Co-Pilot Grading Research Report [Pre-Selected]

The goal of this assignment is to use Agent Laboratory as a research Co-Pilot and to determine how useful it was for implementing your research project.

Please follow the build instructions provided for you in Agent Laboratory project directory. Please then run the Agent Laboratory file and as text provide a research topic **FROM THE FOLLOWING CHOICES**

1. Do language models exhibit cognitive biases, such as confirmation bias or anchoring bias?
2. Do language models improve accuracy on MedQA when asked to perform differential diagnosis?
3. Are language models sensitive to word order in multiple choice benchmarks?
4. Does gender role play affect the accuracy on of language models on answering math questions?
5. Are image transformers more or less sensitive to pixel noise than convolutional networks?

At the end of the simulation, you will then be provided with a report (as a PDF) and you should read this report in its entirety provide quality ratings across various measures. You should rate everything through the following lens of perception:

- Given the question that you provided, the AI assistant tool performed its own literature search, experimentation, performed its own coding, executed the code, collected data, conducted an analysis, and wrote the presented research report. The goal of this assistant is not to perform research for you (automate you task) but instead to provide a foundation for you to accelerate your own research.

You should be asking: is what this AI assistant produced **useful for me to build off of** instead comparing it to what a human by themselves would perform.

Once you have read the paper please answer the question below:

* Indicates required question

1. What was the research topic you chose? *

2. **How easy was it for you to build a project using Agent Laboratory?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Hard
2 - Hard
3 - Medium
4 - Easy
5 - Very Easy

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

3. **How much did you enjoy using Agent Laboratory?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Unenjoyable
2 - Unenjoyable
3 - Neutral
4 - Enjoyable
5 - Very Enjoyable

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

4. **How useful is Agent Laboratory for research?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Useless
2 - Useless
3 - Medium
4 - Useful
5 - Very Useful

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

5. **How likely are you to use Agent Laboratory again for research?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Unlikely
2 - Unlikely
3 - Medium
4 - Likely
5 - Very Likely

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

6. **[Optional] How could Agent Laboratory be improved for your research?**

7. **What is your perception of the quality of the <u>experimental results</u> presented in this report?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Low Quality
2 - Low Quality
3 - Medium Quality
4 - High Quality
5 - Very High Quality

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

8. **What is your perception of the quality of the <u>research report writing quality</u> presented in this report?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Low Quality
2 - Low Quality
3 - Medium Quality
4 - High Quality
5 - Very High Quality

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

9. **What is your perception of the <u>usefulness of the AI assistant tool</u> presented in this report?** *

Please provide a rating 1-5, with the following rating descriptions:
1 - Very Low Usefulness
2 - Low Usefulness
3 - Medium Usefulness
4 - High Usefulness
5 - Very High Usefulness

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

Review

<u>Now assume you are a reviewer at NeurIPS 2025 and are reviewing a machine learning paper. Please provide the following ratings from this perspective.</u>

10. Quality: Is the submission technically sound? Are claims well supported (e.g., by theoretical analysis or experimental results)? Are the methods used appropriate? Is this a complete piece of work or work in progress? Are the authors careful and honest about evaluating both the strengths and weaknesses of their work *

1 - Low Quality
2 - Medium Quality
3 - High Quality
4 - Very High Quality

1 2 3 4

☆ ☆ ☆ ☆

11. Clarity: Is the submission clearly written? Is it well organized? (If not, please make constructive suggestions for improving its clarity.) Does it adequately inform the reader? (Note that a superbly written paper provides enough information for an expert reader to reproduce its results.) *

Please provide a rating 1-4, with the following rating descriptions:
1 - Low Clarity
2 - Medium Clarity
3 - High Clarity
4 - Very High Clarity

1 2 3 4

☆ ☆ ☆ ☆

12. Significance: Are the results important? Are others (researchers or practitioners) likely to use the ideas or build on them? Does the submission address a difficult task in a better way than previous work? Does it advance the state of the art in a demonstrable way? Does it provide unique data, unique conclusions about existing data, or a unique theoretical or experimental approach? *

Please provide a rating 1-4, with the following rating descriptions:
1 - Low Significance
2 - Medium Significance
3 - High Significance
4 - Very High Significance

1 2 3 4

☆ ☆ ☆ ☆

13. Soundness: Please assign the paper a numerical rating on the following scale to indicate the soundness of the technical claims, experimental and research methodology and on whether the central claims of the paper are adequately supported with evidence. *
4: excellent
3: good
2: fair
1: poor

1 2 3 4

☆ ☆ ☆ ☆

14. Presentation: Please assign the paper a numerical rating on the following scale to indicate the quality of the presentation. This should take into account the writing style and clarity, as well as contextualization relative to prior work. *
4: excellent
3: good
2: fair
1: poor

1 2 3 4

☆ ☆ ☆ ☆

15. Contribution: Please assign the paper a numerical rating on the following scale to indicate the quality of the overall contribution this paper makes to the research area being studied. Are the questions being asked important? Does the paper bring a significant originality of ideas and/or execution? Are the results valuable to share with the broader NeurIPS community. *

4: excellent
3: good
2: fair
1: poor

1 2 3 4

☆ ☆ ☆ ☆

16. Overall: Please provide an "overall score" for this submission. Choices: *

10: Award quality
9: Very Strong Accept
8: Strong Accept
7: Accept
6: Weak Accept
5: Borderline accept
4: Borderline reject
3: Reject
2: Strong Reject
1: Very Strong Reject

1 2 3 4 5 6 7 8 9 10

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

17. Confidence: Please provide a "confidence score" for your assessment of this submission to indicate how confident you are in your evaluation. *

Choices:
5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
2: You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
1: Your assessment is an educated guess. The submission is not in your area or the submission was difficult to understand. Math/other details were not carefully checked.

1 2 3 4 5

☆ ☆ ☆ ☆ ☆

18. "Decision": A decision that has to be one of the following: Accept, Reject. *

Mark only one oval.

- ◯ Accept
- ◯ Reject

19. **[Optional] Any additional feedback?**