Analysis of results in the ML research field

Author: Stefan Dood (S.G.dood@student.tudelft.nl)
Supervisor: David Tax, Chenxu Hao

TU Delft Delft University of Technology

Background

Increasing Submission Pressure

- Rapid growth in AI/ML research
- Increasing NeurIPS submissions
- Limited reviewer time

NeurIPS Checklist

- 19 mandatory questions
- Focus on reproducibility and transparency
- Requires manual verification

Research Question

Can an LLM determine whether a paper satisfies the NeurIPS 'Experiment statistical significance' check-list criterion?

- Can a LLM correctly classify papers as Yes/No/NA?
- How closely do LLM judgments align with human reviewers?

NeuRIPS Question

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:

Justification:

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, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

Methodology

Manual evaluation

- 72 papers with Yes, No, or NA classification
- 18 papers with justification

LLM evaluation

- Gemini
- Remove authors' checklist
- Insert paper into the prompt
- Gather the classification as given
- Repeat with second prompt

Conclusion

- Compare ground truth with LLM
- Calculate accuracy and F1-scores
- Compare justifications

Prompt Design

3 Total prompts First Prompt:

- Basic description of tasks
- Used the official guidelines
- Ran on 20 papers
- Too many false negatives

Second Prompt:

- Very similar to prompt 1
- Performed slightly better on 20 papers
- Has been ran on all 72 papers

Third prompt:

- Basic description of tasks
- No more guidelines
- Originally too lenient
- Ran on all 72 papers

Accuracy

prompt 2	precision	recall	F1-score
Yes	0.79	0.76	0.77
No	0.74	0.85	0.79
NA	1.00	0.60	0.75
Accuracy			0.78

prompt 3	precision	recall	F1-score
Yes	0.72	0.90	0.80
No	0.87	0.79	0.83
NA	1.00	0.60	0.75
Accuracy			0.81

Discussion

- 8 papers changed label
- NA has 1.00 precision
- The recall of Yes improved significantly

Justification

Comparison	Cohen's Kappa
Human-Human	0.89
Human-Prompt 2	0.62
Human-Prompt 3	0.67

Discussion

- Cohen's Kappa shows good agreement
- LLM is often more thorough
- LLM lacking in contextual awareness

Conclusion & Limitations

Limitations

- Small data set
- Non-expert ground truth creation
- Single LLM usage

Conclusion

- ~80% accurate
- Can be Gamed
- Tool to assist, not replace