
Alignment Research Critiquer

Nancy Vigil¹
Independent

Ashish Rai²
NYU

With
Jacques Thibodeau & Apart Research

Abstract

There are currently insufficient senior researchers in alignment to provide adequate mentorship for everyone looking to move into this field. Early career and independent alignment researchers may have difficulties identifying valuable directions for research and may not have access to academic reviewers. Given the advances in AI capabilities, it is urgent that we have more people working on AI safety research to reap as many benefits from advanced AI systems while minimizing harm.

To help alleviate these problems, we created the Alignment Research Critiquer to be a tool for early career and independent alignment researchers to have access to high-quality feedback loops. As of now, our Alignment Research Critiquer provides high-quality peer reviews for mechanistic interpretability. We have designed a set of prompts that, given a research paper, will produce an overview of the paper's strengths, weaknesses, areas for improvement, and potential directions for future research.

Keywords: Research tooling, AI safety, alignment, mechanistic interpretability

1. Tool overview

The tool is live at: <https://huggingface.co/spaces/raishish/paper-reviewer> (prompts gpt-4o-mini).

It has OpenAI and Anthropic integrations. It has two sets of prompts:

- Alignment (more general, detailed, provides reasoning and evidence)
- Mech Interp (specific, concise, actionable)

- this generates two reviews - high and low quality (inspired by Neel Nanda's [twitter post](#))

Spaces raishish / **papex-reviewer** like 0 Running

App Files Community Settings

Alignment Research Critiquer

An LLM-based app to critique AI alignment research papers.

79.8 KB ↓

Research domain

alignment (general)

Clear

Submit

Critique

Paper Overview and Claims

The paper "INTERPRETABILITY IN THE WILD: A CIRCUIT FOR INDIRECT OBJECT IDENTIFICATION IN GPT-2 SMALL" by Kevin Wang et al. presents several key claims regarding the mechanistic interpretability of the GPT-2 small model, specifically focusing on how it performs the task of indirect object identification (IOI). Below is an overview of the paper's claims, along with extracted text as evidence and reasoning for each step.

1. **"Claim of Bridging the Gap in Interpretability Research"**:

The authors assert that their work bridges the gap between previous research that either focuses on simple behaviors in small models or describes complex behaviors in larger models. They state:

> "However, most previous work either focuses on simple behaviors in small models or describes complicated behaviors in larger models with broad strokes."

****Reasoning****: This claim sets the stage for the authors' contribution, indicating that they aim to provide a detailed mechanistic understanding of a specific behavior in a well-known model, GPT-2 small.

2. **"Identification of a Circuit for IOI"**:

The authors claim to have identified a circuit composed of 26 attention heads that are responsible for performing the IOI task. They write:

> "We discover a circuit of 26 attention heads- 1.1% of the total number of (head, token position) pairs- that completes the bulk of this task."

****Reasoning****: This claim is significant as it suggests that the authors have successfully reverse-engineered a specific behavior of the model, providing a concrete example of how attention heads contribute to a defined task.

Use via API · Built with Gradio

Spaces raishish / **papex-reviewer** like 0 Running

App Files Community Settings

Alignment Research Critiquer

An LLM-based app to critique AI alignment research papers.

79.8 KB ↓

Research domain

mech_interp

Clear

Submit

Critique

High Quality Review

****Reviewer Comments on "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 Small**"**

****Summary****

The authors present a comprehensive study on the mechanistic interpretability of the GPT-2 small model, focusing specifically on the task of indirect object identification (IOI). They identify a circuit of attention heads that contribute to the model's performance on this task, employing a combination of interpretability techniques, including path patching and knockout experiments. The paper is well-structured, and the findings are significant, providing insights into the internal workings of transformer-based language models.

****Strengths****

1. **"Novelty and Relevance"**: The paper addresses a critical area in machine learning—mechanistic interpretability of large language models. The focus on a specific task (IOI) allows for a detailed exploration of the model's behavior, which is often obscured in broader analyses.

2. **"Methodological Rigor"**: The authors employ a systematic approach to identify and validate the circuit responsible for IOI. The use of path patching and knockout experiments is well-justified and effectively demonstrates the contributions of different attention heads.

3. **"Comprehensive Analysis"**: The paper provides a thorough examination of the identified circuit, categorizing attention heads into distinct classes and elucidating their roles. This level of detail is commendable and contributes to a deeper understanding of the model's behavior.

Use via API · Built with Gradio

2. Why this tool should exist

Responses from SoTA LLMs are often of not much value to an inexperienced alignment researcher seeking feedback on their work or research questions. The base model tends to generate vague, and “fluff-filled” recommendations. There is a lack of specific, actionable insights needed to improve the research. Because of this, any feedback received has limited value.

Furthermore, academic reviews are a critical step in research. However, it takes a lot of time, effort, and connections to receive high-quality reviews. Many independent researchers do not have access to timely high-quality academic reviews.

The Alignment Research Critiquer allows junior researchers without mentors to access high-quality critiques of any paper, including their own work and research directions. Additionally, this tool may reduce the time a mentor may need to spend with individual mentees. For those with formal mentors, this tool can help researchers arrive more prepared for meetings by teaching mentees how to begin identifying weaknesses in papers and potential research directions, reducing the cognitive load needed for each mentee. This could allow some mentors to take on more mentees.

3. Results and next steps

3.1 Results

We crafted a set of prompts that produces a reasonable critique of a paper and subsequently identifies further directions for research using GPT-4o, a model with fewer reasoning skills than Sonnet 3.5 (from empirical results).

Our testing phase consisted of creating prompts and iterating to achieve better results. We explored the effectiveness of prompt chaining and chain-of-thought (CoT) methods within GPT-4o. These techniques were used to elicit the model's reasoning capabilities and increase the specificity of its outputs.

These tests allowed us to arrive at the final set of prompts. We came to understand the potential of a model with limited reasoning capability and the limitations of using just a single model.

We encountered the following limitations. In prompt set 1 (the prompt sets can be found in the Appendix), we were able to produce a strong critique of the paper but received vague research questions and directions for further research. This may be due to the initialization prompt, a lack of access to the newest mechanistic interpretability research, guidelines for what makes a good paper in this specific field, and common mistakes in mechanistic interpretability research. A comparative analysis using a prompt from a third party showed that Sonnet 3.5 gave more meaningful, specific, and actionable feedback than GPT-4o. This needs to be tested further.

Additionally, parsing PDF files proved more challenging than initially anticipated, adding unexpected technical hurdles in data extraction and interpretation.

3.2 Next Steps

Our long-term vision is to create a capable "personalized" tool that provides preliminary feedback for early-career and independent alignment researchers and supplements formal mentorship. We've identified key areas for improvement, such as incorporating more up-to-date knowledge and enhancing the AI's understanding

of specific research fields. We aim to have the Alignment Research Critiquer draw insights from large databases of academic reviews, and understand the nuances of additional areas of alignment research. We will also tackle another challenge we faced: tokenizing complex PDF documents. Our tool has the potential to accelerate research in technical alignment, reduce barriers that prevent junior researchers from producing meaningful research, and provide a space to receive actionable feedback for researchers who lack access to academic reviews.

Our plan is as follows for the coming three months:

1. Add access to more research in mechanistic interpretability, including recent research and the content of papers referenced in the research article being analyzed.
2. Use few-shot prompting to produce higher-quality answers
3. Glean insights into what makes a mechanistic interpretability “good” by interviewing top mechanistic interpretability researchers.
4. Compile the insights to provide guidelines for Sonnet 3.5
5. Adding support for parsing PDF files
6. Introduce critiques and research direction for other alignment research domains such as evaluations or scalable oversight.

4. References

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., ... & Zou, J.

(2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, AIOa2400196.

Prompt Engineering Guide – Nextra. (n.d.). www.promptingguide.ai.

<https://www.promptingguide.ai/>

Zaumanis, M. (n.d.). *How to ML Review - A brief Guide*. Docs.google.com.

Retrieved July 29, 2024, from

https://docs.google.com/document/d/1ApQjLIP29vHhs3uOWO3V6sn0GLJnPPfAPdeoz5_Kwys/mobilebasic

Zaumanis, M. (2023, November 30). *ChatGPT for Research Paper Review*. Peer Recognized.

<https://peerrecognized.com/chatgpt-for-research-paper-review/>

5. Appendix

The set of prompts tried in this experiment and various iterations can be accessed at this [link](#).