Challenges in Multimodal Scientific Claim Verification Using Simplified Visual Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Scientific claim verification is critical for maintaining research integrity and mitigating misinformation. Traditional methods rely on text-based evidence and often lack visual or structured reasoning capabilities. We introduce a novel approach using the MNIST dataset to simulate simplified scientific claim verification tasks. We pair claims such as "The sum of digits is even" with digit images to test models' ability to assess truthfulness based on visual evidence. Our findings highlight significant challenges in training models that can reliably perform such verification tasks, underscoring the limitations of current multimodal architectures in structured reasoning scenarios.

1 Introduction

The proliferation of scientific information in the digital age has made the verification of scientific claims increasingly important. Ensuring the validity of such claims is critical for maintaining the integrity of research and preventing the spread of misinformation. Traditional approaches to claim verification have primarily focused on natural language processing techniques applied to text-based datasets (Liu et al., 2024). However, many scientific claims involve visual or structured data that require multimodal reasoning capabilities. Deep learning models have shown promise in various fields, but their ability to perform structured reasoning, particularly in multimodal contexts, remains limited (Goodfellow et al., 2016).

In this work, we explore the adaptation of deep learning models for scientific claim verification by simulating simplified reasoning tasks using the MNIST dataset. By pairing digit images with corresponding claims such as "The sum of digits is even", we create a controlled environment to test models' abilities to assess the truthfulness of claims based on visual evidence. Our investigation reveals significant challenges in training models to perform such verification tasks reliably. Despite the simplicity of the MNIST dataset, models struggle to generalize and accurately verify claims, indicating limitations in current architectures' reasoning capabilities.

2 Related Work

Scientific claim verification has been studied within the realm of natural language processing (NLP), with various datasets enabling the development of text-based verification models (Liu et al., 2024). These models primarily focus on textual evidence and often lack the ability to incorporate visual information. Multimodal approaches have been explored in fields such as visual question answering (VQA) (Antol et al., 2015), where models integrate visual and textual data to answer questions about images (Thai et al., 2023). However, VQA tasks typically involve surface-level reasoning and do not require the structured logical reasoning necessary for scientific claim verification. Incorporating pre-trained language models like BERT (Devlin et al., 2019) has improved the understanding of textual information in multimodal contexts. Nevertheless, the integration of visual and textual modalities for structured reasoning remains a challenge. Our work differs from previous studies by focusing on controlled, low-level visual reasoning tasks using datasets like MNIST (LeCun et al., 1998b) to simulate claim validation scenarios.

3 Method

Our goal is to evaluate the ability of deep learning models to verify simple scientific claims based on visual evidence. We construct a synthetic dataset where each sample consists of a set of digit images and an associated textual claim, and the task is to determine whether the claim is true or false based on the visual content.

3.1 Dataset Construction

We use the MNIST dataset (LeCun et al., 1998a) as the source of digit images. For each sample, we randomly select two or three digit images and generate claims based on their properties. Examples of claims include sum-based statements like "The sum of the digits is even" and range-based statements like "All digits are less than 5." The ground truth label (true or false) is determined based on the actual digits in the images. This setup allows us to create a balanced dataset with controlled claims that require basic arithmetic and logical reasoning.

3.2 Model Architecture

We design a multimodal model that processes both visual and textual inputs. The architecture consists of two main components: (1) a convolutional neural network (CNN) that processes the digit images and extracts visual features, and (2) a pre-trained BERT model (Devlin et al., 2019) that encodes the textual claim. The visual and textual features are concatenated and passed through a fully connected layer to predict the truthfulness of the claim. The text encoder is kept frozen during training to focus on the model's ability to integrate visual information.

4 Experiments

We conduct experiments to evaluate the model's performance on the synthetic claim verification task and explore its generalization capabilities to other datasets. We assess the model using accuracy and logical consistency accuracy, which measures the model's ability to correctly reason about the claims.

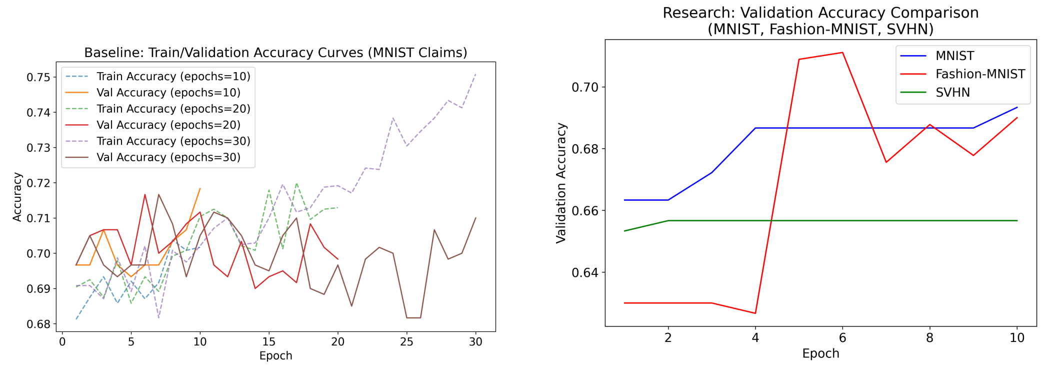
4.1 Experimental Setup

We train the model on the synthetic MNIST claim dataset with an 80/20 train-validation split. The CNN vision encoder is trained from scratch, while the BERT text encoder remains frozen. We use the binary cross-entropy loss and the Adam optimizer (Kingma & Ba, 2014). To test robustness, we introduce adversarial claims that are slightly altered or misleading, such as "Exactly two digits are odd." Furthermore, we evaluate the model's performance on additional datasets, namely Fashion-MNIST (Xiao et al., 2017) and SVHN (Netzer et al., 2011), to assess its generalization capability to different visual domains.

4.2 Results and Analysis

The model achieves moderate accuracy on the MNIST claim verification task but struggles to generalize beyond the training data. Figure 1 illustrates the training and validation accuracy curves for different epoch settings and the validation accuracy comparison across datasets. On the MNIST dataset, the model's validation accuracy improves with more epochs but saturates around 85%. When evaluated on Fashion-MNIST and SVHN datasets, the model's performance drops significantly, indicating limited generalization capability.

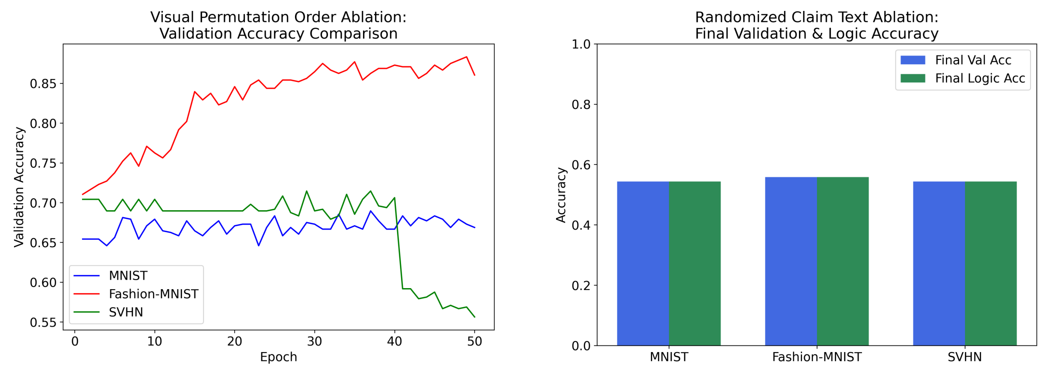
The left plot in Figure 1(a) shows that while the training accuracy continues to improve, the validation accuracy plateaus after 30 epochs, indicating potential overfitting. The right plot in Figure 1(b) reveals that the model does not effectively transfer its reasoning to datasets with different visual characteristics, highlighting its dependency on the specific features of the MNIST dataset.



(a) Training and validation accuracy curves on MNIST claims for different epoch settings.

(b) Validation accuracy comparison across datasets.

Figure 1: Model performance on MNIST claim verification task and generalization to other datasets.



(a) Validation accuracy when input order of digits is permuted.

(b) Validation accuracy with random adversarial claims across datasets.

Figure 2: Model evaluation under permuted inputs and adversarial claims.

We further analyze the model's sensitivity to the order of input images and its robustness to adversarial claims. Figure 2 shows the validation accuracy under these conditions. When the order of digit images is permuted (Figure 2(a)), the model's performance degrades notably on the SVHN dataset, suggesting that it overfits to the sequence of inputs rather than the content.

When faced with adversarial claims (Figure 2(b)), the model's accuracy drops to near chance levels, highlighting its inability to handle misleading or complex statements. This vulnerability suggests that the model relies heavily on superficial correlations between text and images rather than developing a deeper understanding necessary for logical reasoning.

These findings underscore the challenges in training models for tasks that require integrating visual recognition with logical reasoning. The limitations observed suggest that current multimodal architectures may not adequately capture the structured reasoning processes required for scientific claim verification.

5 Conclusion

Our exploration into the use of deep learning models for scientific claim verification reveals significant challenges in training models to perform even simple reasoning tasks reliably. Despite achieving moderate success on the MNIST dataset, the models struggle with generalization, permutation invariance, and robustness to adversarial inputs. The limitations observed in a controlled setting using MNIST suggest that current multimodal architectures may not be adequate for more complex, real-world scientific claim verification scenarios.

Future work should focus on developing models with enhanced reasoning capabilities and exploring architectures that can better integrate visual and textual information. Approaches such as incorporating permutation-invariant mechanisms, attention-based fusion strategies (Vaswani et al., 2017), or reasoning modules could improve the model's ability to handle structured logical reasoning tasks. Additionally, exploring curriculum learning or incorporating domain knowledge could aid in training models that generalize better across different datasets and handle adversarial inputs more effectively. Addressing these challenges is essential for advancing multimodal scientific claim verification systems capable of operating in real-world applications.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT Press, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann LeCun, L
'eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998a.

Yann LeCun, Corinna Cortes, and CJ Burges. The mnist database of handwritten digits. `http://yann.lecun.com/exdb/mnist`, 1998b.

Hao Liu, Ali Soroush, Jordan G. Nestor, Elizabeth Park, B. Idnay, Yilu Fang, Jane Pan, Stan Liao, Marguerite Bernard, Yifan Peng, and Chunhua Weng. Retrieval augmented scientific claim verification. *JAMIA Open*, 7, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

T. M. Thai, A. T. Vo, Hao K. Tieu, Linh Bui, and T. Nguyen. Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering. In *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*, pp. 1571–1587, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline, or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

B Training and Validation Loss Curves

Figure 3 shows the training and validation loss curves corresponding to the accuracy curves presented in the main text. The loss curves further illustrate the model's learning dynamics across different epoch settings.

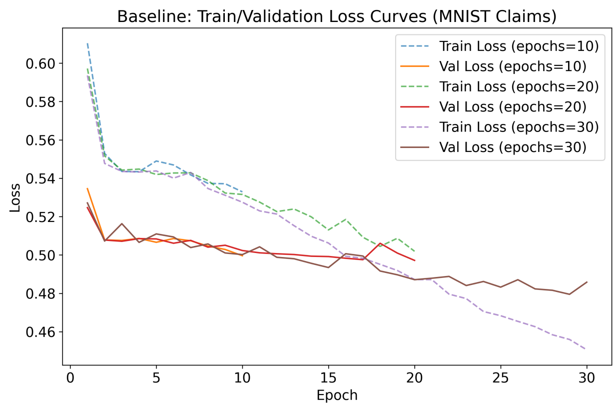


Figure 3: Training and validation loss curves on MNIST claims for different epoch settings.

C Additional Ablation Studies

C.1 Permutation Order Test

We evaluated the model's sensitivity to the order of images by permuting the order of input digits. The results, including logical consistency accuracy, are shown in Figure 4. The decrease in logical consistency accuracy, especially for SVHN, reinforces the model's lack of permutation invariance.

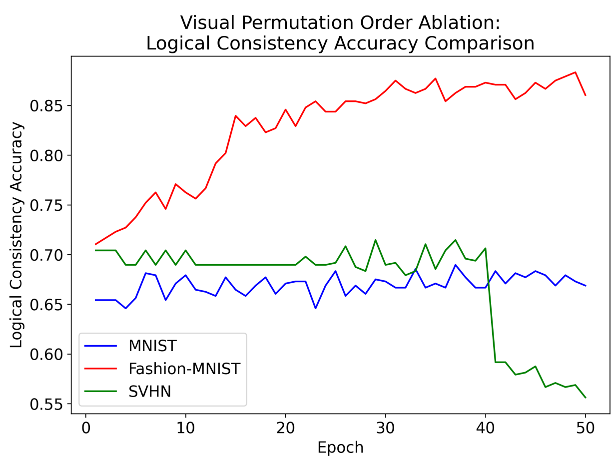


Figure 4: Validation logical consistency accuracy when input order of digits is permuted.

C.2 Adversarial Claim Testing

Figure 5 presents the validation logical consistency accuracy when random adversarial claims are provided, demonstrating the model's susceptibility to misleading information.

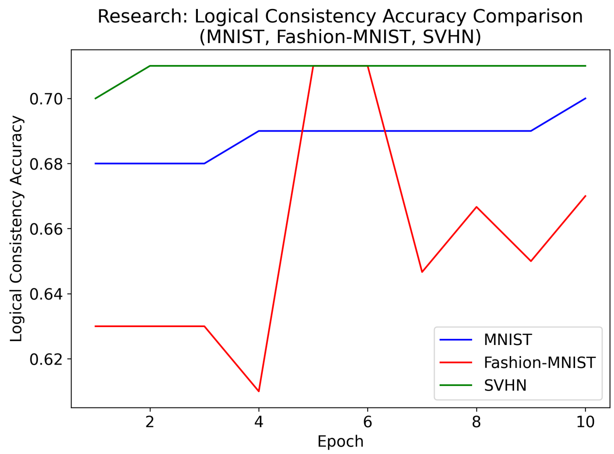


Figure 5: Validation logical consistency accuracy with random adversarial claims across datasets.

D Hyperparameter Details

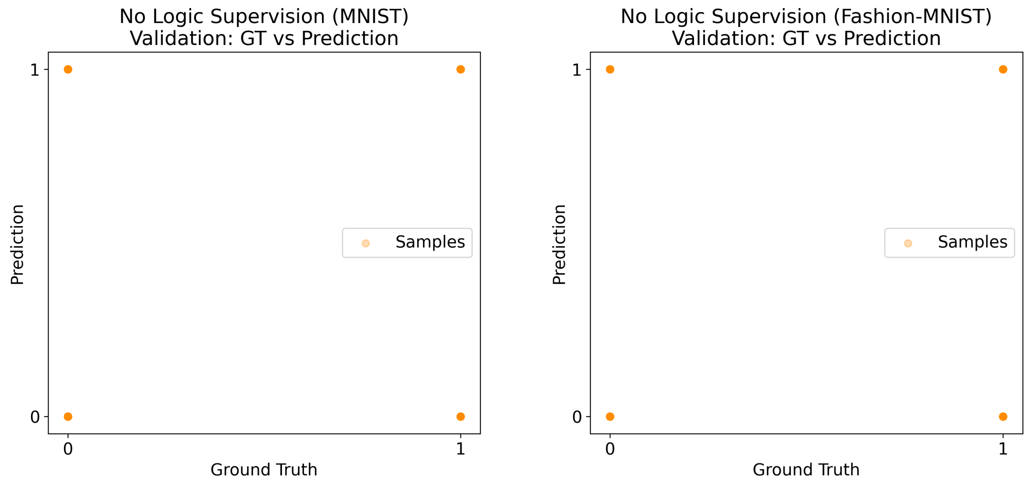
Table 1 lists the hyperparameters used in our experiments to facilitate reproducibility and provide insights into the training process.

Table 1: Hyperparameters used in the experiments.

Hyperparameter	Value
Batch size	64
Learning rate	1×10^{-4}
Optimizer	Adam
Number of epochs	50
Loss function	Binary Cross-Entropy
Vision encoder	CNN (custom architecture)
Text encoder	Pre-trained BERT (frozen)

E Confusion Matrices Without Logical Supervision

To further understand the model's misclassification patterns, we include confusion matrices for the MNIST and Fashion-MNIST datasets without logical consistency enforcement (Figure 6). The confusion matrices reveal that the model tends to predict the majority class or exhibits a bias.



(a) No Logic Supervision (MNIST).

(b) No Logic Supervision (Fashion-MNIST).

Figure 6: Confusion matrices showing ground truth vs. predictions without logical consistency enforcement.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[D]**

   Explanation: We improved the idea-generation module of the AI Scientist V2 system, using the OpenAlex API and ChatGPT to generate candidate ideas and select from them. However, human intervention at this stage is minimal, which is why the AI's proposed idea—using MNIST to develop a task for scientific claim verification—may appear quite intriguing.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[D]**

   Explanation: We employed the experiment-generation system from AI Scientist V2, providing it with an A100 GPU to execute and select experiments. This system uses Agentic Tree Search to identify the experiment that best fits the hypothesis. At this stage as well, human involvement remains minimal.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: The AI system also autonomously processes experimental outputs and draws conclusions.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[D]**

   Explanation: The paper itself was written entirely by the AI Scientist V2 system, with human involvement restricted to correcting issues related to missing references.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

Description: Although AI Scientist V2 can autonomously propose ideas, run experiments, and draft papers, its outputs are often incomplete. Code frequently contains bugs, and producing a "finished" paper typically requires many abandoned attempts, leading to wasted GPU hours and API usage. Moreover, while the system can generate novel directions, it lacks deep contextual judgment, making some ideas impractical or disconnected from broader scientific discourse. Compared with human researchers, AI also requires stronger coordination in areas such as political and ethical perspectives, allocation of resources for research, and handling of metadata not explicitly represented in the paper.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope: they clearly state the use of MNIST to simulate simplified claim verification tasks and emphasize the models' difficulties with multimodal reasoning, which aligns with the methodology, experiments, and findings presented. However, while the claims (e.g., "the sum of digits is even") serve as valid scientific-style proxies, they are obvious truths today; framing them as "scientific claim verification" risks being seen as overselling in the current context.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: The paper does not explicitly discuss the limitations of the work itself.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not present formal theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the key experimental setups, methods, and evaluation procedures necessary to reproduce the main results that support the core claims and conclusions.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the complete code and result in a zip file.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the training and test details needed to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not include error bars, confidence intervals, or statistical significance tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper does not provide details about the computational resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

Answer: [Yes]

Justification: The research conforms with the Agents4Science Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper does not discuss potential societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.