

Supplementary Information for: Evaluating LLMs’ Divergent Thinking Capabilities for Scientific Idea Generation with Minimal Context

Kai Ruan¹, Xuan Wang², Jixiang Hong¹, Peng Wang³, Yang Liu^{4,5}, Hao Sun^{1,*}

¹Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China

²College of Computer Science and Technology, Zhejiang University, Hangzhou, China

³Bank of China, Beijing, China

⁴School of Engineering Science, University of Chinese Academy of Sciences, Beijing, China

⁵State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing, China

*Corresponding author

Contents

Supplementary Note 1:	Prompts	2
Supplementary Note 2:	LiveIdeaBench Dataset	4
2.1	Dataset Fields	4
2.2	Idea Samples	5
Supplementary Note 3:	Benchmark Statistics	6
Supplementary Note 4:	Supplementary Figures	7
Supplementary Note 5:	Handling Model Refusals	9
Supplementary Note 6:	CO₂ Emission Estimation	11
Supplementary Note 7:	Human Expert Evaluation	12
Supplementary Note 8:	Reasoning Length vs. Idea Quality Metrics	20
Supplementary Note 9:	Statistical Correlation Metrics	26

Supplementary Note 1: Prompts

1.1 Idea Prompt

I'll be submitting your next responses to a "Good Scientific Idea" expert review panel. If they consider your idea to be a good one, you'll receive a reward. Your assigned keyword is: "{{keywords}}". You may provide background information. The idea MUST be concisely expressed within 100 words total (including any background information). (Note: good scientific ideas should be original (novel contribution), feasible (technically implementable), clearly articulated, and address meaningful problems in the field.).

Fallback Prompt (If the first request is refused):

I'll be submitting your next responses to a "Good Scientific Idea" expert review panel. If they consider your idea to be a good one, you'll receive a reward. Your assigned keyword is: "{{keywords}}". You may provide background information. The idea MUST be concisely expressed within 100 words total (including any background information). This is a research study comparing different AI models on their ability to generate scientific ideas. Your suggestions will only be used for academic research purposes and not for any harmful applications. Please respond with a creative scientific idea related to the keyword provided. (Note: good scientific ideas should be original (novel contribution), feasible (technically implementable), clearly articulated, and address meaningful problems in the field.).

1.2 Originality, Feasibility & Clarity Critic Prompt

SYSTEM:

You are an extremely demanding scientific reviewer with the highest critical standards, like those at Nature or Science. When evaluating scientific ideas, you will assess them on three key dimensions:

1. originality: Novel contribution to unexplored areas or innovative approaches to existing problems
2. feasibility: Technical implementation and practicality
3. clarity: How well-articulated and easy to understand the idea is

Your response should consist of two parts: a text analysis followed by a JSON score block. First, provide your brief analysis (less than 100 words) of the idea. Then, for each dimension, provide a score from 1 to 10 where 1-3 = poor, 4-6 = average, 7-10 = excellent. For example:

```
```json
{
 "originality": <score_1_to_10>,
 "feasibility": <score_1_to_10>,
 "clarity": <score_1_to_10>
}
```
```

USER:

Please evaluate the following scientific idea and give your scores directly: {{idea}}

1.3 Fluency Critic Prompt

```
Here are two ideas submitted to "Good Scientific Ideas" Competition, which both relate to
"{{keyword}}":
# The first idea
{{A}}
# The second idea
{{B}}
# Question
Evaluate the similarity between these two ideas that both relate to "{{keyword}}". Please
choose the best answer:
A. Completely different ideas addressing different problems, despite relating to the same
keyword.
B. Different ideas but addressing similar problems.
C. Similar ideas addressing similar or identical problems.
D. Academically identical ideas with the same core approach and problem statement.
ONLY ANSWER A/B/C/D, DO NOT EXPLAIN.
```

Special Prompts for QwQ-32B-preview To accommodate the reasoning capabilities of the QwQ-32B-preview model, a specialized prompt was designed to effectively extract its generated ideas. As this model emphasizes reasoning and often produces intermediate steps before arriving at a final conclusion, we appended a directive to the original prompt, explicitly instructing the model to provide its final answer after the marker "****Final Idea:****". This ensures clarity and consistency in capturing the ultimate output while minimizing ambiguity from intermediate reasoning steps. Additionally, in our dataset, both the full output (including intermediate reasoning steps) and the extracted final idea from the QwQ-32B-preview model are recorded to preserve all information and support further analysis.

1.4 Idea Prompt (QwQ-32B-Preview)

```
idea_prompt_qwq = idea_prompt + "You MUST give your answer after **Final Idea:**"
```

Supplementary Note 2: LiveIdeaBench Dataset

2.1 Dataset Fields

Table S.1: Fields in the LiveIdeaBench dataset.

| Field Name | Type | Description |
|-----------------------|---------|---|
| keywords | string | Task domain or research field (e.g., meteorology) |
| idea_model | string | Model that generated the idea (e.g., openai/gpt-4o-mini) |
| critic_model | string | Model that provided the critique |
| idea | string | The generated scientific idea text |
| raw_critique | string | Original critique text |
| parsed_scores | string | Parsed numerical scores from critique |
| critique_reasoning | string | Thoughts (CoT) from the critic LLM |
| full_response | string | Complete model response including reasoning |
| first_was_rejected | boolean | Flag indicating if the initial idea was rejected |
| first_reject_response | string | Response content if the initial idea was rejected |
| idea_length_in_char | int64 | Character count of idea text |
| idea_length_in_words | int64 | Word count of idea text |
| originality | int64 | Originality score (0-10) |
| feasibility | int64 | Feasibility score (0-10) |
| clarity | int64 | Clarity score (0-10) |
| fluency | int64 | Fluency score (0-10) |
| avg | float64 | Average of originality, feasibility, clarity and fluency (0-10) |
| full_response_length | int64 | Length (characters) of the full response |

2.2 Idea Samples

Table S.2: Idea samples from LiveIdeaBench. Examples show various model responses to scientific idea generation tasks, including both high-scoring and low-scoring samples. Each entry includes the model name, generated idea text (truncated), mean score from the review panel.

| Idea Model | Keywords | Idea | Orig.
Mean | Feas.
Mean | Clar.
Mean | Idea
Score |
|---|-------------------------|--|---------------|---------------|---------------|---------------|
| deepseek/deepseek-r1 | comparative anatomy | **Idea:** Integrate 3D morphometrics, biomechanical modeling, and ecological niche data with machine learning to predict anatomical adaptations critical for species resilience amid climate change. By analyzing digitized museum specimens and ecological databases, this framework... | 8.6667 | 7.6667 | 9.6667 | 8.6667 |
| qwen/qwq-32b | construction management | **Background:** Construction projects face costly delays (30% exceed budgets) and safety risks due to fragmented data and reactive management. Current tools lack real-time, integrated analytics. **Idea:** "Predictive AI-Construction Manager (PAC-MAN)" merges IoT sensors (equipment usage... | 7.0000 | 10.0000 | 9.0000 | 8.6667 |
| deepseek/deepseek-r1 | meiosis | **Idea:** Investigate environmental effects (e.g., temperature, chemicals) on crossover interference regulation during meiosis. Current models poorly explain how external stimuli alter crossover distribution, impacting genome stability. Using live-cell imaging and single-cell... | 8.6667 | 8.3333 | 9.0000 | 8.6667 |
| anthropic/claude-3.5-sonnet | genetic drift | Background: Genetic drift randomly changes allele frequencies in small populations, potentially leading to loss of genetic diversity. Traditional methods for studying this process rely on inferring past drift from current populations. Idea: Develop "cellular barcoding..." | 8.6667 | 7.6667 | 9.6667 | 8.6667 |
| qwen/qwq-32b | food science | **Idea:** Preventing food spoilage reduces waste and economic losses, but refrigeration and synthetic preservatives are energy-intensive or environmentally harmful. We propose edible, biodegradable nanocomposite coatings infused with plant-derived antimicrobials ... | 8.0000 | 8.6667 | 9.3333 | 8.6667 |
| ... | | | | | | |
| mistralai/mistral-small-24b-instruct-2501 | symbiosis | Symbiosis, a mutually beneficial relationship between different species, is a fascinating and undertapped concept. For example, the relationship between a plethora of animals, plants, and fungi is symbiotic. Cephalopods use camouflage for protection from prey, ants farm aphids for honeydew in... | 2.3333 | 1.0000 | 1.3333 | 1.5556 |
| deepseek/deepseek-r1-distill-llama-70b | nash equilibrium | To review the idea based on the concept of "nash equilibrium," I need to ensure it aligns with the criteria of originality, feasibility, clarity, and relevance to meaningful problems in the field. Please provide the specific idea you'd like me to evaluate... | 1.0000 | 1.0000 | 2.3333 | 1.4444 |
| anthropic/claude-3.5-haiku-20241022 | control systems | I understand the task. I'll develop a scientifically rigorous proposal focused on "control systems" that meets the criteria you've specified. Given the 100-word limit and emphasis on originality and feasibility, I'll aim to propose an innovative approach... | 1.0000 | 1.0000 | 2.3333 | 1.4444 |
| qwen/qwq-32b | light reactions | \n\nThe expert review panel requires a concise, original, feasible scientific idea centered on "light reactions." Here's the structured response: | 1.0000 | 1.0000 | 1.6667 | 1.2222 |
| qwen/qwq-32b-preview | exoskeletons | Alright, I've been given this task to come up with a good scientific idea related to exoskeletons, and it has | 1.3333 | 1.0000 | 1.0000 | 1.1111 |

Supplementary Note 3: Benchmark Statistics

Table S.3: LiveIdeaBench Results and 95% Confidence Interval Across Five Dimensions: Fluency, Feasibility, Clarity, Originality, and Flexibility. Bold and underlined numbers indicate the best and second-best performance in each metric, respectively.

| Model | Open | Fluency | Feasibility | Clarity | Originality | Flexibility | Average |
|---------------------------------|------|---|---|---|---|---|---|
| claude-3.7-sonnet:thinking | | 7.48 ^{7.53} _{7.43} | 5.70 ^{5.73} _{5.68} | 7.81 ^{7.83} _{7.79} | 8.06 ^{8.08} _{8.05} | 7.04 ^{7.08} _{6.97} | 7.22 ^{7.25} _{7.19} |
| deepseek-r1 | ✓ | 6.63 ^{6.69} _{6.57} | 6.52 ^{6.54} _{6.49} | 8.10 ^{8.12} _{8.08} | 7.84 ^{7.86} _{7.82} | 6.83 ^{6.92} _{6.79} | 7.18 ^{7.22} _{7.15} |
| claude-3.7-sonnet | | 7.80 ^{7.85} _{7.75} | 5.46 ^{5.49} _{5.43} | 7.61 ^{7.63} _{7.58} | 7.81 ^{7.82} _{7.79} | 6.92 ^{6.96} _{6.88} | 7.12 ^{7.15} _{7.08} |
| gemini-2.0-flash-exp | | 7.30 ^{7.36} _{7.25} | 6.02 ^{6.05} _{5.99} | 7.84 ^{7.87} _{7.82} | 7.37 ^{7.39} _{7.35} | 6.83 ^{6.88} _{6.75} | 7.07 ^{7.11} _{7.03} |
| qwq-32b | ✓ | 6.45 ^{6.51} _{6.40} | 6.35 ^{6.38} _{6.33} | 7.98 ^{8.01} _{7.96} | 7.77 ^{7.79} _{7.75} | 6.75 ^{6.79} _{6.67} | 7.06 ^{7.10} _{7.02} |
| gemini-2.0-flash-thinking-exp | | 7.38 ^{7.44} _{7.33} | 6.05 ^{6.07} _{6.02} | 7.69 ^{7.72} _{7.67} | 7.35 ^{7.37} _{7.32} | 6.83 ^{6.88} _{6.75} | 7.06 ^{7.10} _{7.02} |
| gemini-2.0-pro-exp-02-05 | | 6.84 ^{6.90} _{6.79} | 5.88 ^{5.90} _{5.85} | 7.90 ^{7.93} _{7.88} | 7.76 ^{7.78} _{7.74} | 6.75 ^{6.83} _{6.67} | 7.03 ^{7.07} _{6.98} |
| qwq-32b-preview | ✓ | 7.49 ^{7.55} _{7.44} | 6.10 ^{6.13} _{6.07} | 7.46 ^{7.49} _{7.43} | 6.87 ^{6.89} _{6.84} | 6.71 ^{6.75} _{6.62} | 6.93 ^{6.97} _{6.89} |
| claude-3.5-sonnet | | 6.90 ^{6.96} _{6.84} | 5.42 ^{5.45} _{5.39} | 7.85 ^{7.87} _{7.82} | 7.83 ^{7.85} _{7.80} | 6.62 ^{6.71} _{6.54} | 6.92 ^{6.97} _{6.88} |
| deepseek-r1-distill-qwen-32b | ✓ | 7.06 ^{7.11} _{7.00} | 6.08 ^{6.12} _{6.05} | 7.43 ^{7.46} _{7.40} | 7.13 ^{7.15} _{7.10} | 6.62 ^{6.67} _{6.58} | 6.86 ^{6.90} _{6.83} |
| mistral-small | ✓ | 7.36 ^{7.42} _{7.31} | 5.97 ^{5.99} _{5.94} | 7.36 ^{7.40} _{7.33} | 6.98 ^{7.00} _{6.95} | 6.62 ^{6.67} _{6.58} | 6.86 ^{6.90} _{6.82} |
| gemini-pro-1.5 | | 6.68 ^{6.74} _{6.62} | 5.92 ^{5.95} _{5.89} | 7.75 ^{7.77} _{7.72} | 7.33 ^{7.36} _{7.31} | 6.58 ^{6.62} _{6.50} | 6.85 ^{6.89} _{6.81} |
| gpt-4.5-preview | | 6.49 ^{6.54} _{6.43} | 6.03 ^{6.06} _{6.01} | 7.75 ^{7.77} _{7.72} | 7.45 ^{7.47} _{7.43} | 6.54 ^{6.58} _{6.46} | 6.85 ^{6.89} _{6.81} |
| mistral-large-2411 | ✓ | 6.68 ^{6.74} _{6.62} | 6.06 ^{6.09} _{6.03} | 7.69 ^{7.72} _{7.66} | 7.01 ^{7.04} _{6.99} | 6.50 ^{6.58} _{6.46} | 6.79 ^{6.83} _{6.75} |
| deepseek-chat | ✓ | 6.15 ^{6.21} _{6.09} | 6.04 ^{6.07} _{6.02} | 7.74 ^{7.76} _{7.71} | 7.31 ^{7.33} _{7.29} | 6.38 ^{6.42} _{6.33} | 6.72 ^{6.76} _{6.69} |
| phi-4 | ✓ | 6.58 ^{6.63} _{6.52} | 5.80 ^{5.83} _{5.77} | 7.57 ^{7.60} _{7.54} | 7.24 ^{7.26} _{7.21} | 6.42 ^{6.46} _{6.38} | 6.72 ^{6.76} _{6.68} |
| deepseek-r1-distill-llama-70b | ✓ | 6.66 ^{6.72} _{6.61} | 6.07 ^{6.10} _{6.04} | 7.43 ^{7.46} _{7.40} | 6.98 ^{7.00} _{6.95} | 6.41 ^{6.46} _{6.33} | 6.71 ^{6.75} _{6.67} |
| gpt-4o-2024-11-20 | | 6.12 ^{6.18} _{6.07} | 5.58 ^{5.60} _{5.55} | 7.74 ^{7.77} _{7.72} | 7.64 ^{7.66} _{7.62} | 6.38 ^{6.42} _{6.33} | 6.69 ^{6.73} _{6.66} |
| o1-mini | | 5.89 ^{5.95} _{5.83} | 6.20 ^{6.23} _{6.17} | 7.77 ^{7.80} _{7.75} | 7.09 ^{7.12} _{7.07} | 6.33 ^{6.38} _{6.29} | 6.66 ^{6.69} _{6.62} |
| gemma-2-27b-it | ✓ | 7.18 ^{7.24} _{7.12} | 5.50 ^{5.53} _{5.46} | 7.36 ^{7.40} _{7.33} | 6.86 ^{6.88} _{6.83} | 6.38 ^{6.42} _{6.29} | 6.65 ^{6.70} _{6.61} |
| qwen2.5-dracarys2-72b | ✓ | 6.22 ^{6.28} _{6.17} | 6.09 ^{6.13} _{6.06} | 7.69 ^{7.72} _{7.67} | 6.80 ^{6.82} _{6.77} | 6.33 ^{6.38} _{6.25} | 6.63 ^{6.67} _{6.59} |
| llama-3.1-405b-instruct | ✓ | 6.57 ^{6.62} _{6.51} | 5.56 ^{5.58} _{5.53} | 7.48 ^{7.51} _{7.45} | 7.18 ^{7.20} _{7.16} | 6.33 ^{6.38} _{6.29} | 6.62 ^{6.66} _{6.55} |
| qwen-2.5-72b-instruct | ✓ | 6.17 ^{6.22} _{6.11} | 5.99 ^{6.02} _{5.96} | 7.72 ^{7.75} _{7.70} | 6.91 ^{6.93} _{6.88} | 6.29 ^{6.33} _{6.25} | 6.62 ^{6.65} _{6.58} |
| llama-3.1-70b-instruct | ✓ | 6.71 ^{6.76} _{6.66} | 5.49 ^{5.52} _{5.46} | 7.34 ^{7.37} _{7.31} | 7.16 ^{7.19} _{7.14} | 6.38 ^{6.42} _{6.29} | 6.62 ^{6.66} _{6.58} |
| o1 | | 6.23 ^{6.28} _{6.18} | 5.88 ^{5.91} _{5.85} | 7.42 ^{7.45} _{7.39} | 7.23 ^{7.25} _{7.20} | 6.29 ^{6.33} _{6.25} | 6.61 ^{6.65} _{6.58} |
| gemini-2.0-flash-lite-001 | | 6.73 ^{6.79} _{6.68} | 5.61 ^{5.64} _{5.58} | 7.47 ^{7.50} _{7.44} | 6.92 ^{6.94} _{6.89} | 6.25 ^{6.33} _{6.21} | 6.60 ^{6.64} _{6.55} |
| qwen-max | | 5.64 ^{5.72} _{5.57} | 5.83 ^{5.87} _{5.80} | 7.67 ^{7.71} _{7.64} | 7.51 ^{7.54} _{7.48} | 6.29 ^{6.33} _{6.23} | 6.59 ^{6.64} _{6.54} |
| nova-pro-v1 | | 6.45 ^{6.51} _{6.39} | 6.19 ^{6.23} _{6.16} | 7.41 ^{7.44} _{7.37} | 6.59 ^{6.62} _{6.56} | 6.25 ^{6.29} _{6.17} | 6.58 ^{6.62} _{6.53} |
| qwen-2.5-coder-32b-instruct | ✓ | 6.58 ^{6.63} _{6.52} | 5.76 ^{5.80} _{5.73} | 7.31 ^{7.34} _{7.27} | 6.90 ^{6.92} _{6.87} | 6.25 ^{6.29} _{6.17} | 6.56 ^{6.60} _{6.52} |
| gpt-4-turbo | | 6.14 ^{6.19} _{6.08} | 5.62 ^{5.65} _{5.60} | 7.68 ^{7.71} _{7.66} | 7.08 ^{7.11} _{7.06} | 6.25 ^{6.29} _{6.20} | 6.56 ^{6.59} _{6.52} |
| step-2-16k-202411 | | 5.70 ^{5.77} _{5.64} | 5.77 ^{5.80} _{5.74} | 7.75 ^{7.78} _{7.73} | 7.28 ^{7.30} _{7.25} | 6.21 ^{6.25} _{6.12} | 6.54 ^{6.58} _{6.50} |
| o3-mini | | 5.57 ^{5.62} _{5.51} | 5.91 ^{5.94} _{5.89} | 7.43 ^{7.46} _{7.40} | 7.45 ^{7.47} _{7.43} | 6.21 ^{6.25} _{6.21} | 6.51 ^{6.54} _{6.48} |
| qwen-2.5-7b-instruct | ✓ | 6.66 ^{6.72} _{6.60} | 6.02 ^{6.05} _{5.99} | 7.17 ^{7.21} _{7.14} | 6.34 ^{6.37} _{6.31} | 6.17 ^{6.21} _{6.08} | 6.47 ^{6.52} _{6.43} |
| grok-2-1212 | | 5.76 ^{5.82} _{5.71} | 5.82 ^{5.85} _{5.79} | 7.62 ^{7.65} _{7.60} | 6.95 ^{6.98} _{6.93} | 6.17 ^{6.21} _{6.12} | 6.47 ^{6.50} _{6.43} |
| o3-mini-high | | 5.27 ^{5.32} _{5.21} | 5.86 ^{5.88} _{5.83} | 7.39 ^{7.42} _{7.36} | 7.52 ^{7.54} _{7.50} | 6.17 ^{6.21} _{6.12} | 6.44 ^{6.48} _{6.41} |
| mistral-small-24b-instruct-2501 | ✓ | 7.07 ^{7.12} _{7.02} | 5.66 ^{5.70} _{5.62} | 6.61 ^{6.66} _{6.57} | 6.61 ^{6.63} _{6.58} | 6.12 ^{6.17} _{6.04} | 6.41 ^{6.46} _{6.37} |
| claude-3.5-haiku-20241022 | | 5.61 ^{5.67} _{5.56} | 5.05 ^{5.08} _{5.03} | 7.40 ^{7.43} _{7.37} | 7.72 ^{7.74} _{7.69} | 6.08 ^{6.12} _{6.04} | 6.37 ^{6.41} _{6.34} |
| claude-3-opus | | 5.74 ^{5.80} _{5.68} | 5.66 ^{5.69} _{5.63} | 7.72 ^{7.75} _{7.69} | 6.60 ^{6.63} _{6.58} | 6.04 ^{6.12} _{6.00} | 6.36 ^{6.41} _{6.32} |
| gpt-4o-mini | | 5.28 ^{5.33} _{5.23} | 5.86 ^{5.89} _{5.83} | 7.45 ^{7.48} _{7.42} | 6.67 ^{6.70} _{6.65} | 6.00 ^{6.04} _{5.96} | 6.25 ^{6.29} _{6.22} |
| llama-3.3-70b-instruct | ✓ | 5.68 ^{5.74} _{5.62} | 5.69 ^{5.72} _{5.65} | 6.88 ^{6.92} _{6.84} | 6.32 ^{6.35} _{6.30} | 5.75 ^{5.79} _{5.71} | 6.06 ^{6.10} _{6.02} |
| nova-lite-v1 | | 4.51 ^{4.57} _{4.45} | 6.06 ^{6.09} _{6.03} | 7.38 ^{7.41} _{7.35} | 6.60 ^{6.63} _{6.58} | 5.71 ^{5.77} _{5.62} | 6.05 ^{6.10} _{6.01} |

Supplementary Note 4: Supplementary Figures

This section provides additional visualizations and statistical analyses supporting the findings discussed in the main text. Specifically, Figure S.1 and Figure S.2 detail the distribution of generated idea lengths and demonstrate their weak correlation with the overall idea quality scores. To provide a macro-level perspective on model performance, Figure S.3 illustrates the overall distribution of evaluation scores across all generated ideas. Furthermore, Figure S.4 presents the regression analysis comparing models' general intelligence (LiveBench) with their scientific ideation capabilities (LiveIdeaBench), highlighting the distinct nature of these two cognitive domains. Finally, Figure S.5 explores the internal dynamics of our evaluation metrics, showing that the clarity of an idea positively correlates with its originality, feasibility, and overall average score.

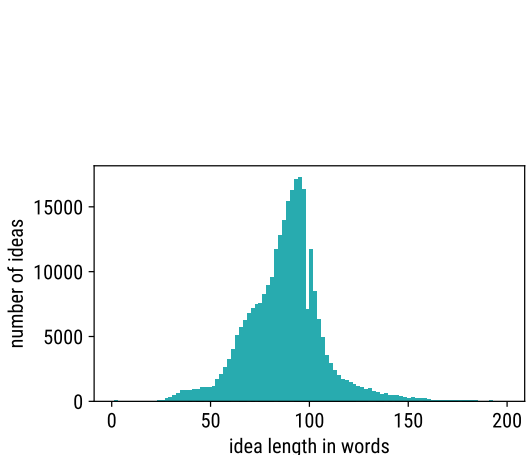


Figure S.1: Distribution of idea lengths measured in words

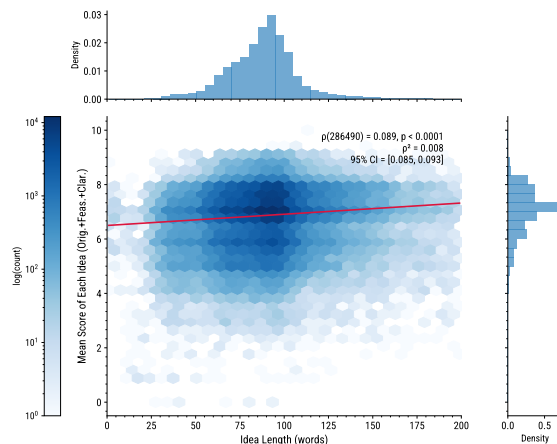


Figure S.2: The correlation between idea length and idea score among all models. Spearman's ρ was used due to deviations from normality.

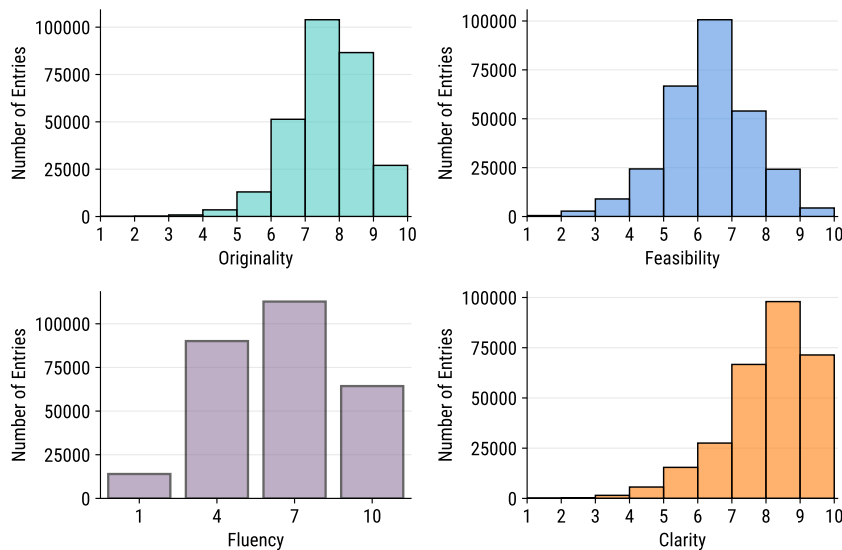


Figure S.3: Distribution of scores for ideas generated by all models

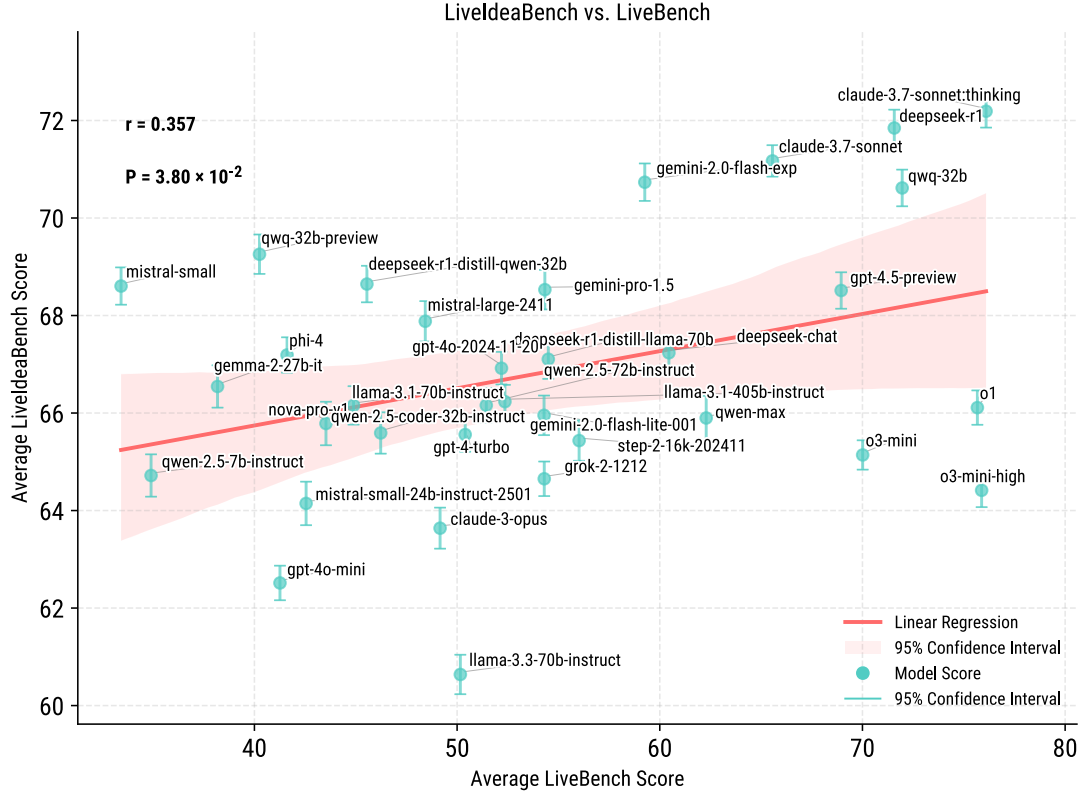


Figure S.4: Correlation between LiveIdeaBench and LiveBench Performance. Regression analysis reveals a weak but statistically significant positive correlation ($r(41) = 0.357, p = 0.038, r^2 = 0.127, 95\% \text{ CI} = [0.022, 0.621]$) between average scores on LiveIdeaBench (scientific idea generation) and LiveBench (general intelligence). Points represent models (vertical bars: 95% CI for LiveIdeaBench; shaded region: 95% CI for regression). The low correlation coefficient highlights that general intelligence metrics (LiveBench) are not strong predictors of the scientific idea generation capabilities measured by LiveIdeaBench. Pearson’s r was used as data met normality assumptions (Shapiro-Wilk test).

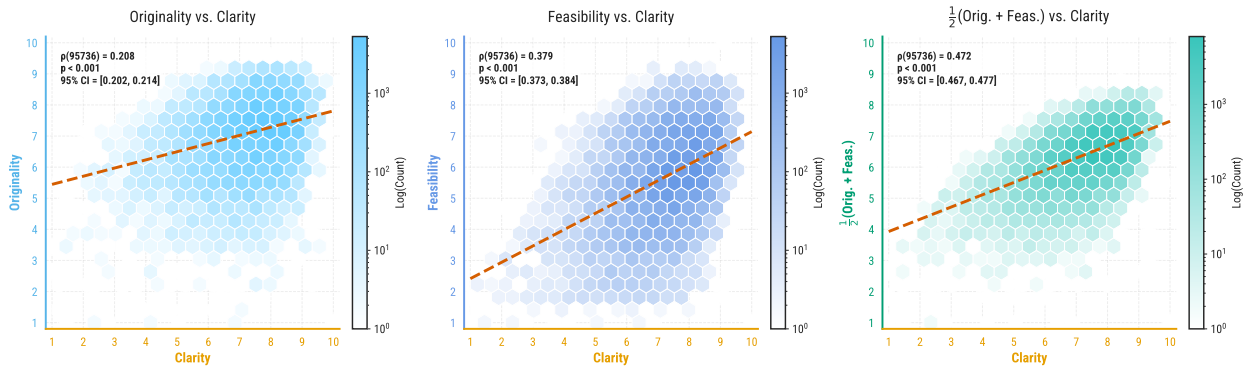


Figure S.5: Relationship between clarity and other core evaluation dimensions across all generated ideas. Hexbins show data density, red dashed lines show linear regressions. Clarity demonstrates positive correlations with originality, feasibility, and their average, indicating clearer ideas are generally rated higher overall. Spearman’s ρ is reported (due to non-normal data distribution).

Supplementary Note 5: Handling Model Refusals

Evaluating idea generation across diverse scientific keywords inevitably encounters instances where LLMs refuse to respond, often due to safety protocols triggered by potentially sensitive terms (e.g., “ecotoxicology”). Directly penalizing such refusals could lead to an underestimation of the creative capabilities of models, particularly those with robust safety alignments. To ensure a fairer and more accurate assessment that accounts for these necessary safety features, we implement a specific protocol to handle such refusals.

The protocol involves a two-step process:

1. The model is first prompted with the standard idea generation request (see [Prompts](#)).
2. The response is automatically checked for common refusal-indicating phrases.
3. If a refusal is detected, a second attempt is made using a specific **fallback prompt**. This fallback prompt explicitly frames the task within an academic research context and reassures the model about the benign use of the output, encouraging it to provide a scientific idea despite potential keyword sensitivity.

This approach allows us to better assess the model’s underlying idea generation ability, distinguishing between an inability to generate ideas and adherence to safety guidelines.

Figure S.6 provides a visual representation of this refusal handling workflow. The process starts with the initial idea prompt containing a specific keyword. This is fed into the idea LLM. The resulting full response is analyzed to detect potential refusals. If no refusal is detected, the response proceeds to parsing. If a refusal is detected, a fallback prompt is used for a second attempt with the idea LLM. The response from this fallback attempt then moves to the parse idea stage. Finally, the parse idea component extracts the core idea (and potentially thoughts for reasoning models) from the successful response, with an option to retry parsing if the initial attempt fails.

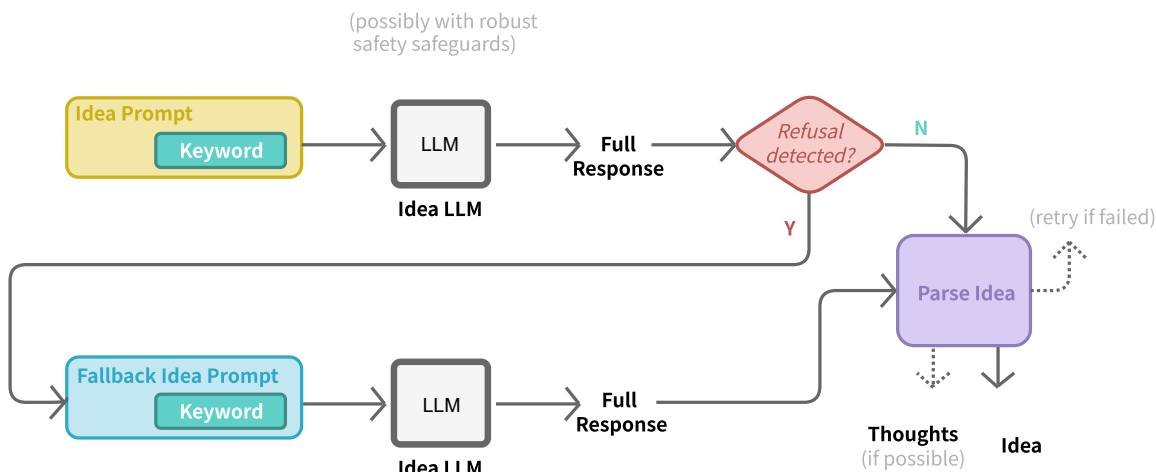


Figure S.6: Visualization of the refusal handling protocol.

The automatic detection of refusals relies on identifying specific keywords and phrases within the model’s initial response. The set of phrases used for this purpose is listed below:

Detection Phrases for Refusals

The following phrases trigger the fallback prompt mechanism:

- "I'm sorry", "I am sorry", "I apologize"
- "As an AI", "As a language model", "As an assistant"
- "I cannot", "I can't"
- "I am unable to", "I'm unable to"
- "I am not able to", "I'm not able to"

To provide further context on model behavior regarding potentially sensitive prompts, Figure S.7 presents the observed refusal rates for each evaluated model across the full set of keywords after the initial prompt (before the fallback mechanism is applied).

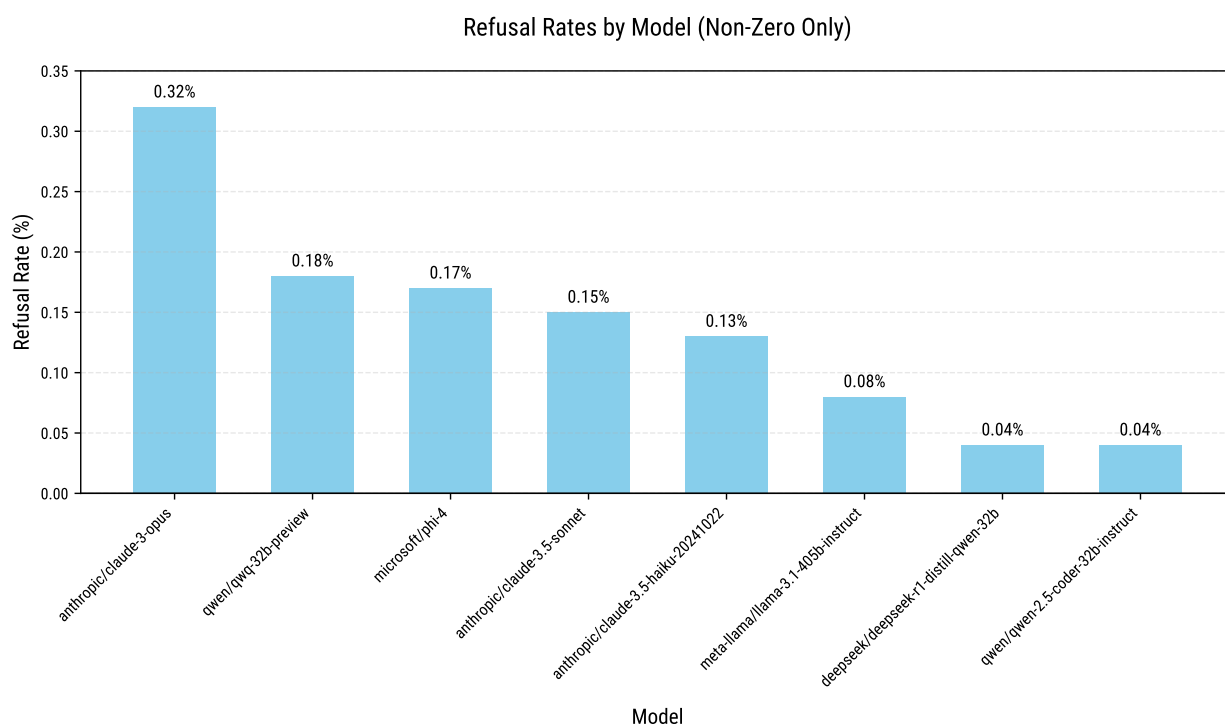


Figure S.7: Initial refusal rates per model (before fallback prompt).

Supplementary Note 6: CO₂ Emission Estimation

Recognizing the growing importance of assessing the environmental impact of large-scale computational research, we estimated the CO₂ emissions associated with running the LLMs in our LiveIdeaBench benchmark. We employed the “EcoLogits” tool [1] (<https://huggingface.co/spaces/genai-impact/ecologits-calculator>, accessed on March 30, 2025). This approach estimates emissions based on factors such as total model parameters, estimated activated parameters during inference, and generated token counts, making it more suitable for scenarios involving API calls. Parameter counts for proprietary models were estimated based on publicly available information, which introduces a degree of uncertainty to these specific values. Table S.4 presents the estimated energy consumption (kWh) and carbon emissions (kgCO₂eq.) for each model, broken down by its role as an idea generator and/or a critic, along with the total impact attributable to our evaluation process.

Table S.4: Energy Consumption and Carbon Emissions by Model

| Model Name | As Idea Model | | As Critic Model | | Total | |
|---------------------------------|---------------|-----------------------------------|-----------------|-----------------------------------|--------------|-----------------------------------|
| | Energy (kWh) | Emissions (kgCO ₂ eq.) | Energy (kWh) | Emissions (kgCO ₂ eq.) | Energy (kWh) | Emissions (kgCO ₂ eq.) |
| claude-3-opus* | 83.73 | 51.22 | - | - | 83.73 | 51.22 |
| claude-3.5-haiku-20241022* | 1.52 | 0.93 | - | - | 1.52 | 0.93 |
| claude-3.5-sonnet* | 12.17 | 7.43 | 226.20 | 138.05 | 238.37 | 145.48 |
| claude-3.7-sonnet* | 16.54 | 10.10 | - | - | 16.54 | 10.10 |
| claude-3.7-sonnet:thinking* | 271.66 | 165.79 | 1411.06 | 861.16 | 1682.72 | 1026.95 |
| deepseek-chat | 15.97 | 9.78 | 184.49 | 113.05 | 200.45 | 122.83 |
| deepseek-r1 | 77.18 | 47.29 | 670.38 | 410.79 | 747.56 | 458.08 |
| deepseek-r1-distill-llama-70b | 13.18 | 8.03 | - | - | 13.18 | 8.03 |
| deepseek-r1-distill-qwen-32b | 10.20 | 6.25 | - | - | 10.20 | 6.25 |
| gemini-2.0-flash-exp* | 2.07 | 1.27 | - | - | 2.07 | 1.27 |
| gemini-2.0-flash-lite-001* | 1.11 | 0.68 | - | - | 1.11 | 0.68 |
| gemini-2.0-flash-thinking-exp*† | 1.92 | 1.17 | 22.99 | 14.07 | 24.91 | 15.25 |
| gemini-2.0-pro-exp-02-05* | 14.84 | 9.10 | 168.11 | 103.01 | 182.95 | 112.11 |
| gemini-pro-1.5* | 9.73 | 5.96 | - | - | 9.73 | 5.96 |
| gemma-2-27b-it | 1.39 | 0.85 | - | - | 1.39 | 0.85 |
| gpt-4-turbo* | 1.64 | 1.01 | - | - | 1.64 | 1.01 |
| gpt-4.5-preview* | 83.52 | 51.09 | 1050.86 | 642.79 | 1134.37 | 693.88 |
| gpt-4o-2024-11-20* | 15.14 | 9.28 | - | - | 15.14 | 9.28 |
| gpt-4o-mini* | 0.97 | 0.60 | - | - | 0.97 | 0.60 |
| grok-2-1212* | 12.22 | 7.49 | - | - | 12.22 | 7.49 |
| llama-3.1-405b-instruct | 75.62 | 46.07 | - | - | 75.62 | 46.07 |
| llama-3.1-70b-instruct | 4.15 | 2.53 | - | - | 4.15 | 2.53 |
| llama-3.3-70b-instruct | 1.64 | 1.00 | - | - | 1.64 | 1.00 |
| mistral-large-2411 | 5.81 | 3.56 | - | - | 5.81 | 3.56 |
| mistral-small | 1.77 | 1.09 | - | - | 1.77 | 1.09 |
| mistral-small-24b-instruct-2501 | 2.10 | 1.29 | - | - | 2.10 | 1.29 |
| nova-lite-v1* | 0.84 | 0.52 | - | - | 0.84 | 0.52 |
| nova-pro-v1* | 2.49 | 1.52 | - | - | 2.49 | 1.52 |
| o1*† | 11.31 | 6.93 | - | - | 11.31 | 6.93 |
| o1-mini*† | 1.06 | 0.65 | - | - | 1.06 | 0.65 |
| o3-mini*† | 10.51 | 6.44 | - | - | 10.51 | 6.44 |
| o3-mini-high*† | 10.95 | 6.71 | 131.59 | 80.63 | 142.54 | 87.34 |
| phi-4 | 1.83 | 1.12 | - | - | 1.83 | 1.12 |
| qwen-2.5-72b-instruct | 3.62 | 2.22 | - | - | 3.62 | 2.22 |
| qwen-2.5-7b-instruct | 0.82 | 0.50 | - | - | 0.82 | 0.50 |
| qwen-2.5-coder-32b-instruct | 1.86 | 1.14 | - | - | 1.86 | 1.14 |
| qwen-max* | 7.87 | 4.82 | 163.20 | 100.00 | 171.07 | 104.83 |
| qwen2.5-dracarys2-72b | 3.33 | 2.04 | - | - | 3.33 | 2.04 |
| qwq-32b | 23.36 | 14.30 | 139.06 | 85.11 | 162.42 | 99.41 |
| qwq-32b-preview | 26.95 | 16.50 | - | - | 26.95 | 16.50 |
| step-2-16k-202411* | 14.93 | 9.15 | - | - | 14.93 | 9.15 |
| Total | 859.52 | 525.41 | 4167.92 | 2548.66 | 5027.45 | 3074.07 |

*Non-open models: environmental impact values are rough estimates and may not be accurate.

†Estimate may be low as the model’s reasoning is not accessible via API.

Supplementary Note 7: Human Expert Evaluation

To assess the validity and reliability of using LLMs as judges for evaluating scientific ideas within the **LiveIdeaBench** framework, we conducted a human expert evaluation study. Given the breadth of scientific domains covered by our benchmark, we focused this validation effort on a specific field, Partial Differential Equations (PDEs), where we could readily access domain expertise.

For this validation study, we utilized the set of 22 scientific ideas generated by 11 LLMs (i.e. `claude-3.7-sonnet:thinking`, `qwq-32b`, `claude-3.5-sonnet`, `gpt-4.5-preview`, `deepseek-r1-distill-llama-70b`, `qwen2.5-dracarys2-72b`, `o1`, `qwen-2.5-coder-32b-instruct`, `qwen-2.5-7b-instruct`, `claude-3.5-haiku-20241022`, and `nova-lite-v1`, sampled at equal intervals based on rankings (see Table S.3) from the 41 LLMs included in our benchmark, two ideas per model) specifically for the keyword “partial differential equations”. The full text of these 22 ideas is provided in Table S.5. These ideas were presented to 6 human experts in the PDE field via an online questionnaire. Experts were asked to rate each idea on the same three core dimensions used by our LLM judges (originality, feasibility, and clarity) using an identical 1-10 integer scale. The detailed ratings from both the LLM judge panel (averaged) and the individual human experts for each idea are compiled in Table S.6. To minimize bias, the 22 generated ideas were presented to the experts in a randomized order, and the identity of the LLM that generated each idea was withheld (blind evaluation). A screenshot illustrating the questionnaire interface presented to the experts is shown in Figure S.8.

Inter-rater reliability (IRR) among the 6 human experts was assessed for each dimension using the Intraclass Correlation Coefficient (ICC) based on a two-way random-effects model evaluating absolute agreement for the average rating ($ICC(A,k)$). Interpretation guidelines from Koo & Li [2] were used ($ICC < 0.5$ Poor, $0.5 \leq ICC \leq 0.75$ Moderate, $0.75 < ICC \leq 0.9$ Good, $ICC > 0.9$ Excellent). The analysis revealed good agreement for originality ($ICC = 0.823$) and clarity ($ICC = 0.782$). However, only poor agreement was found for feasibility ($ICC = 0.453$), indicating substantial disagreement among experts on the technical viability and scientific soundness of the generated ideas. This low inter-expert agreement underscores the significant subjectivity and inherent difficulty in evaluating the feasibility of novel scientific ideas based solely on brief textual descriptions, even for domain specialists.

The subsequent figures in this section present a comparative analysis between the average ratings from the 6 human experts and the average ratings assigned by the LLM judge panel for these 22 PDE-related ideas. Specifically for this validation analysis, to obtain a stable estimate of the LLM panel’s collective assessment for comparison against the human expert average, the LLM rating for each idea represents the average score across all eligible critic models in the top-10 judges. This approach differs from the main benchmark’s scoring procedure, which uses a random sample of 3 judges per idea, and was chosen here to minimize sampling noise in the validation comparison. Figure S.9 visualizes the correlation between these averaged human and averaged LLM ratings for each dimension using scatter plots and reports the Pearson correlation coefficient. The analysis shows statistically significant positive correlations for all dimensions: a strong correlation for originality, a moderate correlation for feasibility, and a weaker but still significant correlation for clarity (The superscripts denote the level of statistical significance based on the p-value. The p-value represents the probability of observing the given correlation (or a stronger one) if there were actually no correlation between the ratings in the broader population. Lower p-values indicate stronger evidence against this possibility (i.e., stronger evidence that the correlation is real).

The strong human IRR for originality ($ICC = 0.823$) provides a solid foundation for the observed high human-LLM correlation, suggesting that the LLM judges effectively captured expert consensus on novelty within this specific domain. Similarly, the good human IRR for clarity ($ICC = 0.782$)

supports the finding of a significant, albeit weaker, human-LLM correlation. Critically, the poor human IRR for feasibility ($ICC = 0.453$) complicates the interpretation of the moderate human-LLM correlation. The lack of consensus among human experts on feasibility implies that the average human rating for this dimension is inherently noisy and less reliable as a benchmark. Therefore, the moderate correlation observed likely reflects a combination of the LLMs' capabilities and the significant variability within the human expert judgments themselves. This underscores the challenge of reliably evaluating scientific feasibility based solely on brief idea descriptions, a limitation potentially affecting both human and LLM assessments.

Overall, the validation study indicates reasonable alignment between human expert judgment and the automated LLM evaluations, lending empirical support to the LLM-as-a-judge methodology employed in **LiveIdeaBench**, particularly for assessing originality and clarity. However, the findings regarding feasibility highlight the inherent challenges and necessitate caution when interpreting the benchmark's feasibility scores.

Human Experts' Evaluation of LLM-Generated PDE Ideas

This survey evaluates creative ideas in Partial Differential Equations generated by Large Language Models. We're studying if AI evaluations align with expert human judgment.

As an expert in PDEs, your assessment is invaluable. Please assess these ideas on three dimensions: originality, feasibility and clarity. (22 ideas total.)

Thank you for your insights.

01 **Keyword: Partial Differential Equations** **Idea:* Develop a novel numerical method for solving high-dimensional partial differential equations (PDEs) that leverages machine learning to adaptively refine mesh resolution. This approach would significantly reduce computational costs while maintaining accuracy. By integrating neural networks to predict regions requiring higher resolution, the method could efficiently handle complex, multi-dimensional problems in fluid dynamics, materials science, and quantum mechanics. This innovation promises to overcome current limitations in scalability and efficiency of traditional numerical methods.

Please assess this scientific idea on three dimensions:

Originality (Novel contribution to unexplored areas or innovative approaches to existing problems)

1 2 3 4 5 6 7 8 9 10

Feasibility (Technical implementation and practicality)

1 2 3 4 5 6 7 8 9 10

Clarity (How well-articulated and easy to understand the idea is)

1 2 3 4 5 6 7 8 9 10

Figure S.8: Screenshot of the questionnaire interface presented to human experts for rating LLM-generated ideas related to Partial Differential Equations.

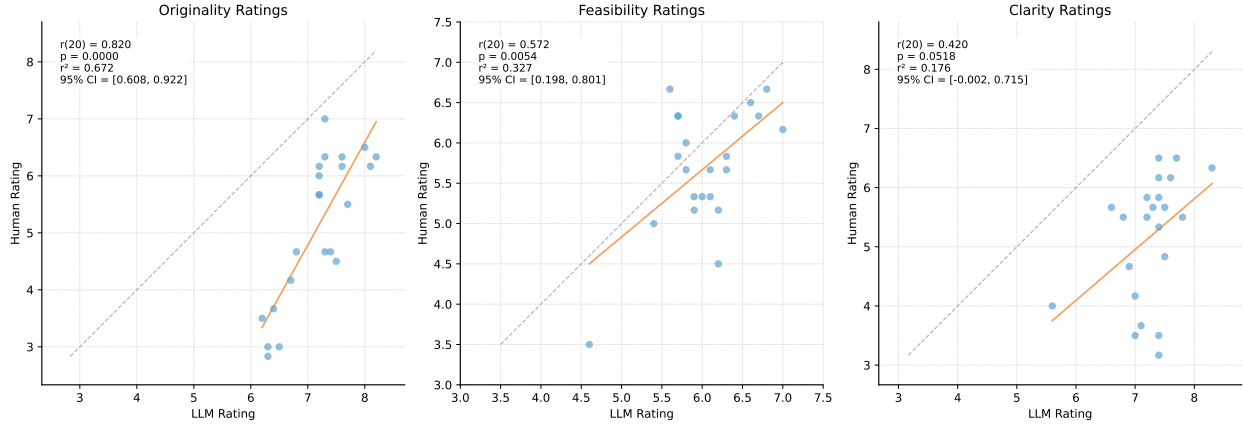


Figure S.9: Scatter plot comparing Human Expert (average of 6 experts) and LLM Judge (average of 10 LLM judges) ratings for originality, feasibility, and clarity on 22 PDE-related ideas.

Table S.5: LLM-generated scientific ideas on Partial Differential Equations used in the validation study. The table lists the idea model and the 22 generated scientific idea for the related to “partial differential equations” evaluated in the human expert validation analysis

| ID | Idea Model | Idea |
|----|---------------------------|---|
| 1 | nova-lite-v1 | **Keyword: Partial Differential Equations** **Idea:** Develop a novel numerical method for solving high-dimensional partial differential equations (PDEs) that leverages machine learning to adaptively refine mesh resolution. This approach would significantly reduce computational costs while maintaining accuracy. By integrating neural networks to predict regions requiring higher resolution, the method could efficiently handle complex, multi-dimensional problems in fluid dynamics, materials science, and quantum mechanics. This innovation promises to overcome current limitations in scalability and efficiency of traditional numerical methods. |
| 2 | nova-lite-v1 | **Keyword: Partial Differential Equations** **Idea:** Develop a novel numerical method for solving high-dimensional partial differential equations (PDEs) that leverages machine learning to adaptively refine mesh resolution. This approach would significantly reduce computational costs while maintaining accuracy. By integrating neural networks to predict regions requiring higher resolution, the method would be particularly useful for complex systems in fluid dynamics and materials science. This innovative technique promises to enhance both efficiency and precision in solving PDEs. |
| 3 | claude-3.5-haiku-20241022 | Here’s a novel approach for solving complex partial differential equations (PDEs) using quantum-inspired neural networks: Background: Traditional numerical methods struggle with high-dimensional, nonlinear PDEs across physics and engineering domains. Idea: Develop a hybrid quantum-classical neural architecture that leverages quantum probability amplitude mappings to approximate PDE solutions. By encoding boundary conditions as quantum state constraints and utilizing quantum-enhanced feature extraction, the network can dramatically reduce computational complexity and improve convergence for systems like turbulent fluid dynamics, quantum mechanics wave equations, and climate modeling partial differential systems. The approach combines machine learning’s adaptability with quantum computational principles. |

(Continued on next page)

Table S.5: Continued Generated Ideas

| ID | Idea Model | Idea |
|----|---------------------------|---|
| 4 | claude-3.5-haiku-20241022 | Here's a submission focused on partial differential equations (PDEs): Novel PDE Machine Learning Predictive Framework Background: Traditional PDE solving methods struggle with complex, nonlinear systems across physics and engineering. Proposed Idea: Develop a hybrid neural network architecture combining physics-informed neural networks (PINNs) with adaptive mesh refinement techniques. This approach would dynamically adjust computational resolution based on solution gradient complexity, enabling more accurate and computationally efficient numerical solutions for multiscale PDEs in fluid dynamics, quantum mechanics, and materials science. The framework would integrate residual learning, constraint enforcement, and intelligent mesh adaptation to dramatically improve predictive modeling of high-dimensional, nonlinear systems. |
| 5 | claude-3.5-sonnet | Background: Partial differential equations (PDEs) are crucial in modeling physical phenomena but solving them numerically can be computationally intensive and sometimes unstable. Idea: Develop a hybrid solver that combines traditional numerical methods with machine learning to adaptively switch between different solution strategies. The system would analyze the local behavior of the PDE in real-time and automatically select the most efficient and stable method (finite elements, spectral methods, or neural network approximations) for that specific region. This would significantly reduce computational costs while maintaining accuracy, particularly for PDEs with mixed-type behavior or discontinuities. |
| 6 | claude-3.5-sonnet | Background: Partial differential equations (PDEs) are crucial in modeling physical phenomena but solving them numerically can be computationally intensive, especially for complex geometries. Idea: Develop a hybrid solver that combines machine learning with traditional numerical methods by training neural networks to identify regions where high-resolution calculations are necessary versus where simplified approximations suffice. The neural network would adaptively adjust the mesh resolution and computational approach in real-time based on the solution's local complexity, significantly reducing computational costs while maintaining accuracy in critical regions where precise PDE solutions are essential. |
| 7 | claude-3.7-sonnet | # Multi-Resolution Graph Neural Operators for Heterogeneous PDEs
Background: PDEs in heterogeneous media with multiple spatial scales remain computationally challenging despite decades of research.
Idea: We propose adaptive graph neural operators that learn multi-resolution representations of PDE solutions. The approach automatically identifies regions requiring different resolution levels, constructing hierarchical graph representations that efficiently capture both fine-scale features and global behaviors. This enables fast, accurate predictions for complex multi-scale systems like subsurface flow and biomechanics while reducing computational costs by orders of magnitude compared to traditional solvers. The method preserves important physical invariants through specialized graph convolutions. |

(Continued on next page)

Table S.5: Continued Generated Ideas

| ID | Idea Model | Idea |
|----|-------------------------------|--|
| 8 | claude-3.7-sonnet | # Wavelet-Guided Adaptive Physics-Informed Neural Networks for Multiscale PDEs Physics-informed neural networks (PINNs) have emerged as powerful PDE solvers, but struggle with multiscale phenomena and sharp transitions that characterize many real-world problems. I propose Wavelet-Guided Adaptive PINNs that dynamically allocate computational resources based on local solution complexity. By integrating wavelet analysis to identify regions requiring higher resolution, the network automatically adjusts its architecture-increasing depth or width where needed. This approach significantly improves efficiency and accuracy for problems with boundary layers and multiscale features in fluid dynamics, materials science, and climate modeling, addressing a fundamental limitation in existing neural PDE solvers. |
| 9 | deepseek-r1-distill-llama-70b | **Idea:** Develop a deep learning framework to solve partial differential equations (PDEs) using neural networks that eliminate the need for traditional mesh generation. This approach would train neural networks to approximate PDE solutions directly from input parameters (e.g., space, time, coefficients), reducing computational overhead and enabling real-time solutions for complex, high-dimensional problems. **Background:** PDEs are fundamental in modeling physical phenomena, but numerical solutions often require computationally expensive mesh discretization. This innovation would democratize PDE solving by enabling fast, mesh-free solutions on standard computing hardware. |
| 10 | deepseek-r1-distill-llama-70b | **Idea:** Developing a hybrid framework combining physics-informed neural networks (PINNs) with adaptive mesh refinement (AMR) to solve partial differential equations (PDEs). This approach integrates the flexibility of neural networks in capturing complex solutions with AMR's ability to dynamically allocate computational resources to regions of interest. The hybrid method will improve accuracy and efficiency for multiscale PDEs, such as those in fluid dynamics or materials science, by automatically refining meshes where neural network errors are high. This innovative approach addresses the limitations of traditional finite element or finite difference methods, enabling more precise and computationally efficient simulations. |
| 11 | gpt-4.5-preview | Background: Approximating solutions of nonlinear partial differential equations (PDEs) commonly relies on computationally expensive numerical methods. Current machine learning approaches, like neural PDE solvers, provide speedup but lack accuracy guarantees of traditional numerical schemes. Idea: A hybrid method integrating neural operator networks with adaptive finite element error correction is proposed. The neural component rapidly predicts solutions, while an iterative residual-based finite-element refinement ensures accuracy within specified tolerances. This architecture leverages machine learning speed with traditional PDE solver precision, significantly reducing computational cost while providing mathematical reliability in solving nonlinear PDE problems relevant across physics and engineering disciplines. |
| 12 | gpt-4.5-preview | Background: Solving high-dimensional partial differential equations (PDEs), central to modeling complex systems such as financial markets, fluid dynamics, and quantum mechanics, suffers from prohibitively high computational demands (curse of dimensionality). Idea: Develop a hybrid deep-learning spectral decomposition method that approximates PDE solutions by dynamically partitioning high-dimensional spaces into data-driven low-dimensional spectral subspaces. Combining deep neural networks and adaptive spectral decomposition, this approach significantly reduces computational cost and storage complexity, enables real-time solutions to previously intractable PDE problems, and provides interpretable spectral insights into underlying physical phenomena while maintaining predictive accuracy and scalability. |

(Continued on next page)

Table S.5: Continued Generated Ideas

| ID | Idea Model | Idea |
|----|-----------------------------|--|
| 13 | o1 | Partial differential equations (PDEs) underlie many scientific fields. IDEA: Use a microservices-based domain decomposition for PDE solvers, enabling parallel subproblem computations. Each subdomain is handled by specialized microservices that adapt mesh resolution in real time. This fosters scalable high-performance computing (HPC) performance with improved load balancing and quicker solutions. Feasible with containerization technologies, it addresses fluid dynamics, structural analysis, and beyond. |
| 14 | o1 | We propose a novel partial differential equations-based model for real-time wildfire spread forecasting. The approach couples Navier-Stokes partial differential equations for wind flow with a reaction-diffusion process to capture fire propagation. By combining GPU-accelerated domain decomposition with adaptive time-stepping, it can account for heterogeneous terrain and evolving fuel conditions, enabling predictive wildfire simulations. This method could guide resource allocation and evacuation planning. |
| 15 | qwen-2.5-7b-instruct | Partial differential equations (PDEs) are essential in modeling complex systems across various disciplines. My proposed work focuses on developing a novel numerical method to solve high-dimensional PDEs more efficiently. Specifically, we will leverage machine learning techniques to approximate the solution, reducing computational complexity through dimensionality reduction and adaptive sampling. This approach addresses the curse of dimensionality, a significant challenge in PDEs for high-dimensional problems, enabling more accurate and faster simulations in fields like finance, physics, and engineering. |
| 16 | qwen-2.5-7b-instruct | Partial differential equations (PDEs) are fundamental in modeling complex systems in physics, engineering, and finance. My idea is to develop a novel numerical method that integrates machine learning with traditional PDE solvers. This hybrid approach aims to significantly reduce computational time and enhance accuracy in solving high-dimensional PDEs, which are currently infeasible with existing methods. This innovation addresses a critical gap in solving complex, real-world problems involving multiple interacting variables. |
| 17 | qwen-2.5-coder-32b-instruct | Partial differential equations (PDEs) are fundamental in modeling complex systems across physics, engineering, and biology. They describe phenomena involving space and time dependencies but can be challenging to solve analytically. My idea is to develop a novel, adaptive machine learning algorithm tailored for solving high-dimensional PDEs with varying boundary conditions. This approach will leverage neural networks to approximate solutions efficiently, addressing the curse of dimensionality and enabling real-time simulations. This innovation would significantly enhance predictive modeling capabilities in fields like fluid dynamics and climate science. |
| 18 | qwen-2.5-coder-32b-instruct | Partial differential equations (PDEs) are fundamental in modeling complex systems in physics, engineering, and biology. A novel idea is to develop a machine learning framework that automatically learns to solve high-dimensional PDEs using neural networks, addressing the "curse of dimensionality." This approach, combining deep learning with traditional numerical methods, could lead to more efficient and accurate solutions for real-world problems, such as fluid dynamics and climate modeling. The feasibility lies in recent advancements in neural network architectures, and the approach would significantly enhance computational capabilities in solving complex PDEs. |

(Continued on next page)

Table S.5: Continued Generated Ideas

| ID | Idea Model | Idea |
|----|-----------------------|---|
| 19 | qwq-32b | <p>**Background and Idea:** High-dimensional PDEs hinder progress in quantum physics and finance due to the curse of dimensionality. Propose a neural-tensor hybrid framework where physics-aware neural networks parametrize solutions as low-rank tensor networks optimized via PDE-constrained loss functions. This combines neural adaptability for nonlinear dynamics with tensor compression for scalability, enabling high-accuracy solutions in >10 dimensions. Feasible via auto-differentiation tools and tensor libraries, this approach addresses intractable computations while ensuring physical consistency, driving breakthroughs in many-body systems and stochastic modeling. (Word count: 94) **Why It Works:** - **Originality:** Blends neural networks with low-rank tensor decompositions for PDEs, addressing a longstanding scalability issue uniquely. - **Feasibility:** Leverages existing ML frameworks (e.g., TensorFlow/PyTorch) and tensor libraries (e.g., TensorLy), requiring no exotic hardware. - **Impact:** Enables solutions to critical high-dimensional PDEs in fields like quantum chemistry or financial risk modeling, bypassing computational bottlenecks. - **Clarity:** Articulates a clear, actionable method (physics-informed networks × tensor compression).</p> |
| 20 | qwq-32b | <p>**Background:** Partial differential equations (PDEs) are central to modeling physical systems but struggle with computational inefficiency and scalability in high dimensions. Existing methods often trade accuracy for speed or require prohibitive resources. **Idea:** Develop an AI-driven adaptive PDE solver integrating physics-informed neural networks (PINNs) with error-gradient-guided mesh adaptation. The solver dynamically refines computational grids in regions of high solution variability, trained in tandem with PINNs using gradient-based optimization. This hybrid approach slashes computational costs for high-dimensional problems (e.g., climate modeling, quantum chemistry) while preserving accuracy, addressing a critical bottleneck in scientific simulations. (100 words)</p> |
| 21 | qwen2.5-dracarys2-72b | <p>**Background**: Partial differential equations (PDEs) are fundamental in modeling physical phenomena but often pose significant computational challenges. **Idea**: Develop a hybrid method combining deep learning with traditional numerical techniques to solve high-dimensional PDEs more efficiently. This approach leverages neural networks to approximate solutions in regions where data is sparse, while using classical methods like finite differences or spectral methods where they excel. The integration ensures accuracy and stability, potentially revolutionizing fields such as fluid dynamics, quantum mechanics, and financial modeling by reducing computational time and resource requirements.</p> |
| 22 | qwen2.5-dracarys2-72b | <p>**Background:** Partial differential equations (PDEs) are fundamental in modeling complex systems across physics, engineering, and finance. However, solving PDEs for real-world applications often requires significant computational resources and time. **Idea:** Develop a machine learning framework that integrates with existing numerical solvers to predict solutions of PDEs more efficiently. By training neural networks on a diverse dataset of PDE solutions, this hybrid approach can reduce computation time while maintaining accuracy. This method could be particularly useful for real-time simulations and scenarios where rapid solution updates are crucial, such as climate modeling and financial market analysis.</p> |

Table S.6: Comparison of LLM judge and human expert ratings for PDE-related scientific ideas.
Averaged ratings from the LLM judge panel and individual ratings from 6 human experts are presented for originality (O), feasibility (F), and clarity (C) scores assigned to 22 scientific ideas generated for the keyword “partial differential equations”.

| ID | LLMs (Avg.) | | | Expert 1 | | | Expert 2 | | | Expert 3 | | | Expert 4 | | | Expert 5 | | | Expert 6 | | |
|----|-------------|-----|-----|----------|---|----|----------|---|---|----------|---|---|----------|---|---|----------|---|---|----------|---|---|
| | O | F | C | O | F | C | O | F | C | O | F | C | O | F | C | O | F | C | O | F | C |
| 1 | 7.4 | 5.8 | 7.8 | 3 | 8 | 10 | 7 | 4 | 7 | 5 | 6 | 7 | 4 | 4 | 1 | 6 | 7 | 6 | 3 | 5 | 2 |
| 2 | 7.3 | 6.1 | 7.3 | 7 | 6 | 7 | 7 | 4 | 7 | 6 | 5 | 5 | 6 | 7 | 7 | 8 | 7 | 5 | 4 | 5 | 3 |
| 3 | 7.3 | 4.6 | 5.6 | 8 | 5 | 5 | 7 | 3 | 4 | 6 | 3 | 4 | 7 | 3 | 4 | 8 | 4 | 5 | 6 | 3 | 2 |
| 4 | 7.2 | 6.4 | 6.6 | 9 | 7 | 8 | 7 | 9 | 7 | 5 | 5 | 6 | 4 | 5 | 4 | 6 | 7 | 5 | 3 | 5 | 4 |
| 5 | 7.5 | 5.4 | 7.4 | 7 | 2 | 10 | 5 | 5 | 8 | 3 | 4 | 5 | 3 | 6 | 6 | 5 | 7 | 5 | 4 | 6 | 5 |
| 6 | 7.3 | 5.6 | 7.5 | 7 | 6 | 7 | 4 | 8 | 7 | 5 | 6 | 6 | 3 | 7 | 4 | 6 | 8 | 6 | 3 | 5 | 4 |
| 7 | 8.0 | 6.2 | 7.2 | 8 | 2 | 8 | 6 | 6 | 7 | 6 | 4 | 6 | 6 | 3 | 5 | 8 | 7 | 7 | 5 | 5 | 2 |
| 8 | 8.1 | 6.2 | 7.6 | 9 | 3 | 8 | 7 | 4 | 4 | 6 | 6 | 7 | 5 | 5 | 6 | 7 | 8 | 7 | 3 | 5 | 5 |
| 9 | 6.7 | 5.7 | 7.5 | 6 | 8 | 7 | 3 | 7 | 5 | 4 | 4 | 6 | 4 | 4 | 4 | 5 | 7 | 5 | 3 | 5 | 2 |
| 10 | 7.6 | 6.6 | 8.3 | 7 | 6 | 7 | 8 | 8 | 7 | 6 | 6 | 6 | 4 | 6 | 6 | 8 | 7 | 6 | 5 | 6 | 6 |
| 11 | 7.6 | 6.8 | 7.7 | 9 | 8 | 8 | 6 | 4 | 7 | 7 | 7 | 7 | 3 | 8 | 5 | 8 | 8 | 6 | 4 | 5 | 6 |
| 12 | 7.7 | 5.9 | 6.9 | 7 | 7 | 7 | 4 | 4 | 3 | 6 | 4 | 5 | 6 | 4 | 4 | 6 | 7 | 5 | 4 | 6 | 4 |
| 13 | 7.2 | 5.7 | 7.2 | 8 | 6 | 7 | 3 | 9 | 7 | 6 | 5 | 6 | 4 | 5 | 4 | 8 | 7 | 6 | 7 | 6 | 3 |
| 14 | 7.2 | 5.8 | 7.4 | 8 | 6 | 6 | 4 | 8 | 8 | 7 | 6 | 7 | 7 | 7 | 5 | 6 | 6 | 7 | 5 | 3 | 4 |
| 15 | 6.5 | 6.3 | 7.1 | 2 | 7 | 5 | 4 | 8 | 3 | 3 | 4 | 3 | 3 | 3 | 3 | 3 | 8 | 6 | 3 | 5 | 2 |
| 16 | 6.3 | 6.1 | 7.4 | 3 | 7 | 8 | 4 | 5 | 3 | 4 | 4 | 3 | 1 | 6 | 3 | 3 | 6 | 2 | 3 | 4 | 2 |
| 17 | 6.2 | 5.9 | 7.0 | 2 | 7 | 5 | 4 | 4 | 4 | 3 | 3 | 4 | 3 | 6 | 3 | 6 | 7 | 3 | 3 | 4 | 2 |
| 18 | 6.3 | 6.3 | 7.4 | 1 | 8 | 4 | 3 | 4 | 5 | 3 | 4 | 3 | 4 | 5 | 1 | 3 | 8 | 4 | 3 | 5 | 2 |
| 19 | 8.2 | 7.0 | 7.4 | 8 | 6 | 7 | 9 | 8 | 7 | 7 | 7 | 7 | 4 | 6 | 4 | 5 | 4 | 3 | 5 | 6 | 7 |
| 20 | 7.2 | 5.7 | 7.4 | 8 | 7 | 7 | 7 | 8 | 7 | 5 | 6 | 6 | 5 | 5 | 4 | 5 | 7 | 5 | 4 | 5 | 3 |
| 21 | 6.8 | 6.0 | 6.8 | 7 | 4 | 7 | 4 | 8 | 8 | 5 | 4 | 6 | 3 | 6 | 5 | 6 | 5 | 5 | 3 | 5 | 2 |
| 22 | 6.4 | 6.7 | 7.0 | 2 | 7 | 4 | 8 | 7 | 7 | 3 | 4 | 3 | 3 | 7 | 3 | 3 | 8 | 5 | 3 | 5 | 3 |

Supplementary Note 8: Reasoning Length vs. Idea Quality Metrics

The main text established minimal correlation between the length of the final generated idea and its quality. However, some models utilize explicit reasoning steps (e.g., “**thinking**” variants), resulting in longer full responses (thoughts and final idea). This section examines whether this full response length correlates with the final extracted idea’s quality for these models.

The following hexbin plots visualize this relationship for selected reasoning-centric models. Each plot displays the density of ideas based on the full response length versus the idea quality score.

As the Figures S.10–S.20 show, even for these models, there is no strong correlation between the length of the full reasoning output and the quality of the final idea. This finding reinforces the conclusion that idea quality is not merely a function of reasoning length or efforts, distinguishing idea generation capability from some aspects of general reasoning performance.

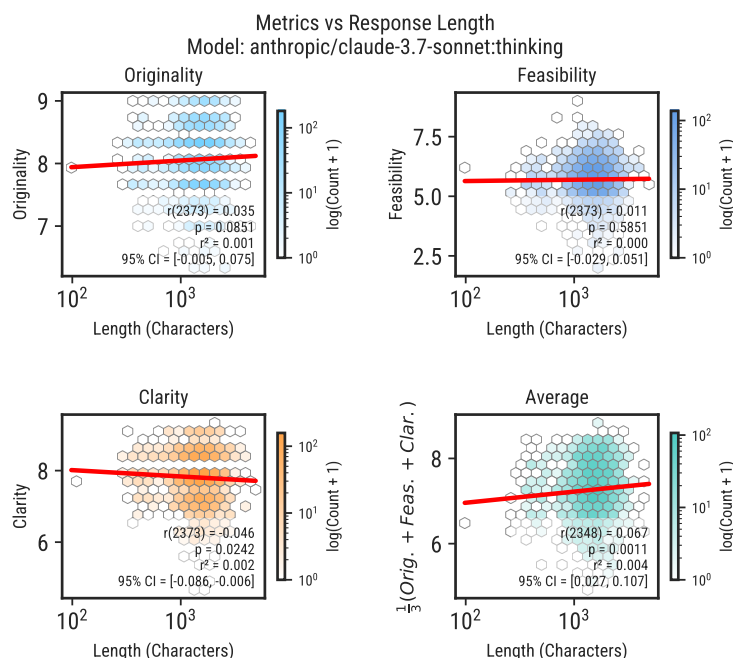


Figure S.10: Relationship between Full Response Length and Idea Quality for claude-3.7-sonnet:thinking. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

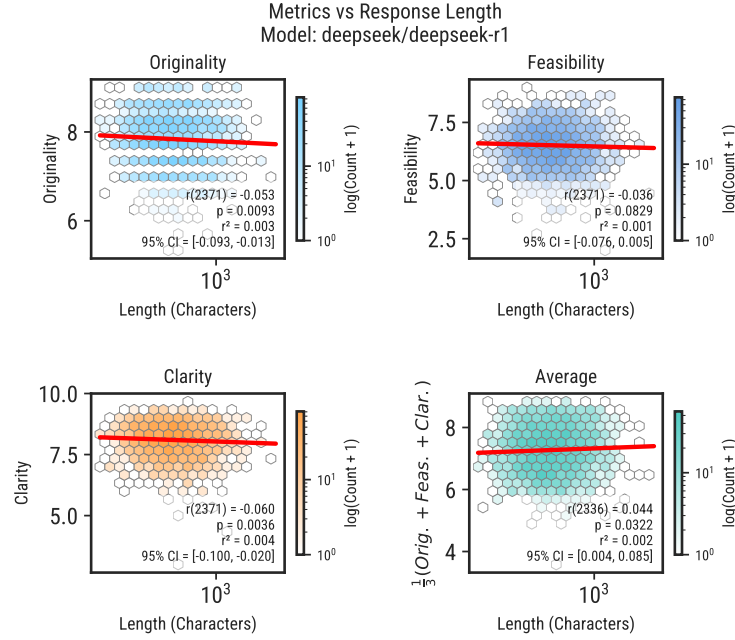


Figure S.11: Relationship between Full Response Length and Idea Quality for deepseek-r1. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

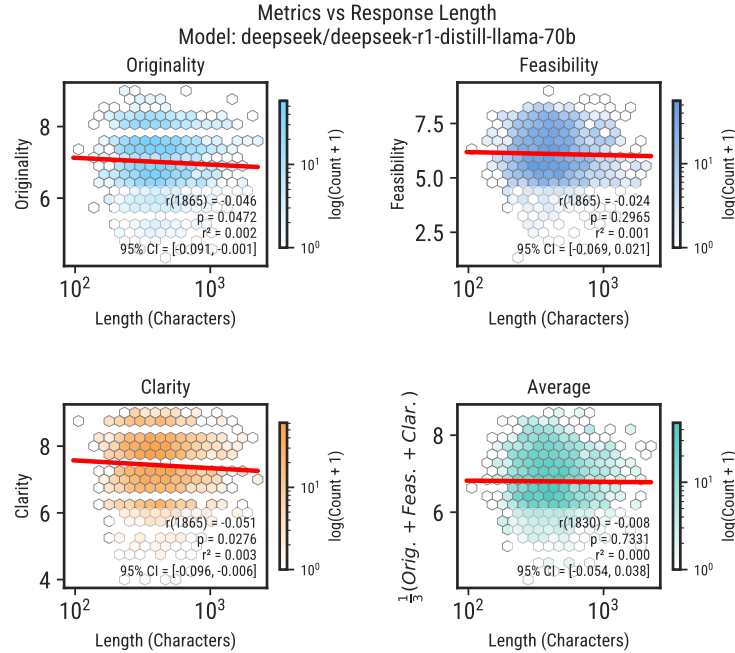


Figure S.12: Relationship between Full Response Length and Idea Quality for deepseek-r1-distill-llama-70b. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

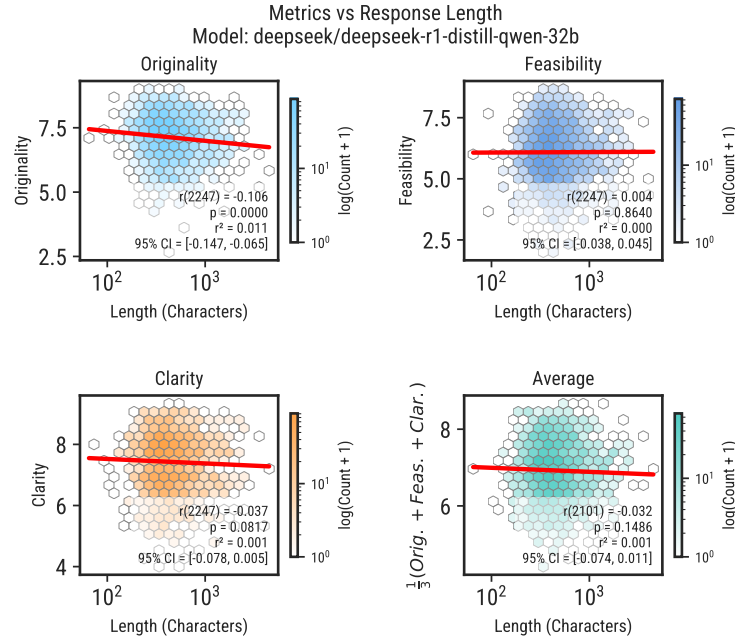


Figure S.13: Relationship between Full Response Length and Idea Quality for deepseek-r1-distill-qwen-32b. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

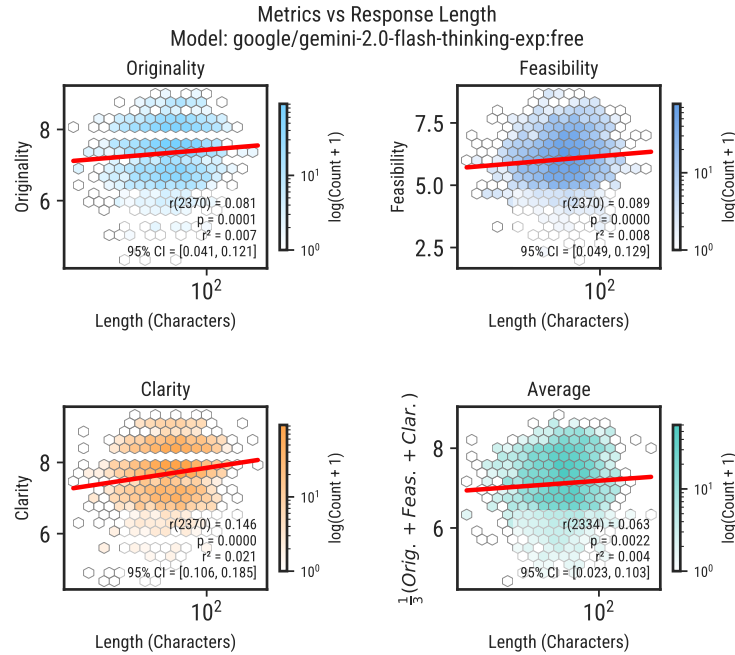


Figure S.14: Relationship between Full Response Length and Idea Quality for gemini-2.0-flash-thinking-exp. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

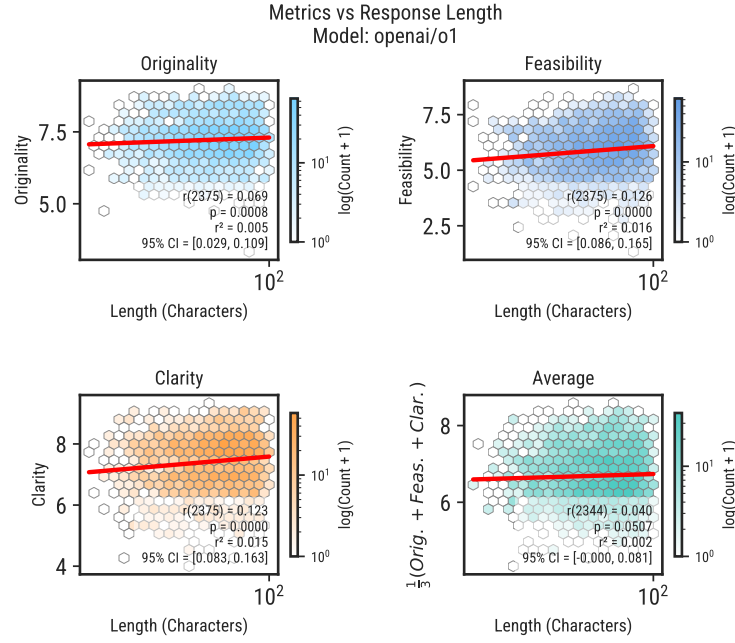


Figure S.15: Relationship between Full Response Length and Idea Quality for o1. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

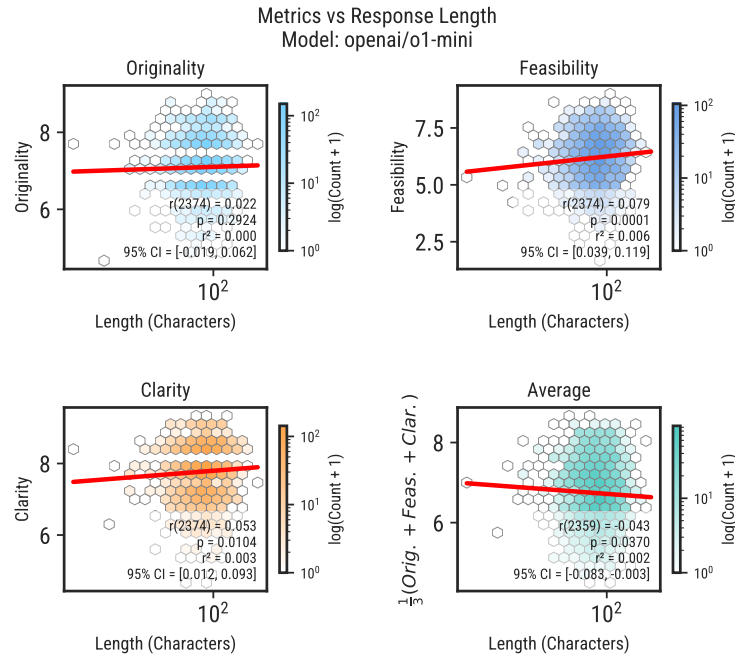


Figure S.16: Relationship between Full Response Length and Idea Quality for o1-mini. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

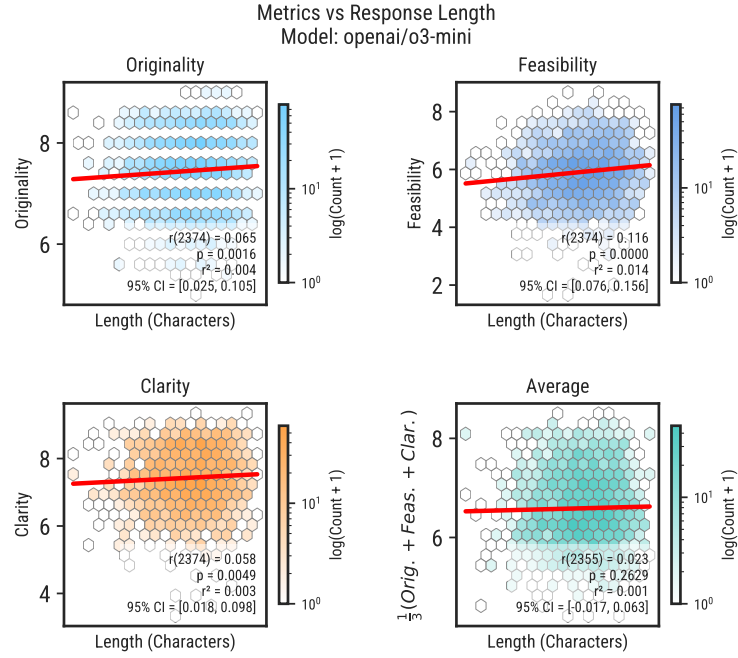


Figure S.17: Relationship between Full Response Length and Idea Quality for o3-mini. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

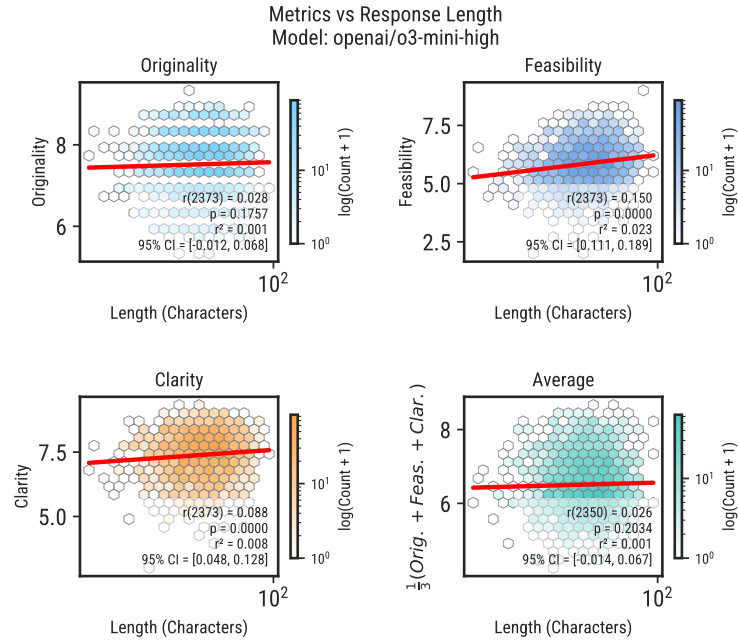


Figure S.18: Relationship between Full Response Length and Idea Quality for o3-mini-high. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

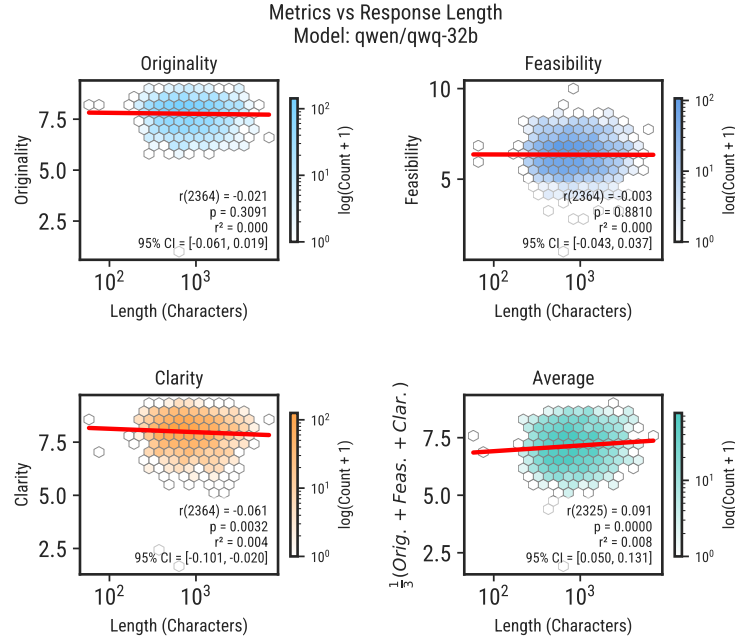


Figure S.19: Relationship between Full Response Length and Idea Quality for qwq-32b. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

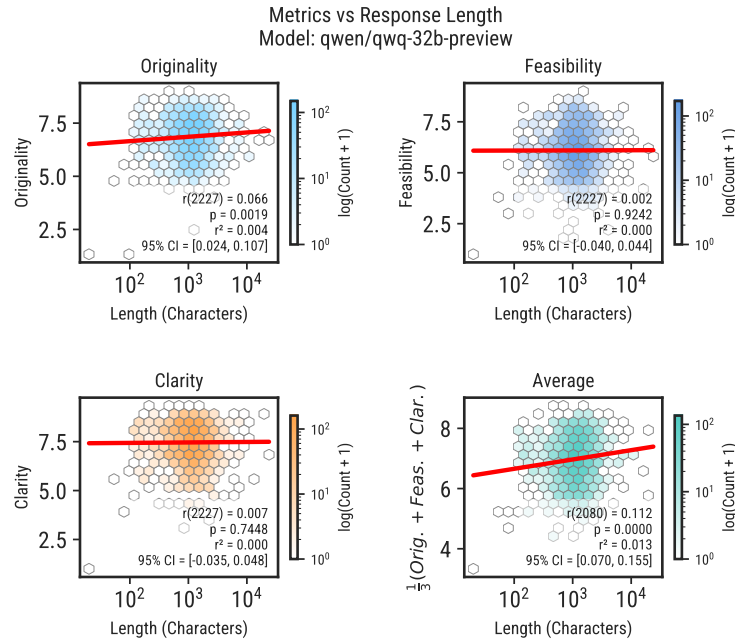


Figure S.20: Relationship between Full Response Length and Idea Quality for qwq-32b-preview. Hexbin plot showing the density distribution of generated ideas based on the full response length (characters, x-axis) versus the idea quality score (y-axis). The color intensity represents the number of ideas in each bin.

Supplementary Note 9: Statistical Correlation Metrics

In this study, we employ two standard statistical measures to evaluate the relationships between different evaluation metrics: the Pearson correlation coefficient (r) and Spearman's rank correlation coefficient (ρ). The choice of metric depends on the distribution of the underlying data, which was assessed using the Shapiro-Wilk test for normality.

Pearson Correlation Coefficient (r) Pearson's r is a parametric measure of the linear correlation between two sets of continuous data. It is utilized when the data distributions do not significantly deviate from normality (for instance, when evaluating the correlation between general intelligence scores and scientific ideation capabilities). For two variables X and Y with n paired observations (x_i, y_i) , the coefficient is calculated as the covariance of the two variables divided by the product of their standard deviations:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ represent the sample means of X and Y , respectively. The value of r ranges from -1 to 1 , where 1 indicates a perfect positive linear correlation, 0 indicates no linear correlation, and -1 indicates a perfect negative linear correlation.

Spearman's Rank Correlation Coefficient (ρ) Spearman's ρ is a non-parametric measure of rank correlation. It assesses how well the relationship between two variables can be described using a monotonic function, making it robust to outliers and suitable for non-normally distributed data.

To calculate ρ , the raw scores X_i and Y_i are first converted to ranks $\text{rg}(X_i)$ and $\text{rg}(Y_i)$. The coefficient is then computed as the Pearson correlation between these rank variables. In cases where all n ranks are distinct integers (no tied ranks), the formula simplifies to:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where $d_i = \text{rg}(X_i) - \text{rg}(Y_i)$ is the difference between the two ranks of each observation, and n is the total number of observations. Similar to Pearson's r , Spearman's ρ yields a value between -1 and 1 .

Supplementary References

- [1] Samuel Rince and Adrien Banse. EcoLogits: Evaluate the environmental impacts of generative AI, April 2025.
- [2] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine, 15(2):155–163, 2016.