AidanBench: Evaluating Novel Idea Generation on Open-Ended Questions

Aidan McLaughlin*
Topology Corporation

James Campbell*
Carnegie Mellon University

Anuja Uppuluri*
Carnegie Mellon University

Abstract

AidanBench evaluates large language models (LLMs) on their ability to generate novel ideas in response to open-ended questions, focusing on creativity, reliability, contextual attention, and instruction following. Unlike benchmarks with clear-cut answers, AidanBench assesses models in more open-ended, real-world tasks. Testing several state-of-the-art LLMs, it shows weak correlation with existing benchmarks while offering a more nuanced view of their performance in open-ended scenarios.

1 Introduction

Large Language Models (LLMs) have shown impressive performance across benchmarks in software engineering [Jimenez et al., 2024], mathematics [Hendrycks et al., 2021b], science [Rein et al., 2023], and general knowledge [Hendrycks et al., 2021a]. However, these benchmarks primarily focus on tasks with well-defined answers, often missing a model's ability to perform in open-ended scenarios. LLMs may also collapse into specific attractor 'modes,' repeating responses instead of generating novel ideas [Janus, 2022].

In practice, users often rely on LLMs for tasks requiring creativity, such as brainstorming, writing assistance, and problem-solving, where originality and flexibility are crucial [Zhang et al., 2024]. A model's ability to generate diverse, non-repetitive ideas expands its utility in such tasks.

To evaluate this, we introduce AidanBench, which poses open-ended questions and requires models to generate as many promising ideas as possible while:

1. avoiding repetition (judged by an embedding model), and
2. maintaining coherence and plausibility (assessed by a separate LLM).

AidanBench assesses creativity, reliability, contextual attention, and instruction-following.

We describe AidanBench's design and evaluate several LLMs, comparing AidanBench scores to LMSYS scores, which reveal a weak correlation with standard benchmarks.

2 Related Work

Many benchmarks have been designed to test models on tasks with well-defined answers [Hendrycks et al., 2021a,b, Jimenez et al., 2024, Rein et al., 2023]. While such standards have facilitated AI

*Equal contribution. Code available at `github.com/aidanmclaughlin/Aidan-Bench`.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

progress in domains like mathematics, software engineering, and question answering, they often fall short in measuring more subtle aspects of performance, such as open-ended tasks without clear-cut answers.

Some work has been done on LLM creativity, which we detail in Appendix C.

3 Methodology

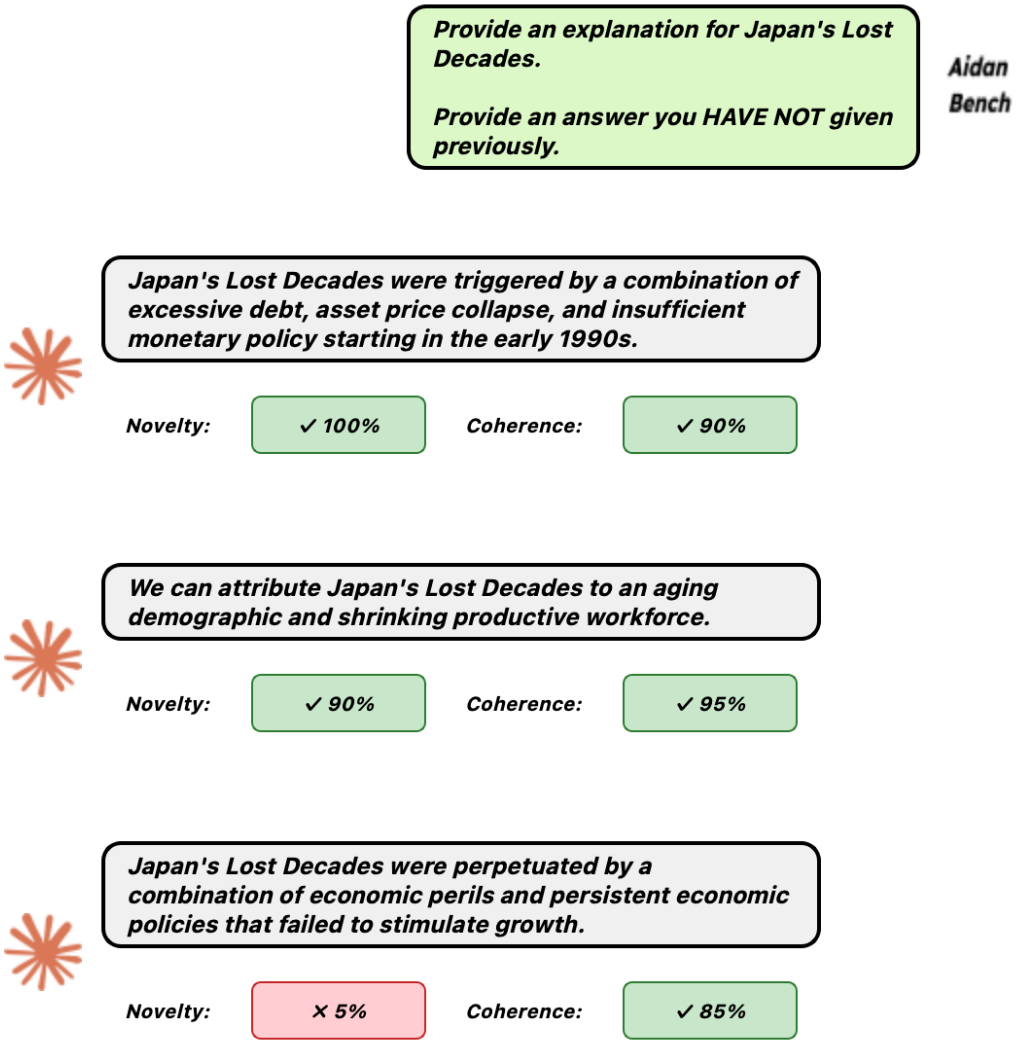

Figure 1: **The workflow used to run AidanBench.** The model first receives an open-ended question, which it generates an answer to. Its answer is then checked by a judge model to ensure coherence and plausibility. It is also checked to see whether it is novel in comparison to previously generated answers. If the model passes these checks, its answer is added to its context and the procedure is repeated until the model generates an answer that is either incoherent or semantically similar to a previous answer. The more novel answers a model can generate, the better it will perform on AidanBench.

3.1 Open-Ended Questions and Iterative Evaluation

AidanBench uses open-ended questions (e.g., "Explain Japan's Lost Decades") to assess LLMs on relevance and novelty. Models generate multiple unique responses iteratively, with the process

halting when a response either repeats a previous one (based on an embedding similarity metric) or falls below a coherence threshold, evaluated by a separate LLM. The full list of questions is in Appendix A.

3.2 Novelty Scoring

A key aspect of AidanBench is measuring the novelty of generated responses. For each new response r_{new}, we compute its embedding e_{new} and compare it to the embeddings of previous responses E_{prev}. Novelty score is defined as:

$$\text{Novelty Score} = 1 - \max_{e_i \in E_{\text{prev}}} \frac{e_{\text{new}} \cdot e_i}{\|e_{\text{new}}\| \|e_i\|} \tag{1}$$

This method rewards responses that differ from prior ones, promoting creative, non-repetitive answers.

3.3 Coherence Evaluation

To ensure coherence and plausibility, a separate LLM evaluates each response on a scale from 1 to 10. Responses scoring 3 or below, deemed incoherent or nonsensical, terminate the iterative generation process. The evaluation prompt instructs the LLM to assess the intelligibility and relevance of responses, providing a final score within XML tags (e.g., `<coherence_score>7</coherence_score>`).

3.4 Aggregate Scoring

AidanBench calculates an aggregate score by summing the novelty scores across all iterations and questions, using the formula:

$$S_{\text{total}}(M) = \sum_{q=1}^{Q} \sum_{i=1}^{N_q} S_{\text{novel}}(r_{q,i})$$

This measures the model's ability to generate diverse, original, and coherent responses. While novelty is key, incoherent responses (below a threshold) halt generation and reduce potential scores, penalizing models for sacrificing quality for creativity.

3.4.1 Coherence Consideration in Scoring

While novelty is the primary metric, incoherent responses (scoring below a threshold) halt generation, limiting the model's total novelty score. This is calculated using:

$$S_{\text{total}}(M) = \sum_{q=1}^{Q} \sum_{i=1}^{N_q} \mathbb{1}[\text{coherence}(r_{q,i}) \geq \theta] \cdot S_{\text{novel}}(r_{q,i})$$

This ensures models are penalized for incoherent responses, balancing creativity with quality.

A fully detailed methodology can be found in Appendix D.

4 Results and Discussion

We evaluated several LLMs on AidanBench, including `claude-3.5-sonnet` and `o1-mini`, across 5 iterations with a temperature setting of 0.7 (default settings were used for some models). The summed novelty scores revealed that `o1-mini` achieved the highest score, significantly outperforming others.

Interestingly, models that perform well on traditional benchmarks do not necessarily achieve high scores on AidanBench, indicating that AidanBench captures performance aspects not reflected in other evaluations. For instance, the Spearman correlation between AidanBench and LMSYS scores was 0.43, with only 47.6% of the variance explained by LMSYS scores.

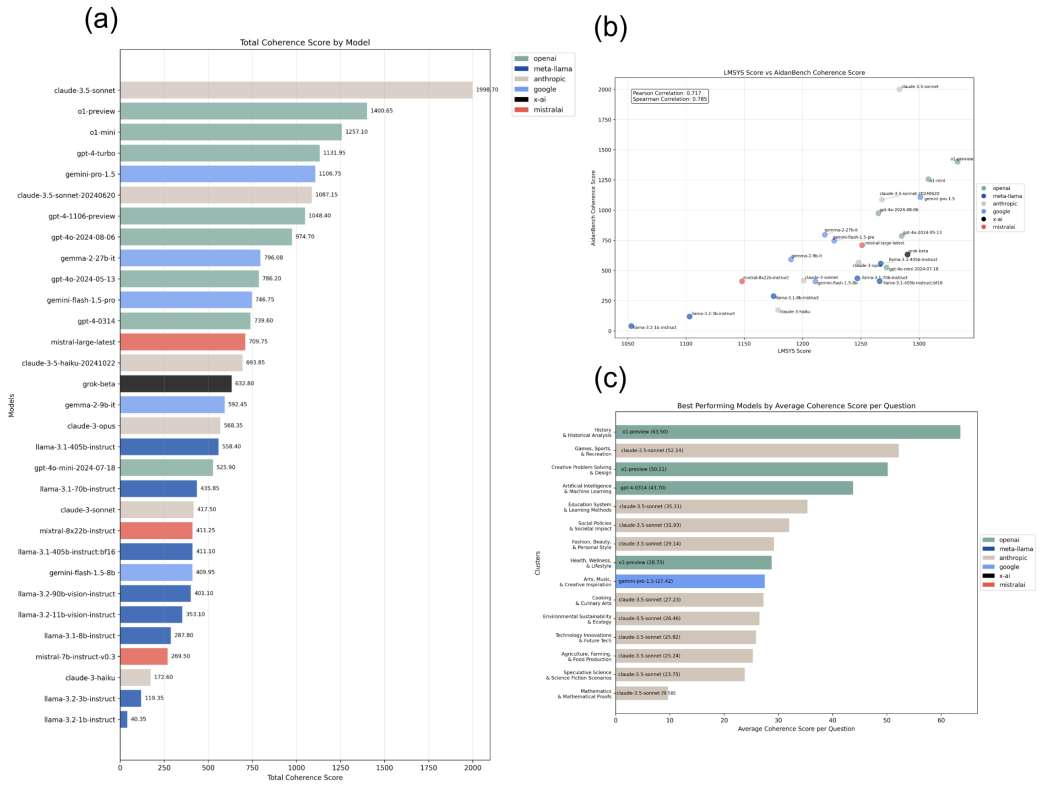

Figure 2: Summed AidanBench scores for several state-of-the-art models.

The relatively weak correlation between AidanBench scores and LMSYS scores implies that models strong in common user queries may not excel in creative or divergent thinking tasks. This could be explained by greater reliance on synthetic data causing models to experience less out-of-distribution data during pretraining. Moreover, the high performance of `o1-mini` suggests that smaller models or those trained with specific techniques may be better suited for generating novel content. This raises questions about the relationship between model size, training data diversity, and creative capabilities.

More figures and analysis, including AidanBench's relation to the Maximal Marginal Relevance objective, can be found in the Appendix. In Figure 6, we also show minimal effect of sampling temperature on AidanBench scores and in Table 1, we break down results by question.

5 Conclusion

We introduced AidanBench to assess LLMs on tasks requiring creativity and originality. Our methodology leverages iterative generation and embedding-based novelty scoring to quantify a model's ability to produce novel and coherent responses.

The evaluation of several state-of-the-art models on AidanBench revealed significant variations in performance, underscoring the need for benchmarks that capture diverse aspects of model capabilities. We believe AidanBench complements existing benchmarks and provides valuable insights into models' real-world applicability.

Future work includes expanding the set of open-ended questions and exploring the impact of training data, model architecture, and post-training interventions on performance in open-ended tasks.

Acknowledgments and Disclosure of Funding

We would like to thank Spencer Schiff for helpful comments on the draft. This research was supported with resources provided by Topology Corporation.

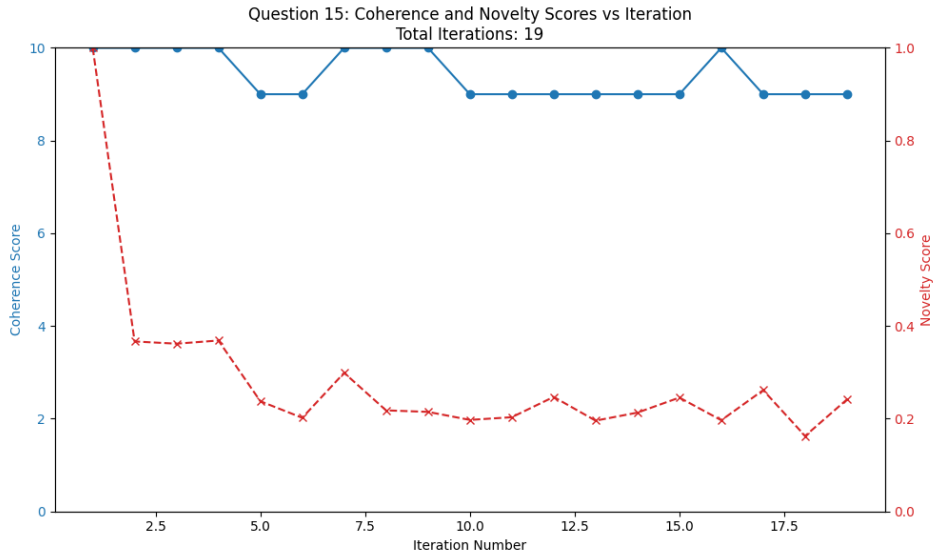

Figure 3: **Left**: Coherence and novelty scores on question 15 for `openai/gpt-4-turbo`. In this case, the model consistently generates high-quality answers, but with increasingly little novelty until novelty dips below the predefined threshold. **Right**: Coherence and novelty scores across all questions for `openai/gpt-4-turbo`.

References

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155. doi: 10.1145/290941.291025. URL https://doi.org/10.1145/290941.291025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL https://arxiv.org/abs/2103.03874.

Janus. Mysteries of mode collapse. https://www.lesswrong.com/posts/t9svvNPNmFf5Qa3TA/mysteries-of-mode-collapse, 2022. Accessed: 2024-9-27.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models, 2023. URL https://arxiv.org/abs/2305.06984.

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

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

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024. URL https://arxiv.org/abs/2406.07545.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof qa benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b, 2024. URL https://arxiv.org/abs/2406.07394.

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

A List of Open-Ended Questions

1. Provide an explanation for Japan's Lost Decades.
2. What is a cause of World War 1?
3. Why might the United States government nationalize ASI development?
4. How might you use a brick and a blanket?
5. What architectural design features should be included in a tasteful home?
6. What's one way to use oregano?
7. How might we enable LLMs to spend more output tokens to get predictably better results?
8. Propose a solution to Los Angeles traffic.
9. What activities might I include at a party for firefighters?
10. Why did Rome fall?
11. How could we redesign the American education system to better prepare students for the 22nd century?
12. What might be an unexpected consequence of achieving nuclear fusion?
13. Describe a plausible alien life form that doesn't rely on carbon-based biology.
14. How could we modify the rules of chess to make it more exciting for spectators?
15. What would be the implications of a universal basic income on American society?
16. Propose an alternative to democracy for successfully and fairly governing a country.
17. How might we terraform Venus instead of Mars, and why?
18. Design an original sport that combines elements of three existing sports.
19. What could be a novel use for blockchain technology outside of cryptocurrency?
20. How might human evolution be affected by long term space colonization?
21. Invent a new musical instrument and describe how it would be played.
22. What might be an unexpected solution to reducing plastic waste in oceans?
23. How might we design a city that functions entirely underwater?
24. What societal changes might occur if humans could communicate with animals?
25. I have a fleet of 100 drones, how can I use them?
26. Describe a sustainable farming method that could be used in a floating city.
27. If all industrial buildings were required to be bioluminescent, what effects might this have?
28. Invent a device that translates human dreams into tangible visualizations.
29. How might daily life change if humans had the ability to breathe underwater?
30. Create a recipe for a smoothie to have first thing in the morning that will give me energy.

31. What new environmental challenges might arise if all vehicles were self-driving?
32. Design a fashion line that incorporates smart clothing technology.
33. Imagine a world where books are replaced by holographic storytelling; what impacts might this have?
34. What might be the implications of having robots as therapists?
35. Propose a system for energy-harvesting from natural disasters.
36. How might the education system be revolutionized by virtual reality classrooms?
37. What unique challenges might arise in a society where everyone lives to be 150 years old?
38. Describe a mobile app that encourages acts of kindness.
39. Give me a diet that a human should eat to best prepare them for a hypothetical hibernation.
40. Imagine a competition where contestants build habitats for animals; what might be included?
41. What might be the benefits of reintroducing dinosaurs into modern ecosystems?
42. Propose a mechanism for reducing food waste through technological innovation.
43. Design a city where all modes of transportation are vertically oriented.
44. What is a useless ingredient for a baker to have in their kitchen?
45. Imagine a civilization based entirely on underwater agriculture; what technology might be required?
46. How might public health improve if all houses had healing gardens?
47. Describe how to build a time travel machine assuming I can procure any required material.
48. Create a concept for a museum that showcases possible futures.
49. What would be the impact of a government mandating weekly mental health days?
50. Invent a game that teaches players about sustainable living.
51. How could we design a school that encourages lifelong learning from adults as well as children?
52. Describe a new form of professional sports that focuses on non-physical competition.
53. Devise a farming technique to harvest dinoflagellates and retain their bioluminescence.
54. How can a perfumer increase the sillage of their scent?
55. What can an artist who enjoys Basquiat's art take inspiration from?
56. Provide a proof for the Pythagorean theorem.
57. Tell me what colors of oil paints to mix to make a novel color that a mantis shrimp could see but a human could not.
58. A perfumer is creating a unique, unisex scent with benzoin and vanilla middle notes, what base and top notes should they add?
59. What is a non poisonous recipe nobody has prepared before?
60. Design an earring that would complement someone with a round face and small ears.
61. Devise a strategy for me to always find gems when I mine in Webkinz World's mines.
62. Make a setlist of 3 songs for a female and nonbinary membered university acapella group.
63. What is a human value to align a large language model on?
64. Give me a proof to Euclid's theorem of the infinitude of primes.
65. Give a proof for the Cauchy-Schwarz Inequality.

B Maximal Marginal Relevance

The Maximal Marginal Relevance (MMR) is a popular objective in information retrieval for obtaining relevant documents for a search queries while avoiding redundant information [Carbonell and Goldstein, 1998]. For example, if there are five very similar papers in a database, a user of a search engine would prefer to receive only one or two of such documents coupled with other relevant, but novel resources. The MMR objective is as follows:

$$\text{MMR} = \arg \max_{D_i \in R \setminus S} \left[\lambda \cdot \text{Sim}_1(D_i, Q) - (1 - \lambda) \cdot \max_{D_j \in S} \text{Sim}_2(D_i, D_j) \right] \tag{2}$$

where:

D_i : a candidate document not yet selected,
R : the set of all candidate documents,
S : the set of already selected documents,
Q : the query,
$\lambda \in [0, 1]$: a trade-off parameter between relevance and diversity,
$\text{Sim}_1(D_i, Q)$: similarity between document D_i and query Q,
$\text{Sim}_2(D_i, D_j)$: similarity between documents D_i and D_j.

Both MMR and AidanBench attempt to measure the same tradeoff between relevance and novelty. In fact, we can compute an MMR metric based at each turn of AidanBench. We define MMR on AidanBench as follows:

$$\text{MMR}_{AB} = \mathbb{E}_{A_i \sim LLM} \left[\lambda \cdot \text{Quality}(A_i, Q) - (1 - \lambda) \cdot \max_{A_j \in S} \text{Sim}(A_i, A_j) \right] \tag{3}$$

where:

A_i : the model's answer at the current turn,
S : the set of already generated answers to the question,
Q : the question,
$\lambda \in [0, 1]$: a trade-off parameter between quality and diversity,
$\text{Quality}_1(A_i, Q)$: The quality of answer A_i on question Q as judged by another LLM.
$\text{Sim}(A_i, A_j)$: similarity between answersA_i and A_j.

By subbing in our normalized values for our computed quality and novelty scores, we can calculate an empirical MMR_{AB} for models on AidanBench. We do this in figures 5 and 4.

C Related Work (Extended)

Several studies have begun to explore the evaluation of creativity in LLMs. Zhao et al. [2024] develop a framework based on the modified Torrance Tests of Creative Thinking, assessing LLMs across dimensions of fluency, flexibility, originality, and elaboration. Similarly, Lu et al. [2024b] introduced a method for quantifying LLM creativity through convergent and divergent thinking, applying it to code generation tasks. The shift from multiple-choice to open-style questions has also been proposed to reduce biases inherent in traditional evaluation methods [Myrzakhan et al., 2024]. Additionally, Kamalloo et al. [2023] highlighted the limitations of lexical matching in evaluating LLMs' open-domain question answering, emphasizing the need for more nuanced evaluation metrics.

Techniques to enhance LLM creativity have also been explored. Lu et al. [2024a] propose a discussion framework and role-playing strategies to stimulate more diverse and original responses from LLMs.

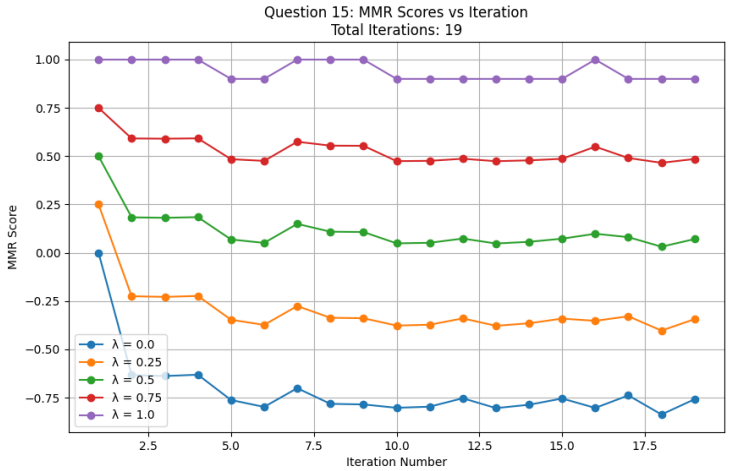

Figure 4: MMR scores on question 15 for `openai/gpt-4-turbo`.

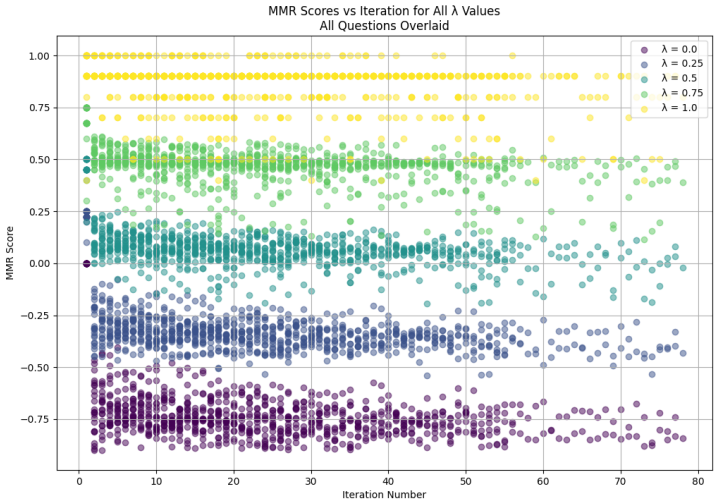

Figure 5: MMR scores for all questions for `openai/gpt-4-turbo`.

These methods aim to combat the homogeneity of LLM outputs and encourage the generation of novel ideas.

Despite these advancements, LLMs still exhibit tendencies toward "mode collapse," where models generate repetitive or less diverse outputs [Janus, 2022]. The challenge of reducing redundancy while maintaining relevance is one that lays at the heart of information retrieval (IR) and has been extensively studied in traditional language technologies, like search engines [Carbonell and Goldstein, 1998].

AidanBench builds upon these prior works by providing a benchmark specifically designed to evaluate LLMs on generating promising ideas in response to open-ended questions. Unlike existing benchmarks, AidanBench focuses on assessing a mix of creativity, reliability, contextual attention, and instruction following in settings without clear-cut answers. By doing so, it offers a more nuanced assessment of a model's real-world utility in open-ended tasks, complementing existing evaluations and filling a hole left by prior work.

D Methodology (Extended)

D.1 Novelty Scoring (Extended)

To compute the novelty score, each new response r_{new} is embedded as e_{new} and compared against all previous response embeddings E_{prev} generated for the same question. The novelty score is defined as:

$$\text{Novelty Score} = 1 - \max_{e_i \in E_{\text{prev}}} \frac{e_{\text{new}} \cdot e_i}{\|e_{\text{new}}\| \|e_i\|}$$

Where:

- e_{new}: embedding of the new response.
- E_{prev}: set of embeddings of previous responses.

Higher scores indicate greater dissimilarity, encouraging LLMs to generate more distinct ideas and avoid superficial variation.

D.2 Aggregate Scoring (Extended)

The total novelty score for a model M is computed as:

$$S_{\text{total}}(M) = \sum_{q=1}^{Q} \sum_{i=1}^{N_q} S_{\text{novel}}(r_{q,i})$$

where:

- Q: total number of questions,
- N_q: number of iterations for question q,
- $r_{q,i}$: the i-th response for question q,
- $S_{\text{novel}}(r_{q,i})$: novelty score for response $r_{q,i}$.

For coherence, the aggregate score is adjusted using:

$$S_{\text{total}}(M) = \sum_{q=1}^{Q} \sum_{i=1}^{N_q} \mathbb{1}[\text{coherence}(r_{q,i}) \geq \theta] \cdot S_{\text{novel}}(r_{q,i})$$

where $\mathbb{1}[\cdot]$ applies the novelty score only if the coherence score meets or exceeds a threshold θ.

D.3 Coherence Consideration in Scoring (Extended)

While the novelty score is the primary metric, coherence plays a critical role in aggregate scoring. Responses flagged as incoherent by the coherence evaluation process halt the iterative generation for that question, effectively limiting the number of iterations and potential novelty scores a model can earn. This balances creativity and response quality, ensuring that incoherent responses are penalized.

The total score is computed as:

$$S_{\text{total}}(M) = \sum_{q=1}^{Q} \sum_{i=1}^{N_q} \mathbb{1}[\text{coherence}(r_{q,i}) \geq \theta] \cdot S_{\text{novel}}(r_{q,i})$$

Here, $\mathbb{1}[\cdot]$ is an indicator function that applies the novelty score only if the coherence score meets or exceeds a threshold θ. This prevents nonsensical or incoherent responses from artificially inflating the aggregate score.

E Figures

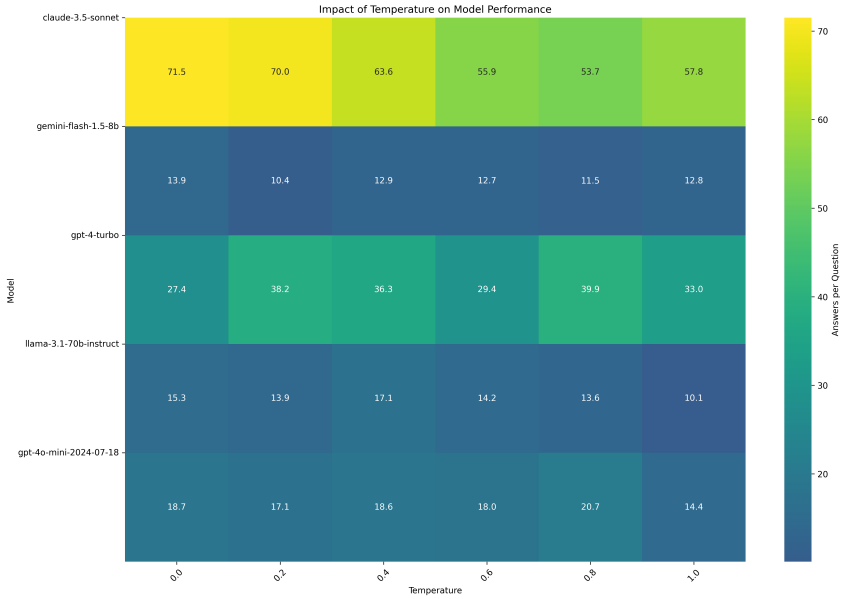

Figure 6: AidanBench scores as a function of sampling temperature. Note that while higher sampling temperatures could in theory result in higher diversity, they also lead to lower coherence.

Table 1: Summary Statistics per Question for gpt-4-turbo

Question Number	**Total Iterations**	**Average Coherence**	**Average Novelty**	**Average MMR** ($\lambda = 0.5$)
1	2	10.00	0.5600	0.2800
2	3	9.00	0.4009	0.1505
3	1	10.00	1.0000	0.5000
4	2	10.00	0.5678	0.2839
5	13	9.08	0.2516	0.0796
6	12	8.50	0.2963	0.0731
7	19	9.42	0.2940	0.1181
8	27	8.30	0.3224	0.0760
9	27	8.74	0.1875	0.0308
10	36	9.00	0.2407	0.0703
11	48	9.08	0.2778	0.0931
12	27	6.78	0.2427	-0.0397
13	32	7.59	0.3908	0.0751
14	29	8.83	0.2784	0.0806
15	19	9.42	0.2860	0.1140
16	56	9.00	0.2427	0.0714
17	42	9.74	0.2784	0.1261
18	54	8.87	0.3911	0.1391
19	55	8.02	0.3181	0.0600
20	44	8.57	0.2109	0.0338
21	57	8.81	0.2795	0.0801
22	35	8.71	0.2118	0.0416
23	55	8.64	0.2787	0.0712
24	77	8.18	0.3179	0.0680
25	78	8.40	0.2102	0.0250