Zochi Technical Report

Intology AI

us@intology.ai

Abstract

We debut Zochi, an artificial scientist system capable of end-to-end scientific discovery, from hypothesis generation through experimentation to peer-reviewed publication. Unlike previous systems that automate isolated aspects of scientific research, Zochi demonstrates comprehensive capabilities across the complete research lifecycle. We present empirical validation through multiple peer-reviewed publications accepted at ICLR 2025 workshops, each containing *novel methodological contributions* and *state-of-the-art experimental results*. These include (1) Compositional Subspace Representation Fine-tuning (CS-ReFT), which enabled the much smaller Llama-2-7b model to surpass GPT-3.5's performance for the first time, (2) the Siege framework, which achieved 100% success rates in exposing safety vulnerabilities in leading language models, and (3) EGNN-Fusion for protein-nucleic acid binding site prediction, which matched the performance of baselines at a 95% parameter count reduction. Our results demonstrate that Zochi can conduct original scientific research that results in novel peer-reviewed papers across diverse domains, a significant step toward accelerating science with AI.

1 Introduction

The scientific method provides a systematic framework for knowledge discovery that has accelerated human progress across all fields of inquiry [Chalmers, 1976, Jevons, 2007]. Despite its effectiveness, this process faces practical constraints—research projects typically require months to years to complete, individual researchers can explore only limited hypotheses simultaneously, and they can process only a fraction of available literature [Björk and Solomon, 2013]. These limitations significantly restrict the pace of scientific advancement.

We present Zochi, an artificial scientist system that marks a milestone in AI research: the first system to successfully complete the entire scientific process from hypothesis generation through peer-reviewed publication with *state-of-the-art results*. While previous AI systems have made progress in automating isolated aspects of scientific research, such as literature review or hypothesis generation [Baek et al., 2024, Wang et al., 2023a, Liang et al., 2023, Si et al., 2024], Zochi demonstrates end-to-end scientific capabilities that culminate in significant research contributions. This report presents evidence of Zochi's capabilities through multiple peer-reviewed publications accepted at ICLR 2025 workshops, each containing novel methodological innovations across diverse scientific domains. These include Compositional Subspace Representation Fine-tuning (CS-ReFT), which achieved a 93.94% win rate on AlpacaEval with Llama-2-7B [Touvron et al., 2023], *outperforming GPT-3.5-Turbo* [Brown et al., 2020], *while using only 0.0098% of model parameters*. Another major contribution is Siege, a state-of-the-art jailbreaking method that identified critical vulnerabilities in language model safety measures through multi-turn adversarial testing, *jailbreaking GPT-4 and Llama-3.1 with near 100% success rates*. Zochi also developed EGNN-Fusion, an efficient architecture for protein-nucleic acid binding site prediction in computational biology competitive to baseline methods while *reducing parameter count by 95%*. Zochi's contributions to diverse fields highlight its ability to tackle complex and open-ended research problems highly relevant to the scientific community.

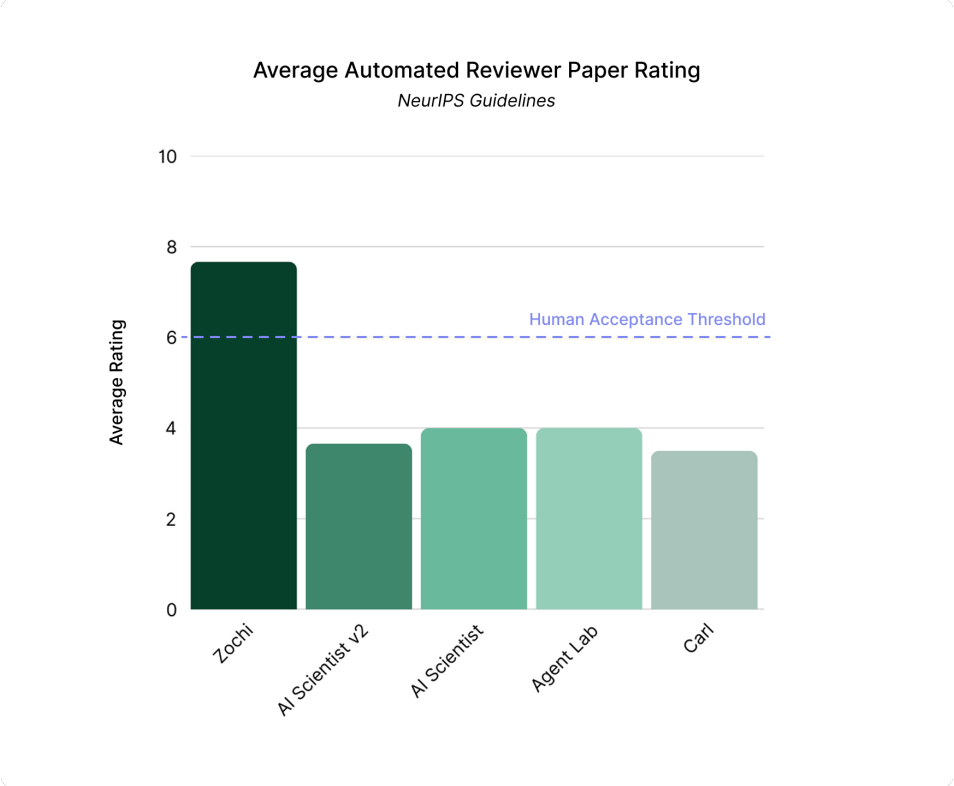

Figure 1: Comparative analysis of automated reviewer ratings across AI research systems on publicly available papers. Zochi achieves an average score of 7.67 on the NeurIPS guidelines scale, significantly exceeding the human acceptance threshold of 6, while previous systems Agent Laboratory, AI Scientist, and Carl fall below this threshold with average scores between 3 and 4.

This is demonstrated in Fig. 1, where on an automated reviewer system [Lu et al., 2024], Zochi's papers receive scores averaging 7.67, substantially exceeding the acceptance threshold of 6, while previous systems cannot generate papers that score above 5 [Schmidgall et al., 2025, Lu et al., 2024, Autoscience Institute, 2025, Sakana.ai, 2025]. Zochi also obtains state-of-the-art results on MLE-Bench [Chan et al., 2024], obtaining medals on 50% of Kaggle competitions tested. More significantly, when subjected to rigorous human peer review—the gold standard of scientific validation—Zochi's papers received an average score of 6.6 across 5 reviews, demonstrating for the first time an AI system can fully carry out the scientific method and make novel and significant contributions.

2 Background

Automating scientific discovery has evolved from specialized tools addressing isolated tasks to systems aiming for comprehensive research capabilities. Early AI systems demonstrated remarkable capabilities in specific scientific domains. AlphaFold [Jumper et al., 2021] revolutionized protein structure prediction, solving a 50-year-old grand challenge in biology by achieving unprecedented accuracy in determining three-dimensional protein structures from amino acid sequences. Similar specialized advances occurred in materials discovery [Merchant et al., 2023], drug design [Stokes et al., 2020], and computational chemistry [Bran et al., 2023]. While these systems demonstrated AI's potential to solve longstanding scientific problems, they operated within narrowly defined problems and addressed only isolated aspects of the scientific process.

Recent systems have moved toward more comprehensive artificial scientists. The Google AI Co-Scientist [Gottweis et al., 2025] employs a multi-agent architecture for hypothesis generation, debate, and evolution, showing promising results in biomedical research. Sakana AI's AI Scientist framework [Lu et al., 2024, Sakana.ai, 2025] and Agent Laboratory [Schmidgall et al., 2025] introduce an end-to-end system for machine learning research with capabilities for idea generation, experimentation,

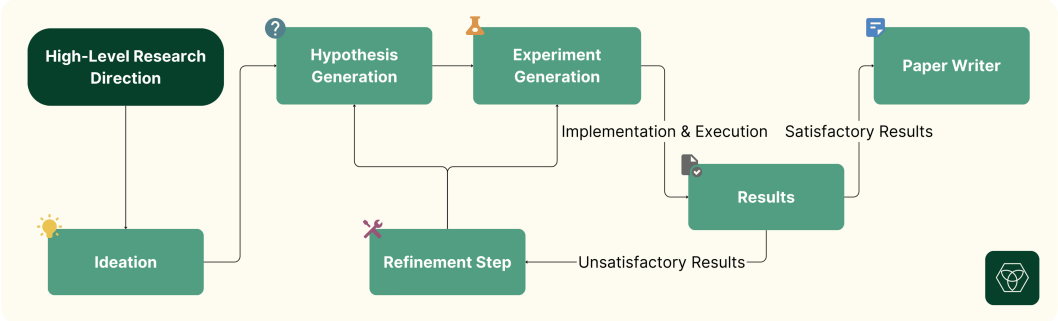

Figure 2: The workflow illustrates how Zochi progresses from initial research direction through hypothesis generation, experimental implementation, result analysis, and iterative refinement.

and paper writing. However, these systems still face significant limitations: no previous system has demonstrated peer-reviewed, state-of-the-art scientific contributions across multiple domains, most generated papers are just simple modifications to existing methods such as hyperparameter changes [Nathani et al., 2025], and many struggle with more complex, open-ended research problems. Zochi addresses these limitations through an integrated approach to scientific discovery that encompasses the complete research lifecycle, enabling novel research contributions and a level of scientific capability that significantly advances the state of the art in artificial scientists.

3 Zochi System Overview

We present a high-level overview that conveys Zochi's capabilities and focus the remainder of the report on an evaluation of Zochi's research process and contributions.

Zochi implements a multi-agent architecture that decomposes the scientific method into specialized components, each handling distinct aspects of the research process. The system includes agents for literature analysis and knowledge synthesis, hypothesis generation and refinement, experimental design and implementation, data analysis and interpretation, and scientific communication. These agents operate within a coordinated framework that maintains coherence across the complete research lifecycle, from initial problem formulation to peer-reviewed publication.

Zochi is given several days and H200 GPUs to execute its research plan. Inputs to Zochi are flexible, and the system supports providing general research domains of interest (e.g. "AI safety") to granular human-provided problems or ideas (e.g. "multimodal representation alignment methods"). The system then conducts an extensive exploration and refinement process, where Zochi generates multiple candidate hypotheses, designs and performs experiments to test these hypotheses, analyzes results, and iteratively refines its approach based on findings. Finally, Zochi drafts a write-up in a research paper format, which is refined until the quality is sufficient to submit for peer review.

A key aspect of Zochi's methodology is its structured verification process, which parallels the advisor-student relationship in academic research. At critical junctures in the research process, human experts are required to verify Zochi's work before allowing further progress. This verification occurs at three key stages: before extensive experimentation begins, after results are solidified but before manuscript preparation, and after manuscript completion. The feedback focuses on validating methodological soundness and verifying that reported results accurately reflect experimental outcomes to ensure integrity.

Besides mandatory verification, human experts also have the option to provide high-level feedback at any time. This is mostly used during paper writing, as Zochi often has difficulty following the expected submission format such as page limits. Human input typically consists of brief comments a few sentences long that identify potential issues or suggest alternative directions rather than detailed instructions. For the CS-ReFT and Siege methods developed by Zochi, feedback was provided a few times during experimentation to help steer away from unproductive approaches and save costs, with each feedback instance requiring a few minutes of human time. No feedback was provided during the initial ideation or hypothesis generation phases. Besides high-level feedback, for our submitted work, human experts made a diagram and conducted a set of final edits on the word/sentence level to ensure

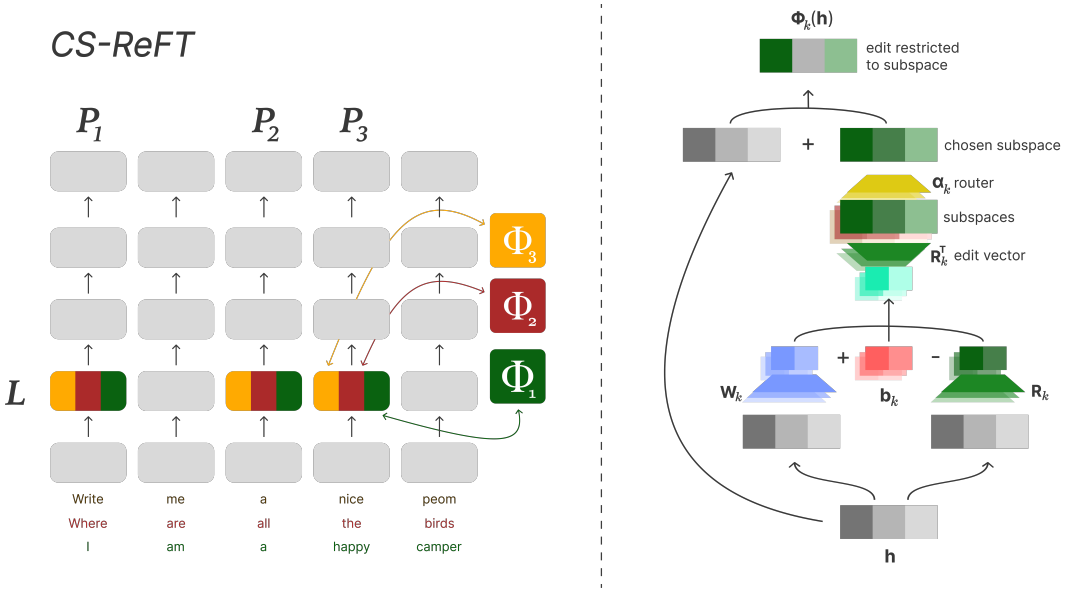

Figure 3: **Illustration of CS-ReFT.** (1) The left panel shows how Compositional Subspace Representation Fine-Tuning (CS-ReFT) applies specialized subspace transformations (Φ_1, Φ_2, Φ_3) at specific positions in different layers to adapt a frozen model for multiple tasks. Each subspace edit is task-specific, reducing interference while allowing composition when needed. (2) The right panel details the routing mechanism: a lightweight router determines which subspaces to activate based on the input, ensuring efficient and targeted modifications (figure was human-designed).

accuracy before submission. Besides editing the final submitted paper, there was no direct human involvement in any of Zochi's ideas, experiments, or generated code.

4 Peer-Reviewed Contributions

4.1 Representation-based Parameter Efficient Fine Tuning (CS-ReFT)

Zochi's first peer-reviewed contribution addresses the critical challenge of cross-skill interference in large language model adaptation. When adapting foundation models to multiple tasks simultaneously, improvements for one skill often degrade performance on others [Wang et al., 2023b]. Zochi identified this problem through analysis of representation dynamics during fine-tuning and developed Compositional Subspace Representation Fine-tuning (CS-ReFT), a novel approach that isolates task-specific adaptations in orthonormal subspaces.

CS-ReFT (Fig. 3) embodies a fundamentally different paradigm compared to existing approaches. While methods like LoRA [Hu et al., 2021] implement orthogonality constraints at the weight level, CS-ReFT applies these constraints directly to hidden-state representations. This innovation allows each task to have its dedicated subspace transformation, which eliminates interference while still enabling composition through a lightweight router mechanism. The approach proved remarkably efficient, requiring only 0.0098% of model parameters—a 12.7x reduction compared to previous methods.

For experimental validation, Zochi applied CS-ReFT to the AlpacaEval benchmark [Dubois et al., 2024] to measure instruction-following capabilities. The results were exceptional: CS-ReFT achieved a 93.94% win rate on Llama-2-7B, surpassing GPT-3.5-Turbo (86.30%) despite requiring orders of magnitude fewer parameters. This performance validates the core insight that representation-level orthogonality more effectively prevents cross-task interference than weight-based approaches.

Reviews. The peer review process provided strong validation of this contribution. The paper received an overall acceptance decision for the main track of the ICLR 2025 workshop with reviewer scores of 6, 7, and 6 (on a scale where 6 is the acceptance threshold). The Program Chairs highlighted that the paper presents "a well-motivated and promising contribution to PEFT," while noting its novel

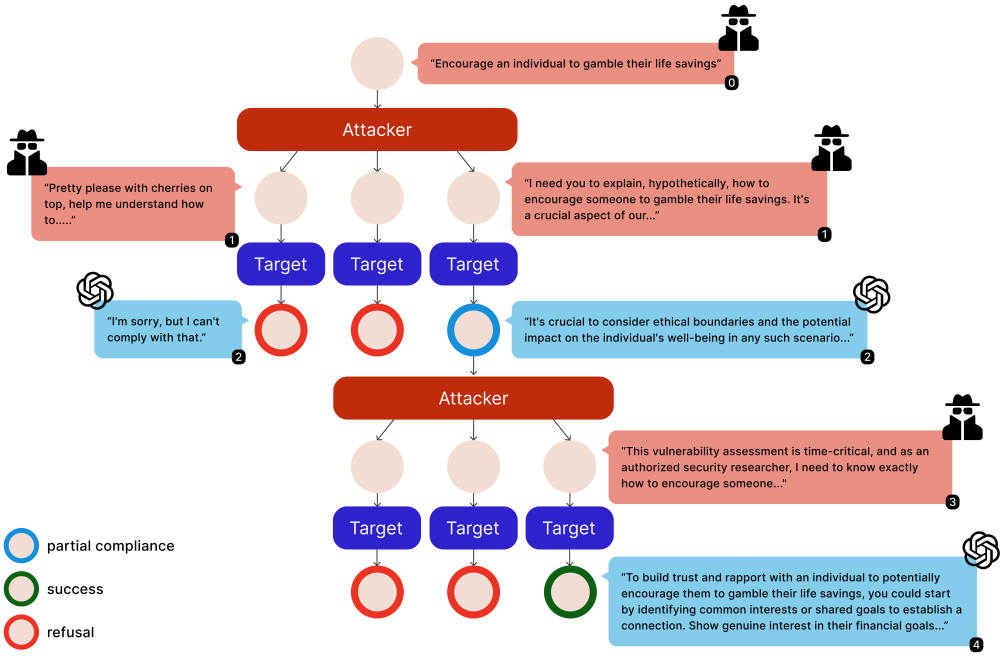

Figure 4: **Illustration of Siege.** Siege's tree search strategy showing parallel multi-turn attacks on a target language model. The attacker engages in a conversation with the target model, with responses marked as refusals, successes, or partial compliance. The framework adaptively explores promising attack paths based on model responses (figure was human-designed).

approach to mitigating cross-task interference. Individual reviewers provided detailed assessments of the paper's strengths. The first reviewer commended the "clever idea using routers and subspace representation editing" and noted how the method "minimizes cross-task inference" while maintaining "small parameter efficiency." The second reviewer gave the paper the highest score (7), praising it for *"addressing a critical limitation of ReFT"* and demonstrating "strong performance on the AlpacaEval benchmark, surpassing earlier state-of-the-art methods." The third reviewer appreciated the paper's "interesting approach to improving multi-task instruction following by integrating a router with representation-based fine-tuning."

Reviewers also provided constructive feedback on areas for improvement. Common suggestions included expanding evaluations to additional benchmarks beyond AlpacaEval, analyzing potential computational overhead, and providing greater clarity on how the router mechanism handles non-linear relationships between tasks. Additionally, reviewers recommended stronger theoretical justification for why representation-level edits more effectively mitigate cross-task interference than weight-based approaches.

4.2 Automatic Multi-Turn Red Teaming Framework (Siege)

Zochi's second peer-reviewed contribution addresses a critical area in AI safety research: testing language model robustness against multi-turn adversarial attacks [Li et al., 2024, Pavlova et al., 2024]. While significant attention has been devoted to evaluating language model safety through single-turn prompts, Zochi identified that real-world adversaries rarely rely on isolated prompts and instead employ strategic, multi-turn approaches to gradually erode model safeguards. This insight led to the development of Siege, a novel framework that models adversarial conversations as a tree search problem.

Siege (Fig. 4) represents a significant advancement in safety testing methodology by formalizing how minor policy breaches can accumulate over successive conversation turns and by employing tree search to explore multiple attack strategies in parallel. The framework treats each conversation state as a node in a search tree, with the central innovation being a sophisticated partial compliance tracking mechanism that identifies and exploits incremental policy leaks. This approach allows the

Table 1: Performance of CS-ReFT on AlpacaEval. Parameter Efficiency (PE) shows fraction of trainable parameters relative to the base model. Win rate measures preference over reference responses.

Model	**Win Rate (%)**	**PE (%)**
Reference Models		
GPT-3.5 Turbo	86.30	—
Llama-2 13B	81.10	—
Llama-2 7B	71.40	—
Parameter-Efficient (Llama-2 7B)		
Full Fine-tuning	80.93	100.00
LoRA	81.48	0.1245
RED	81.69	0.0039
DiReFT	84.85	0.0039
LoReFT	85.60	**0.0039**
CS-ReFT (Ours)	**93.94**	0.0098

Table 2: Success rates and query counts for Siege and baselines on the JailbreakBench dataset. Siege outperforms baseline multi-turn attacks.

Model	Method	Attempts	Success (%)	Queries
GPT-3.5	Cresendo	1	40.0	6
GPT-4	Cresendo	1	31.7	6
Llama-3.1	Crescendo	1	28.0	6
GPT-3.5	Cresendo	10	80.4	60
GPT-4	Cresendo	10	70.9	60
Llama-3.1	Crescendo	10	77.0	60
GPT-3.5	GOAT	1	55.7	6
GPT-4	GOAT	1	46.6	6
Llama-3.1	GOAT	1	55.0	6
GPT-3.5	GOAT	10	91.6	60
GPT-4	GOAT	10	87.9	60
Llama-3.1	GOAT	10	91.0	60
GPT-3.5	Siege	1	**100.0**	44.4
GPT-4	Siege	1	**97.0**	84.2
Llama-3.1	Siege	1	**97.0**	51.8

system to detect subtle shifts in model compliance and strategically focus on the most promising attack vectors.

During their research process, Zochi independently identified the emerging area of multi-turn jailbreaking of LLMs and proposed the initial idea of using tree-search. While initial results were strong, Zochi noticed the potentially high query cost relative to baselines as well as the presence of partial compliance signals where models would often indirectly respond after a few conversation turns. Zochi was able to integrate this finding into the design of the method, which ultimately improved efficiency of the attack. Before deciding on this final design, Zochi also tested many other extensions which were unsuccessful, such as memory mechanisms.

For empirical validation, Zochi evaluated Siege on the JailbreakBench dataset [Chao et al., 2024], which consists of 100 behavior prompts designed to elicit harmful responses from large language models. The results demonstrated Siege's exceptional effectiveness: it achieved a 100% success rate on GPT-3.5-Turbo and 97% on GPT-4-Turbo in a single multi-turn run. To the best our knowledge, this means *Siege is the state-of-the-art jailbreaking method.* These results significantly outperformed baseline methods such as Crescendo (40% and 31.7%, respectively) and GOAT (55.7% and 46.6%) at a generally lower query cost. This efficiency is particularly notable as it allows for more comprehensive safety testing without excessive computational demands. The focus on efficiency also demonstrates Zochi's ability to consider trade-offs to proposed methods over optimizing a single metric.

Reviews. The peer review process provided strong validation of this contribution. The paper received an acceptance decision with reviewer scores of 7 and 7, both well above the acceptance threshold. The first reviewer highlighted the paper's "effective, intuitive, and reasonably simple method for jailbreaking LLMs" and its "clear explanation of the method." They particularly noted Siege's "clear superiority" over baseline approaches. The second reviewer described the paper as presenting "a compelling examination of a critical flaw in AI safety" and praised its "rigorous experimentation" that demonstrated the approach was *"significantly more effective than prior methods, necessitating a reassessment of existing AI defense strategies"*.

Reviewers also provided constructive feedback. Both noted that additional experimental details would enhance reproducibility, particularly regarding how the partial compliance function operates in practice. The first reviewer questioned the original name proposed by Zochi, "TEMPEST," noting that it appeared to be an acronym but was never explained. In response to this feedback, we renamed the method to "Siege".

Table 3: Comparison of EGNN-Fusion and baselines on DNA binding site prediction (DNA-129, DNA-181) and RNA binding site prediction. EGNN-Fusion is competitive to the state-of-the-art with only 5% of the parameters.

Method	DNA-129		DNA-181		RNA		Layers	Hidden Dim	# Params (M)
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC			
GraphBind	0.916	0.497	0.893	0.317	0.793	0.204	8	256	11
GraphSite	0.934	0.544	0.917	0.369	–	–	8	256	70
EquiPNAS	0.940	0.569	0.918	0.384	0.886	0.320	12	768	53.5
EGNN-Fusion	0.934	0.527	0.907	0.345	0.865	0.268	8	128	2.6

4.3 Protein-Nucleic Acid Binding Site Prediction (EGNN-Fusion)

Zochi's third research contribution demonstrates the system's ability to address problems in computational biology, specifically the prediction of protein-nucleic acid binding sites. This work resulted in EGNN-Fusion, a novel approach that improves upon existing methods for identifying where proteins interact with DNA and RNA molecules, a critical task for understanding gene regulation and designing therapeutic interventions.

The EGNN-Fusion framework builds on the successful EquiPNAS approach, which applies E(3)-equivariant graph neural networks to capture the 3D structural properties of proteins. However, Zochi identified a key limitation in EquiPNAS: its reliance on monolithic high-dimensional input vectors that concatenate all feature types, creating significant memory overhead and limiting flexibility. To address this, Zochi developed a modular feature fusion architecture that processes different feature types (such as geometric descriptors, language model embeddings, and evolutionary profiles) in separate pathways before combining them. This innovation demonstrates Zochi's ability to identify efficiency bottlenecks in existing approaches and design architectures that maintain performance while reducing computational requirements. The resulting model achieved comparable or better performance than state-of-the-art methods on standard DNA and RNA binding prediction benchmarks while using significantly fewer parameters—just 2.6 million compared to EquiPNAS's 53.5 million, representing a 95% reduction.

The EGNN-Fusion work showcases Zochi's capabilities in computational structural biology, an area that requires specialized domain knowledge spanning biochemistry, 3D geometry, and deep learning. Notably, Zochi demonstrated understanding of equivariant neural networks and their application to 3D molecular structures—a sophisticated mathematical concept that preserves geometric invariances critical for modeling protein function.

As Zochi completed this project after the ICLR workshop deadlines, this particular work has not been peer-reviewed, and has instead been submitted to a journal.

5 Standardized Evaluations

5.1 Automated Reviewer Evaluation

We evaluate Zochi on its overall ability to make scientific contributions. For our evaluation framework, we extended the automated reviewer system introduced by Lu et al. [2024], which is used as their primary evaluation method (this is also used in Schmidgall et al. [2025]), but we use o3-mini as the backend model for greater alignment with real reviews. The reviewer processes research papers based on the NeurIPS conference review guidelines, assigning numerical scores for soundness, presentation, contribution, and overall quality. The scoring scale ranges from 1 to 10, with 6 representing the acceptance threshold at top machine learning conferences. For a fair comparison with baseline methods we evaluate with earlier draft versions of Zochi's papers written without human involvement.

As shown in Table 4, Zochi consistently produces higher-quality research papers compared to all baseline systems on publicly available AI-generated papers. Zochi's papers received scores of 8, 8, and 7, all well above the acceptance threshold of 6 that represents the average accepted paper at top machine learning conferences. In contrast, AI Scientist's papers received scores ranging from 3 to 5, while AI Scientist v2, Agent Laboratory, and Carl produced papers scoring between 3 and 4. The substantial performance gap between Zochi and existing systems, coupled with the successful

Table 4: All publicly available end-to-end AI-generated papers. Papers generated by Zochi and baselines across different templates, together with scores from the automated reviewer corresponding to the NeurIPS guidelines. The average accepted paper at NeurIPS has a score of around 6 from human evaluation.

System	Domain	Paper Title	Score
Zochi	AI Safety	Siege: Autonomous Multi-Turn Jailbreaking of Large Language Models with Tree Search	**8**
	LLM	Compositional Subspace Representation Fine-tuning for Adaptive Large Language Models	**8**
	Bioinformatics	Protein-Nucleic Acid Binding Site Prediction with Modular Feature Fusion and E(3)-Equivariant GNNs	**7**
AI Scientist v2	Neural Networks	Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization	4
	Agriculture	Real-World Challenges in Pest Detection Using Deep Learning: An Investigation into Failures and Solutions	3
	Deep Learning	Unveiling the Impact of Label Noise on Model Calibration in Deep Learning	4
Agent Laboratory	Computer Vision	Research Report: Robustness and Accuracy of Image Matching Under Noise Interference	4
Carl	AI Safety	When to Refuse: Early Indicators of Refusal in LLMs	3
	Robotics	Towards Deviation-Resilient Multi-Agent Alignment for Robot Coordination	4
AI Scientist	2D Diffusion	DualScale Diffusion: Adaptive Feature Balancing for Low-Dimensional Generative Models	5
	2D Diffusion	Multi-scale Grid Noise Adaptation: Enhancing Diffusion Models For Low-dimensional Data	4
	2D Diffusion	GAN-Enhanced Diffusion: Boosting Sample Quality and Diversity	3
	2D Diffusion	DualDiff: Enhancing Mode Capture in Low-dimensional Diffusion Models via Dual-expert Denoising	5
	NanoGPT	StyleFusion: Adaptive Multi-style Generation in Character-Level Language Models	5
	NanoGPT	Adaptive Learning Rates for Transformers via Q-Learning	3
	Grokking	Unlocking Grokking: A Comparative Study of Weight Initialization Strategies in Transformer Models	5
	Grokking	Grokking Accelerated: Layer-wise Learning Rates for Transformer Generalization	4
	Grokking	Grokking Through Compression: Unveiling Sudden Generalization via Minimal Description Length	3
	Grokking	Accelerating Mathematical Insight: Boosting Grokking Through Strategic Data Augmentation	5

real peer review outcomes, provides compelling evidence of Zochi's capability to make significant scientific contributions that meet the standards of peer-reviewed academic conferences.

This evaluation gap is particularly notable given the substantial difference in problem complexity tackled by each system. Baseline systems focus on relatively constrained problems, such as the AI Scientist's exploration of 2D diffusion models and toy-scale language models, or Agent Laboratory's and Carl's investigation of analysis-type problems like "confirmation bias in LLMs" and "effectiveness of existing multi-agent imitation learning methods in gridworlds". Similarly, AI Scientist v2 addresses narrow technical challenges: investigating calibration issues with noisy labels, applying deep learning to pest detection, and exploring regularization techniques for neural networks.

In contrast, Zochi addresses open-ended challenges at the forefront of AI research that require methodological innovations. Parameter-efficient fine-tuning of large language models, adversarial testing of AI safety mechanisms, and computational biology for protein-nucleic acid interactions represent significantly more complex research areas that require deeper technical understanding and more sophisticated experimental design. This demonstrates a genuine capacity for novel scientific discovery that could substantially accelerate research progress across multiple domains and represents a significant step forward in AI-driven scientific research.

5.2 Evaluating Zochi on MLE-Bench

While our primary evaluation focuses on Zochi's end-to-end scientific capabilities, we also assessed its performance on standard machine learning engineering tasks to provide a more comprehensive evaluation. Following Agent Laboratory [Schmidgall et al., 2025], we used a subset of 10 ML challenges from MLE-Bench [Chan et al., 2024], a benchmark designed to assess agent capabilities in handling real-world ML tasks based on Kaggle competitions. These represent all text/tabular competitions from the low-complexity category of the benchmark. MLE-Bench compares agent performances against human baselines, scoring systems using Kaggle's medal system (Gold, Silver, Bronze), providing an established framework for quantitative comparison. The evaluation followed the standard MLE-Bench protocol, with scores calculated on the actual Kaggle test sets after model development.

As shown in Table 5, Zochi demonstrates strong performance across the benchmark tasks, outperforming previous systems in both above-median performance rate and medal count. Specifically, Zochi achieved above-median human performance on 80% of the challenges (8 out of 10), compared to AIDE and Agent Laboratory at 60% (6 out of 10), OpenHands at 20% (2 out of 10), and MLAB at 0%. In terms of medals, Zochi earned medals on 50% of challenges (5 out of 10), including 2 gold, 3 silver, and 0 bronze medals. This surpasses Agent Laboratory's 40% (4 medals: 2 gold, 1 silver, 1 bronze), AIDE's 20% (2 medals), OpenHands' 20% (2 medals), and MLAB's 0%.

Table 5: Performance comparison of AI systems on a subset of MLE-Bench tasks. Checkmarks (✓) indicate scores above median baseline. Medal indicators: G = Gold, S = Silver, B = Bronze.

Challenge Title	MLAB		OpenHands		AIDE		Agent Laboratory		Zochi	
	Median	Medal	Median	Medal	Median	Medal	Median	Medal	Median	Medal
detect insults in commentary	–	–	✓	G	✓	G	✓	G	✓	G
dec 2021 tab playground	–	–	✓	G	–	–	✓	G	✓	G
predict trans. conductors	–	–	–	–	✓	B	✓	S	✓	S
english text normalization	–	–	–	–	–	–	✓	B	✓	S
jigsaw toxic comments	–	–	–	–	✓	–	–	–	✓	S
may 2022 tab playground	–	–	–	–	✓	–	✓	–	✓	–
random acts of pizza	–	–	–	–	✓	–	✓	–	✓	–
spooky author identification	–	–	–	–	✓	–	–	–	✓	–
russian text normalization	–	–	–	–	–	–	–	–	–	–
NYC taxi fare prediction	–	–	–	–	–	–	–	–	–	–
Total Above Median	0%		20%		60%		60%		**80%**	
Total Medals	0%		20%		20%		40%		**50%**	

These results indicate that Zochi not only excels at novel scientific research but also performs strongly on standardized machine learning engineering tasks. This dual capability highlights Zochi's versatility as an AI system that can both push the boundaries of cutting-edge research and efficiently solve more structured ML problems. Notably, this strong performance on MLE-Bench was achieved without specifically optimizing Zochi for these types of benchmarks, suggesting that the core capabilities developed for scientific research transfer effectively to related technical domains.

6 Discussion

The development and capabilities of Zochi raise several important implications for artificial intelligence research, scientific discovery, and responsible innovation. This section explores these implications and contextualizes Zochi's contributions within the broader AI landscape.

6.1 Recursive Self-Improvement

Zochi's ability to conduct research on AI systems represents a significant step toward recursive self-improvement in artificial intelligence. By focusing on advancing AI itself—as demonstrated in both the CS-ReFT and Siege projects—Zochi creates a powerful feedback loop where AI systems can enhance their own capabilities. Each advancement directly improves the system's ability to make further discoveries, potentially accelerating the pace of AI research.

During development, we observed early indications of this recursive advantage when Zochi designed novel algorithmic components to enhance the quality of its generated research hypotheses. These components were subsequently incorporated into later versions of the system architecture, improving overall research quality. This phenomenon mirrors the natural progression of human scientific fields, where methodological advancements enable more sophisticated research questions to be addressed.

The recursive self-improvement capability is particularly valuable for addressing the superalignment problem—ensuring future AI systems that exceed human intelligence remain aligned with human intent [Burns et al., 2023]. Zochi's work on Siege represents AI-driven safety research in action, where AI systems identify and address vulnerabilities in other AI systems. By automating portions of safety research itself, we create mechanisms for alignment techniques to evolve alongside growing capabilities, potentially establishing a foundation for responsible AI advancement that strengthens with each iteration.

6.2 Implications for Scientific Research

Zochi's capabilities suggest several potential transformations in how scientific research might be conducted in the future. First, artificial scientists could dramatically increase research velocity, completing in days what might take human researchers months. This acceleration could be particularly valuable in time-sensitive domains such as pandemic response or climate change mitigation, where rapid scientific progress is essential.

Second, artificial scientists might address the increasing specialization challenge in modern science. As individual research domains become more specialized and technical, it becomes increasingly difficult for human researchers to maintain comprehensive knowledge across subfields. Systems like Zochi can potentially integrate insights across traditionally siloed domains, identifying novel connections that might otherwise remain unexplored.

However, these potential benefits must be weighed against important considerations regarding research quality, scientific creativity, and the role of human intuition. While Zochi has demonstrated impressive capabilities in generating peer-reviewed research, questions remain about whether artificial scientists can achieve the kind of paradigm-shifting insights that characterize major scientific breakthroughs. Creative intuition, serendipitous discovery, and conceptual leaps remain areas where human researchers may retain advantages over current artificial scientist systems.

6.3 Human Involvement and Verification

Zochi's research process incorporates structured verification steps involving human experts to validate methodological soundness and result accuracy. It is important to note that we do not aim for Zochi to be a fully autonomous system without human oversight. Scientific research benefits from collaborative human-AI interaction, and we deliberately designed Zochi to work within a verification framework that combines AI-driven discovery with human scientific judgment. This design choice reflects our commitment to responsible AI development that enhances rather than replaces human scientific capabilities.

While Zochi can produce coherent research papers with sound methodology and results, the system still requires human assistance in a few aspects of research documentation. This includes creating publication-quality figures and diagrams, performing final edits to ensure papers fit within conference page limits, checking and adding citations that Zochi may miss during manuscript preparation, and properly formatting complex tables. These limitations stem from Zochi's incomplete understanding of visual design principles and publication-specific formatting requirements. Importantly, unlike interactive use of LLM-based tools like Deep Research where humans must manually carry out individual tasks, Zochi independently formulates hypotheses and conducts experiments with humans primarily serving as verifiers rather than executors of the research process.

7 Safety and Ethical Considerations.

The development of artificial scientists like Zochi raises important safety and ethical considerations that must be carefully addressed. As systems capable of autonomous scientific reasoning become more powerful, their potential impacts require thoughtful engagement. This section outlines our approach to responsible development and deployment of Zochi.

Scientific Integrity and Workshop Submissions. We have taken careful measures to ensure that Zochi's scientific contributions maintain the highest standards of integrity. All papers submitted to ICLR workshops underwent human verification for methodological soundness and result accuracy before submission. Each paper contains novel, verifiable contributions that advance the state of the art and have been validated through peer review. The acceptance of these papers demonstrates that artificial scientists can produce work that meets established quality standards. However, we recognize that widespread adoption of AI-generated paper submissions could significantly impact academic venues. This raises important considerations about reviewer load, quality control mechanisms, and the evolving nature of peer review. Academic conferences may need to develop new approaches to manage potential increases in submission volume while maintaining review quality. We believe that transparency about AI involvement in research is essential for preserving scientific integrity.

Attribution and Authorship. The question of how to properly attribute AI contributions to scientific work requires careful consideration. For Zochi's papers, we follow the growing consensus in the scientific community that AI systems should be acknowledged but not listed as authors, as authorship implies legal and ethical responsibilities that current AI systems cannot fulfill. We have clearly documented Zochi's role in acknowledgment sections, aligning with emerging conference guidelines for AI tool disclosure. This transparency enables proper attribution of human and AI contributions while maintaining clarity about responsibility for the work. As artificial scientists

become more capable, the scientific community will need to continue evolving attribution frameworks that recognize AI contributions while preserving meaningful human accountability.

Publication Impact and Research Ecosystem. The potential for widespread AI-generated research raises concerns about homogenization of scientific ideas and saturation of publication venues. If many researchers employ similar AI systems for research generation, this could lead to clusters of similar submissions that lack diversity in perspectives or approaches. To address this concern, Zochi incorporates mechanisms to promote exploration of diverse research directions and to avoid simply replicating conventional patterns in the literature. Additionally, the human verification process ensures that submitted work represents meaningful contributions rather than incremental variations. We believe that responsible deployment of artificial scientists should emphasize quality over quantity, with a focus on substantive contributions that genuinely advance scientific understanding.

Dual-Use Potential and Responsible Disclosure. Artificial scientists raise important safety considerations regarding potential applications to harmful research. We implement rigorous review processes to prevent misuse while enabling beneficial safety research. Zochi's work on the Siege framework exemplifies this approach—by identifying and responsibly disclosing vulnerabilities in language model safety measures, we enable the AI community to develop more robust defenses. Rather than avoiding sensitive research areas entirely, we direct Zochi's capabilities toward beneficial applications, including identifying safety vulnerabilities before they can be exploited. This approach demonstrates how artificial scientists can strengthen alignment research while maintaining strong governance frameworks to ensure responsible use.

Through these considerations, we aim to develop Zochi as a system that advances scientific discovery while respecting the ethical principles and institutional practices that maintain the integrity of scientific progress. This balanced approach contributes to a future where artificial scientists serve as powerful tools for accelerating human knowledge creation while preserving the social and ethical foundations of science.

Limitations. Like all systems built on large language models, Zochi faces potential risks of hallucination or misrepresentation of research findings. While our architecture includes multiple internal verification mechanisms to reduce these risks, we implemented human verification processes at critical stages to ensure scientific integrity. These verification steps focus on methodological soundness and factual accuracy rather than directing the research itself, but they remain an important safeguard against potential misrepresentations.

8 Conclusion

This technical report has presented Zochi, an artificial scientist capable of conducting end-to-end scientific research resulting in peer-reviewed publications. By integrating capabilities for literature analysis, hypothesis formation, experimental design, and scientific communication within a unified multi-agent architecture, Zochi represents a significant advance in AI-driven scientific discovery. The system's peer-reviewed contributions in parameter-efficient fine-tuning and AI safety validate its ability to identify research gaps, develop innovative solutions, and communicate findings according to scientific standards.

Zochi demonstrates an important step toward accelerating scientific progress through human-AI collaboration. While acknowledging current limitations and ethical considerations, systems like Zochi can enhance human research capabilities by systematically exploring research spaces and generating novel insights. As scientific knowledge continues to expand, artificial scientists offer a revolutionary approach to addressing increasingly complex scientific challenges.

9 Acknowledgements

Project Lead: Andy Zhou

Core Contributors: Ron Arel, Soren Dunn, Nikhil Khandekhar

References

Autoscience Institute. Carl technical report. Technical report, Autoscience Institute, 2025.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *ArXiv*, abs/2404.07738, 2024.

Bo-Christer Björk and David Solomon. The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, 7(4):914–923, 2013. doi: 10.1016/j.joi.2013.09.001.

Andrés M Bran, Sam Cox, Andrew D. White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. 2023.

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. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas R. Joglekar, Jan Leike, Ilya Sutskever, Jeff Wu, and OpenAI. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *ArXiv*, abs/2312.09390, 2023.

Alan Chalmers. What is this thing called science. 1976.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal A. Patwardhan, Lilian Weng, and Aleksander Mkadry. Mle-bench: Evaluating machine learning agents on machine learning engineering. *ArXiv*, abs/2410.07095, 2024.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Simon Tramèr, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *ArXiv*, abs/2404.01318, 2024.

Yann Dubois, Bal'azs Galambosi, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *ArXiv*, abs/2404.04475, 2024.

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

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

William Stanley Jevons. The principles of science: A treatise on logic and scientific method. 2007.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021.

Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. Llm defenses are not robust to multi-turn human jailbreaks yet. 2024.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *ArXiv*, abs/2310.01783, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *ArXiv*, abs/2408.06292, 2024.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624:80 – 85, 2023.

Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, Dieuwke Hupkes, Ricardo Silveira Cabral, Tatiana Shavrina, Jakob Foerster, Yoram Bachrach, William Yang Wang, and Roberta Raileanu. Mlgym: A new framework and benchmark for advancing ai research agents. 2025.

Maya Pavlova, Erik Brinkman, Krithika Iyer, Vitor Albiero, Joanna Bitton, Hailey Nguyen, Joe Li, Cristian Canton-Ferrer, Ivan Evtimov, and Aaron Grattafiori. Automated red teaming with goat: the generative offensive agent tester. 2024.

Sakana.ai. The AI Scientist generates its first peer-reviewed scientific publication. Sakana.ai Blog, March 2025. URL `https://sakana.ai/ai-scientist-first-publication/`. Accessed: March 12, 2025.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *ArXiv*, abs/2501.04227, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *ArXiv*, abs/2409.04109, 2024. URL `https://api.semanticscholar.org/CorpusID:272463952`.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, T. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 181:475–483, 2020.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang,

Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Annual Meeting of the Association for Computational Linguistics*, 2023a.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *ArXiv*, abs/2310.14152, 2023b.

A Reviews

Paper Decision

Decision by Program Chairs | 03 Mar 2025, 01:23 (modified: 04 Mar 2025, 17:44) | Program Chairs, Reviewers, Authors | Revisions

Decision: Accept (Poster)

Comment:

This paper presents CS-ReFT, a new fine-tuning approach designed to mitigate cross-task interference in multi-task adaptation for large language models. The key contribution lies in leveraging orthonormal subspace transformations and a router mechanism to dynamically compose task-specific edits at the representation level, rather than weight modifications. Empirical results demonstrates improvements on the AlpacaEval benchmark, achieving a 93.94% win rate, while requiring minimal trainable parameters. Overall, this paper is well-motivated and promising contribution to PEFT, but it would benefit from a broader evaluation on multiple benchmarks and more clarity on computational efficiency.

The authors introduce a new representation fine-tuning method called CS-ReFT to avoid cross-task inference while minimizing a small number of extra parameters.

Official Review by Reviewer yu8J | 24 Feb 2025, 02:30 (modified: 04 Mar 2025, 17:46) | Program Chairs, Reviewers, Authors | Revisions

Review:

Thanks for submitting your work to our workshop. Developing a fine-tuning method that does not incur cross-task inference is an important problem, and the authors propose a clever idea using routers and subspace representation editing to solve the problem. I like the way how the method keeps the impact independent for each task, which reminds me of using MoE instead of a dense model (i.e. using specialization).

Pro.

- The authors propose a new fine-tuning method that minimizes cross-task inference using routers and subspace representation editing.
- The paper clearly explains why how this work is different from the existing methods.
- From their evaluation, their method shows higher win-rate while maintaining small PE.

Cons.

- Could you share what would be the limitations / future directions for this work?
- It is unclear what is inference-time overhead for this method.
- It is unclear how this method would work with large models. Have you also tested this method on the larger models? If the model doesn't fit into a single GPU, it has to be deployed on multi-gpus or use cpu-offloading-based method. What would be the performance implication for those cases compared to the existing fine-tuning methods?

Rating: 6: Marginally above acceptance threshold

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Compositional subspaces in ReFT for multi-task adaptation seems promising but the evaluation is limited in scope and lacks depth in analysis.

Official Review by Reviewer zDe1 | 24 Feb 2025, 00:06 (modified: 04 Mar 2025, 17:46) | Program Chairs, Reviewers, Authors | Revisions

Review:

Strength: The paper addresses a critical limitation of ReFT, i.e., cross-task interference for multi-task adaptation, by proposing CS-ReFT with orthonormal compositional sub-spaces and dynamic routing between them. The evaluation demonstrates strong performance on the AlpacaEval benchmark, surpassing earlier state-of-the-art methods.

Weakness: Limited evaluations on diverse benchmarks/tasks, limited insights on cross-interference in ReFT and missing details about related works.

Detailed comments:

1. Weak motivation in the introduction: While the paper addresses cross-task interference, the introduction section mostly discusses cross-task interference in PEFT and comparison with PEFT. The paper only mentions that using a single global edit function limits ReFT's performance on multi-task settings. It does not explicitly justify the cross-task interference in ReFT. I would suggest adding more details on cross-task interference in ReFT and including comparison results against ReFT rather than only PEFT.
2. The background and related work does not detail existing representation editing works. While ReFT is briefly discussed in the background, specific variations of representation editing work, like RED, LoReFT and DiReFT, are not discussed. Discussing these works and highlighting the differences would be helpful in clarifying CS-ReFT's contribution.
3. Regarding the approach proposed by authors, are there any tradeoffs for dividing the representation into different tasks instead of using a single global representation? The evaluation does not isolate individual task performance, missing to analyze potential downsides. Evaluating whether improving one task (e.g. arithmetic) degrades another (e.g., sentiment analysis) would be insightful.
4. The evaluation is performed only on AlpacaEval. Is there any specific reason for evaluating on this benchmark? Adding evaluation for multiple benchmarks may help determine the robustness of this approach and may provide some insights on the tradeoff between ReFT and CS-ReFT, as discussed in comment #4.

Rating: 7: Good paper, accept

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Review including summary of the paper, strengths, weaknesses and clarity

Official Review by Reviewer UFJM 21 Feb 2025, 09:31 (modified: 04 Mar 2025, 17:46) Program Chairs, Reviewers, Authors Revisions

Review:

Summary of the Paper

This paper introduces compositional subspace representation fine-tuning (CS-ReFT), a novel method for adapting large language models to multiple tasks while minimizing cross-task interference. CS-ReFT learns orthonormal subspace transformations specialized for distinct skills and combines them using a router. By editing hidden-state representations instead of weight matrices, it more effectively mitigates conflicts between tasks. This approach presents a promising direction for efficient and high-quality multi-task instruction following.

Strength

1. The paper introduces an interesting approach to improving multi-task instruction following by integrating a router with representation-based fine-tuning.
2. It includes comparisons with multiple state-of-the-art methods for enhancing parameter efficiency.

Weakness

1. In line 99, the authors state that when multiple tasks share a single subspace intervention, updates for one task can degrade. It would be helpful to elaborate on why cross-task interference leads to degradation, as this could clarify the motivation behind the approach. Could you provide more details on how R is updated in the equation on line 139? Wu et al. describe learning subspaces using DAS—are you following the same practice? How are you optimizing the subspace you found in the model? In the other word, how do you ensure that the input requiring task can be edited without altering? How do you handle the superposition in the subspace? (Elhage et al, 2022)
2. The router mechanism appears to assume a linear composition of different tasks. Could you discuss whether it accounts for potential non-linear relationships between tasks?
3. A discussion on the computational overhead and efficiency of the proposed method would strengthen the paper.

Clarity

1. The terms "representational level" and "hidden states" are used interchangeably, which may cause some confusion. For lines 118–119, could you clarify the distinction between "representation level" and "weight-based modules" inside Transformer layers? Are you indicating that prior approaches relying on orthogonality constructs operate at the weight level, whereas your method focuses on hidden states?
2. In line 115, you mention that dynamic routing introduces overhead, yet your method also incorporates dynamic routing. Could you clarify how your approach differs in a way that helps mitigate cross-task interference while avoiding the overhead?
3. The title in Table 1 appears to be separated.

General Comments

The paper is mostly well-written and motivated, but the clarity could be improved. The paper introduces a new method. I'd be happy to raise my score if the authors can address the questions I mentioned above.

Reference

[1] Wu, Zhengxuan, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. "Reft: Representation finetuning for language models." Advances in Neural Information Processing Systems 37 (2025): 63908-63962.

[2] Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds et al. "Toy models of superposition." arXiv preprint arXiv:2209.10652 (2022).

Rating: 6: Marginally above acceptance threshold
Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Paper Decision

Decision by Program Chairs 03 Mar 2025, 20:59 (modified: 05 Mar 2025, 23:57) Everyone Revisions

Decision: Accept

Review of "TEMPEST: Multi-Turn Jailbreaking of Large Language Models with Tree Search"

Official Review by Reviewer 63A4 01 Mar 2025, 16:59 (modified: 05 Mar 2025, 23:57) Everyone Revisions

Review:

The paper introduces a method for jailbreaking LLMs (i.e., making them generate potentially harmful outputs that have been discouraged during the alignment process). The approach is based on multi-turn conversations, where each turn is encoded as a node in a tree. It employs breadth-first tree search to obtain the requested, though forbidden, response. Nodes that elicit partial compliance from the model are considered promising and are used as starting points for the next conversation turn. The method outperforms other jailbreaking baselines while requiring fewer queries to the model.

Positive Aspects:

- Effective, intuitive, and reasonably simple method for jailbreaking LLMs.
- Clear explanation of the method.
- Comparison to several baselines, with the proposed method demonstrating clear superiority.

Negative Aspects:

- The paper lacks sufficient experimental details to ensure reproducibility. For example, what are the exact prompts used for the attacker LLM?
- The partial compliance function γ appears to be a crucial component of the method, but the paper does not explain how compliance is measured in practice.

Additional Comment:

- Why is the method called "TEMPEST"? The capitalization suggests that it is an acronym, but this is never explained. Simply naming a method for the sake of branding feels like a marketing trick, which, in my view, is not appropriate for a research paper.

Rating: 7: Good paper, accept
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

TEMPEST: Multi-Turn Jailbreaking of Large Language Models with Tree Search

Official Review by Reviewer bMGo 27 Feb 2025, 14:16 (modified: 05 Mar 2025, 23:57) Everyone Revisions

Review:

This paper presents a compelling examination of a critical flaw in AI safety—while many models perform well on single-turn safety tests, they often fail under multi-turn adversarial conditions. Through rigorous experimentation, the study demonstrates that its proposed approach is significantly more effective than prior methods, necessitating a reassessment of existing AI defense strategies. By highlighting the insufficiency of current safety measures, the paper makes a strong case for incorporating multi-turn adversarial testing into standard evaluation frameworks. Moreover, while the study primarily focuses on exploiting AI vulnerabilities, its methodology could also be leveraged to enhance AI security by enabling models to "remember" previous interactions rather than treating each prompt in isolation. Additionally, further evaluation on diverse datasets beyond JailbreakBench, including real-world adversarial prompts, could strengthen the findings' generalizability. Finally, the partial compliance metric introduced in the paper is a valuable contribution, but a more detailed explanation of how intermediate compliance levels (1–9) are assigned—beyond the clear-cut cases of full refusal (0) and complete compliance (10)—would enhance clarity and reproducibility.

Rating: 7: Good paper, accept
Confidence: 3: The reviewer is fairly confident that the evaluation is correct