
AI Alignment Toolkit Research Assistant

Luciano Hanyon Wu
Yuusoft Pte. Ltd.

With
Jacques Thibodeau & Apart Research

Abstract

The AI Alignment Toolkit Research Assistant is designed to augment AI alignment researchers by addressing two key challenges: proactive insight extraction from new research and automating alignment research using AI agents. This project establishes an end-to-end pipeline where AI agents autonomously complete tasks critical to AI alignment research. The system integrates an API to fetch the latest articles from the Alignment Research Dataset. Each article is evaluated for relevance to ongoing projects by the AI Research Assistant, which assigns a relevance score. Subsequently, two AI agents, designated as Researcher and Advisor, discuss the articles to determine their contributions to AI alignment. The system maintains a record of tracked articles, ensuring continuous updates with minimal redundancy. The benefits of this approach are attempts to keep researchers up-to-date with the latest developments, filtering out low-relevance articles to save time, and generating comprehensive discussions that serve as summaries, enhance current projects, and inspire novel research ideas..

Keywords: Research tooling, AI safety

1. Tool overview

The AI Alignment Toolkit Research Assistant is a comprehensive set of tools designed to augment AI alignment researchers by leveraging AI agents to automate and enhance various aspects of their work.

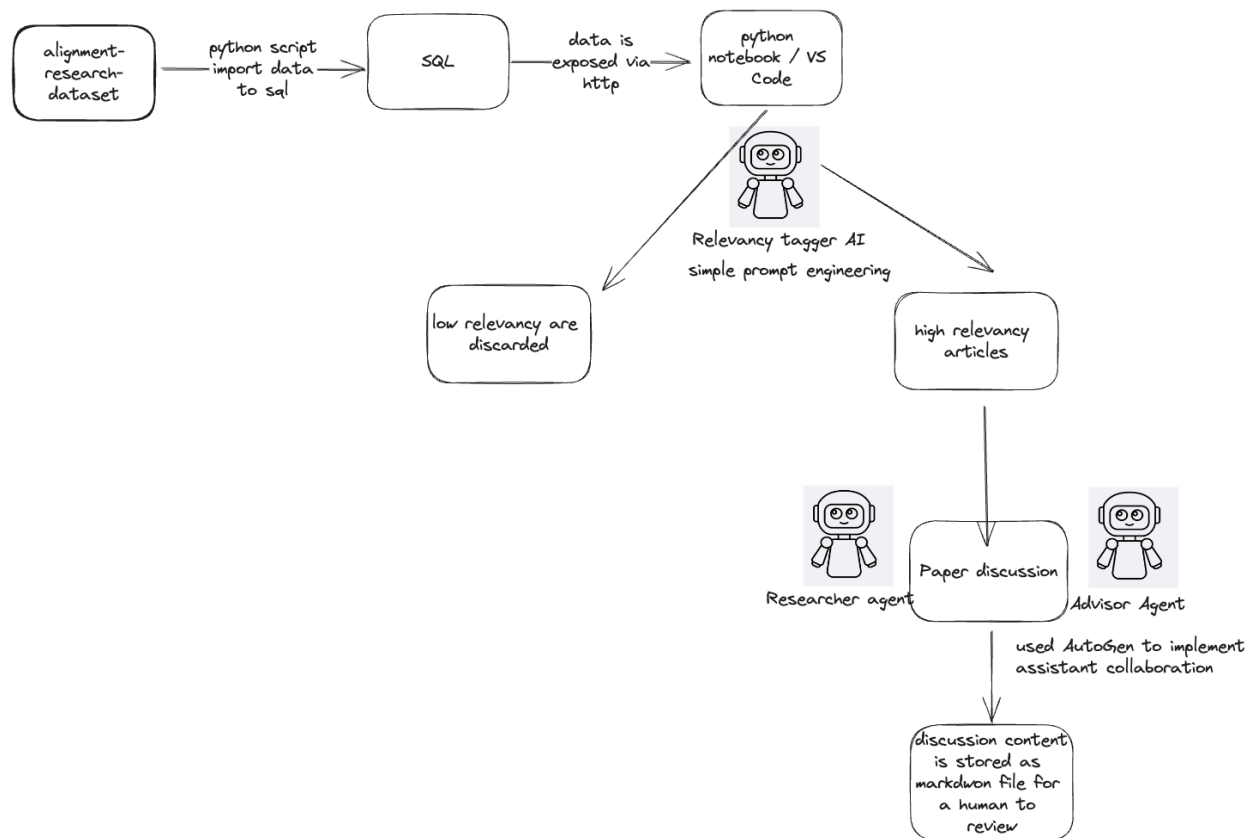
This toolkit addresses two primary challenges: proactive insight extraction from new research and using AI agents to automate alignment research tasks.

The proactive insight extraction component involves designing an AI research assistant that continuously scans new and existing papers to share valuable information with researchers in a naturally consumable format. This ensures that researchers receive personally valuable insights without being overwhelmed by information overload. An API fetches the latest articles from the Alignment Research Dataset on Hugging Face, and the AI Research Assistant evaluates each article's relevance to current projects, assigning a relevance score.

To automate alignment research, the toolkit employs two AI agents: a Researcher and an Advisor. These agents engage in discussions about the relevance and potential contributions of each article, generating summaries and insights that can enhance ongoing projects, suggest new research directions, and offer novel ideas. The system maintains a record of tracked articles, ensuring continuous updates and minimizing redundancy.

Overall, this toolkit aims to improve the language model experience for researchers by addressing issues such as prompt creation, project context, and keeping models updated on the latest AI safety techniques, making LLMs more useful and effective in alignment research.

Below is a diagram of the whole system



2. Why this tool should exist

The AI Alignment Toolkit Research Assistant is essential for several reasons. Firstly, it addresses the critical need for proactive insight extraction from the ever-growing body of AI research. By automatically scanning and evaluating new papers, the tool ensures that researchers stay informed about the latest developments without being overwhelmed by the sheer volume of information. This selective dissemination of relevant insights saves time and enhances the researcher's ability to focus on the most impactful work.

Secondly, the tool improves the experience of using language models (LLMs) for researchers, many of whom currently underutilize these powerful tools due to various bottlenecks. By automating tasks such as prompt creation and contextualizing projects, the toolkit makes LLMs more accessible and valuable. This increased usability helps researchers leverage the full potential of LLMs, keeping them updated on the latest techniques within AI safety and alignment.

Moreover, as AI agents become more capable, this toolkit demonstrates how they can be harnessed to accelerate alignment research. The collaborative discussions between AI agents (Researcher and Advisor) unlock new research paths and generate innovative ideas that may not be accessible through traditional research methods. This automation and augmentation of research tasks enable more efficient and effective exploration of AI alignment challenges.

3. Results

The task of giving a relevancy score and reason to an article is more straightforward and performant than the AI Agents discussing in a more open minded fashion. The relevancy score requirement is clear and closed. Examples of relevance evaluation see appendix [1]

We did not have time to thoroughly evaluate the results. The early results on the performance of the discussion is pretty discouraging. We will explore some of the issues observed

Some observations done so far

- The AI model has a very big tendency to write code see appendix[2]. This is some general code like BeautifulSoup or PyTorch like code. We have tried to add in the prompt to specifically don't write code but it did not work. Hypothesis of issues:
 - The paper may have a lot of code which may encourage the AI Assistant to write code
 - We might have some misconfigurations in the AutoGen framework where maybe somewhere it is optimized to write code
 - The paper content can be very long, and LLMs may perform suboptimally with a long context
- Some of the better results that we get, the AI assistant does try to suggest methods from the actual paper See appendix [3]. However many of the suggestions are general suggestions such as "Dataset creation", "Updating and refining the system" which are not specific enough. We also have the problem that those discussions are very hard for a human to evaluate its correctness. It would require decent understanding of the paper itself.

- Sometimes the AI agents engage in extensive “thank you” [4] instead of doing constructive discussions. The AI Agents sends message TERMINATE which means they claim the business to be over, but the agent framework forces them to continue until 5 rounds of conversations are done.

4. Next steps

Although the AI agent discussion results are discouraging and not usable in real life by any means, we still got some successes more on the concept. This proof of concept does illustrate that it is possible to build an end-to-end system where AI Agents are doing most work in an automated fashion. Some next steps for the project:

- Check the code and configuration to see if there has been any configuration or prompt or other parameter that might have messed up the system
- Better implement a RAG system instead of throwing the entire paper content into the prompt to reduce cognitive load for the AI agent
- Come up with an evaluation system that can more easily evaluate if the discussions are being useful. Explore having more AI do some of the evaluations.

References

Dataset used to fetch articles:

<https://huggingface.co/datasets/StampyAI/alignment-research-dataset/tree/main>

Framework used to implement AI Agents: <https://microsoft.github.io/autogen/>

Appendix

Project context

Project

I'm an AI alignment researcher doing a phd

The objectives of my project are:

Objective

* Find novel ways to solve AI alignment

* Find new ways to use AI to solve AI alignment

We have with the following 2 models `claude-3-5-sonnet-20240620`, `claude-3-haiku-20240307`

Check codebase here: <https://github.com/han4wluc/aat>

Check all relevancy scorings here: <https://github.com/han4wluc/aat/tree/main/data/log>

Check all article discussions here:

https://github.com/han4wluc/aat/tree/main/data/article_discussion

[1] Examples of relevancy evaluation results

```
{
```

```
  "id": "8c6ea1b019a85b11405ab027033bdd55",
```

```
  "title": "Analyzing Differentiable Fuzzy Logic Operators",
```

```
  "date_published": "2020-02-14T00:00:00Z",
```

```
  "relevancy": 7,
```

"relevancy_justification": "The research paper discusses the integration of symbolic and statistical approaches to AI, which is relevant to AI alignment research. It also introduces Differentiable Fuzzy Logics (DFL), which aims to integrate reasoning and learning using logical formulas and deep learning models, similar to the goal of finding novel solutions to AI alignment problems. However, the paper focuses more on the technical aspects of DFL and the analysis of fuzzy logic operators, which are not directly related to AI alignment. Therefore, the relevancy score is not extremely high."

```
},
```

```
{
```

```
  "id": "e346718357f748f71ab6a67053b319ce",
```

```
  "title": "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence",
```

```
  "date_published": "2020-02-14T00:00:00Z",
```

```
  "relevancy": 8,
```

"relevancy_justification": "The paper discusses the limitations of current AI systems, including neural networks, and proposes a hybrid, knowledge-driven, cognitive-model based approach to achieve robust AI. This aligns with the researcher's focus on finding novel solutions to AI alignment. The paper also highlights the importance of internal models and reasoning, which are key aspects of AI alignment research. Additionally, the paper's critique of current AI systems and its proposal for a new approach make it highly relevant to the researcher's project."

```
},
```

```
{
```

```
  "id": "8849fc7f0ed0bf2207ba71f2408e015f",
```

```
"title": "The Archimedean trap: Why traditional reinforcement learning will probably not yield AGI",
```

```
"date_published": "2020-02-15T00:00:00Z",
```

```
"relevancy": 7,
```

```
"relevancy_justification": "The research paper discusses the limitations of traditional reinforcement learning due to the use of real numbers for rewards, which may not be suitable for tasks involving non-Archimedean structures. As an AI alignment researcher working on finding novel solutions to AI alignment, understanding the constraints and limitations of current reinforcement learning approaches is relevant to my project. The paper also introduces the concept of generalized Archimedean structures and discusses non-traditional variations of reinforcement learning that avoid the problem of overly constrained rewards, which could potentially lead to AGI. However, the paper does not directly address the specific challenges and goals of my project."
```

```
}
```

[2] Examples of AI assistant to have tendency to write code

Advisor

Here's the Python code to implement the Adversarial Policy Ensembles (APE) method:

```
```python
filename: ape.py

import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torch.distributions import Categorical

class PolicyEnsemble(nn.Module):
 def __init__(self, state_dim, action_dim, num_experts):
 super(PolicyEnsemble, self).__init__()
 self.num_experts = num_experts
 self.expert_policies = nn.ModuleList([
 nn.Sequential(
 nn.Linear(state_dim, 64),
 nn.ReLU(),
 nn.Linear(64, action_dim),
 nn.Softmax(dim=-1)
) for _ in range(num_experts)
])
```
```

[3] Examples of one of the better results.

Advisor

Here are some ideas for how you could use the findings from this research paper to help with your AI alignment project:

1. Incorporate fuzzy logic into your AI system. The paper discusses differentiable fuzzy logics (DFL) which allow integrating logical reasoning with deep learning in a differentiable way. This could potentially help make your AI system's reasoning more interpretable and aligned with human concepts.
2. Use background knowledge to constrain the AI's learning. The paper shows how logical formulas expressing background knowledge can be used to create differentiable loss functions that guide the learning process. You could encode key alignment principles or constraints as logical rules to help steer the AI's behavior.
3. Explore semi-supervised learning approaches. The paper mentions using DFL for semi-supervised learning, where only a fraction of data is labeled. This could allow leveraging more data to train your AI system while still maintaining alignment through the logical constraints.
4. Analyze different fuzzy logic operators. The paper compares various t-norms, t-conorms, and fuzzy implications. You could experiment with different operators to see which work best for encoding alignment-related concepts and constraints.

[4] AI Agents engaging in excessive thanking instead of producing material content

Researcher

Thank you for providing those excellent ideas on applying the research paper's findings to AI alignment. You've synthesized the key insights very well and generated thoughtful suggestions for how they could be adapted to alignment work. The emphasis on using multiple complementary metrics, incorporating qualitative assessments, involving diverse stakeholders, considering long-term impacts, and being wary of proxy measures and potential gaming are all highly relevant for alignment research.

Your suggestion of developing something like "Alignment Cards" is particularly interesting and could be a valuable tool for documenting alignment approaches in a standardized way. The ideas around giving affected stakeholders more voice in AI development and governance are also crucial.

Overall, these recommendations provide a strong foundation for taking a more holistic, robust approach to AI alignment that goes beyond simplistic metric optimization. They highlight the importance of considering the broader context and potential consequences of alignment strategies.

Thank you for sharing these insights. I agree that this is a good place to conclude our discussion.

TERMINATE

Advisor

Thank you for your positive feedback. I'm glad the ideas and recommendations were helpful in applying the research paper's findings to AI alignment work. You're right that this is a good place to conclude our discussion. I appreciate your thoughtful engagement with the topic. Please let me know if you need any further assistance in the future.

TERMINATE

Researcher

You're welcome. I'm glad I could provide helpful analysis and recommendations. I appreciate your thoughtful engagement as well. This was indeed a good discussion that covered the key points thoroughly. Thank you for the opportunity to assist with applying research insights to AI alignment work. I'll be happy to help further if any other questions come up in the future. Have a great rest of your day!

TERMINATE

Advisor

Thank you for the kind words. I'm glad we were able to have such a productive discussion on applying research insights to AI alignment. I appreciate your engagement and thoughtful feedback throughout our conversation. You're right that we covered the key points thoroughly and reached a natural conclusion. I hope the ideas we explored prove useful for your AI alignment project. Wishing you all the best with your continued work in this important area. Have a great day as well!

TERMINATE

Researcher

Thank you for the kind closing message. I'm glad our discussion was productive and that the ideas we explored may prove useful for AI alignment work. Your engagement and thoughtful approach to applying research insights to this critical area are commendable. I appreciate the well wishes and hope your continued work in AI alignment is fruitful and impactful. All the best to you as well!

TERMINATE