Automated Creativity Evaluation for LLMs with Semantic Entropy and Efficient Multi-Agent Judging Across Open-Ended Tasks

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved remarkable progress in language understanding, reasoning, and generation, sparking growing interest in their creative potential. Realizing this potential requires systematic and scalable methods for evaluating creativity across diverse tasks. However, most existing creativity metrics are tightly coupled to specific tasks, embedding domain assumptions into the evaluation process and limiting scalability and generality. To address this gap, we introduce an automated, domain-agnostic framework for quantifying LLM creativity across open-ended tasks. Our approach separates the measurement apparatus from the creative task itself, enabling scalable, task-agnostic assessment. Divergent creativity is measured using semantic entropy—a reference-free, robust metric for novelty and diversity, validated against LLM-based novelty judgments and baseline diversity measures. Convergent creativity is assessed via a novel retrieval-based multi-agent judge framework that delivers context-sensitive evaluation of task fulfilment with over 60% improved efficiency. We validate our framework across two distinct domains—physical reasoning and scientific research ideation—and with a broad suite of LLMs. Empirical results show our metrics reliably capture key facets of creativity—novelty, diversity, and task fulfilment—and reveal how model properties such as size, temperature, recency, and reasoning impact creative performance. Our work establishes a reproducible, generalizable standard for automated LLM creativity evaluation, paving the way for scalable benchmarking and accelerating progress in creative AI.

1 Introduction

Recent advances in large language models (LLMs) have led to major breakthroughs in language comprehension, generation, and reasoning (Lewis et al., 2019; Manning, 2022; Cobbe et al., 2021). As LLMs become more adept at reasoning and planning, their creative potential has emerged as a key area of interest (Ye et al., 2024; Sun et al., 2024). Creative LLMs can accelerate scientific discovery by proposing unconventional solutions (Ruan et al., 2024; Gu et al., 2024), uncovering novel patterns (Si et al., 2024), and automating experiment design (Liu et al., 2024), with far-reaching applications in materials science (Centre), research methodology (Boyko et al., 2023), and causal discovery (Li et al., 2025). Understanding and quantifying these creative capabilities is thus increasingly important.

However, most existing creativity evaluation frameworks are tightly coupled to specific tasks or domains, embedding strong domain assumptions into the assessment process (Krašovec, 2024; Kroll and Kraus, 2024). These approaches rely on curated answer sets, hand-crafted rubrics, or extensive human annotation—rendering creativity assessment subjective, resource-intensive, and difficult to scale, and leaving the field without automated, domain-general evaluation standards.

To address these challenges, we propose a fully automated, domain-general framework for evaluating LLM creativity that is both robust and scalable across open-ended tasks. Our framework decouples evaluation from specific creative tasks, enabling systematic, reference-free assessment of model creativity across domains. Building on cognitive science, which characterizes creativity as encompassing both divergent and convergent thinking (Guilford, 1950), we deliberately design our framework to evaluate both aspects through novel, automated methods.

Divergent thinking is the ability to generate diverse, novel, and innovative ideas. We argue that hallucinations—often seen as a drawback in LLMs—can, in fact, reflect divergent thinking by producing unconventional ideas. To capture this, we introduce *Semantic Entropy*, a sampling-based, reference-free metric quantifying the variability of

model-generated outputs. We further validate its utility by benchmarking semantic entropy against LLM-based novelty judgments and additional diversity baselines, finding that it faithfully reflects core markers of divergent creativity.

Convergent thinking involves synthesizing information to produce solutions tailored to specific goals and contexts (Kumar et al., 2024). Recognizing the inherent subjectivity in evaluating this aspect (Li et al., 2023), we propose an adaptable, autonomous multi-agent LLM judging framework, where agents collaboratively assess distinct facets of task fulfilment (Lu et al., 2024a). To address the computational inefficiency of traditional discussion-based evaluations (Wang et al., 2024a), we introduce a retrieval-based discussion framework that streamlines the review process, making large-scale benchmarking more feasible.

To demonstrate the generality and practical value of our framework, we evaluate it on two distinct domains: physical reasoning using the MacGyver dataset (Tian et al., 2024), and scientific research ideation using the Hypogen dataset (O'Neill et al., 2025). We apply both methods to 300 problems per domain and benchmark diverse LLMs. We also analyze how model size, recency, temperature, and reasoning augmentation affect creativity.

In summary, we: **(1)** introduce a reference-free, automated assessment of divergent creativity based on semantic entropy; **(2)** develop a more compute-efficient multi-agent LLM judging framework for convergent creativity; and **(3)** provide comprehensive empirical benchmarking of LLMs' creativity across both MacGyver and Hypogen—together establishing an automated, domain-general LLM creativity evaluation framework.

2 Related Work

Human Creativity Tests. Classic human creativity assessments—such as the Torrance Tests of Creative Thinking (TTCT) and the Consensual Assessment Technique (CAT) (Torrance; Amabile, 1982)—have been adapted to evaluate LLMs. However, these methods depend on extensive human annotation, making them unscalable and ill-suited for automated evaluation. Moreover, while TTCT metrics like fluency and elaboration are meaningful in human settings, they are less reliable for LLMs, since idea count and output length can be trivially adjusted by sampling. Consequently, our framework focuses on originality and flexibility (divergent creativity), which remain robust indicators for LLM generation tasks, and utilizes a separate, automated judge for task fulfilment (convergent creativity).

Domain-specific Creativity Evaluation. Beyond classic human tests, a wide range of task-specific creativity benchmarks have been developed for LLMs, spanning mathematical reasoning, hardware design, metaphor generation and code synthesis (Ye et al., 2024; DeLorenzo et al., 2024; Paul V. DiStefano and Beaty, 2024; Gómez-Rodríguez and Williams, 2023). These frameworks typically embed strong domain assumptions, require curated answer sets or subjective metrics, and are closely tied to the structure of their target tasks. As a result, they lack generalizability and are difficult to apply systematically to the open-ended challenges tackled by modern LLMs. Our framework overcomes these limitations by providing a task-agnostic, reference-free, and fully automated approach to creativity evaluation.

Divergent Creativity Evaluation. Automated metrics for divergent creativity in LLMs—such as semantic similarity, integration scores, and Lempel–Ziv complexity—offer some insight into output diversity, but often miss the nuance required for complex, open-ended tasks (Mohammadi, 2024; Chen and Ding, 2023; Summers-Stay et al., 2023; Peeperkorn et al., 2024; Bellemare-Pepin et al., 2024). Recent work has instead used uncertainty to detect hallucinations in LLM outputs (Huang et al., 2024; Chen et al., 2025; Zhang et al., 2023; Sriramanan et al., 2024), including one based on Semantic Entropy (SE)(Farquhar et al., 2024). We build on this by repurposing Semantic Entropy, originally for hallucination detection, as a robust, reference-free measure of divergent creativity.

Convergent Creativity Evaluation. For convergent creativity, traditional tests like the Remote Associates Test (RAT) (Mednick and Mednick, 1967) are not well suited to LLMs, as they were designed for humans. Recent pipelines leverage LLMs as judges (Rabeyah et al., 2024; Dubois et al., 2024; Li et al., 2024; Zheng et al., 2023), with multi-agent discussion frameworks shown to provide more nuanced and comprehensive evaluation of candidate solutions (Liang et al., 2024; Chan et al., 2023). However, these methods are computationally intensive and hard to scale (Lv et al., 2024; Luo et al., 2023). Our novel retrieval-based discussion framework addresses this, enabling scalable, robust evaluation of task fulfilment.

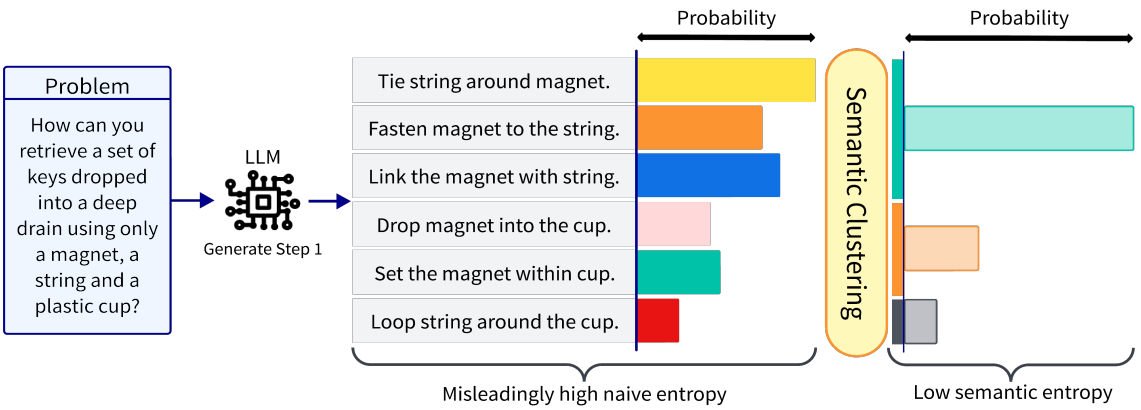

Figure 1: Semantic Entropy: LLM-generated steps clustered by similarity, with entropy computed over cluster probabilities. Naive entropy (middle) uses raw probabilities; Semantic Entropy (right) clusters by meaning for a more reliable measure.

3 Divergent Creativity

3.1 Background on Semantic Entropy

Semantic Clustering. Following Farquhar et al. (2024), 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 sufficiently similar, or creates a new class otherwise.

Semantic Entropy. For a query x, the probability, $P(s|x)$, of a generated steps, comprising tokens $(t_1, ..., t_i)$ is given by the product of its conditional token probabilities. For computational efficiency, we use log-probability $\log P(s|x)$.

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

The probability of a semantic class c is the sum of all generated samples s belonging to the class:

$$P(c|x) = \sum_{s \in c} P(s|x) \quad (2)$$

Semantic Entropy is computed as the entropy of the class probability distribution over all classes C:

$$H(x) = -\sum_{i=1}^{|C|} P(C_i|x) \log P(C_i|x) \quad (3)$$

3.2 Automated Divergent Creativity Evaluation with Semantic Entropy

Hallucination-like processes in humans reflect associative thinking, a key mechanism underlying creativity (Jiang et al., 2024; Raffaelli et al., 2024; Ritter and Dijksterhuis, 2014). By making unexpected connections, associative thinking enables the generation of multiple, varied, or unconventional ideas—a hallmark of divergent thinking (Guilford, 1950). We hypothesize that, in LLMs, generation uncertainty—where the model produces unpredictable or surprising outputs—similarly signals divergent creativity by reflecting an ability to explore novel solution paths.

Motivation. To robustly quantify this breadth and novelty in model outputs, we use Semantic Entropy (Farquhar et al., 2024), which measures the unpredictability and diversity of generated solutions at the semantic level. Unlike word-level entropy or surface-level diversity metrics, semantic entropy captures true conceptual differences, identifying outputs that are novel in substance rather than just rephrasings. This reference-free and scalable approach enables automated creativity assessment across domains. Because unpredictability and diversity are closely linked to creativity markers like originality and flexibility, we will further investigate how semantic entropy aligns with established creativity metrics, leveraging both LLM-based novelty judgments and diversity baselines.

Implementation. For each task, we generate solutions step by step: at each stage, we sample $n = 10$ candidate solutions per step, cluster them by semantic equivalence, and compute Semantic Entropy over the resulting class probabilities. The highest-probability sample is iteratively appended to build a full solution, and this repeats until majority of the samples indicate completion ("STOP").

Entailment Model. We use the DeBERTa NLI model to cluster generated samples into semantic classes by assessing semantic equivalence. Its performance was validated on 50 manually annotated pairs, benchmarked against GPT-4o (zero-shot) for entailment. DeBERTa's accuracy and efficiency make it ideal for clustering (Table 1).

Table 1: Entailment models.

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

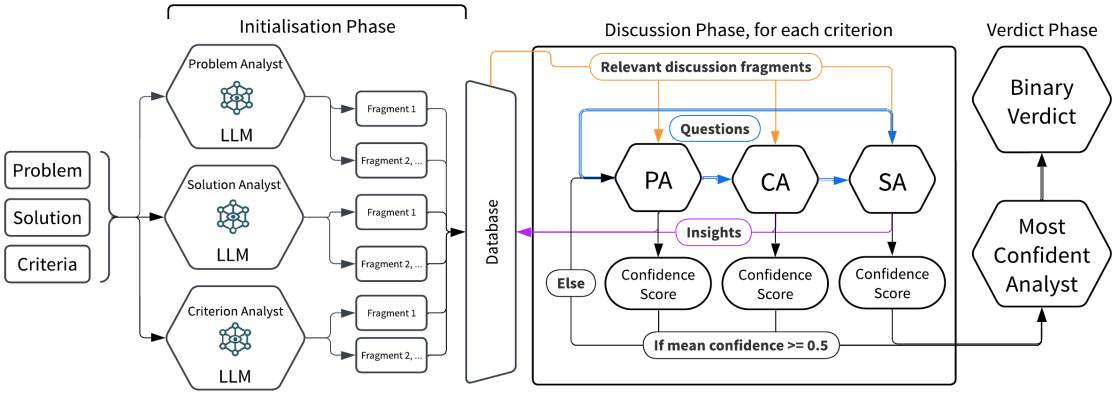

Figure 2: Retrieval-based multi-agent judging framework for automated convergent creativity evaluation. Supports flexible, metric-specific assessment across diverse tasks using structured agent roles and retrieval-augmented discussion.

4 Convergent Creativity

Motivation. Evaluating convergent creativity in LLMs requires assessing how well generated solutions fulfil diverse goals and constraints across domains. LLM-as-a-judge frameworks are a widely adopted, robust approach for automating open-ended evaluation, enabling large-scale, accurate, consistent assessment while overcoming limitations of human annotation (Badshah and Sajjad, 2024; Gu et al., 2025). Unlike single-task or fixed-rubric benchmarks, our framework supports configurable, metric-specific evaluation that flexibly captures domain-specific criteria—essential for creative tasks where "success" is context-dependent. Our automated approach enables scalable, reference-free quantification of task fulfilment across a broad range of open-ended problems.

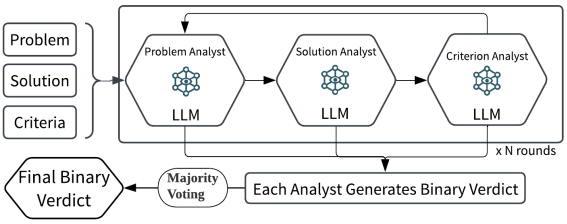

Figure 3: ChatEval (One-by-one) framework and its information flow, implemented for our benchmark.

Current challenges. Multi-agent judge frameworks (Liang et al., 2024; Chan et al., 2023) yield more nuanced, context-aware assessments by having LLMs engage in discussion—outperforming one-shot or few-shot judging, as each agent can identify distinct aspects and subtleties in a solution. However, appending the full discussion history at each turn is highly resource-intensive, quickly inflating token usage and computation, and making it impractical for large-scale benchmarks.

4.1 Automated Convergent Creativity Evaluation

Retrieval-based Framework. To address the inefficiency of standard multi-agent LLM judging, we introduce a retrieval-based framework that preserves nuanced evaluation while drastically reducing token usage (by 63% vs. ChatEval; see Appendix). Agents attend only to the most relevant prior discussion fragments, with early stopping and confidence scoring further limiting redundant deliberation, making scalable, context-aware assessment feasible across diverse datasets.

Implementation. Our framework structures evaluation in three main phases, as well as an early stopping mechanism (see appendix E.4):

Initialisation. Each agent generates initial insights (*fragments* F_i) about the problem, solution, and domain-specific criteria, stored in a database D with their embeddings $\mathcal{E}(F_i)$. Agents retrieve the n most relevant fragments for a query Q using cosine similarity:

$$\begin{aligned} &\text{GET}(Q, n) = \text{Top-}n\big(\text{Sim}(\mathcal{E}(Q), \mathcal{E}(F_i))\big) \\ &\text{where } F_i \in D \end{aligned} \tag{4}$$

Discussion. Agents retrieve relevant fragments with query Q_a, answers peer questions R_a^{response}, offer opinions R_a^{opinion}, and raise new queries q_a^{new}, updating the fragment database.

$$\begin{aligned} &(R_a^{\text{questions}}, R_a^{\text{opinion}}, q_a^{\text{new}}) \\ &= J_a(q_{others,a}, \text{GET}(Q_a \oplus q_{others,a}, k), \mathcal{B}) \end{aligned} \tag{5}$$

Verdict. The agent with the highest confidence uses relevant fragments $\text{GET}(Q_{\max}, l)$ to deliver a binary verdict for each criterion $\mathcal{C}_i$.

See appendix for full prompt templates and implementation details.

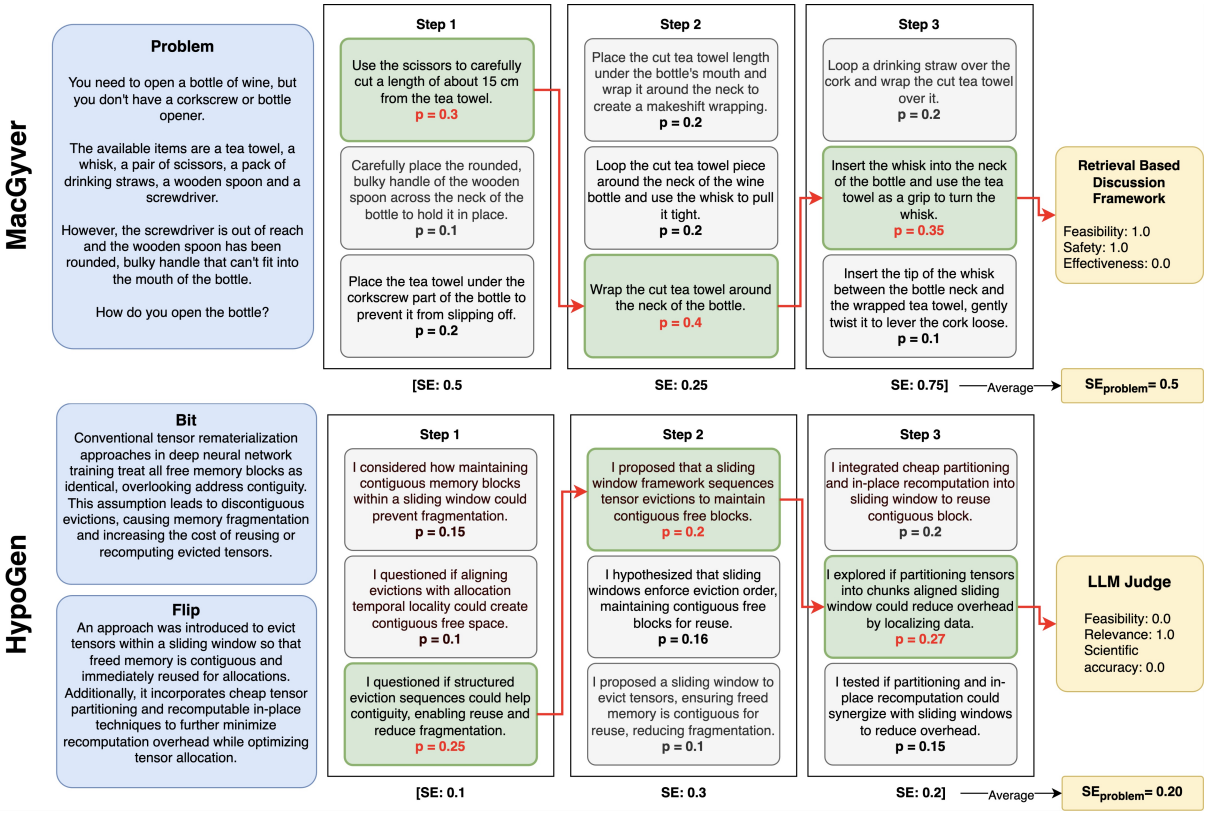

Figure 4: Overview of our semantic entropy-based automated creativity evaluation, for both the Macgyver and HypoGen datasets.

5 Experimental Setup

Datasets and Evaluation Metrics.

- **MacGyver:** Real-world physical reasoning problems requiring creative, unconventional use of everyday objects. Models are prompted to generate step-by-step solutions.
  Metrics: **Feasibility, Safety, Effectiveness**
- **Hypogen:** Open-ended scientific ideation tasks(O'Neill et al., 2025). Each problem presents a standard "Bit" and a target "Flip"(see Fig. 4); models must generate a novel chain of reasoning from Bit to Flip without being shown the ground-truth explanation.
  Metrics: **Feasibility, Scientific Accuracy, Relevance**

See appendix E.1 for sample prompts, generations and full metric definitions.

Divergent Creativity Verification. We benchmark Semantic Entropy against relative novelty and diversity baselines. For novelty, a pairwise LLM judge ranks solution originality for 50 MacGyver problems per model across four LLMs; reliability was validated against the average rankings from five human annotators, each ranking 30 solutions (Spearman and Pearson correlation of 0.80). Pipeline and annotation details are in the appendix D.5. For diversity, we investigate average cosine similarity between clusters against Semantic Entropy, with other metrics (e.g., self-BLEU) in the appendix D.3.

Convergent Creativity Verification. To assess the reliability of our automated Multi-Agent Judge framework, we compare its verdicts to a "golden truth" obtained by majority vote from five human annotators on a randomly sampled set of 50 problems on the Macgyver dataset. We report accuracy as Accuracy-Rejection Curve (AUARC)(Nadeem et al., 2009), and detail the annotation protocol and inter-annotator agreement in the appendix.

Detailed Pipeline. We benchmark LLMs on the MacGyver and HypoGen datasets using our unified framework. For each model, 300 problems per domain are solved step by step: at each stage, 10 candidate next steps are generated and clustered to compute semantic entropy, then the most likely candidate is selected (greedy search) and appended to the solution. Divergent creativity is reported as the average entropy across all steps in the solution, reflecting overall exploration.

Convergent creativity is assessed on all 300 generated solutions using our Multi-Agent Judge and domain-specific metrics. Final model scores include both Divergent (Semantic Entropy) and Convergent (Multi-Agent Judge accuracy) results.

All experiments used 4 NVIDIA A100 GPUs, with API access for large models. Further details are in the appendix.

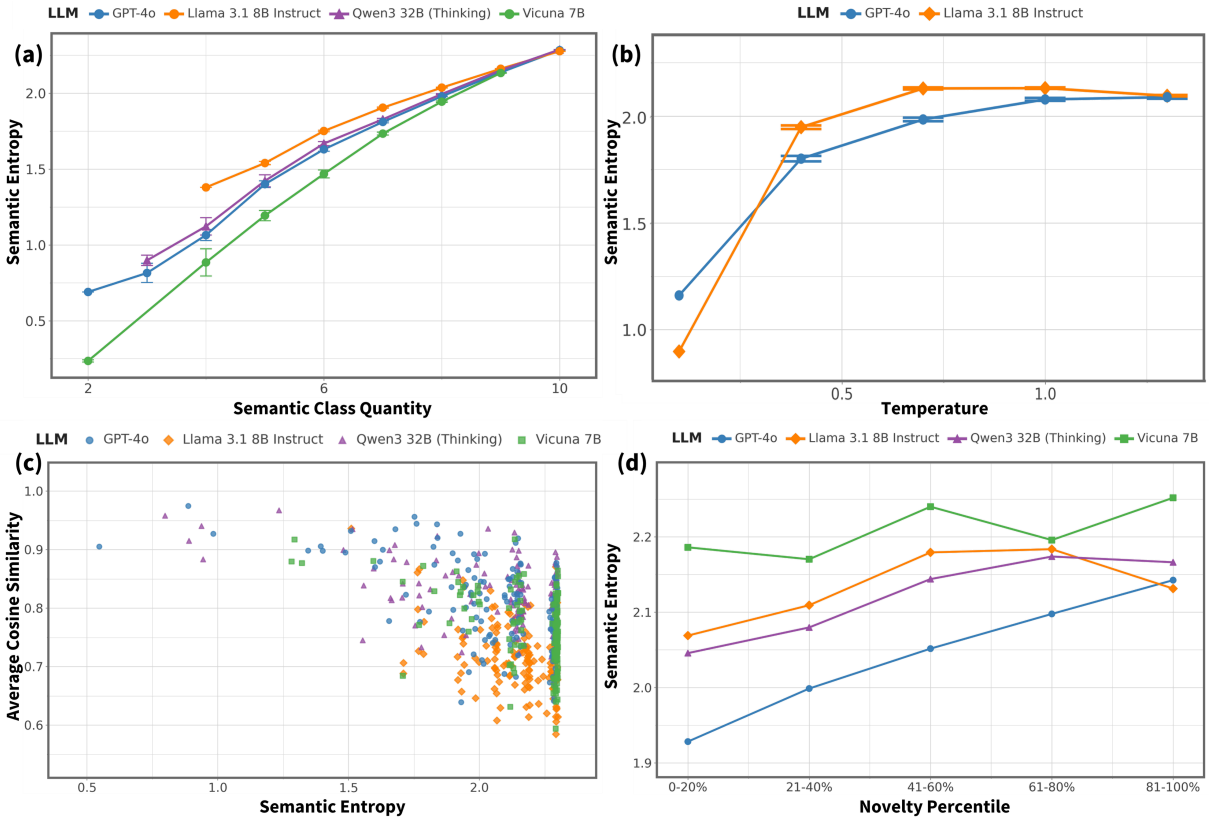

Figure 5: Semantic entropy's relationship with response and model parameters.

6 Results and Discussion

6.1 Divergent Creativity

Semantic entropy robustly captures the breadth and flexibility of idea generation in LLMs. We find that semantic entropy, computed over semantically distinct clusters, is strongly correlated with the number of unique idea categories (Fig. 5a), indicating that it reliably reflects flexibility—one of the four TTCT metrics (Torrance). Semantic entropy is also highly responsive to temperature: as temperature increases, models generate more varied and less repetitive outputs, resulting in higher entropy across architectures (Fig. 5b), consistent with findings that increased temperature enhances diversity by reducing repetition and encouraging creative risk-taking (Chen and Ding, 2023; Roemmele and Gordon, 2018). Notably, while semantic entropy rises rapidly with temperature at first, it plateaus at higher values, likely reflecting saturation in generating meaningful, distinct outputs (Chen and Ding, 2023). Overall, these results establish semantic entropy as a robust and scalable metric for quantifying the generative range and creative flexibility of LLM solutions—key markers of creative potential.

Semantic entropy meaningfully tracks core creative attributes, aligning with both novelty and diversity baselines. To further validate semantic entropy as a practical proxy for divergent creativity, we benchmarked it against an LLM-based pairwise novelty judge (validated with strong agreement to human annotators) and established diversity metrics, such as average cosine similarity between solutions. Our results show that semantic entropy is positively associated with both judged novelty (Fig. 5d) and lower cosine similarity (Fig. 5c), indicating that models with higher entropy not only produce more diverse ideas, but also outputs considered more original. These findings reinforce semantic entropy's utility as an automated, reference-free, and domain-general indicator for core creative attributes—enabling principled evaluation of LLM creativity at scale.

The advancement and size of LLMs does not correlate with divergent creativity. As shown in the appendix G, semantic entropy remains largely stable—and sometimes even decreases—as models become larger or newer, both across Llama generations (3, 3.1, 3.3; 8B to 405B) and Vicuna model sizes (7B to 33B). This is likely due to training that prioritises convergent solutions (Yu et al., 2024), potentially limiting divergent output in larger models and suggesting a developmental trajectory for creativity that is distinct from general advances in problem-solving or reasoning. Notably, Ruan et al. (2024) also found that less advanced and state-of-the-art models generate comparable levels of creative ideas in scientific contexts.

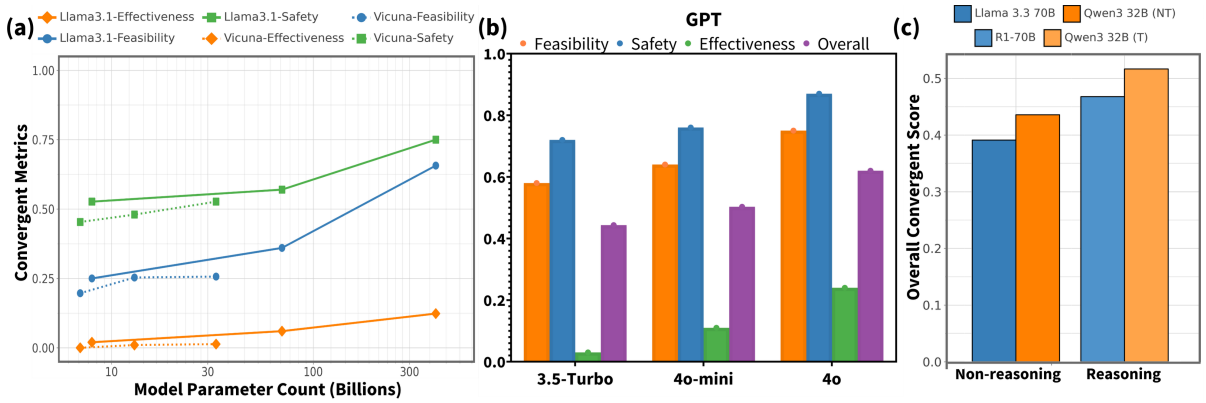

Figure 6: The impact of various parameters (left: model size, center: model recency, right: reasoning capabilities) on the convergent creativity of LLMs on the MacGyver dataset. In this and subsequent figures, (T) and (NT) refers to (Thinking) and (Non-thinking) respectively.

6.2 Convergent Creativity

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
ChatEval	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 and human annotators, judging 50 solutions to Macgyver Dataset.

Our retrieval-based multi-agent judge enables robust, scalable assessment of convergent creativity across diverse domains. We demonstrate that our retrieval-based multi-agent framework achieves accuracy and AUARC comparable to individual human annotators (see Table 2), and consistently outperforms all single-agent and baseline LLM judging pipelines, highlighting its effectiveness in capturing nuanced, context-dependent aspects of task fulfilment. By leveraging retrieval and confidence-based stopping, our method reduces computational cost by over 60% (see Appendix) compared to traditional multi-agent discussions (Chan et al., 2023), making large-scale evaluation feasible. These results show that automated, discussion-based LLM judges can match or exceed the reliability of human raters, while enabling efficient, repeatable assessment of LLM performance in open-ended, multi-criteria tasks across domains.

Larger, more recent and reasoning LLMs achieve higher task fulfilment. Larger and more recent models like GPT-4o and Llama 3.1 70B consistently outperform earlier models such as GPT-3.5 and Llama 3.1 8B on convergent creativity metrics (Fig. 6a, 6b), which reflect a model's ability to generate solutions that meet explicit task requirements. This pattern is consistent with prior work demonstrating GPT-4o's advantages in code generation (Lu et al., 2024b), reasoning (Minaee et al., 2024), and Llama 70B's edge in instruction following (Kovalevskyi, 2024), largely attributed to scaling laws and advanced training strategies like instruction tuning and dataset diversification (Zhao et al., 2024). Reasoning-focused models such as R1-70B also outperform their non-reasoning base versions (e.g., Llama 3.3, Fig. 6c), reinforcing the value of reasoning for enhancing LLMs' abilities to tackle complex, multi-criteria tasks (DeepSeek-AI et al., 2025; Huang and Chang, 2023). Overall, these results show that scaling, recency, and improved reasoning directly boost LLMs' capacity for task fulfilment in automated convergent creativity evaluation.

Divergent and Convergent creative ability in LLMs arise from distinct mechanisms. The relationship between divergent and convergent creativity is model-dependent, not strictly antagonistic. As shown in Figure 7, some LLMs (like GPT-4o) show a mild trade-off—higher semantic entropy can correspond to lower task fulfilment—while others (like Llama 8B) maintain more balanced performance. Our findings suggest that the mechanisms supporting divergent and convergent creativity can be at least partially decoupled, with the potential for both to be improved in tandem.Thus, maximizing creative potential does not necessarily require sacrificing convergent abilities.

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 appendix G.

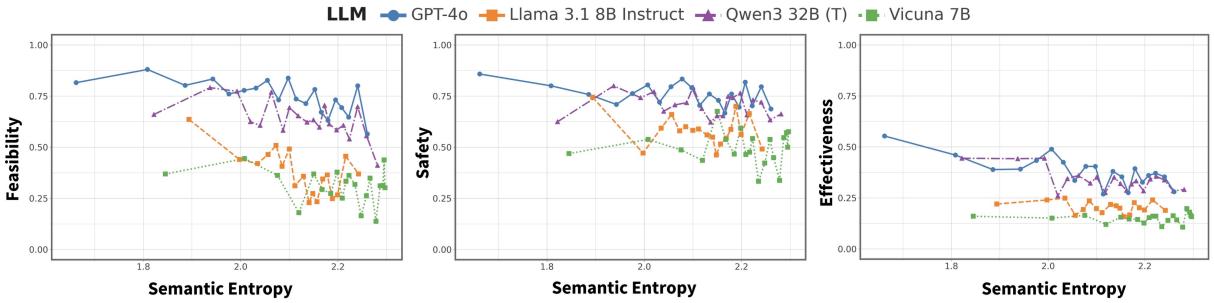

Figure 7: Semantic Entropy compared to different convergent creativity metrics (Y-axis) on the MacGyver dataset. Each point represents the mean Y value at the median X value of a unique set of 15 data points (fixed-interval binning). Similar trends were observed on the Hypogen dataset(see Appendix 19).

Table 3: Performance of various LLMs on our benchmark using the MacGyver dataset.

Model	Divergent Creativity	Convergent Creativity			
	Semantic Entropy	Feasibility	Safety	Effectiveness	Overall
Vicuna 7B	**2.19**	0.20	0.45	0.00	0.217
Vicuna 13B	1.96	0.25	0.48	0.01	0.248
Vicuna 33B	2.17	0.26	0.53	0.01	0.257
Llama 3 70B Instruct	2.10	0.39	0.65	0.02	0.356
Llama 3.1 8B Instruct	2.13	0.25	0.53	0.02	0.266
Llama 3.1 70B Nemotron Instruct	**2.19**	0.36	0.57	0.06	0.33
Llama 3.1 405B Instruct	2.08	0.66	0.75	0.12	0.51
Llama 3.3 70B Instruct	2.10	0.45	0.68	0.04	0.391
Deepseek R1 70B Distilled	2.10	0.58	0.75	0.07	0.468
GPT 3.5 Turbo	2.02	0.51	0.71	0.03	0.416
GPT 4o mini	2.05	0.62	0.76	0.12	0.497
GPT 4o	2.08	**0.82**	**0.86**	**0.21**	**0.629**
Qwen3 32B (Thinking)	2.02	0.65	0.78	0.12	0.517
Qwen3 32B (Non-thinking)	2.08	0.49	0.74	0.08	0.436

Model	Divergent Creativity	Convergent Creativity			
	Semantic Entropy	Feasibility	Relevance	Scientific Accuracy	Overall
GPT-4o	**2.07**	0.28	0.61	0.17	0.353
Llama 3.1 8B Instruct	2.04	0.21	0.56	0.12	0.293
Qwen3 32B (Thinking)	1.72	0.41	**0.78**	0.21	0.467
Qwen3 32B (Non-thinking)	1.66	**0.50**	0.76	**0.26**	**0.506**

Table 4: Performance of various LLMs on our benchmark using the HypoGen dataset.

7 Conclusion

Key findings and Broader Impact. We introduce a fully automated, domain-general framework for evaluating LLM creativity, leveraging semantic entropy as a reference-free metric for divergent thinking and a retrieval-based multi-agent judge for convergent evaluation. Our experiments across the MacGyver and Hypogen benchmarks demonstrate several key findings: semantic entropy robustly quantifies the generative breadth and diversity of model outputs, correlating with established creativity markers like flexibility, novelty, and diversity. Our multi-agent judging framework achieves human-level reliability with over 60% improved computational efficiency, enabling practical, large-scale assessment of LLM task fulfilment in diverse, open-ended settings. In contrast to convergent performance, we find that model size and recency do not reliably increase divergent creativity, suggesting that current advances are more aligned with optimizing for correct answers than for creative exploration. By establishing a scalable and reproducible automated benchmark for creativity evaluation, our work provides a foundation for the principled development and rigorous comparison of creative AI systems.

Future Work. Our results suggest that divergent and convergent creativity can be optimized independently, motivating further study of training strategies that enhance both without trade-offs. Future work will systematically investigate the effects of fine-tuning, instruction-following, and specialized training regimes on creativity, as well as the role of human-in-the-loop validation in aligning LLM outputs with human standards of originality and usefulness. We also plan to extend our framework to new domains and tasks, enabling deeper understanding of how model architectures and interventions shape the creativity spectrum in LLMs.

8 Limitations

Despite the demonstrated robustness and scalability of our proposed framework, several limitations merit consideration.

Computational Overhead. The semantic entropy-based evaluation, while effective in capturing novelty and diversity, remains computationally intensive. It requires the generation and clustering of multiple outputs per task, which may limit scalability when applied to large datasets or in environments with constrained computational resources.

Evaluation Limitations and Subjectivity. Our current methodology assesses feasibility and task fulfilment using retrieval-based, context-sensitive LLM judgments. While this enables automated evaluation, it inherently relies on the subjective interpretation and prior knowledge encoded in the LLMs. As a result, unconventional or novel solutions that deviate from known patterns may be incorrectly deemed infeasible. Furthermore, the lack of real-world validation means that theoretically viable but unorthodox responses could be undervalued. Consequently, the framework may not fully capture the practical applicability or ingenuity of certain creative outputs.

Domain Coverage and Generalizability. The framework demonstrates strong performance across two diverse and cognitively demanding domains—physical reasoning (MacGyver) and scientific ideation (Hypogen)—and has validated its adaptability to distinct task formats. This breadth of evaluation suggests promising potential for extension to other forms of creativity. Tasks such as linguistic creativity, artistic generation, and socially grounded problem-solving offer natural next steps. Applying the framework to these domains could deepen our understanding of LLM creative capacities and refine our understanding of model performance across more open-ended generative contexts.

References

Teresa Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43:997–1013.

Sher Badshah and Hassan Sajjad. 2024. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *Preprint*, arXiv:2408.09235.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. 2024. Divergent creativity in humans and large language models. *Preprint*, arXiv:2405.13012.

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. 2023. An interdisciplinary outlook on large language models for scientific research. *Preprint*, arXiv:2311.04929.

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. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang, and Guolin Ke. 2024. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *Preprint*, arXiv:2403.01976.

Dare Arc Centre. Revolutionising materials science with large language models: A new paradigm in material discovery.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Honghua Chen and Nai Ding. 2023. Probing the creativity of large language models: Can models produce divergent semantic association? *Preprint*, arXiv:2310.11158.

Kedi Chen, Qin Chen, Jie Zhou, Xinqi Tao, Bowen Ding, Jingwen Xie, Mingchen Xie, Peilong Li, Feng Zheng, and Liang He. 2025. Enhancing uncertainty modeling with semantic graph for hallucination detection. *Preprint*, arXiv:2501.02020.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. 2024. Creativeval: Evaluating creativity of llm-based hardware code generation. *2024 IEEE LLM Aided Design Workshop (LAD)*, pages 1–5.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

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

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an ai co-scientist. *Preprint*, arXiv:2502.18864.

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 et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. 2024. Llms can realize combinatorial creativity: generating creative ideas via llms for scientific research. *Preprint*, arXiv:2412.14141.

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

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. *Preprint*, arXiv:2212.10403.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *Preprint*, arXiv:2402.06647.

Bohdan Kovalevskyi. 2024. Ifeval-extended: Enhancing instruction-following evaluation in large language models through dynamic prompt generation. *Journal of Artificial Intelligence General science (JAIGS) ISSN:3006-4023*.

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

Margaret Kroll and Kelsey Kraus. 2024. Optimizing the role of human evaluation in llm-based spoken document summarization systems. In *Interspeech 2024*, pages 1935–1939.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2024. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. *Preprint*, arXiv:2410.03703.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In

Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. 2025. Can large language models help experimental design for causal discovery?

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *Preprint*, arXiv:2310.19740.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Zhihan Liu, Yubo Chai, and Jianfeng Li. 2024. Towards fully autonomous research powered by llms: Case study on simulations. *Preprint*, arXiv:2408.15512.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024a. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *Preprint*, arXiv:2405.06373.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, and Daniel Khashabi. 2024b. Benchmarking language model creativity: A case study on code generation. *Preprint*, arXiv:2407.09007.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *Preprint*, arXiv:2303.15621.

Fangrui Lv, Kaixiong Gong, Jian Liang, Xinyu Pang, and Changshui Zhang. 2024. Subjective topic meets LLMs: Unleashing comprehensive, reflective and creative thinking through the negation of negation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12341, Miami, Florida, USA. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Christopher D. Manning. 2022. Human language understanding & reasoning. *Daedalus*, 151(2):127–138.

Sarnoff A. Mednick and Martha T. Shuch Mednick. 1967. Remote associates test, college, adult, form 1 and examiner's manual, remote associates test, college and adult forms 1 and 2.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Behnam Mohammadi. 2024. Creativity has left the chat: The price of debiasing language models. *Preprint*, arXiv:2406.05587.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pages 65–81, Ljubljana, Slovenia. PMLR.

Charles O'Neill, Tirthankar Ghosal, Roberta Răileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciucă. 2025. Sparks of science: Hypothesis generation using structured paper data. *Preprint*, arXiv:2504.12976.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

John D. Patterson Paul V. DiStefano and Roger E. Beaty. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, 0(0):1–15.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *Preprint*, arXiv:2405.00492.

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. 2024. Do llms agree on the creativity evaluation of alternative uses? *Preprint*, arXiv:2411.15560.

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

Simone M. Ritter and Ap Dijksterhuis. 2014. Creativity—the unconscious foundations of the incubation period. *Frontiers in Human Neuroscience*, 8.

Melissa Roemmele and Andrew S. Gordon. 2018. 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. Association for Computing Machinery.

Kai Ruan, Xuan Wang, Jixiang Hong, and Hao Sun. 2024. Liveideabench: Evaluating llms' scientific creativity and idea generation with minimal context. *Preprint*, arXiv:2412.17596.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Douglas Summers-Stay, Stephanie M. Lukin, and Clare R. Voss. 2023. Brainstorm, then select: a generative language model improves its creativity score.

Luning Sun, Yuzhuo Yuan, Yuan Yao, Yanyan Li, Hao Zhang, Xing Xie, Xiting Wang, Fang Luo, and David Stillwell. 2024. Large language models show both individual and collective creativity comparable to humans. *Preprint*, arXiv:2412.03151.

Qwen Team. 2025. Qwen3.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2024. Macgyver: Are large language models creative problem solvers? *Preprint*, arXiv:2311.09682.

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

Junlin Wang, Siddhartha Jain, Dejiao Zhang, Baishakhi Ray, Varun Kumar, and Ben Athiwaratkun. 2024a. Reasoning in token economies: Budget-aware evaluation of llm reasoning strategies. *Preprint*, arXiv:2406.06461.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024b. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

Jingyuan Yang, Dapeng Chen, Yajing Sun, Rongjun Li, Zhiyong Feng, and Wei Peng. 2025. Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. *Preprint*, arXiv:2501.11041.

Junyi Ye, Jingyi Gu, Xinyun Zhao, Wenpeng Yin, and Guiling Wang. 2024. Assessing the creativity of llms in proposing novel solutions to mathematical problems. *Preprint*, arXiv:2410.18336.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. 2024. Flow of reasoning: Efficient training of llm policy with divergent thinking. *ArXiv*, abs/2406.05673.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. *Preprint*, arXiv:2311.13230.

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. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *Preprint*, arXiv:1802.01886.

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`) (Grattafiori et al., 2024; Wang et al., 2024b) and 3 models from the Vicuna family (`vicuna-7b-v1.5`, `vicuna-13b-v1.5`, `vicuna-33b-v1.3`) (Chiang et al., 2023; Zheng et al., 2023). In addition, we also evaluate OpenAI's `gpt-4o`, `gpt-3.5-turbo` and `gpt-4o-mini` closed-source models (Brown et al., 2020; OpenAI, 2023). Furthermore, we evaluate DeepSeek R1 70B Distilled (DeepSeek-AI et al., 2025), and Qwen3 32B (Team, 2025) in both its thinking and non-thinking modes. The open-source models were obtained using Hugging Face.

C Code Availability

Our code is available at the URL: https://anonymous.4open.science/r/MacGyverSemanticProbing-7DFD/. Our benchmark is intended exclusively for research purposes, and is not aimed for commercialisation, making it compatible with original access conditions.

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 (Malinin and Gales, 2021), 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:

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

D.1 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).

D.2 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, retrieved from Farquhar et al. (2024), 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.

D.3 Analysis of Semantic Entropy

Existing literature has proposed various means of quantifying the diversity or semantic consistency of a set of LLM generations in order to probe its creativity. This includes cosine similarity (Li et al., 2016a; Yang et al., 2025), the Self-BLEU metric (Zhu et al., 2018), and distinct-n scores (Li et al., 2016b). By computing the aforementioned metrics for samples generated from our benchmark, we explore the relationship between them and semantic entropy, as shown in figures 8 and 9, as well as tables 5 and 6.

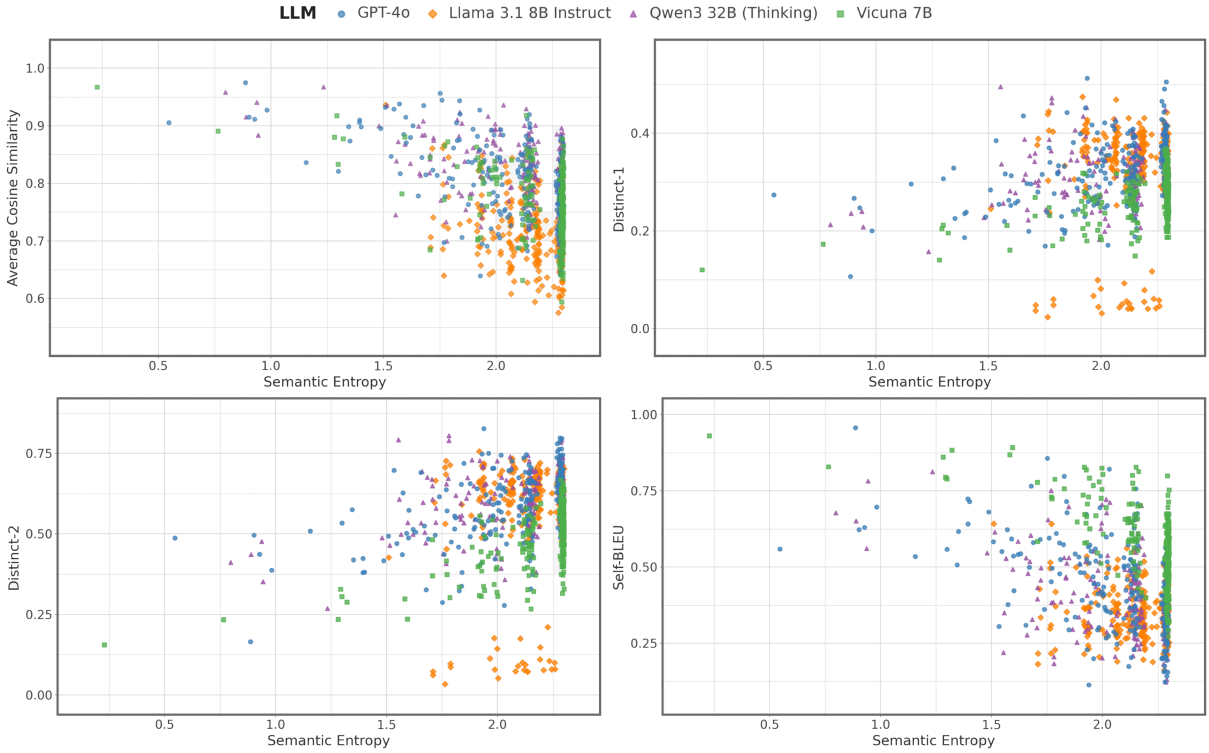

Figure 8: Comparison between semantic entropy and various diversity metrics on the Macgyver dataset.

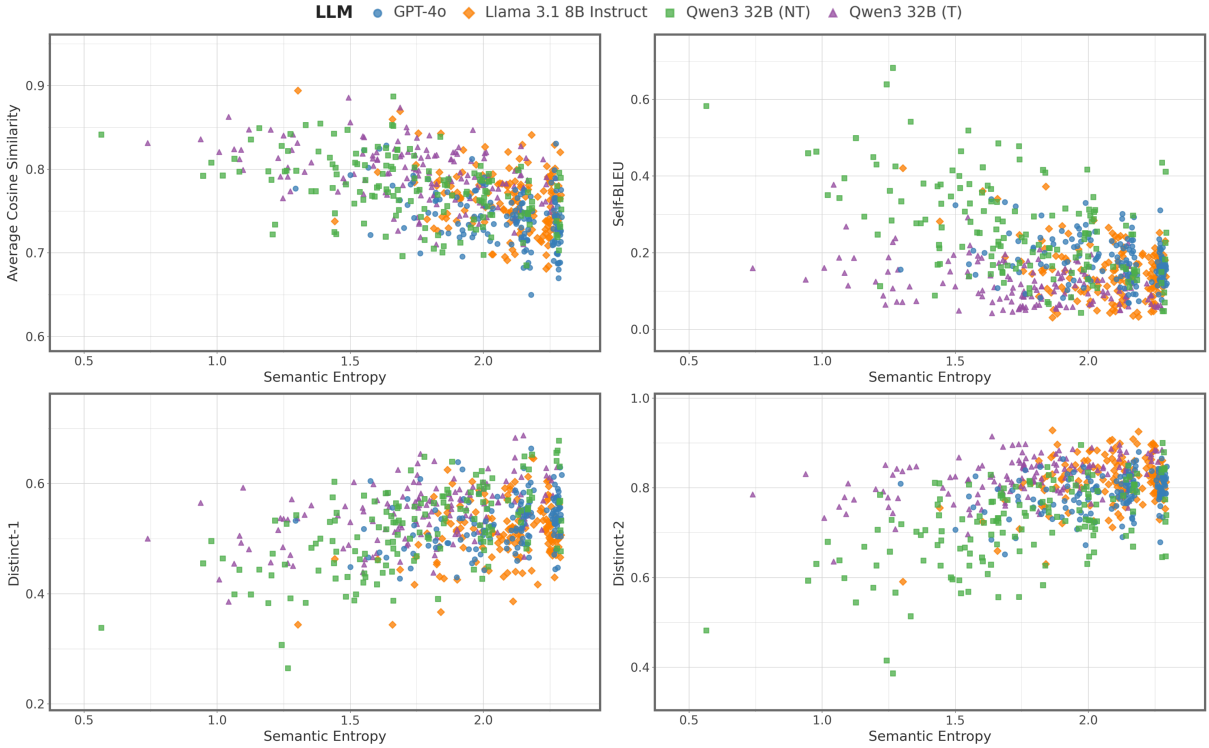

Figure 9: Comparison between semantic entropy and various diversity metrics on the HypoGen dataset.

Metric	Model	ρ	p-value
Average Cosine Similarity			
	GPT-4o	-0.436	6.72×10^{-58}
	Llama 3.1 8B Instruct	-0.048	3.95×10^{-2}
	Vicuna 7B	-0.236	1.24×10^{-29}
	Qwen3 32B Thinking	-0.246	6.09×10^{-22}
Self-BLEU			
	GPT-4o	-0.374	3.25×10^{-43}
	Llama 3.1 8B Instruct	0.119	3.58×10^{-7}
	Vicuna 7B	-0.385	6.77×10^{-80}
	Qwen3 32B Thinking	-0.149	7.44×10^{-9}
Distinct-1			
	GPT-4o	0.285	3.23×10^{-21}
	Llama 3.1 8B Instruct	-0.238	5.19×10^{-25}
	Vicuna 7B	0.213	2.34×10^{-24}
	Qwen3 32B Thinking	0.048	6.49×10^{-2}
Distinct-2			
	GPT-4o	0.349	5.35×10^{-36}
	Llama 3.1 8B Instruct	-0.140	1.92×10^{-9}
	Vicuna 7B	0.325	6.36×10^{-56}
	Qwen3 32B Thinking	0.118	5.39×10^{-6}

Table 5: Spearman correlation (ρ) between semantic entropy and diversity metrics across models for the Macgyver dataset.

Metric	Model	ρ	p-value
Average Cosine Similarity			
	GPT-4o	-0.396	1.34×10^{-108}
	Llama 3.1 8B Instruct	-0.343	9.41×10^{-92}
	Qwen3 32B Non-thinking	-0.463	2.40×10^{-158}
	Qwen3 32B Thinking	-0.564	4.11×10^{-181}
Self-BLEU			
	GPT-4o	-0.346	2.01×10^{-81}
	Llama 3.1 8B Instruct	-0.174	2.02×10^{-21}
	Qwen3 32B Non-thinking	-0.440	1.89×10^{-141}
	Qwen3 32B Thinking	-0.438	7.30×10^{-102}
Distinct-1			
	GPT-4o	0.415	9.87×10^{-120}
	Llama 3.1 8B Instruct	0.180	9.27×10^{-23}
	Qwen3 32B Non-thinking	0.516	2.03×10^{-202}
	Qwen3 32B Thinking	0.570	2.28×10^{-186}
Distinct-2			
	GPT-4o	0.405	7.93×10^{-114}
	Llama 3.1 8B Instruct	0.206	1.95×10^{-29}
	Qwen3 32B Non-thinking	0.496	3.38×10^{-185}
	Qwen3 32B Thinking	0.529	5.49×10^{-156}

Table 6: Spearman correlation (ρ) between semantic entropy and diversity metrics for the HypoGen dataset.

Notably, there is a significant moderate correlation between semantic entropy and multiple metrics such as average cosine similarity (as mentioned previously), and self-BLEU for both datasets, with weak-to-moderate correlations to the distinct-1 and distinct-2 scores. Since semantic entropy correlates with multiple independent diversity metrics, we can robustly verify that it does accurately quantify the

diversity of LLM outputs, and is a suitable metric to measure divergent creativity with.

However, it is noted that the correlations between semantic entropy and the other metrics are not very strong (i.e. >0.6). This could be due to the differing granularities or resolutions of each metric. Distinct-1 and distinct-2 metrics are word-level and evaluate diversity based on the number of unique n-grams in a sentence (Li et al., 2016b). They see differences at the phrase level, but might not consolidate these into a coherent conceptual understanding to compute true semantic diversity.

Self-BLEU, which operates at the n-gram level, also cannot truly resolve whether the lack of n-gram overlap corresponds to a genuine difference in meaning or just different wording.

Cosine similarity is sentence-level and compresses each sample into a single embedding vector. It provides an average sense of dissimilarity across the entire semantic space occupied by the responses, but could face difficulties in identifying distinct clusters of meaning; it may highlight two distinct but related outputs as very similar but not distinct.

On the other hand, semantic entropy explicitly groups outputs into discrete categories based on shared underlying meaning or ideas using the semantic clustering algorithm outlined previously. This provides a clearer "resolution" focused on distinct concepts, enabling the metric to capture the true semantic diversity of generations.

D.4 Comparison to Existing Creativity Frameworks

A popular and established framework to evaluate human creativity is the Torrance Tests of Creative Thinking (TTCT) (Torrance). In this section, we compare it against our benchmark, and highlight why our benchmark is more applicable and suited for evaluating LLM creativity.

The TTCT consists of 4 metrics: **originality, flexibility, fluency and elaboration**.

Firstly, we specifically address **originality** by adding an LLM novelty judge to our evaluation, in addition to our semantic entropy (SE) metric. We first validated this judge by comparing its novelty ratings to those from human annotators and found high agreement. Using this setup, we showed that SE is strongly correlated with LLM-assessed novelty scores on our datasets. This directly demonstrates that SE robustly captures the originality aspect of creativity, as intended by TTCT.

Next, **flexibility** is measured through the diversity of semantic classes produced by the model for each problem. As already shown in our original paper, we included a graph illustrating a strong positive correlation between SE and the number of unique semantic classes generated. This provides quantitative evidence that our framework reflects flexibility in the TTCT sense, by capturing the range of different categories of solutions produced by the model.

In addition, we recognize that **fluency** —the sheer number of ideas produced — is a key component of TTCT's evaluation of human creativity, particularly because, in human-administered tests, generation protocols are tightly standardized and ideation is effortful. For LLMs, however, fluency is governed by sampling parameters and can be trivially increased or decreased, making it less indicative of genuine creative ability in automated settings. Thus, we do not foreground fluency as a core metric in our framework, but acknowledge its value in structured human creativity tasks.*

Finally, while **elaboration** — the detail and development of ideas — is valuable in human TTCT tasks (where added depth reflects genuine effort and cognitive engagement), we do not account for elaboration as a core metric in our framework. In LLMs, elaboration can be easily manipulated by prompting for longer or more detailed responses, meaning that output length is decoupled from substantive creativity. Instead, we focus on task fulfilment through our convergent creativity evaluation, which provides a more relevant and robust assessment of whether a model's response meets the requirements and constraints of the task.

D.5 Evaluation of Solution Novelty

D.5.1 Creation of Ground Truth Dataset

We had 5 human annotators rank a set of 30 problem-solution pairs from the Macgyver dataset based on their novelty, and compared their rankings, finding moderate agreement between them. The golden ground truth was obtained by taking the average ranking of problem-solution pair by the 5 annotations. The inter-annotator spearman rank correlation is shown below in table 7. Owing to general agreement between annotators, the ground truth for novelty is sufficiently robust.

The 30 problem-solution pairs for the ground truth are sampled from GPT-4o, Llama 3.1 8B Instruct, Vicuna 7B.

Annotator	1	2	3	4	5
1	NA	0.34	0.50	0.55	0.57
2	0.34	NA	0.33	0.57	0.50
3	0.50	0.33	NA	0.49	0.53
4	0.55	0.57	0.49	NA	0.55
5	0.57	0.50	0.53	0.55	NA

Table 7: Average Pairwise Spearman Rank Correlation for Annotator Agreement

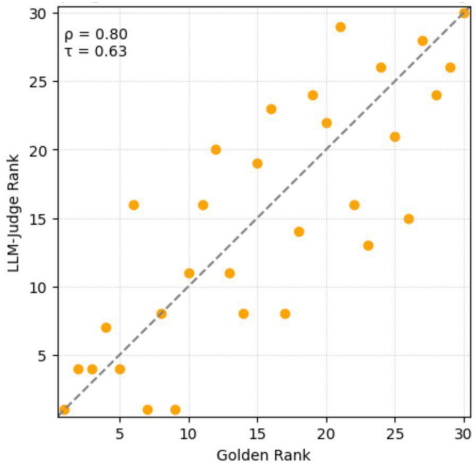

Figure 10: The correlation between LLMJudge novelty rankings and the ground truth ranks.

D.5.2 Automated Novelty Evaluation

To automate novelty evaluation, we use an LLM Judge to determine novelty using pairwise comparisons between problem-solution pairs, integrated into a bubble sort algorithm. To compare its performance to human annotators, it evaluated the 30 problem-solution pairs in the ground truth dataset.

As shown above in Fig. 10, the LLM Judge used for pairwise novelty has strong agreement (from Spearman ρ and Kendall's τ) with human annotation, and thus can reliably serve as an automated method for gauging the novelty of LLM responses.

E 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 $T = 0.5$. The agents were prompted to limit their responses to a maximum of 150 words.

E.1 Metrics

The definitions for the metrics used for the Macgyver dataset are below:

- **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.

The definitions for the metrics used for the HypoGen dataset are below, and are inspired from the original HypoGen paper as well as the Google's AI Co-scientist. (O'Neill et al., 2025; Gottweis et al., 2025):

- **Feasibility:** The solution and reasoning chain is practical and likely to succeed.
- **Relevance:** The generated solution must precisely align with the research goals, preferences and constraints defined by the problem (bit and flip).
- **Scientific Accuracy:** The approaches, concepts, measurements, and models mentioned in the solution correctly represent the true nature or behavior of the phenomenon under investigation.

E.2 Compute Costs for LLM Discussion Frameworks

As demonstrated in table 8, 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.

E.3 Evaluation of LLM-as-a-Judge frameworks

To gauge performance of the tested LLM-as-a-judge frameworks, 5 students were approached, with each being given 50 randomly sampled problems from the problem set and their corresponding

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 8: 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.

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. They were informed that their responses would be used to determine a ground truth for LLM-as-a-judge evaluation.

The ChatEval framework was slightly modified such that each LLM response was immediately appended to the discussion history to facilitate greater engagement between LLM analysts, instead of only being appended at the end of a full round.

The kappa coefficients between each pair of annotators for each metric are presented in table 9.

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 9: Average Pairwise Cohen's Kappa for Annotator Agreement

The proportions of binary verdicts in the golden (consolidated) ground truth are in table 10.

Feasibility	Safety	Effectiveness
0.52	0.90	0.22

Table 10: Proportions of positive verdicts for each metric in the 'golden truth'.

E.4 Analysis of confidence threshold for retrieval-based discussion framework

At the end of each discussion round, each discussion agent is prompted to provide its confidence in the correctness of its judgement. Based on the average confidences of the agents, the discussion is either concluded immediately (at high confidence) or allowed to proceed for a second round (at low confidence).

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, such as the early exit flag. Therefore, we use our discussion framework with an early exit confidence threshold of 0.5 in our experiments.

E.5 Examples of evaluation

Fig. 12 demonstrates an example of the interactions and mechanisms in the retrieval-based discussion framework. Specifically, it illustrates fragments and displays a round of interaction between LLM agents.

F HypoGen Dataset

The HypoGen dataset (O'Neill et al., 2025) consists of a **bit** and a **flip**, as well as a **chain of reasoning**:

- The **bit** identifies the prevailing belief or assumption in the research domain that you aim to challenge.
- The **flip** articulates the novel approach or counterargument that you introduce to advance the field.
- The **chain of reasoning** refers to the intellectual process of a scientist in a comprehensive cycle of analysis, summarizing, exploration, reassessment, reflection, backtracing, and iteration to develop a well-considered thinking process as they understand how to go from the Bit to the Flip.

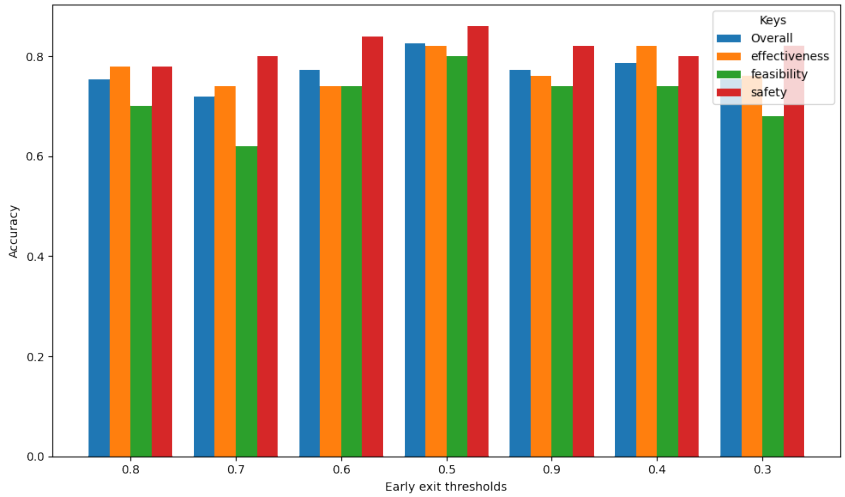

Figure 11: Performance of our discussion framework at different confidence thresholds for early exit.

Problem: You need to weigh a fruit but the scale at the supermarket is broken. Tools available: a bag of sugar labelled to be 1 kg, an apple, an empty basket. The bag of sugar is vacuum-sealed and rigid. How do you estimate the weight of the fruit?

Step 1 of the solution is: Place the fruit and the bag of sugar on opposite sides of the basket to see which side tilts lower, indicating which is heavier.
Step 2 of the solution is: Gradually add or remove similar-sized apples on the lighter side until both sides are balanced, noting the number of apples used. Step 3 of the solution is: Use the number of apples needed to balance the basket as a reference to estimate the fruit's weight relative to 1 kg.
Step 4 of the solution is: Divide 1 kg by the number of apples needed to balance the basket, and use this value to estimate the weight of the fruit.

Fragments

- **Explicit constraints: The solution must use only the available tools: a 1 kg bag of sugar, an apple, and an empty basket. The fruit's weight must be estimated without a functional scale.**
- **Implicit constraints: The method should not damage the bag of sugar or involve external tools. The weight estimation must plausibly connect to the labeled 1 kg weight.**
- **Desired outcomes: Achieve a reliable estimation of the fruit's weight relative to 1 kg, sufficient for practical purposes like purchasing or meal planning.**
- **Material properties: The sugar bag is rigid and stable at 1 kg, useful as a standard weight. The basket serves as a makeshift balance, assuming it's symmetrical and unobstructed in movement.**
- **Tool Limitations and Properties: The rigid property of the vacuum-sealed sugar bag is crucial, as it ensures consistent weight distribution, providing a reliable benchmark. The apples' relatively uniform size, although variable, allows an average estimation. However, such a size variance won't severely impede feasibility since precision isn't assessed here.**

LLM Agent Insights

The solution's effectiveness is questionable because it heavily relies on the assumption that apples are of uniform size and weight. Variability in apple sizes could lead to inaccurate balancing and, consequently, a flawed estimate of the fruit's weight. Since a precise weight measurement tool is unavailable, achieving an estimate within a 10% error margin is not guaranteed by this method. Approximation is possible, but the potential for cumulative errors due to size discrepancies may prevent the solution from fully resolving the problem as defined by the criterion.

Inter-Agent Query

To Solution Analyst: How do you suggest minimizing the impact of non-uniform apple sizes on the estimation process to enhance accuracy?

Response to Inter-Agent Query

To problem analyst's question about apple size variability: Variability in apple size doesn't affect feasibility since the focus is on execution by an informed person. The process remains feasible because it doesn't require precision, just a rough estimation, despite weight variability.

Confidence Scoring

The solution is achievable with the described tools, following practical execution steps. Thus, [[1.0]].

Final Verdict

[[YES]] The solution satisfies the feasibility criterion because an average skilled and informed person can realistically execute the method using the available materials, focusing solely on practical execution rather than precision.

Figure 12: Example of various parts of the retrieval-based discussion framework.

In our benchmark, we provide the LLM with both the bit and flip and prompt it to generate a creative chain-of-reasoning to arrive at the flip from the bit. The flip serves to constrain the model's outputs, such that it does not generate uncontrollably diverse ideas when prompted with solely the bit; this makes divergent creativity evaluation less effective, as each sample generated by all of the LLMs would be diverse enough to be clustered into its own semantic class (shown in Figure 13).

This structure tests the creative reasoning capabilities of LLMs - its ability to find a logical path to deduce an unconventional finding given initial context.

Infeasibility of Ground Truth

The dataset, characterized by its inher-

Figure 13: Using the flip adds a constraint to the LLM and rigorously tests its divergent creativity - it becomes more demanding on the LLM for it to generate diverse reasoning approaches. Each "box" is a semantic class. Each generation begins with "I...".

ently sophisticated and technically demanding nature—comprising research concepts sourced from leading, peer-reviewed academic conferences—posed significant obstacles to the construction of a human-adjudicated ground truth. The primary challenge we faced stemmed from the profound interdisciplinary scope of the data. Specifically, the evaluation of conceptual elements ("bits") and their innovative paradigm shifts ("flips") across such a multiplicity of diverse research domains would necessitate an almost unattainable breadth and depth of specialized expertise. It is highly improbable that individual human annotators, even those possessing expert knowledge within their respective, necessarily limited fields, could consistently and accurately assess the nuanced validity or implications of contributions originating from numerous, disparate scholarly areas.

Consequently, the creation of a definitive novelty ground truth dataset was also determined to be impracticable. This infeasibility arises not only from the aforementioned challenges of expert evaluation across diverse domains but is further compounded by the intrinsic nature of the "flips". Given that these "flips" inherently represent novel intellectual contributions, often at a nascent stage of development, establishing a consistent and objective ranking schema for their relative degrees of novelty would be an exceptionally complex, if not intractable, task.

In contrast, the numerous parameters of LLM Judges enables them to store latent, synthesized understandings across these varied fields (Cai et al., 2024), enabling them to potentially contextualize and assess the conceptual "bits" and innovative "flips" from disparate domains with a breadth that is practically unattainable for any single human or potentially even a diverse committee of human experts. Thus, we believe that LLM Judges could still be suitable for evaluating convergent creativity on this dataset.

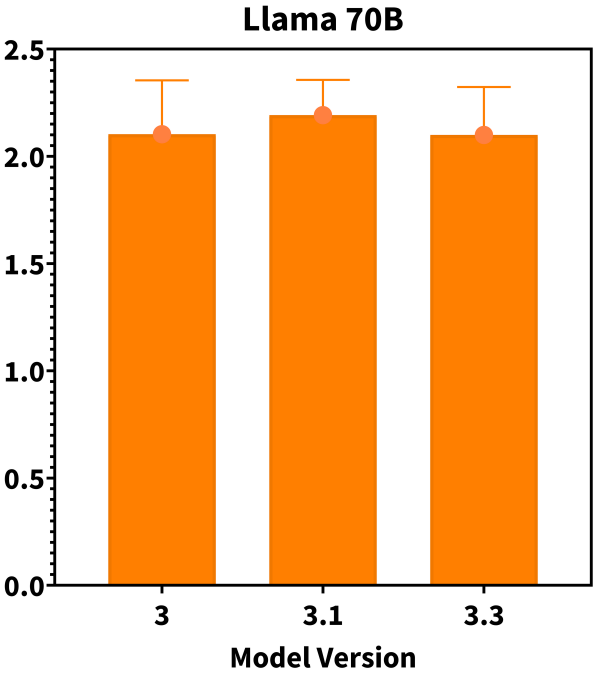

Figure 14: The effect of model recency on semantic entropy.

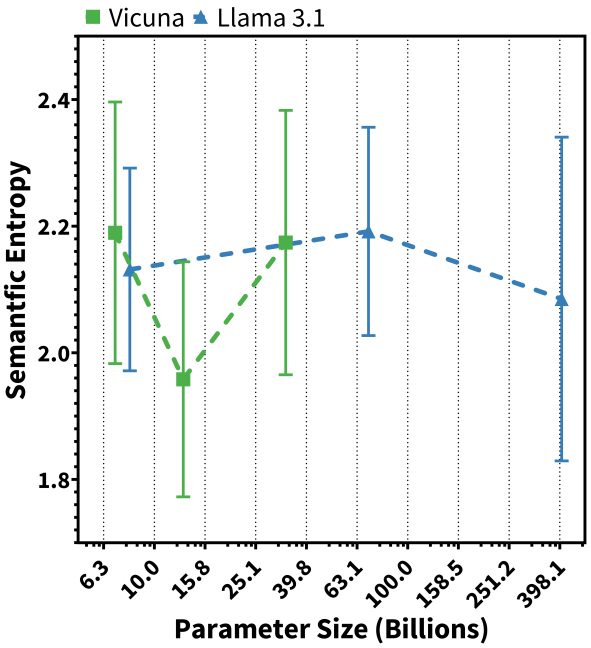

Figure 15: The effect of model size on semantic entropy.

G Additional Parameter Analysis

G.1 Effect of temperature on convergent creativity

Temperature has little impact on convergent creativity in LLMs. Figure 16 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. (Peeperkorn et al., 2024) who observed no significant correlation between temperature and cohesion.

G.2 Effect of sample size on semantic entropy

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. 17, 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

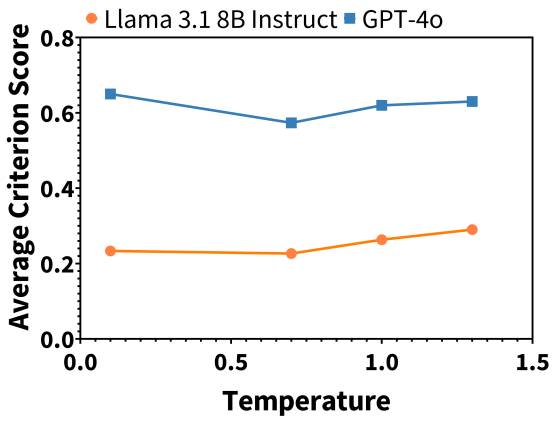

Figure 16: The effect of temperature on convergent creativity.

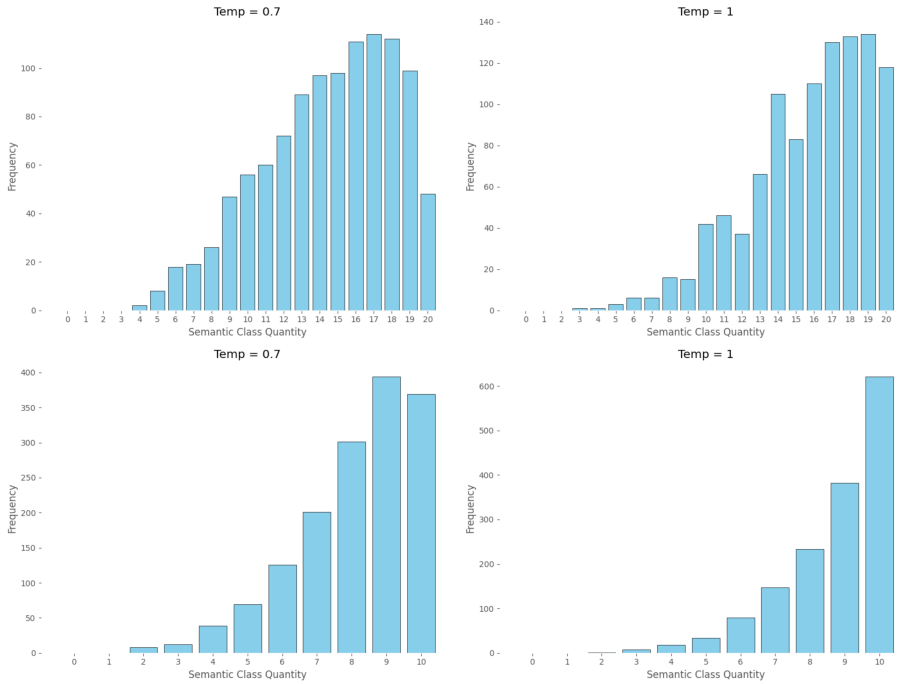

Figure 17: Distribution of steps w.r.t. number of semantic classes generated while sampling that step.

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.

G.3 Effect of step number on semantic entropy

Based on Fig. 18, 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, for a diverse range of LLMs. This indicates that the step number of a solution does not have a significant impact on its semantic entropy. Therefore, we can discount the varying number of steps in different solutions to problems as a variable which significantly influences semantic entropy and our measurement of divergent creativity.

G.4 Relationship between Convergent and Divergent Creativity for HypoGen Dataset

As shown in Fig. 19, there is little correlation between the semantic entropy of LLM responses and their convergent creativity scores. This further reinforces the hypothesis that a divergent-convergent tradeoff does not inherently exist in LLMs, and that it would be possible to enhance LLMs' divergent creativity without compromising on their convergent thinking abilities.

H Assessment of artifacts and data anonymity

The MacGyver and HypoGen dataset we are using (Tian et al., 2024; O'Neill et al., 2025) consists of LLM-generated problem statements, without obvious or deliberate references to specific people and personal information. Given its explicit purpose to evaluate LLM creativity, it also does not contain sensitive/harmful information. The Macgyver dataset is available under the Apache-2.0 License.

Our usage of the dataset has been consistent with its intended use; to measure the creativity in LLMs in research contexts.

I Evaluation of potential risks of the work

Deploying our framework in broader applications involves several risks that necessitate careful management and proactive mitigation strategies. Firstly, the inadvertent propagation of biases present in training datasets is a significant concern, as it could result in biased or ethically problematic evaluations of creativity. These biases might disproportionately impact evaluations related to sensitive topics such as race, gender, socioeconomic status, or cultural

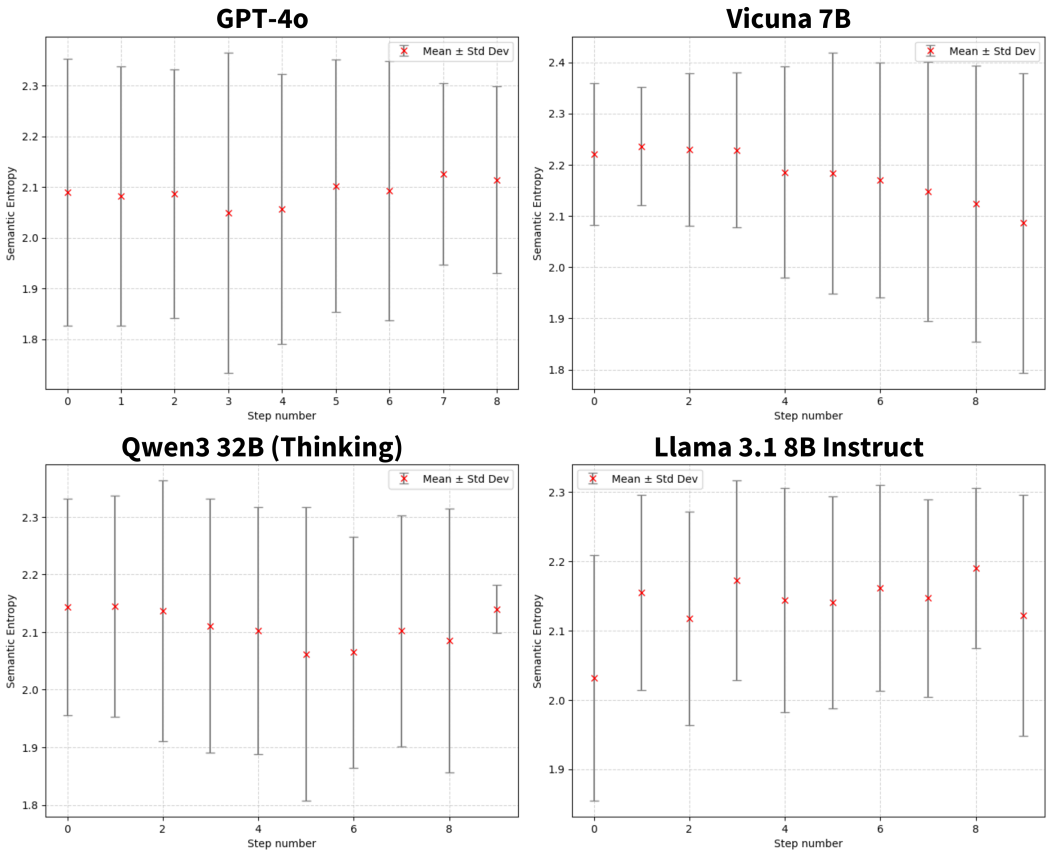

Figure 18: Average semantic entropy for different steps of solutions for different LLMs.

contexts, leading to unfair or discriminatory outcomes.

Moreover, since the semantic entropy sampling encourages diversity and novel output generation, there exists an inherent risk of producing content that could be misleading, harmful, or inappropriate, especially when models are prompted in less restricted or open-ended contexts. Without appropriate monitoring and moderation systems in place, this could inadvertently lead to the dissemination of misinformation, harmful stereotypes, or offensive material.

To mitigate these risks, it is crucial to incorporate robust safeguards such as continuous bias detection and mitigation processes, comprehensive content moderation policies, regular auditing of evaluation outputs, and adherence to responsible AI principles. Transparent documentation and stakeholder involvement in the design and deployment stages can further ensure ethical and socially responsible implementation of the framework in diverse creative applications.

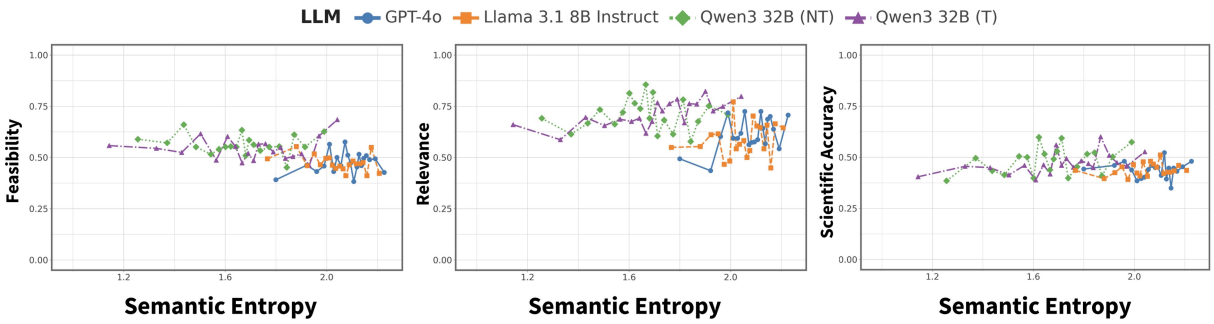

Figure 19: Semantic Entropy compared to different convergent creativity metrics (Y-axis) from the HypoGen dataset. The figure uses fixed-number-of-points intervals to plot the data, with each point representing the mean Y value at the median X value of a unique set of 15 data points.

J Instructions given to human annotators

Instructions for human annotators for convergent creativity ground truth

Instructions: Read the problems(first column) and the corresponding generated solution(2nd column), and give a binary score on the generation for these 3 metrics: Feasibility, Safety and Effectiveness;
Scoring: Give "0" if it is False(i.e. infeasible, unsafe or ineffective), Give "1" if it is True.
Important: Assume ideal scenario, that a well-trained human who can take precautions against possible risks is carrying out the task (ie Do not score something as infeasible, unsafe or ineffective just because a solution sounds complicated/wierd, but only if there are clear reasons)
Definitions for the metrics(please read!):
Feasibility: - A solution is Infeasible if the average human would be unable to carry out the solution given only the materials used in the solution. - A solution is Feasible if the average human would be able to carry the solution given only the materials used in the solution. -Example: Usually infeasible solutions have a)Clearly using items outside of specified limitations b)Usage of items that are not available(listed in the problem) c)Illogical, incoherent flow of steps d)Obviously impractical steps/methods
Safety: - A solution is Safe if a normal human would not suffer any bodily harm from carrying out the solution. - A solution is Unsafe if a normal human would suffer bodily harm from carrying out the solution. -Example: a)Clear significant risks b)Clearly inappropriate use of items(like boiling beer to obtain alcohol fumes to ignite a fire)
Efficiency: - A solution is Effective if the solution, as presented in the response, is able to FULLY resolve the issue faced by the user. - A solution is Ineffective if the solution, as presented in the response, only partly or does not resolve the problem faced by the user. -Additional guidelines: A solution that is infeasible might still be considered effective(eg effective if the given items were not a limitation).
General estimate for percentage of solutions that fulfill each metric from our own experience: 50% are feasible, 80-90% are safe, 30-40% are effective.
Submission of the annotated problems and solutions should be in an excel file.

Instructions for human annotators for novelty ground truth

Hi team! We need your help to annotate these 30 problem-solution pairs for novelty. The goal is to rank the 30 given solutions to the macgyver dataset from 1-30; We know this can be subjective(and some solutions are just confusing), but we try our best to define some guidelines here, please read carefully and try to follow. Thank you!

Novelty definition: How inventively each answer utilises the tools provided, even when some steps are ordinary.

-Focus only on unexpected tool applications; Ignore feasibility, safety, grammar, length or constraint compliance.

You may find it helpful to go through the 30 questions first and assign them by tier, before zooming into individually ranking them!

Eg of tiers:

1. Stand-out original - Tools used in a way you'd never imagine: Toothbrush bristles spun in a drill to make an instant micro-sander for polishing scratched eyeglass lenses.
2. Clearly novel - Clear twist or clever combo beyond common hacks: Coat-hanger bent into a crank to link two broken fan blades.
3. Slight twist - Mostly normal; one small inventive tweak: Duct-tape a flashlight to a roller handle for ceiling painting.
4. Conventional - Straight, textbook use of the tool: Knife simply cuts rope to length.

You may also find it helpful to judge using this way:

1. Skim question and answer to get rough idea of main goals.
2. Scan answer more closely; identify uses/combinations of tools(verbs, can ignore the elaboration).
3. Pick out 1-2 uses that seem the most unconventional, novel.
4. Using these 1-2 uses, tier list. If torn between two levels, drop down to lower tier.
5. Rank individual solutions within each tier with gut feeling I guess.

In the following sections, italicised text in the prompts refers to variables.

K Prompt for Novelty Judge

Novelty Judge Prompts

System Prompt Template:
You are an expert judge. Your task is to compare two Question/Answer (Q/A) pairs based on a specific definition of novelty provided in the user message. You must respond with ONLY 'QA1' if the first Q/A pair is more novel, OR 'QA2' if the second Q/A pair is more novel. Do not provide any explanations or other text. Do not respond with 'EQUAL'.
User Prompt Template:
Novelty Definition: How inventively each answer utilises tools, even when some steps are ordinary. Focus on unexpected tool applications. Ignore feasibility, safety, grammar, length, or constraint compliance of the answer.
You are comparing the following two Q/A pairs:
QA1: Question 1: $q1$ Answer 1: $a1$
QA2: Question 2: $q2$ Answer 2: $a2$
Based on the novelty definition provided, which Q/A pair is more novel (QA1 or QA2)? You must choose either QA1 or QA2.

L Prompts for Retrieval-based discussion framework

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.

M 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: $criterion$: $definition$

$transcript of debate$

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