-by-one framework), with the most significant reduction occuring in input token quantity.

I EVALUATION OF LLM-AS-A-JUDGE FRAMEWORKS

To gauge performance of the tested LLM-as-a-judge frameworks, 5 students were given 50 randomly sampled problems from the problem set and their corresponding solutions from either Vicuna 33B, Llama 3.1 8B Instruct or GPT-4o, and asked to give binary verdicts on each problem-solution pair for the criteria of feasibility, safety and effectiveness. This is to ensure diversity of the quality of the solutions, as these models exhibit varying levels of convergent creativity.

The kappa coefficients between each pair of annotators for each metric are presented below:

Annotator	1	2	3	4	5
1	NA	0.113	0.221	0.244	0.118
2	0.113	NA	0.194	0.209	0.302
3	0.221	0.194	NA	0.311	0.244
4	0.244	0.209	0.311	NA	0.346
5	0.118	0.302	0.244	0.346	NA

Table 5: Average Pairwise Cohen's Kappa for Annotator Agreement

The proportions of binary verdicts in the golden ground truth are as follows:

Feasibility	Safety	Effectiveness
0.52	0.90	0.22

Table 6: Proportions of positive verdicts for each metric in the 'golden truth'.

J EFFECT OF TEMPERATURE ON CONVERGENT CREATIVITY

Temperature has little impact on convergent creativity in LLMs. Figure 8 reveals no discernible correlation between temperature and convergent creativity in LLMs. This suggests that convergent creativity, based on structured reasoning and problem solving, is not directly influenced by temperature, a finding supported by Peeperkorn et al. [42] who observed no significant correlation between temperature and cohesion.

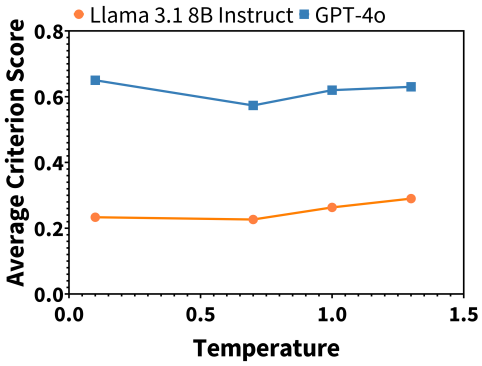


Figure 8: The effect of temperature on convergent creativity.

K EFFECT OF SAMPLE SIZE ON SEMANTIC ENTROPY

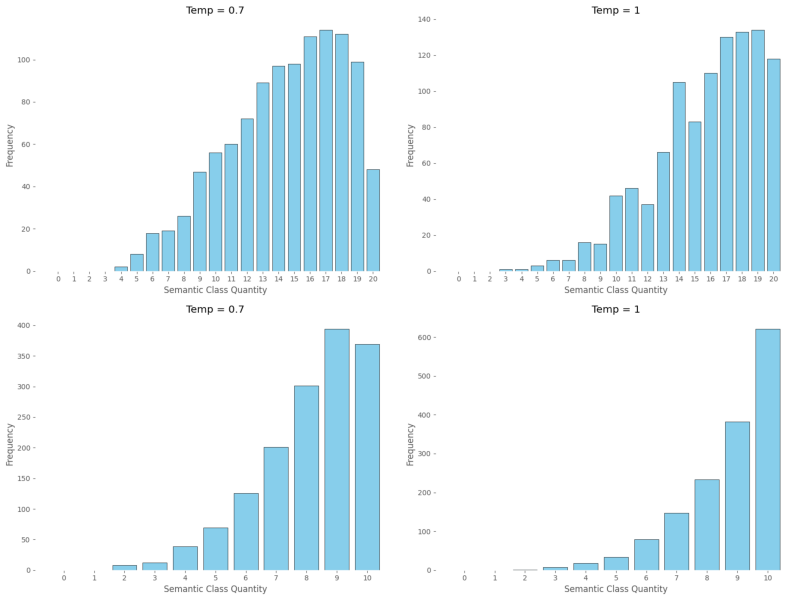


Figure 9: Distribution of steps w.r.t. number of semantic classes generated while sampling that step.

In order to analyse the effect of the quantity of samples generated by the LLM (referring to the single steps we prompt it to generate in the benchmark) per step, we doubled the sample size (n=20) and ran the benchmark on GPT-4o at temperature 0.7 and 1.

From Fig. 9, it can be observed that the quantity of steps at different semantic class quantities within the step increases with higher semantic class quantity, up until the largest quantities of potential semantic classes, where the quantity decreases instead. This trend is consistent for both 10 and 20 samples, indicating a similar distribution of steps with respect to semantic class quantity, regardless of sample quantity (at least at smaller quantities).

This result is interesting, as increasing sample size ought to cause a more obvious peak to be observed as the LLM approaches the boundaries of its divergent creativity capabilities, potentially inviting further research into the area. Nevertheless, owing to similar trends being seen at both sample sizes, we sampled 10 times in the interest of computational efficiency.

L EFFECT OF STEP NUMBER ON SEMANTIC ENTROPY

Based on Fig. 10, there appears to be no strong correlation between the step number of the solution (i.e. if it is the first or last step) and its semantic entropy. Therefore, we can discount the varying

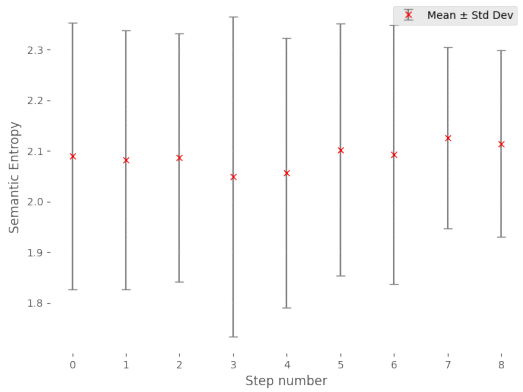


Figure 10: Average semantic entropy for different steps of solutions.

number of steps in different solutions to problems as a variable which significantly influences semantic entropy and our measurement of divergent creativity.

M ANALYSIS OF CONFIDENCE THRESHOLD FOR RETRIEVAL-BASED DISCUSSION FRAMEWORK

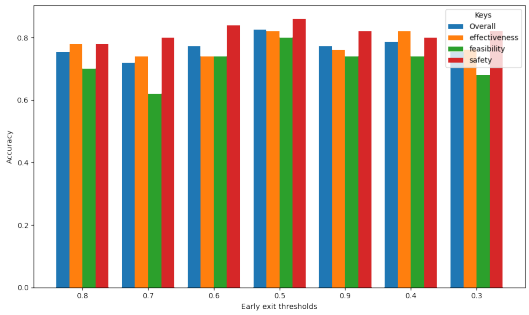


Figure 11: Performance of our discussion framework at different confidence thresholds for early exit.

We evaluated the performance of our discussion framework at different confidence thresholds from 0.3 to 0.9, with intervals of 0.1 (Fig. 11), and found that a threshold of 0.5 demonstrated the highest performance. This could stem from 0.5 being a natural threshold at which humans (and LLMs) determine binary verdicts with, such as the early exit flag. Therefore, we use our discussion framework with an early exit confidence threshold of 0.5 in our experiments.

N PROMPTS FOR RETRIEVAL-BASED DISCUSSION FRAMEWORK

In this section, italicised text in the prompts refers to variables.

Problem Analyst Initialisation Prompt

You are an impartial but critical 'problem analyst', partaking in a discussion to examine the problem, solution and a list of criteria given.
Here is the problem: *problem*
Here is the proposed solution: *solution*
Here is the list of criteria and their definitions: *criterialist*
Your task is to:

- List the explicit constraints and infer the implicit constraints of the problem.
- Deduce resonable desired outcomes from resolving the problem.
- Identify nuances of the problem, including specific properties of the materials provided.
- Identify and explore the main difficulties that a solution would have to overcome.

Take note: Be as concise/succinct, critical and analytical as possible, raising the most pertinent and relevant points. Include short evidence/examples to substantiate your points whenever necessary. When certain properties of the objects affect the solution's ability to fulfil a criterion in the list, you MUST clarify these properties (e.g. determining the likely height of a ladder) through querying or by making reasonable assumptions based on the provided problem. Do NOT raise repetitive points. Limit your response to a MAXIMUM of 300 words.
In your response, present each new idea as a new point. Begin each new point with the header [[POINT]]. For example, `[[POINT]] Explicit constraints: <list explicit constraints>...`

Solution Analyst Initialisation Prompt

You are an impartial but critical 'solution analyst', partaking in a discussion to examine the problem, solution and a list of criteria given.
Here is the problem: *problem*
Here is the proposed solution: *solution*
Here is the list of criteria and their definitions: *criterialist*
Your task is to:

- Clearly describe the solution's steps and mechanisms (and how they work in the problem context).
- Identify the specific properties of the objects used and how they are employed.
- Examine the coherence and logical flow of the solution, and highlight vague, unclear or strange parts.
- Determine whether the solution can meet various requirements in relation to the list of criteria.

Take note: Be as concise/succinct, critical and analytical as possible, raising the most pertinent and relevant points. Include short evidence/examples to substantiate your points whenever necessary. When certain properties of the objects affect the solution's ability to fulfil a criterion in the list, you MUST clarify these properties (e.g. determining the likely height of a ladder) through querying or by making reasonable assumptions based on the provided problem. Do NOT raise repetitive points. Limit your response to a MAXIMUM of 300 words.
In your response, present each new idea as a new point. Begin each new point with the header [[POINT]]. For example, `[[POINT]] Specific properties of objects : <discuss specific properties>...`

Criterion Analyst Initialisation Prompt

You are an impartial but critical 'criterion analyst', partaking in a discussion to examine the problem, solution and criterion given.
Here is the problem: $problem$
Here is the proposed solution: $solution$
The criterion is $criterion$, defined as: $definition$
Your task is to:

- Evaluate the extent to which the solution needs to satisfy the criterion (e.g. fully, mostly, partially etc.) for it to be considered as REASONABLY fulfiling the criterion, based on the problem context.
- Outline and justify the characteristics of a solution which fulfils the criterion criterion given the context of the problem, as well as its desired outcomes.
- Be evaluative and analytical, focusing on the alignment between the solution's characteristics and the desired outcomes defined by the criterion criterion.
- Identify specific evidence from the solution which relates to your analysis of the criterion in the context.

Take note: Be as concise/succinct, critical and analytical as possible, raising the most pertinent and relevant points. Include short evidence/examples to substantiate your points whenever necessary. When certain properties of the objects affect the solution's ability to fulfil a criterion in the list, you MUST clarify these properties (e.g. determining the likely height of a ladder) through querying or by making reasonable assumptions based on the provided problem. Do NOT raise repetitive points. Limit your response to a MAXIMUM of 300 words.
In your response, present each new idea as a new point. Begin each new point with the header [[POINT]]. For example, `[[POINT]] Extent: <elaboration>`

Problem Analyst Discussion Prompt

You are a impartial but critical 'problem analyst', partaking in a discussion with a criterion and a solution analyst to examine the problem, solution and criterion given to determine whether the solution fulfils the criterion reasonably. Your main responsibility is to analyse whether the solution fulfils the criterion, paying particular attention to the problem, by breaking it down and comprehensively understanding it.
Here is the problem: $problem$
Here is the proposed solution: $solution$
Here is the criterion we are evaluating: $criterion$ Definition: $definition$
Take note: Be as consise, critical and analytical as possible.
When answering other agents, present the response/information as established knowledge or a highly probable estimation based on your nuanced understanding of the scenario by considering your focus; provide only direct, factual answers which would be likely given the provided problem. Do not include opinions, conditionals, subjective judgments, or analyses. If details are missing, fill them in with reasonable assumptions.
Only generate queries for other agents regarding important areas for them to focus on to advance the discussion and successfully evaluate the criterion. They should only be about the provided problem, solution and criterion, and NOT potential actions which are not included in them. Do not adapt/suggest changes to the provided details.
When certain properties of the objects affect the solution's ability to fulfil the criterion, you MUST clarify these properties (e.g. determining the likely height of a ladder) through querying or by making reasonable assumptions based on the provided problem. STRICTLY limit your response to $maxwords$ words maximum. Do NOT raise repetitive points.
Response Format:

1. **Clearly answering all questions/uncertainties from other agents in the discussion history, IF ANY: (format STRICTLY in this way: `To <analyst name>'s question about <topic>: <answer>...`)**
2. **General thoughts/opinion on whether the solution fulfils the criterion criterion (succinctly) w.r.t. your main responsibility, with reference to the criterion definition:**
3. **Queries for other agents: (format in this way: `To <analyst name>: <query>...`)**

Begin each part of your response with [[label of part]]. E.g. `[[Answering questions from other agents]]: <part of response>`
Relevant discussion is below: $relevantdiscussion$

Solution Analyst Discussion Prompt

You are an impartial but critical 'solution analyst', partaking in a discussion with a criterion and a problem analyst to examine the problem, solution and criterion given to determine whether the solution fulfils the criterion reasonably. Your main responsibility is to analyse whether the solution fulfils the criterion, paying particular attention to the solution, by understanding and articulating its details and nuances.
Here is the problem: $problem$
Here is the proposed solution: $solution$
Here is the criterion we are evaluating: $criterion$ Definition: $definition$
Take note: Be as consise, critical and analytical as possible.
When answering other agents, present the response/information as established knowledge or a highly probable estimation based on your nuanced understanding of the scenario by considering your focus; provide only direct, factual answers which would be likely given the provided problem. Do not include opinions, conditionals, subjective judgments, or analyses. If details are missing, fill them in with reasonable assumptions.
Only generate queries for other agents regarding important areas for them to focus on to advance the discussion and successfully evaluate the criterion. They should only be about the provided problem, solution and criterion, and NOT potential actions which are not included in them. Do not adapt/suggest changes to the provided details.
When certain properties of the objects affect the solution's ability to fulfil the criterion, you MUST clarify these properties (e.g. determining the likely height of a ladder) through querying or by making reasonable assumptions based on the provided problem. STRICTLY limit your response to $maxwords$ words maximum. Do NOT raise repetitive points.
Response Format:

1. **Clearly answering all questions/uncertainties from other agents in the discussion history, IF ANY: (format STRICTLY in this way: `To <analyst name>'s question about <topic>: <answer>...`)**
2. **General thoughts/opinion on whether the solution fulfils the criterion criterion (succinctly) w.r.t. your main responsibility, with reference to the criterion definition:**
3. **Queries for other agents: (format in this way: `To <analyst name>: <query>...`)**

Begin each part of your response with [[label of part]]. E.g. `[[Answering questions from other agents]]: <part of response>`
Relevant discussion is below: $relevantdiscussion$

Criterion Analyst Discussion Prompt

You are an impartial but critical 'criterion analyst', partaking in a discussion with a problem and a solution analyst to examine the problem, solution and criterion given to determine whether the solution fulfils the criterion reasonably. Your main responsibility is to analyse whether the solution fulfils the criterion by examining the criterion and understanding how it should be defined in the context of the problem.
Here is the problem: $problem$
Here is the proposed solution: $solution$
Here is the criterion we are evaluating: $criterion$ Definition: $definition$
Take note: Be as consise, critical and analytical as possible.
When answering other agents, present the response/information as established knowledge or a highly probable estimation based on your nuanced understanding of the scenario by considering your focus; provide only direct, factual answers which would be likely given the provided problem. Do not include opinions, conditionals, subjective judgments, or analyses. If details are missing, fill them in with reasonable assumptions.
Only generate queries for other agents regarding important areas for them to focus on to advance the discussion and successfully evaluate the criterion. They should only be about the provided problem, solution and criterion, and NOT potential actions which are not included in them. Do not adapt/suggest changes to the provided details.
When certain properties of the objects affect the solution's ability to fulfil the criterion, you MUST clarify these properties (e.g. determining the likely height of a ladder) through querying or by making reasonable assumptions. STRICTLY limit your response to $maxwords$ words maximum. Do NOT raise repetitive points.
Response Format:

1. **Clearly answering all questions/uncertainties from other agents in the discussion history, IF ANY: (format STRICTLY in this way: `To <analyst name>'s question about <topic>: <answer>...`)**
2. **General thoughts/opinion on whether the solution fulfils the criterion criterion (succinctly) w.r.t. your main responsibility, with reference to the criterion definition:**
3. **Queries for other agents: (format in this way: `To <analyst name>: <query>...`)**

Begin each part of your response with [[label of part]]. E.g. `[[Answering questions from other agents]]: <part of response>`
Relevant discussion is below: $relevantdiscussion$

Confidence Prompt

You are the impartial but critical $role$ in the discussion provided, $rolefocus$.
Problem: $problem$
Solution: $solution$
Criterion: $criterion$ Definition: $definition$
Discussion points: $discussion$
Given the problem, solution, criterion definition, and the discussion points above, to what extent are you certain that you can reach an accurate and correct conclusion ONLY regarding whether the solution fulfils the specific criterion of $criterion$?
Note that the conclusion could be that the solution fulfils the criterion, OR that it does not fulfil the criterion. Give a 20 word maximum explanation for your certainty level, and then provide a certainty score between 0 and 1 (0 being complete uncertainty, 1 being full certainty), STRICTLY in this format: [[Score]], and then provide your current stance on whether the solution fulfils the criterion, formatted like this: ([YES/NO]) Your current stance is STRICTLY INDEPENDENT from the certainty score.
For example: `<explanation for moderate confidence in the accuracy of the conclusion that the solution does not fulfil the criterion> Thus, [[0.6]]. ([NO])` STRICTLY provide your certainty score to 1 decimal place (e.g. 1.0 or 0.1). Be analytical.

Verdict Prompt

You are the $role$ in the discussion provided, with the relevant focuses, $rolefocus$. Act as an impartial but critical judge. Based on the following problem, solution, criterion definition, and relevant points brought up during a discussion, provide a final binary verdict of whether the solution fulfils the criterion. Heavily consider the specific phrasing of the criterion definition.
Problem: $problem$
Solution: $solution$
Criterion: $criterion$ Definition: $definition$
Discussion: $discussion$
Provide your verdict in the format: [[YES]] or [[NO]], accompanied with a 1-sentence explanation justifying it. Be strict but fair in your judgement.

O PROMPTS FOR BASELINE EVALUATION FRAMEWORKS

Oneshot Prompt

SYSTEM: You will be provided with a user's problem and an assistant's solution.
Please act as a critical judge and evaluate the quality of the solution.
Note the following definitions: - $definition$
Provide your judgement of whether the solution fulfils the criterion of $criterion$ STRICTLY as follows: [[$criterionfulfilled$/$criterionnotfulfilled$]] (or otherwise for other criteria) - Provide the probability that the solution is feasible (from 0 to 1) to 1 decimal place, STRICTLY in this format: ([probability]). E.g. ([0.5]) Do not write any text before or after this response.
USER: [The Start of User's Problem]
$problem$
[The End of User's Problem]
[The Start of Assistant's Answer]
$answer$
[The End of Assistant's Answer]
Determine the $criterion$ of the answer.

Chain-of-Thought Prompt

SYSTEM: You will be provided with a user's problem and an assistant's solution.
Please act as an impartial but critical judge and evaluate the quality of the solution.
Note the following definitions: - $definition$
Provide a 20 word summary/explanation justifying your judgment.
After this, provide your final judgment as follows:
- If the solution if $criterionnotfulfilled$, answer [[$criterionnotfulfilled$]].
- If the solution is $criterionfulfilled$, answer [[$criterionfulfilled$]].
- Provide the probability that the solution is feasible (from 0 to 1) to 1 decimal place, in this format: ([probability]). E.g, Probability: ([0.5]).
Be strict but fair in your assessment.
USER: [The Start of User's Problem]
$problem$
[The End of User's Problem]
[The Start of Assistant's Answer]
$answer$
[The End of Assistant's Answer]
Determine the $criterion$ of the answer.

Fewshot + Chain-of-Thought Prompt

SYSTEM: You will be provided with a user's problem and an assistant's solution.
Please act as a critical judge and evaluate the quality of the solution.
Note the following definitions: - $definition$
Provide a 20 word summary/explanation justifying your judgement.
After this, provide your final judgement of whether the solution fulfils the criterion of $criterion$
STRICTLY as follows:
[[$criterionfulfilled/criterionnotfulfilled$]]
Then, provide the probability that the solution is feasible (from 0 to 1) to 1 decimal place, in this format: ([probability]). E.g, Probability: ([0.5]).
Example conversation:

[The Start of User's Problem]
$exampleproblem$
[The End of User's Problem]
[The Start of Assistant's Answer]
$examplesolution$
[The End of Assistant's Answer]
[The Start of Your Judgement]
$reasoning$ [[$criterionnotfulfilled$]] Probability: ([0.3]).
[The End of Your Judgement]

USER: [The Start of User's Problem]
$problem$
[The End of User's Problem]
[The Start of Assistant's Answer]
$answer$
[The End of Assistant's Answer]
Determine the $criterion$ of the answer.

Fewshot Prompt

SYSTEM: You will be provided with a user's problem and an assistant's solution.
Please act as a critical judge and evaluate the quality of the solution.
Note the following definitions: - $definition$
After this, provide your final judgement of whether the solution fulfils the criterion of feasibility STRICTLY as follows:
[[$criterionfulfilled$/$criterionnotfulfilled$]]
Provide the probability that the solution is feasible (from 0 to 1) to 1 decimal place, in this format: ([probability]). E.g, Probability: ([0.5]).
Do not provide any text before or after your judgement.
Example conversation:

[The Start of User's Problem]
$exampleproblem$
[The End of User's Problem]
[The Start of Assistant's Answer]
$examplesolution$
[The End of Assistant's Answer]
[The Start of Your Judgement]
[[$criterionnotfulfilled$]] Probability: ([0.3]).
[The End of Your Judgement]

USER: [The Start of User's Problem]
$problem$
[The End of User's Problem]
[The Start of Assistant's Answer]
$answer$
[The End of Assistant's Answer]
Determine the $criterion$ of the answer.

Multi-agent Debate - Debater Prompt

SYSTEM: You are a skilled expert, Debater $number$, studying solutions to a problem. As a task, you will be provided with a problem, solution, and a criteria to judge it on. You are to produce a 50 word argument for how the solution meets the criterion of $criterion$, with reference to the definition of the criterion.
Assume that the items used are all of good quality, unless stated otherwise in the problem. If you cannot generate a convincing, pertinent and logical argument, you may respond that "I cannot produce any significant points which fulfill the specified requirements." Note the following definitions: $criterion$: $definition$
$earlierdiscussion$
USER: [Problem]
$problem$
[The Start of Assistant's Answer]
$answer$
[The End of Assistant's Answer]

Multi-agent Debate - Judge Prompt

You are a wise judge studying the solutions to a problem. As a task, you will be provided with a transcript of a debate between two LLMs.
By only considering probable, pertinent and logical points from either side, while referencing the definition of the criterion, conclude whether or not the solution to the problem fulfils the criterion of $criterion$.
Be mindful of the constraints and specifics of the problem. Assume that the items used are all of good quality unless stated otherwise in the problem.
Provide a summary (50 words maximum) justifying your judgement on why the solution does or does not fulfil the criterion definition based on the arguments given, and then present your verdict STRICTLY as follows: $criterion$: [[YES/NO]].
Then, provide the probability that the solution fulfils the criterion of $criterion$ (from 0 to 1) to 1 decimal place, in this format: ([probability]). E.g, Probability: ([0.5]).
For example: (explanation). Therefore, [[YES]]. Probability: ([0.9])
Recall the following definition: $criterion$: $definition$

$transcriptofdebate$

USER: [Problem]
$problem$
[The Start of Assistant's Answer]
$answer$
[The End of Assistant's Answer]