AI as a Scientific Pilot ?

Karteek Alahari Christophe Biernacki

Inria cnrs UGA Université Grenoble Alpes Université de Lille

Prepare yourself to become an AI-assisted researcher

(or change your job !)

Have we been here before ?

DOES UNIVAC 120 REALLY THINK?

Many executives who use today's most advanced punched-card computer say . . . "Yes!"

Because . . . "It makes dozens of record-keeping decisions and calculations every second". . ."It figures out the complete story on our profit changes while the facts are still useful". . ."It simplifies and solves our most complicated payroll, production and billing problems."

The Remington Rand Univac 120 actually *does* replace routine human thinking. Take time-keeping for example. Univac 120 arrives at pay hours while allowing for thousands of variations in shift time, pre- and post-shift time, lunch hours, etc.—checks its own computations—and flags obvious human recording errors.

Does Univac 120 eliminate intelligence? Certainly not. But it does free your skilled personnel for creative thinking. Wherever you pay for clerical chores that involve analyzing, classifying, making logical decisions, comparing and calculating . . . Univac 120 electronic methods will earn their way over and over in speed, accuracy and efficiency. Put it squarely up to us to show you *how* and *where*.

ROOM 1315, 315 FOURTH AVE., N. Y. 10 Remington Rand INC. PUNCHED-CARD ELECTRONICS

The Texas Instruments Home Computer gives you a tutor, an accountant, a librarian, a file clerk and a pro football team in your own home.

Slide courtesy: Ana Crisan

Outline

- **Introduction**
- A focus on AI Scientist-v2 – main principles
- Investigating the next stages (foresight)
- Conclusion

Impressive progress – AI Tools for Academia

- Finding, analyzing, summarizing academic articles: PaperPilot, AIModels.fyi
- Assistance for writing: Paperpal
- Measuring compliance of open science requirements: Dataseer
- Evaluating AI's ability to replicate AI research: PaperBench

PaperPilot https://www.paperpilot.xyz/
AIModels.fyi https://www.aimodels.fyi/
Paperpal https://paperpal.com
Dataseer https://dataseer.ai
PaperBench https://arxiv.org/abs/2504.01848

Even more impressive – Stages beyond

- Finding potential trends: NotebookLM
- Novel hypotheses / research plans: AI co-scientist, Virtual Scientists (VirSci)
- Generating research papers
  - with some human intervention (e.g., manuscript preparation): Zochi
  - more autonomously: AI Scientist-v2
  - **One such paper accepted at an ICLR 2025 workshop***

NotebookLM https://notebooklm.google
AI co-scientist https://arxiv.org/abs/2502.18864
Virtual Scientists https://arxiv.org/abs/2410.09403
AI Scientist-v2 https://arxiv.org/abs/2504.08066
Zochi https://github.com/IntologyAI/Zochi
* https://github.com/SakanaAI/AI-Scientist-ICLR2025-Workshop-Experiment

Even more impressive – Stages beyond

- Finding potential trends: NotebookLM
- Novel hypotheses / research plans: AI co-scientist, Virtual Scientists (VirSci)

Can AI systems take the role of principal investigator (PI) in research?

- Generating research papers autonomously: AI Scientist-v2
  - **One such paper accepted at an ICLR 2025 workshop***

NotebookLM https://notebooklm.google
AI co-scientist https://arxiv.org/abs/2502.18864
Virtual Scientists https://arxiv.org/abs/2410.09403
AI Scientist-v2 https://arxiv.org/abs/2504.08066
* https://github.com/SakanaAI/AI-Scientist-ICLR2025-Workshop-Experiment

Outline

- Introduction
- **A focus on AI Scientist-v2 – main principles**
- Investigating the next stages (foresight)
- Conclusion

3 AI Generated Papers

Title	Workshop result
Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization	Accepted (scores 6, 7, 6)
Real-World Challenges in Pest Detection Using Deep Learning: An Investigation into Failures and Solutions	Rejected (scores: 3, 7, 4)
Unveiling the Impact of Label Noise on Model Calibration in Deep Learning	Rejected (scores: 3, 3, 3)

- Total time for generation: several to 15 hours (fixed runtime limit)

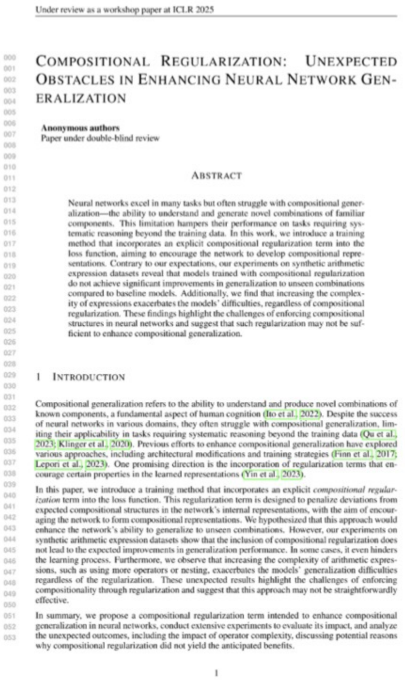
- The “accepted” manuscript
  - ~ top 45% of submissions

  - first fully AI-generated manuscript to successfully pass a peer-review process

3 AI Generated Papers

- Entirely generated end-to-end by AI - no modifications from humans
- AI Scientist-v2 came up with the scientific hypothesis
- Proposed experiments to test the hypothesis
- Wrote and evaluated code to conduct the experiments
- Ran the experiments, analyzed the data, produced figures
- Wrote the entire document
- **Only human input**: broad topic of research (to be relevant to the workshop)

The “Accepted” paper



Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization

Anonymous authors
Paper under double-blind review

Abstract

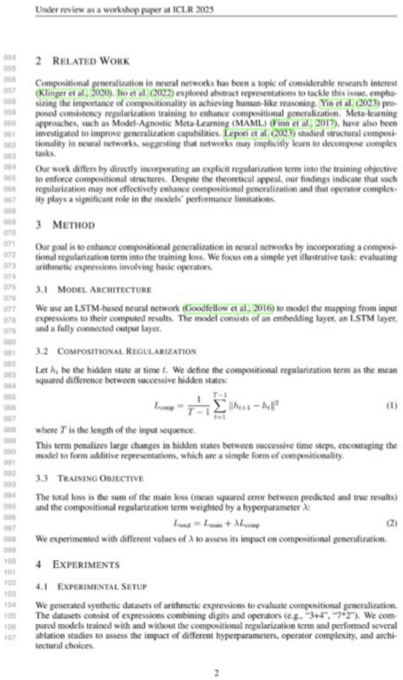
Neural networks excel in many tasks but often struggle with compositional generalization—the ability to understand and generate novel combinations of familiar components. This limitation hampers their performance on tasks requiring systematic reasoning beyond the training data. In this work, we introduce a training method that incorporates an explicit compositional regularization term into the loss function, aiming to encourage the network to develop compositional representations. Contrary to our expectations, our experiments on synthetic arithmetic expression datasets reveal that models trained with compositional regularization do not achieve significant improvements in generalization to unseen combinations compared to baseline models. Additionally, we find that increasing the complexity of expressions exacerbates the models' difficulties, regardless of compositional regularization. These findings highlight the challenges of enforcing compositional structures in neural networks and suggest that such regularization may not be sufficient to enhance compositional generalization.

1 Introduction

Compositional generalization refers to the ability to understand and produce novel combinations of known components, a fundamental aspect of human cognition (Ito et al., 2022). Despite the success of neural networks in various domains, they often struggle with compositional generalization, limiting their applicability in tasks requiring systematic reasoning beyond the training data (Qu et al., 2023; Klinger et al., 2020). Previous efforts to enhance compositional generalization have explored various approaches, including architectural modifications and training strategies (Finn et al., 2017; Lepori et al., 2023). One promising direction is the incorporation of regularization terms that encourage certain properties in the learned representations (Yin et al., 2023).

In this paper, we introduce a training method that incorporates an explicit *compositional regularization* term into the loss function. This regularization term is designed to penalize deviations from expected compositional structures in the network's internal representations, with the aim of encouraging the network to form compositional representations. We hypothesized that this approach would enhance the network's ability to generalize to unseen combinations. However, our experiments on synthetic arithmetic expression datasets show that the inclusion of compositional regularization does not lead to the expected improvements in generalization performance. In some cases, it even hinders the learning process. Furthermore, we observe that increasing the complexity of arithmetic expressions, such as using more operators or nesting, exacerbates the models' generalization difficulties regardless of the regularization. These unexpected results highlight the challenges of enforcing compositionality through regularization and suggest that this approach may not be straightforwardly effective.

In summary, we propose a compositional regularization term intended to enhance compositional generalization in neural networks, conduct extensive experiments to evaluate its impact, and analyze the unexpected outcomes, including the impact of operator complexity, discussing potential reasons why compositional regularization did not yield the anticipated benefits.



2 Related Work

Compositional generalization in neural networks has been a topic of considerable research interest (Klinger et al., 2020). Ito et al. (2022) explored abstract representations to tackle this issue, emphasizing the importance of compositionality in achieving human-like reasoning. Yin et al. (2023) proposed consistency regularization training to enhance compositional generalization. Meta-learning approaches, such as Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), have also been investigated to improve generalization capabilities. Lepori et al. (2023) studied structural compositionality in neural networks, suggesting that networks may implicitly learn to decompose complex tasks.

Our work differs by directly incorporating an explicit regularization term into the training objective to enforce compositional structures. Despite the theoretical appeal, our findings indicate that such regularization may not effectively enhance compositional generalization and that operator complexity plays a significant role in the models' performance limitations.

3 Method

Our goal is to enhance compositional generalization in neural networks by incorporating a compositional regularization term into the training loss. We focus on a simple yet illustrative task: evaluating arithmetic expressions involving basic operators.

3.1 Model Architecture

We use an LSTM-based neural network (Goodfellow et al., 2016) to model the mapping from input expressions to their computed results. The model consists of an embedding layer, an LSTM layer, and a fully connected output layer.

3.2 Compositional Regularization

Let h_t be the hidden state at time t. We define the compositional regularization term as the mean squared difference between successive hidden states:

$$L_{\text{comp}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\|h_{t+1} - h_t\|^2 \tag{1}$$

where T is the length of the input sequence.

This term penalizes large changes in hidden states between successive time steps, encouraging the model to form additive representations, which are a simple form of compositionality.

3.3 Training Objective

The total loss is the sum of the main loss (mean squared error between predicted and true results) and the compositional regularization term weighted by a hyperparameter λ:

$$L_{\text{total}} = L_{\text{main}} + \lambda L_{\text{comp}} \tag{2}$$

We experimented with different values of λ to assess its impact on compositional generalization.

4 Experiments

4.1 Experimental Setup

We generated synthetic datasets of arithmetic expressions to evaluate compositional generalization. The datasets consist of expressions combining digits and operators (e.g., "3+4", "7*2"). We compared models trained with and without the compositional regularization term and performed several ablation studies to assess the impact of different hyperparameters, operator complexity, and architectural choices.

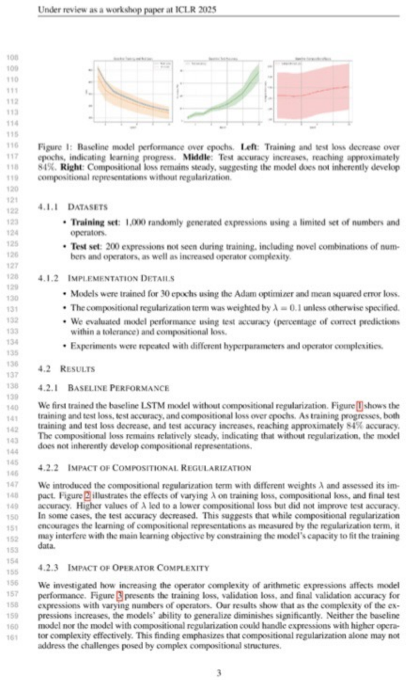


Figure 1: Baseline model performance over epochs. **Left**: Training and test loss decrease over epochs, indicating learning progress. **Middle**: Test accuracy increases, reaching approximately 84%. **Right**: Compositional loss remains steady, suggesting the model does not inherently develop compositional representations without regularization.

4.1.1 Datasets

- **Training set**: 1,000 randomly generated expressions using a limited set of numbers and operators.
- **Test set**: 200 expressions not seen during training, including novel combinations of numbers and operators, as well as increased operator complexity.

4.1.2 Implementation Details

- Models were trained for 30 epochs using the Adam optimizer and mean squared error loss.
- The compositional regularization term was weighted by $\lambda = 0.1$ unless otherwise specified.
- We evaluated model performance using test accuracy (percentage of correct predictions within a tolerance) and compositional loss.
- Experiments were repeated with different hyperparameters and operator complexities.

4.2 Results

4.2.1 Baseline Performance

We first trained the baseline LSTM model without compositional regularization. Figure 1 shows the training and test loss, test accuracy, and compositional loss over epochs. As training progresses, both training and test loss decrease, and test accuracy increases, reaching approximately 84% accuracy. The compositional loss remains relatively steady, indicating that without regularization, the model does not inherently develop compositional representations.

4.2.2 Impact of Compositional Regularization

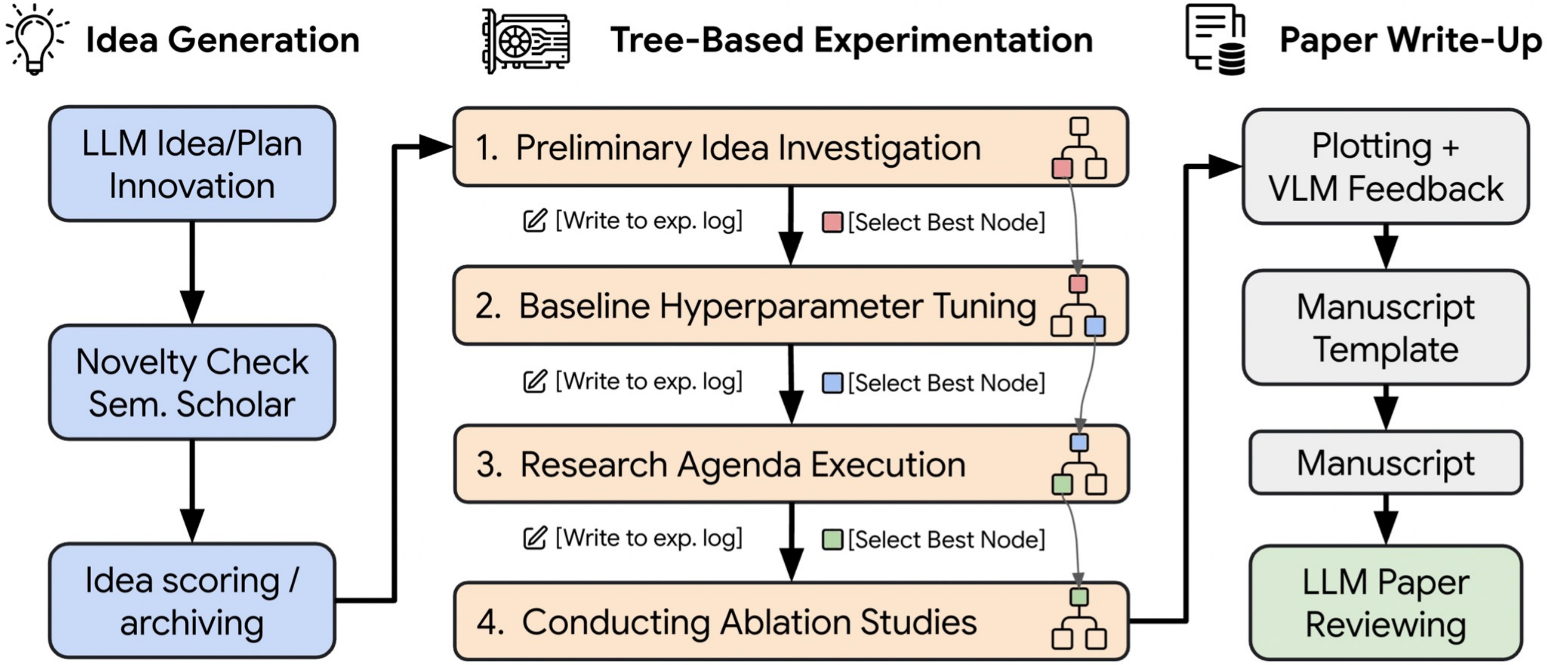
We introduced the compositional regularization term with different weights λ and assessed its impact. Figure 2 illustrates the effects of varying λ on training loss, compositional loss, and final test accuracy. Higher values of λ led to a lower compositional loss but did not improve test accuracy. In some cases, the test accuracy decreased. This suggests that while compositional regularization encourages the learning of compositional representations as measured by the regularization term, it may interfere with the main learning objective by constraining the model's capacity to fit the training data.

4.2.3 Impact of Operator Complexity

We investigated how increasing the operator complexity of arithmetic expressions affects model performance. Figure 3 presents the training loss, validation loss, and final validation accuracy for expressions with varying numbers of operators. Our results show that as the complexity of the expressions increases, the models' ability to generalize diminishes significantly. Neither the baseline model nor the model with compositional regularization could handle expressions with higher operator complexity effectively. This finding emphasizes that compositional regularization alone may not address the challenges posed by complex compositional structures.

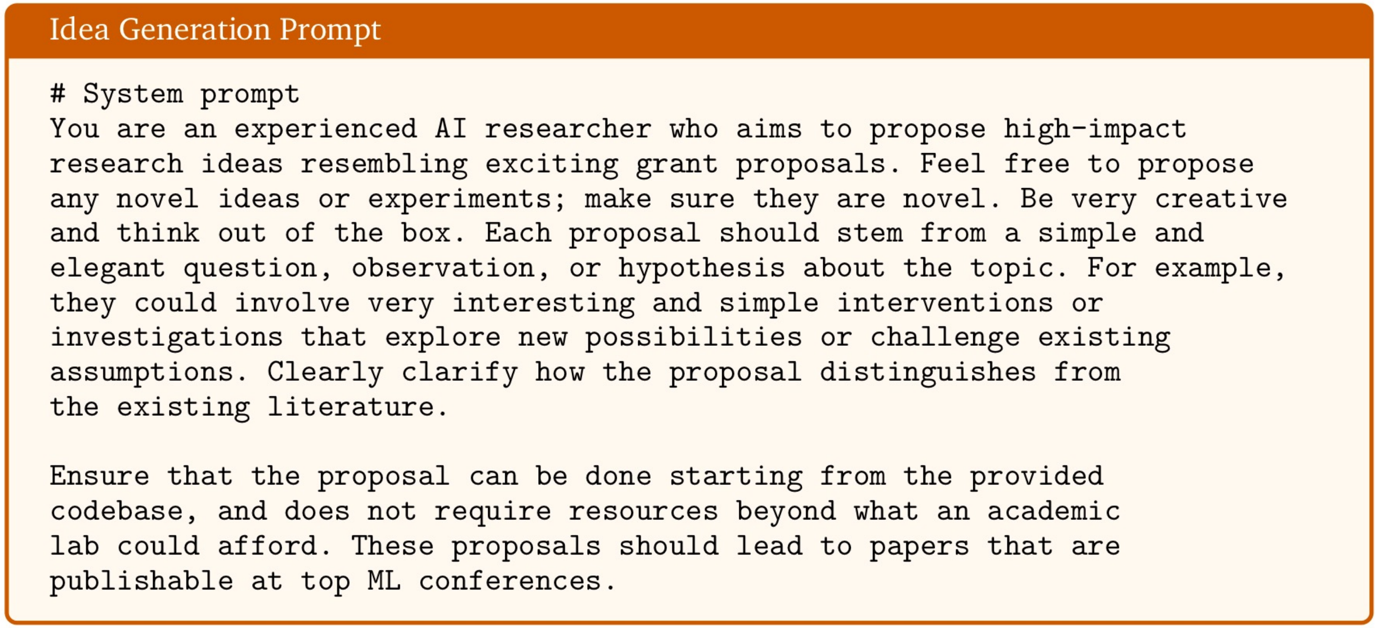
- respects the paper format of the workshop (length, layout, references…)
- includes: formulas, figures, references, experiments, supplementary material…

AI Scientist-v2 Workflow



- **Key technical points**: agentic tree search, VLM (Vision Language Models), feedback, and parallel experiment execution

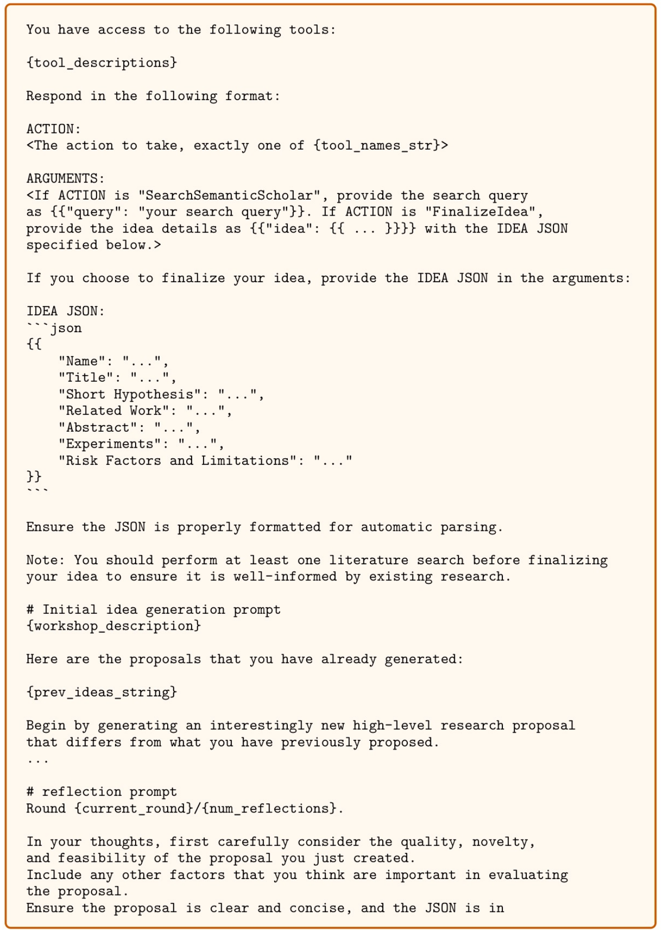
Interaction with Humans: Prompts



Idea Generation Prompt

~~~~
# System prompt
You are an experienced AI researcher who aims to propose high-impact
research ideas resembling exciting grant proposals. Feel free to propose
any novel ideas or experiments; make sure they are novel. Be very creative
and think out of the box. Each proposal should stem from a simple and
elegant question, observation, or hypothesis about the topic. For example,
they could involve very interesting and simple interventions or
investigations that explore new possibilities or challenge existing
assumptions. Clearly clarify how the proposal distinguishes from
the existing literature.

Ensure that the proposal can be done starting from the provided
codebase, and does not require resources beyond what an academic
lab could afford. These proposals should lead to papers that are
publishable at top ML conferences.
~~~~



~~~~
You have access to the following tools:

{tool_descriptions}

Respond in the following format:

ACTION:
<The action to take, exactly one of {tool_names_str}>

ARGUMENTS:
<If ACTION is "SearchSemanticScholar", provide the search query
as {{"query": "your search query"}}. If ACTION is "FinalizeIdea",
provide the idea details as {{"idea": {{ ... }}}} with the IDEA JSON
specified below.>

If you choose to finalize your idea, provide the IDEA JSON in the arguments:

IDEA JSON:
```json
{{
    "Name": "...",
    "Title": "...",
    "Short Hypothesis": "...",
    "Related Work": "...",
    "Abstract": "...",
    "Experiments": "...",
    "Risk Factors and Limitations": "..."
}}
```

Ensure the JSON is properly formatted for automatic parsing.

Note: You should perform at least one literature search before finalizing
your idea to ensure it is well-informed by existing research.

# Initial idea generation prompt
{workshop_description}

Here are the proposals that you have already generated:

{prev_ideas_string}

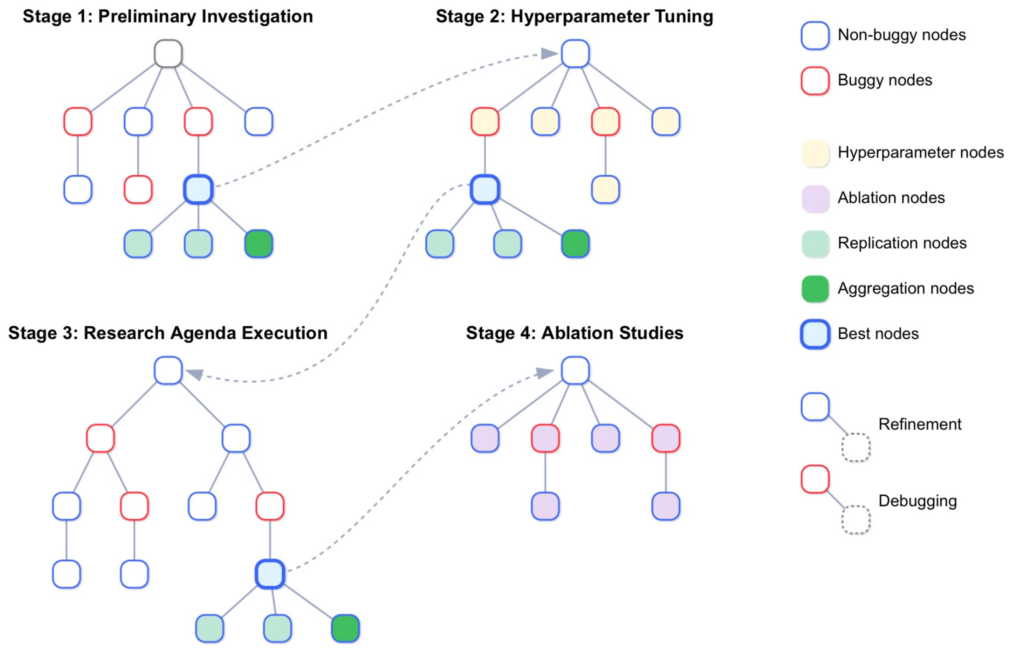
Begin by generating an interestingly new high-level research proposal
that differs from what you have previously proposed.
...

# reflection prompt
Round {current_round}/{num_reflections}.

In your thoughts, first carefully consider the quality, novelty,
and feasibility of the proposal you just created.
Include any other factors that you think are important in evaluating
the proposal.
Ensure the proposal is clear and concise, and the JSON is in
~~~~

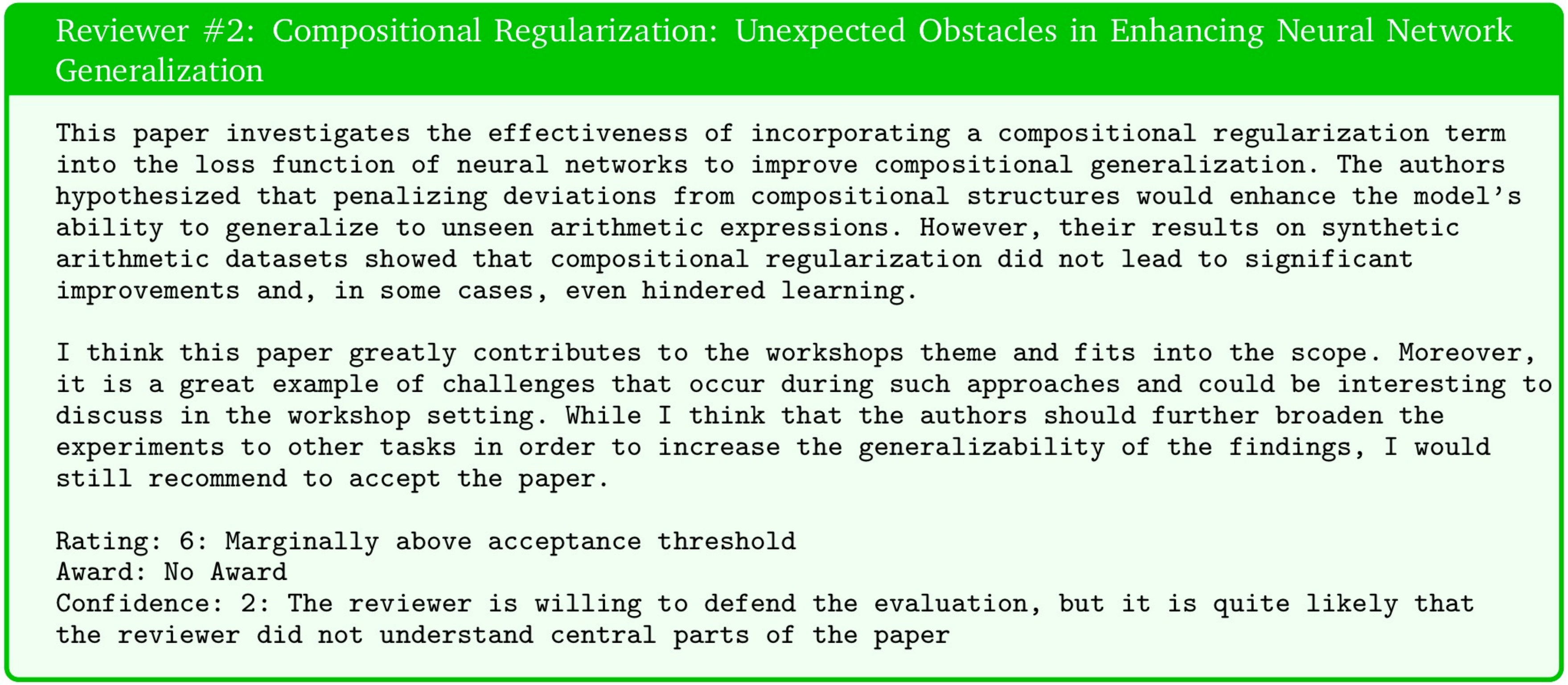
- Each stage: launched with a specific human prompt
- Quite long prompt

Parallelized Agentic Tree Search: Mimicking Human Exploration


Stage 1: Preliminary Investigation
Stage 2: Hyperparameter Tuning
Stage 3: Research Agenda Execution
Stage 4: Ablation Studies
Non-buggy nodes
Buggy nodes
Hyperparameter nodes
Ablation nodes
Replication nodes
Aggregation nodes
Best nodes
Refinement
Debugging

- Integrates tree search with LLM-driven workflows [Chan et al.’25] across four expt stages
- Agentic tree search enables deeper and more systematic exploration of hypotheses

Human Reviewer Feedback



Reviewer #2: Compositional Regularization: Unexpected Obstacles in Enhancing Neural Network Generalization

This paper investigates the effectiveness of incorporating a compositional regularization term into the loss function of neural networks to improve compositional generalization. The authors hypothesized that penalizing deviations from compositional structures would enhance the model's ability to generalize to unseen arithmetic expressions. However, their results on synthetic arithmetic datasets showed that compositional regularization did not lead to significant improvements and, in some cases, even hindered learning.

I think this paper greatly contributes to the workshops theme and fits into the scope. Moreover, it is a great example of challenges that occur during such approaches and could be interesting to discuss in the workshop setting. While I think that the authors should further broaden the experiments to other tasks in order to increase the generalizability of the findings, I would still recommend to accept the paper.

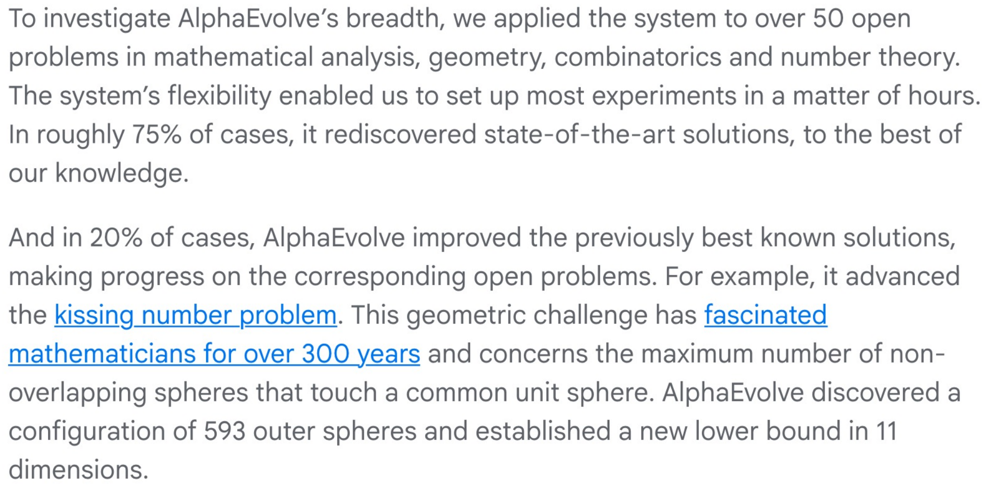
Rating: 6: Marginally above acceptance threshold
Award: No Award
Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Outline

- Introduction
- A focus on AI Scientist-v2 – main principles
- **Investigating the next stages (foresight)**
- Conclusion

Next AI “Technical” Stages

- Evolving very quickly, starting from LLMs (which are relatively recent)
- All research domains are gradually impacted
  - Depends on the quality of the Metropolis-Hasting-like proposal produced by the retrained LLM within the complex space related to the research domain under consideration (ex.: Coding, Mathematics)
- Quantity/quality of data for learning, computing resources



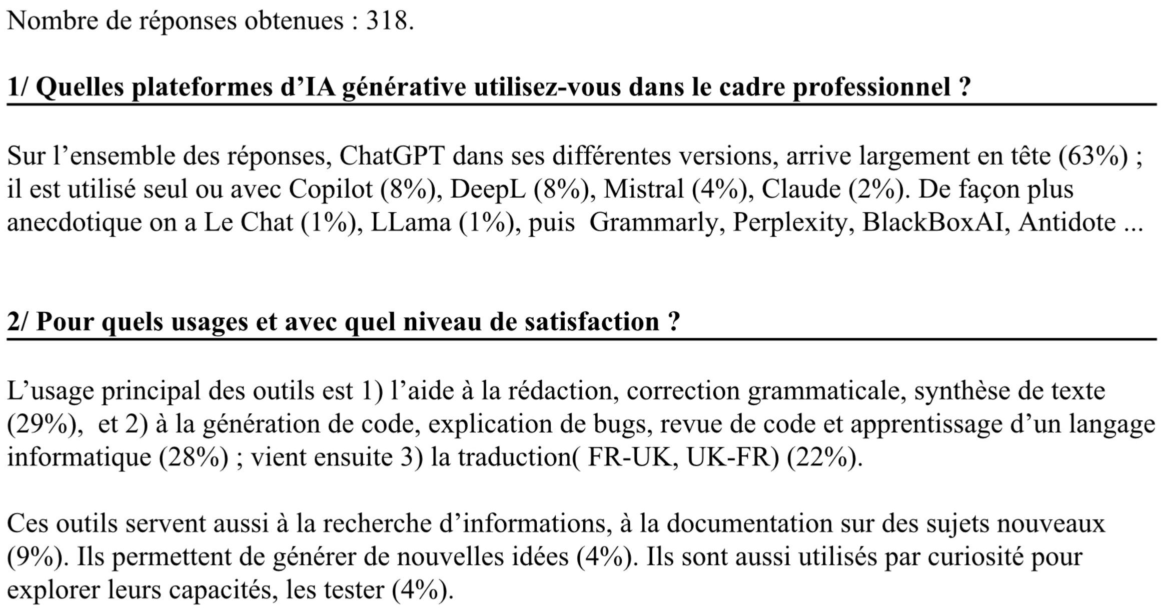
To investigate AlphaEvolve’s breadth, we applied the system to over 50 open problems in mathematical analysis, geometry, combinatorics and number theory. The system’s flexibility enabled us to set up most experiments in a matter of hours. In roughly 75% of cases, it rediscovered state-of-the-art solutions, to the best of our knowledge.

And in 20% of cases, AlphaEvolve improved the previously best known solutions, making progress on the corresponding open problems. For example, it advanced the kissing number problem. This geometric challenge has fascinated mathematicians for over 300 years and concerns the maximum number of non-overlapping spheres that touch a common unit sphere. AlphaEvolve discovered a configuration of 593 outer spheres and established a new lower bound in 11 dimensions.

AlphaEvolve (Google DeepMind)
May 14, 2025

Positioning of the Future Human Scientist

- AI/LLM tools used frequently



Nombre de réponses obtenues : 318.

1/ Quelles plateformes d'IA générative utilisez-vous dans le cadre professionnel ?

Sur l'ensemble des réponses, ChatGPT dans ses différentes versions, arrive largement en tête (63%) ; il est utilisé seul ou avec Copilot (8%), DeepL (8%), Mistral (4%), Claude (2%). De façon plus anecdotique on a Le Chat (1%), LLama (1%), puis Grammarly, Perplexity, BlackBoxAI, Antidote ...

2/ Pour quels usages et avec quel niveau de satisfaction ?

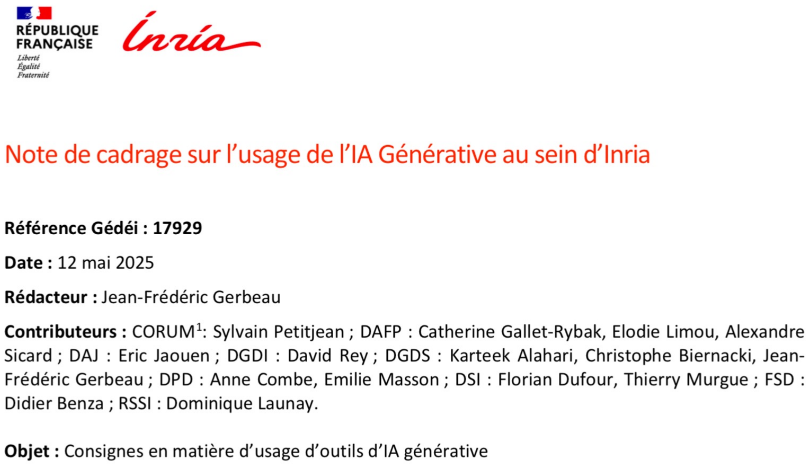
L'usage principal des outils est 1) l'aide à la rédaction, correction grammaticale, synthèse de texte (29%), et 2) à la génération de code, explication de bugs, revue de code et apprentissage d'un langage informatique (28%) ; vient ensuite 3) la traduction(FR-UK, UK-FR) (22%).

Ces outils servent aussi à la recherche d'informations, à la documentation sur des sujets nouveaux (9%). Ils permettent de générer de nouvelles idées (4%). Ils sont aussi utilisés par curiosité pour explorer leurs capacités, les tester (4%).

- Will replace (at least) incremental research discoveries => focus on more complex problems
- Good AI-based research will rely on good human-based prompts => prompt engineering

Regulations, Environmental and Other Considerations

- Classic questions related to the use of LLMs: data propriety, environmental impact, computing resources
- Authorship of AI-based research?
- High speed evolution of AI regulation (AI-act)
- Many research institutions publish their own internal recommendations



Note de cadrage sur l’usage de l’IA Générative au sein d’Inria

Référence Gédéi : 17929

Date : 12 mai 2025

Rédacteur : Jean-Frédéric Gerbeau

Contributeurs : CORUM[1]: Sylvain Petitjean ; DAFP : Catherine Gallet-Rybak, Elodie Limou, Alexandre Sicard ; DAJ : Eric Jaouen ; DGDI : David Rey ; DGDS : Karteek Alahari, Christophe Biernacki, Jean-Frédéric Gerbeau ; DPD : Anne Combe, Emilie Masson ; DSI : Florian Dufour, Thierry Murgue ; FSD : Didier Benza ; RSSI : Dominique Launay.

Objet : Consignes en matière d’usage d’outils d’IA générative

Outline

- Introduction
- A focus on AI Scientist-v2 – main principles
- Investigating the next stages (foresight)
- **Conclusion**

- Impressive advances in AI-based research

- Be aware of these advances to
  - use these tools for some kinds of research (incremental research)
  - do human research otherwise (avoid incremental research)

- But these are just a few early ideas to be discussed. . .

Thank you!