Exploring Scientific Hypothesis Generation with Mamba

Miaosen Chai[1*], **Emily Herron**[2*], **Erick Cervantes**[3], **Tirthankar Ghosal**[2]
[1]University of Southern California
[2]Oak Ridge National Laboratory [3]Texas A&M International University
miaosenc@usc.edu, {herronej, ghosalt}@ornl.gov, Erickcervantes@dusty.tamiu.edu

Abstract

Generating scientifically grounded hypotheses is a challenging frontier task for generative AI models in science. The difficulty arises from the inherent subjectivity of the task and the extensive knowledge of prior work required to assess the validity of a generated hypothesis. Large Language Models (LLMs), trained on vast datasets from diverse sources, have shown a strong ability to utilize the knowledge embedded in their training data. Recent research has explored using transformer-based models for scientific hypothesis generation, leveraging their advanced capabilities. However, these models often require a significant number of parameters to manage long sequences, which can be a limitation. State Space Models, such as Mamba, offer an alternative by effectively handling very long sequences with fewer parameters than transformers. In this work, we investigate the use of Mamba for scientific hypothesis generation. Our preliminary findings indicate that Mamba achieves similar performance w.r.t. transformer-based models of similar sizes for a higher-order complex task like hypothesis generation. We have made our code available here: https://github.com/fglx-c/Exploring-Scientific-Hypothesis-Generation-with-Mamba

1 Introduction

Large language models (LLMs) have emerged as a cornerstone in artificial intelligence, particularly in scientific discovery. These models have been increasingly integrated into scientific hypothesis and idea generation, transforming traditional approaches to research. Traditionally, the process of scientific hypothesis generation has involved a complex interplay of the scientific method and inductive reasoning, requiring meticulous observation, literature review, and identification of knowledge gaps. This process, while crucial, is time-consuming and labor-intensive, relying heavily on researchers' expertise and creativity.

LLMs offer unique capabilities that address many challenges inherent in traditional scientific inquiry. They excel at processing vast amounts of text, identifying intricate patterns, and drawing upon an extensive knowledge base. This allows them to mitigate cognitive biases, efficiently identify research gaps, and generate a broad spectrum of hypotheses, including unconventional and cross-disciplinary ideas. Their ability to handle complexity makes them particularly valuable for addressing intricate, interdisciplinary problems, potentially accelerating the pace of scientific discovery. (Banker et al., 2023; Zhou et al., 2024; Park et al., 2023; O'Brien et al., 2024)

Scientific Inspiration Machines Optimized for Novelty (SciMON) (Wang et al., 2024) represents a leading approach in LLM-based scientific hypothesis generation. It utilizes an LLM-based generation module and a novel iterative novelty boosting mechanism to produce ideas that are both innovative and grounded in existing literature. However, SciMON still faces limitations in generating outputs that match the depth and utility of real scientific papers. To address these challenges, we have integrated a new LLM architecture called Mamba (Gu and Dao, 2023) into SciMON's generation module. Mamba, based on selective state space models, combines the strengths of Transformer and recurrent architectures. It introduces a selection mechanism for content-based reasoning and selective information processing within a simplified neural network design. This integration aims to enhance SciMON's ability to generate more novel, technically sophisticated, and practically useful scientific ideas.

Our work provides a comprehensive comparison of Mamba and Transformer-based models in scientific hypothesis generation tasks. We evaluate

[*]Equal contribution

Proceedings of the 1st Workshop on NLP for Science (NLP4Science), pages 197–207

Mamba's performance on general in-context learning benchmarks and long-context tasks, assess its capabilities in downstream hypothesis generation, and investigate its potential as a baseline model for scientific hypothesis generation. Throughout our study, we ensure reproducibility by providing detailed experimental setup information, including datasets, benchmark versions, and implementation scripts.

2 Related Work

Recent research has explored the potential of Large Language Models (LLMs) in scientific hypothesis and idea generation, employing various approaches from direct prompting to more complex frameworks. (Park et al., 2023) and (Banker et al., 2023) investigated the capabilities of GPT-3 and GPT-4 in generating hypotheses across diverse fields such as materials chemistry, physics, quantum information, and social psychology. While these models demonstrated broad knowledge and interdisciplinary insights, they often produced scientifically inaccurate outputs, highlighting the need for refined approaches.

More sophisticated methods have emerged, integrating inter-domain translation, iterative processes, and adversarial techniques. The FieldSHIFT framework (O'Brien et al., 2024), for instance, utilized GPT-4 to translate concepts between neuroscience and developmental biology, successfully generating novel hypotheses and demonstrating potential for identifying symmetries across scientific domains. HypoGeniC (Zhou et al., 2024) employed a multi-armed bandit-inspired reward function to iteratively improve hypotheses, outperforming few-shot prompting across multiple tasks. In astronomy, (Ciucă et al., 2023) applied adversarial prompting using multiple GPT-4 instances to generate, critique, and refine hypotheses, significantly improving their quality.

Further advancements in LLM-based hypothesis generation have incorporated multi-agent approaches, causal graphs, knowledge graph-based retrieval augmentation, and novelty optimization. Qi et al. (2023) developed a collaborative framework where LLM agents serve different roles (analyst, engineer, scientist, critic) in the hypothesis generation process. Tong et al. (2023) combined causal graphs extracted from psychology articles with LLMs to generate psychological hypotheses matching the novelty of human experts. The SciMON framework (Wang et al., 2024) generates novel research directions based on background contexts and a seed term used to constrain and guide the hypothesis space for the model. It employs an iterative novelty optimization workflow and various retrieval augmentations. GPT-4 produced the best results within this framework, although generated ideas still fell short of scientific literature in terms of depth.

While previous work has primarily utilized Transformer-based models, this study leverages Mamba (Gu and Dao, 2023), a sequence modeling architecture based on selective state space models. Mamba has demonstrated comparable or superior performance to Transformer-based architectures, particularly with long sequences. By implementing our approach within the SciMON framework, we aim to capitalize on Mamba's strengths for improved hypothesis generation in scientific contexts, potentially addressing limitations observed in previous LLM-based approaches.

3 Methodology

As mentioned, our methodology is inspired by the SciMON model. For our benchmarking study with Mamba, we use the similar experimental framework as SciMON.

3.1 SciMON Model and Dataset Description

We make use of the recently released SciMON (Scientific Inspiration Machines Optimized for Novelty) model (Wang et al., 2024), designed to generate novel, literature-informed scientific ideas in the field of Natural Language Processing (NLP). The system begins by extracting problems, motivations, and proposed ideas from scientific papers accessed through the ACL Anthology[1]. The dataset is derived from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020), comprising 67,408 ACL Anthology papers published between 1952 and 2022. Papers were filtered to include only those in English with available abstracts. The dataset is divided temporally: the training set includes papers before 2021, the validation set contains papers from 2021, and the test set comprises papers from 2022. For our experiments, we use model checkpoints trained on data preceding 2022 to avoid the risk of data contamination. The papers are processed using several information extraction (IE) and natural language processing tools:

[1] https://aclanthology.org

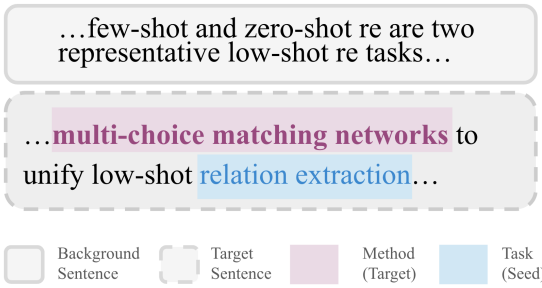


Figure 1: Use of IE to obtain literature data: background, proposed ideas (target), and seed terms.

1. PL-Marker (Ye et al., 2022), pretrained on SciERC (Luan et al., 2018), extracts entities (Task, Method, Evaluation Metric, Material, Other Scientific Terms, and Generic Terms) and their relationships, focusing on used-for relations.

2. SciCo (Cattan et al., 2021) performs coreference resolution for entity normalization.

3. Scispacy (Neumann et al., 2019) expands abbreviations to their full forms.

4. A sentence classification model by Cohan et al. (2019) categorizes abstract sentences into Background, Method, Objective, Other, and Result.

In SciMON, a seed term refers to a key concept or keyword that serves as the starting point for generating hypotheses, while the target sentence is the desired output that articulates a potential scientific idea or goal. SciMON takes a seed term and a background context as inputs and generates a corresponding target sentence as output. To train the model, paper abstracts are categorized into Background sentences (B) and Target sentences (T), forming (B, T) training pairs. The Target sentences are selected from the Methods and Objectives sections of the papers. From these, seed terms (typically Tasks) and target terms (typically Methods) are extracted to form input-output pairs. During evaluation, target information is removed. Figure 1 illustrates this process. To ensure dataset quality, we retain only high-confidence outputs from the IE models. The evaluation indicates high precision rates for most preprocessing steps, except for relation extraction. Overall, 79.7% of instances passed all preprocessing steps, which constitute the challenging dataset. For evaluation, SciMON creates a high-quality gold test set containing 194 instances by removing test cases where models can rely on surface-level background information to infer the ground truth. The remaining instances are then manually annotated to ensure a strong relevance between seed and target terms. At the core of SciMON is its inspiration retrieval module, which retrieves relevant inspirations from three external sources:

1. Semantic Neighbors: Finds similar problems and ideas in the training set based on sentence embeddings.

2. Knowledge Graph (KG) neighbors: Retrieves related concepts from a background knowledge graph built from the text dataset. The background KG has more than 197k nodes and 261k relations.

3. Citation Neighbors: Identifies relevant paper titles from the citation network of the input paper. The citation networks contain 87k paper titles.

SciMON's generation module utilizes either fine-tuned T5 language models or in-context learning with GPT-3.5 or GPT-4 LLMs. When fine-tuning the T5 models, an in-context contrastive objective is employed to discourage the models from simply copying their inputs. The in-context contrastive objective is calculated by taking negative examples from the input text and computing an InfoNCE loss (van den Oord et al., 2019) over the hidden states of the decoder with the objective of maximizing the probability of the ground truth against those of in-text negatives. Both the contrastive loss and cross-entropy loss optimized during fine-tuning. During the generation phase, the input contexts are combined with the inspirations retrieved from the previous module. The next phase in the pipeline is Iterative Novelty Boosting. This process begins with an idea generated by the generation module and retrieves similar ideas from the reference corpus or training dataset. The ideas are compared using a similarity threshold. If the generated ideas are too similar to existing ones, the model is instructed to update the idea to improve its novelty. This process is repeated until a sufficient degree of novelty is achieved. To evaluate the effectiveness of SciMON, both automated metrics such as ROUGE and BERTScore were employed, as well as extensive human evaluation. The human evaluation assessed the relevance, novelty, clarity, and scientific reason-

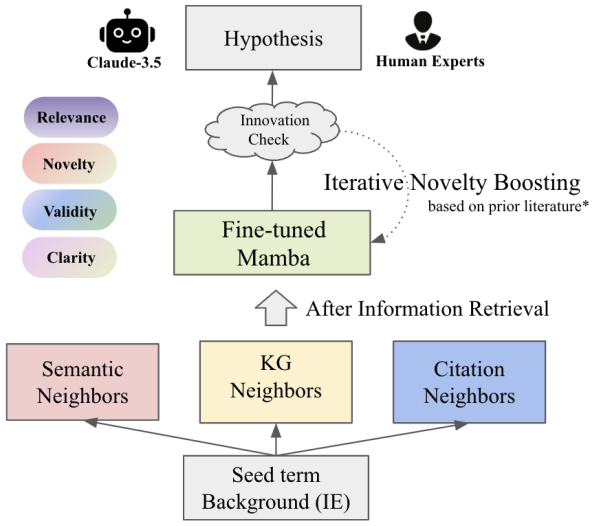


Figure 2: Using the Mamba architecture, the model generates ideas based on background context and literature inspirations, enhancing novelty by repeatedly comparing them to related work.

ableness of the generated ideas, providing a comprehensive assessment of the framework's performance in generating novel scientific ideas. In total, the retrieval dataset includes 59k papers with over 374k sentences, allowing SciMON to ground its idea generation in a broad spectrum of research, enabling it to generate novel and literature-informed scientific ideas in the field of NLP.

3.2 Mamba Architecture

The Mamba architecture (Gu and Dao, 2023) represents a significant advancement in sequence modeling, introducing selective state-space models (SSMs) to achieve linear time processing of long sequences. At the core of Mamba's design is a novel selection mechanism that enables dynamic focusing on or filtering out of inputs, effectively compressing contexts into smaller states. This approach strikes a balance between effectiveness and efficiency in sequence processing, making it particularly suitable for hypothesis generation in scientific contexts. The key innovation in Mamba lies in its selective SSM layer, which modifies traditional SSMs by making multiple parameters (Δ, B, C) functions of its inputs. This feature empowers the model to perform content-based reasoning and selectively propagate or forget information along the sequence length dimension. To implement this mechanism efficiently, Mamba employs a hardware-aware parallel algorithm that leverages the memory hierarchy of GPUs. Structurally, Mamba consists of simplified and heterogeneous blocks. Each block incorporates elements inspired by existing SSM models with MLPs, as found in modern neural networks. A typical Mamba block includes an input linear projection, a convolutional layer, the selective SSM layer, and a linear projection output layer. These blocks are stacked and interleaved with normalization and residual connections throughout the complete architecture, as illustrated in Figure 3. This design represents a simplification of previous SSM architectures by eliminating separate MLP blocks and combining various components into one repeating unit. Mamba distinguishes itself from other state-of-the-art sequence models by avoiding the use of attention mechanisms and standalone MLP blocks. These attributes enable Mamba to achieve state-of-the-art performance across various applications and modalities, including language, audio, and genomics. As demonstrated in Section 4.2, Mamba outperforms other models on language modeling tasks and downstream evaluations. While previous work has primarily utilized Transformer-based models, leveraging Mamba within the SciMON model aims to capitalize on its strengths for improved hypothesis generation in scientific contexts. Mamba's ability to handle long sequences efficiently is particularly advantageous for processing extensive scientific literature and data. Mamba scales better than other models as sequence length increases, potentially addressing limitations observed in previous LLM-based approaches. Furthermore, Mamba boasts inference times up to five times faster than Transformer models and exhibits linear scaling in sequence length (Gu and Dao, 2023). This efficiency is crucial for rapid hypothesis generation and iterative refinement in scientific research. The model's ability to selectively focus on relevant information while filtering out noise could lead to more precise and contextually appropriate hypotheses. By implementing Mamba within the SciMON model, we aim to leverage its unique architecture for enhanced scientific reasoning. The model's demonstrated success in language modeling and its ability to capture long-range dependencies make it a promising approach for efficient and effective hypothesis generation, potentially surpassing the capabilities of previous Transformer-based models in scientific contexts.

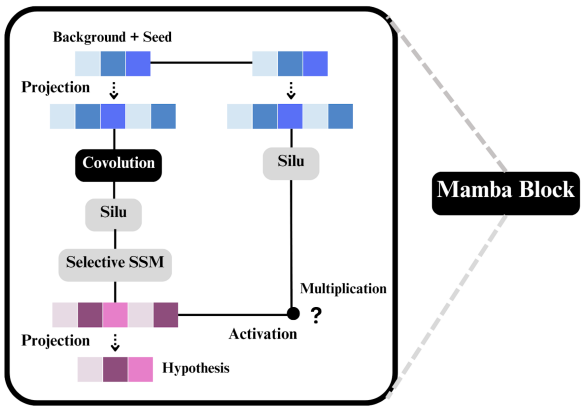


Figure 3: Mamba block we use for SciMON. Background and seed serve as input to the model.

4 Experiments & Discussion

We select T5 (Raffel et al., 2019) and GPT-4 as our baseline models to compare with Mamba. We fine-tune various sizes of T5, Mamba models and use a few short GPT-4 in parallel, with the fine-tuning process taking between 1 to 3 hours using eight H100 GPUs. We present three evaluations: one using the automated metrics and the other with LLM-as-judge (Claude-3.5), following up with a long-text evaluation and finally an evaluation of generated output by a human.

4.1 Automatic Evaluation

It is crucial to recognize that the open-ended nature of scientific hypothesis generation poses challenges for automatic evaluations, as semantically comparing outputs from SciMON to the ground truth can be constrained and shallow. Despite these limitations, automated metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) still offer valuable insights. We conduct an automatic evaluation for the outputs generated through the novelty iteration with the Challenging and Gold datasets(§3)

Results Our findings indicate that both fine-tuned T5 and Mamba models show improved performance with increased model size, as evidenced by higher ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019) metrics in Table 1. Generally, Mamba models perform on par with T5 models of similar sizes, with the Mamba-790M model achieving the highest overall scores for three evaluations. However, Mamba does not show a considerable difference compared to T5, as indicated by the results from the original paper (Gu and Dao, 2023). Additionally, GPT-4 underperformed compared to both T5 and Mamba in few-shot settings, likely because GPT-4 generates longer outputs that do not adhere to the shallow structured templates followed by T5 and Mamba, which are penalized by automatic evaluation metrics. This suggests that human judgment is necessary for a more accurate evaluation.

Model - SciMon	R-L	BERT	R-L (GS)	BERT (GS)
T5 - 60.5 m	0.178	0.514	0.184	0.524
T5 - 223 m	0.197	0.604	0.217	0.627
T5 - 738 m	0.223	0.663	**0.243**	0.684
Mamba - 130 m	0.176	0.523	0.191	0.562
Mamba - 370 m	0.219	0.628	0.237	0.631
Mamba - 790 m	**0.227**	**0.683**	0.242	**0.695**
GPT-4 FS	0.146	0.614	0.143	0.627

Table 1: Automatic results for the challenging (left) and gold (right) subsets. *R-L* denotes ROUGE-L. *BERT* denotes BERTScore with SciBERT as its encoder. *GS* denotes as Gold subsets.

4.2 LLM As a Judge

To address the limitations of automatic evaluation metrics, we incorporate an LLM evaluation to assess the quality of the generated scientific hypotheses. Specifically, we employ Claude-3.5 instead of the more mainstream GPT-4 to mitigate potential self-enhancement bias, which occurs when a model is evaluated using its own framework (Xu et al., 2024b). We utilize a reference-guided style prompt, as introduced by previous research (Zheng et al., 2023), which has been shown to increase the agreement between human and machine evaluations by providing a reference solution.

We design a structured prompt that directs Claude-3.5 to judge the effectiveness of a hypothesis. Specifically, Claude-3.5 is asked to assess the relevance of the hypothesis to the background context and seed term, ensuring sufficient novelty without overly mirroring the provided information, verifying the scientific soundness, and evaluating the clarity and coherence of the language. See the detailed prompt(C). Based on these criteria, Claude-3.5 assigns a rating of either "ineffective" or "effective" to each generated hypothesis, along with a brief justification for its decision, formatted in JSON for consistent analysis. To assist the LLM in making more accurate assessments, we provide it with five examples of both "ineffective" and "effective" hypotheses.

Model	Output
Input:	*Seed Term:* Hierarchical table dataset *Context:* Tables are often created with hierarchies, but existing works on table reasoning mainly focus on flat tables and neglect hierarchical tables. Hierarchical tables challenge numerical reasoning by complex hierarchical indexing, as well as implicit relationships of calculation and semantics.
GPT-4:	In that context, a suitable task for a hierarchical table dataset could be **Hierarchical Table Reasoning**. This task involves understanding, interpreting, and reasoning over complex hierarchical structures present in tables, such as nested rows and columns, hierarchical indexing, and implicit relationships between cells. *Extra generated reasoning*: The task is well-suited for this dataset because it addresses the problem of understanding complex relationships within the hierarchical structure.
T5:	A task that can be used for hierarchical table dataset is **semantic parsing**
Mamba:	**Question answering systems** are an effective tool for reasoning with hierarchical tables
Ground Truth:	Targeting table reasoning, we leverage entity and quantity alignment to explore partially supervised training in QA and conditional generation in NLG, and largely reduce spurious predictions in QA and produce better descriptions in NLG

Table 2: Example of SCIMON outputs with different models. *T5* denotes T5-large. *Mamba* denotes Mamba-790m.

4.3 Human Evaluation

To validate the effectiveness of Claude-3.5, we recruit five NLP experts, each with graduate-level education, to independently rate the outputs. We selected 100 questions at random from the Challenge set and developed a structured questionnaire (Figure 5) for the experts to evaluate the hypotheses. Experts rated each hypothesis as either effective or ineffective based on four key criteria: relevance, novelty, scientific validity, and clarity which is the same as the prompt instruction for Claude-3.5 (C). To ensure objectivity, the raters were blind to the conditions, and the system outputs were randomly shuffled across the instances.

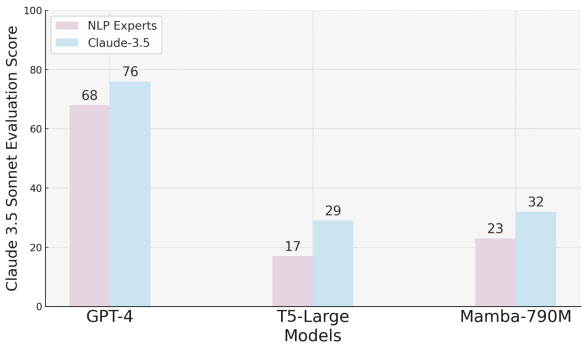


Figure 4: Human and Claude 3.5 Sonnet evaluations of generated scientific hypothesis. The y-axis represents the accuracy(%).

Results We find that both Claude-3.5 and human evaluations yield similar patterns in the performance of the models. GPT-4 achieves the highest scores in both evaluations, with an accuracy of 76% in the Claude-3.5 evaluation and 68% in the human evaluation. This consistency across evaluation methods highlights GPT-4's strong capability in generating hypotheses that align with key criteria such as relevance, novelty, scientific validity, and clarity. Given GPT-4's larger model size, its superior performance is expected. However, Mamba does not significantly outperform the transformer-based T5, likely due to the nature of the SciMON task, which does not fully exploit Mamba's long-context potential. The average input length in this task is less than 10^2 tokens, which favors models with stronger in-context learning abilities like T5. Although we hypothesize that Mamba's strengths would be more apparent in tasks requiring longer contexts, the dataset preprocessing used by the SciMON authors prevents us from directly testing this hypothesis within this context.

To further explore this, we conduct a set of long-context experiments in NLP, ordering tasks by input length: scrolls_narrativeqa (longest), scrolls_quality, and scrolls_contractnli (shortest). Our findings (Table 4) indicate that T5 models excel at tasks with smaller input sizes, with T5-Large achieving the highest accuracy of 35.97% on scrolls_contractnli. Conversely, Mamba models perform significantly better with larger input lengths, as evidenced by Mamba-790M attaining the highest F1 score of 13.81 on scrolls_narrativeqa. However, Mamba models exhibit instability on tasks with smaller inputs, as shown by the non-converging training loss when scaling to large-sized models. Similar instability has been observed in Mamba's performance on the ImageNet dataset

(Xu et al., 2024a), but the underlying cause remains unclear. This issue is likely related to the current instantiation of Mamba, which may suffer from vanishing and exploding gradients. This suggests that while Mamba does not outperform excessively on current tasks, Mamba may be more effective for scientific hypothesis generation under long-input settings. Also, the linear scaling with sequence length benefits Mamba for faster reference. However, future experiments are needed to demonstrate the performance of the Mamba architecture on a large scale.

5 Limitations and Future Work

While this study provides valuable insights, it is important to acknowledge its limitations and potential areas for future research. The architecture of SciMON introduces certain constraints that affect the scope and generalizability of our findings. One key limitation is the data scope, as SciMON's dataset is exclusively composed of ACL Anthology papers from S2ORC. This specialized focus may limit the applicability of our results to other scientific domains, particularly those that rely on multimodal data such as visual representations in biology or chemical structures in materials science.

Our comparative model analysis was restricted to an empirical comparison between Mamba and Transformer-based models under constrained parameter sizes. Future work could benefit from more extensive comparisons involving larger parameter settings, which may reveal additional insights into the relative performance of these models in hypothesis generation tasks.

Furthermore, the rapid pace of development in state space models presents new opportunities for advancing hypothesis generation capabilities. Recent innovations such as Jamba (Lieber et al., 2024), Samba (Ren et al., 2024), and TTT (Sun et al., 2024) were not included in our analysis but represent promising avenues for future research. Investigating these emerging models could potentially uncover novel approaches to improve the efficiency and effectiveness of scientific hypothesis generation.

6 Memorization

Given that LLMs are trained on extensive datasets, including potentially the same sources used for evaluation, there is a risk that the models may reproduce memorized content rather than generating novel hypotheses. So, we conduct a memorization check to ensure the validation of our experiments.

1. (Raffel et al., 2019) shows that T5 is pretrained on C4 which was crawled from web prior to April 2019.

2. Mamba uses the Pile dataset (Gao et al., 2020), and follows the training recipe described in (Brown et al., 2020).

3. The GPT-4 checkpoint used in this study is primarily based on data collected before September 2021, with only a minimal amount of more recent data included during both pretraining and post-training stages (Wang et al., 2024). Given that the evaluation focuses on papers published in 2022, the chance that these papers are part of GPT-4's pretraining dataset is considerably low.

   Furthermore, a manual review of GPT-4's outputs is conducted from SciMON using a gold set composed of 2022 ACL Anthology papers. This review specifically looks for instances where GPT-4 might reproduce detailed information, such as method names, or generate text that closely mirrors the original papers. The findings show no significant evidence of memorization.

7 Conclusion

Our study provides insights into the application of language models, particularly the Mamba architecture, for scientific hypothesis generation within the SciMON model. Comparative analysis reveals that Mamba models perform comparably to T5 models of similar sizes, with Mamba-790M achieving the highest scores in automatic evaluations. GPT-4, however, outperforms both in human and LLM-based evaluations, demonstrating superior capability in generating relevant, novel, and scientifically valid hypotheses. Mamba exhibits strength in processing longer input sequences, suggesting potential for complex scientific reasoning tasks. However, it shows instability with smaller inputs, indicating areas for improvement. These findings highlight the potential of state space models in advancing scientific hypothesis generation, despite limitations such as the use of only ACL Anthology papers and restricted parameter sizes in our analysis. Future research should focus on expanding the dataset to diverse scientific domains, investigating larger

parameter settings and emerging state space models, developing specialized benchmarks for long-sequence processing, and addressing Mamba's instability with smaller inputs. While Mamba shows promise, particularly for long-context tasks, further research is needed to fully harness its potential and address limitations. As language models evolve, their integration into scientific workflows holds great promise for accelerating hypothesis generation and innovation across diverse fields. This research represents a significant step towards leveraging advanced language models to expand the frontiers of scientific inquiry and knowledge generation.

Acknowledgement

This research used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility at the Oak Ridge National Laboratory supported by the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

Sachin Banker, Promothesh Chatterjee, Himanshu Mishra, and Arul Mishra. 2023. Machine-assisted social psychology hypothesis generation.

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, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz 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. *ArXiv*, abs/2005.14165.

Arie Cattan, Sophie Johnson, Daniel S. Weld, Ido Dagan, Iz Beltagy, Doug Downey, and Tom Hope. 2021. Scico: Hierarchical cross-document coreference for scientific concepts. *ArXiv*, abs/2104.08809.

Ioana Ciucă, Yuan-Sen Ting, Sandor Kruk, and Kartheik Iyer. 2023. Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy. *Preprint*, arXiv:2306.11648.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027.

Riccardo Grazzi, Julien N. Siems, Simon Schrodi, Thomas Brox, and Frank Hutter. 2024. Is mamba capable of in-context learning? *ArXiv*, abs/2402.03170.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. In *Conference on Empirical Methods in Natural Language Processing*.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Haim Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avshalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. 2024. Jamba: A hybrid transformer-mamba language model. *ArXiv*, abs/2403.19887.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Thomas O'Brien, Joel Stremmel, Léo Pio-Lopez, Patrick McMillen, Cody Rasmussen-Ivey, and Michael Levin. 2024. Machine learning for hypothesis generation in biology and medicine: exploring the latent space of neuroscience and developmental bioelectricity. *Digital Discovery*, 3:249–263.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Sam Bowman. 2021. Quality: Question answering with long input texts, yes! In *North American Chapter of the Association for Computational Linguistics*.

Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. 2024. Can mamba learn how to learn? a comparative study on in-context learning tasks. *ArXiv*, abs/2402.04248.

Yang Jeong Park, Daniel Kaplan, Zhichu Ren, Chia-Wei Hsu, Changhao Li, Haowei Xu, Sipei Li, and Ju Li. 2023. Can chatgpt be used to generate scientific hypotheses? *Preprint*, arXiv:2304.12208.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. *Preprint*, arXiv:2311.05965.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. 2024. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *ArXiv*, abs/2406.07522.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. 2024. Learning to (learn at test time): Rnns with expressive hidden states.

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. 2023. Automating psychological hypothesis generation with ai: Large language models meet causal graph.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. *Preprint*, arXiv:2305.14259.

Rui Xu, Shu Yang, Yihui Wang, Yu Cai, Bo Du, and Hao Chen. 2024a. Visual mamba: A survey and new outlooks. *Preprint*, arXiv:2404.18861.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024b. Pride and prejudice: Llm amplifies self-bias in self-refinement.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

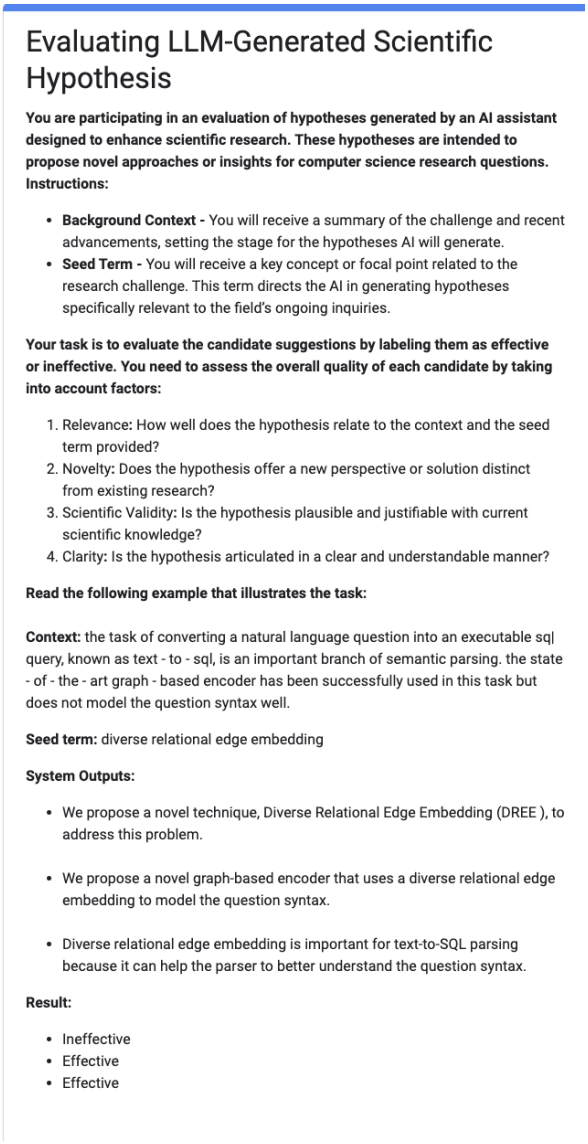
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

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

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. Hypothesis generation with large language models. *Preprint*, arXiv:2404.04326.

A Human Evaluation

To assess the effectiveness of Claude-3.5, we recruit five NLP experts, all of whom have graduate-level education, to independently evaluate the outputs by using the following questionnaire.



Evaluating LLM-Generated Scientific Hypothesis

You are participating in an evaluation of hypotheses generated by an AI assistant designed to enhance scientific research. These hypotheses are intended to propose novel approaches or insights for computer science research questions. Instructions:

- **Background Context -** You will receive a summary of the challenge and recent advancements, setting the stage for the hypotheses AI will generate.
- **Seed Term -** You will receive a key concept or focal point related to the research challenge. This term directs the AI in generating hypotheses specifically relevant to the field's ongoing inquiries.

Your task is to evaluate the candidate suggestions by labeling them as effective or ineffective. You need to assess the overall quality of each candidate by taking into account factors:

1. Relevance: How well does the hypothesis relate to the context and the seed term provided?
2. Novelty: Does the hypothesis offer a new perspective or solution distinct from existing research?
3. Scientific Validity: Is the hypothesis plausible and justifiable with current scientific knowledge?
4. Clarity: Is the hypothesis articulated in a clear and understandable manner?

Read the following example that illustrates the task:

Context: the task of converting a natural language question into an executable sql query, known as text - to - sql, is an important branch of semantic parsing. the state - of - the - art graph - based encoder has been successfully used in this task but does not model the question syntax well.

Seed term: diverse relational edge embedding

System Outputs:

- We propose a novel technique, Diverse Relational Edge Embedding (DREE), to address this problem.
- We propose a novel graph-based encoder that uses a diverse relational edge embedding to model the question syntax.
- Diverse relational edge embedding is important for text-to-SQL parsing because it can help the parser to better understand the question syntax.

Result:

- Ineffective
- Effective
- Effective

Figure 5: Human evaluation instructions

B Additional Experiments

B.1 In-context Learning

Modern attention-based LLMs exhibit remarkable in-context learning (ICL) capabilities, enabling them to learn new tasks effectively with only a few demonstrations. Research indicates that Mamba performs on par with Transformers in standard regression ICL tasks and surpasses them in tasks such as sparse parity learning (Park et al., 2024). Additionally, (Grazzi et al., 2024) found that Mamba incrementally optimizes its internal representations in a manner similar to transformer models, which aids in solving ICL problems. This adaptability suggests that Mamba can be effectively compared to Transformers in few-shot and fine-tuning settings with comparable data and training time due to its ICL, which serves as the basis for our experiment's design.

Model	MATHQA (acc) %	MMLU (acc) %	MMLUSR (acc)%	GPQA (acc)%
Mamba-130M	23.38	22.82	23.05	25.00
Mamba-370M	24.32	22.95	22.96	24.78
Mamba-790M	**25.56**	**23.74**	23.38	25.00
T5-Small	21.64	23.07	**23.49**	24.78
T5-Base	22.18	22.93	22.96	25.00
T5-Large	22.51	22.94	22.94	**25.45**

Table 3: Results for General In-Context Learning Tasks

B.2 Long-Text Evaluation

We selected three datasets, ranging from 10^2 to 10^6 words per input, to test the model's ability in question answering and natural language inference, which are the basic ability for a scientific hypothesis generation model: ContractNLI (10^2 to $10^{3.5}$) (Koreeda and Manning, 2021), QuALITY ($10^{3.3}$ to $10^{3.7}$) (Pang et al., 2021), and Narrative ($10^{3.5}$ to 10^6) (Kociský et al., 2017). The first two tasks use accuracy scores and are designed to answer specific questions based on long science and literature documents, while the latter uses F1 score for evaluation, generating results using the continuation probabilities returned by the model.

Model	Contract NLI (acc) %	QuALITY (acc) %	NarrativeQA (f1)
Mamba-130M	14.46	24.11	8.79
Mamba-370M	10.22	24.88	11.31
Mamba-790M	11.86	24.30	**13.81**
T5-Small	30.76	23.97	2.26
T5-Base	32.88	23.97	0.45
T5-Large	**35.97**	**24.98**	1.63

Table 4: Results for Long-Text Evaluation

C LLM Prompt

This is prompt for Claude: Your goal in this task is to rank idea suggestions written by LLM. The LLM helps its users write paper abstracts by generating sentences with proposals for new ideas or questions to consider. You are first given:

1. A context which describes relevant background in a specific area of interest.
2. A seed term that should be a focus of the generated scientific idea.
3. An idea suggestion generated by LLMs written in the form of a paper abstract (SUGGESTION).

Consider the following factors in your evaluation:

1. Is the suggestion relevant to the context and seed term?
2. Is the suggestion sufficiently novel, not overly copying the context?
3. Is the suggestion scientifically sound?
4. Is the language clear and coherent?

Assign a rating as either "effective" or "ineffective", where:
- "effective" = The SUGGESTION is sufficiently novel, relevant, scientifically sound, and clear.
- "ineffective" = The SUGGESTION lacks novelty, relevance, scientific soundness, or clarity.

Provide your rating and a brief justification for your assessment.

Return your output in JSON format only with the keys "justification" and "rating":

```
{
  "justification": "<your brief justification>",
  "suggestion": "ineffective< / effective>"
}
```