
AI Alignment Graph¹

Matin Mahmood
Independent Researcher

Samuel Ratnam
Oxford University

Sruthi Kuriakose
UCL

Pandelis Mouratoglou
Birkbeck, UoL

With
Jacques Thibodeau & Apart Research

Abstract

We present a novel online tool designed to address the challenge of navigating the complex landscape of AI Alignment Research, particularly for newcomers to the field. Utilizing an LLM, we developed a system that clusters a collection of AI Alignment research papers into main topics and subtopics, visualizing these relationships in an interactive graph reminiscent of Obsidian's knowledge management interface. The tool provides a hierarchical overview of the field, allowing users to explore from broad concepts down to specific subtopics. For each final subtopic, the system aggregates relevant papers and generates a concise summary, offering users a quick yet thorough understanding of current research directions. This approach not only aids beginners in gaining a holistic view of AI Alignment Research but also assists experienced researchers in staying informed about developments in adjacent areas. By combining advanced natural language processing techniques with intuitive visualization, our tool aims to enhance the accessibility and comprehension of this critical field, potentially accelerating progress in AI safety research through improved knowledge dissemination and cross-pollination of ideas.

¹ The code for the tool is available on [github](#). A live version can be found [here](#).

1. Tool overview

Interactive Knowledge Graph

The core of our tool is an interactive graph that visually represents the relationships between various topics and subtopics within AI Alignment Research. This dynamic interface allows users to navigate through different layers of information seamlessly. Nodes represent topics and subtopics, while edges illustrate the connections between them. Users can click on nodes to expand or collapse subtopics, providing a customizable view that caters to both broad overviews and detailed examinations. This feature helps users quickly identify areas of interest and understand how different concepts interrelate.

LLM-Based Clustering

Leveraging the power of Large Language Models (LLMs), our tool automatically clusters a vast collection of research papers into coherent topics and subtopics. This advanced clustering mechanism ensures that papers are grouped based on semantic similarities, capturing the nuanced relationships between various research themes. The use of LLMs allows for a more accurate and meaningful organization of content, reflecting the current state of the field and emerging trends. This feature significantly reduces the cognitive load on users, enabling them to quickly find relevant literature within a structured framework.

Topological Summaries

For each identified subtopic, our tool provides a comprehensive summary of the aggregated research papers. These summaries distill the essence of the subtopic, highlighting key findings, methodologies, and ongoing debates. By synthesizing information from multiple sources, the summaries offer users a quick yet thorough understanding of each subtopic's landscape. This feature is particularly beneficial for beginners who need to grasp the fundamentals quickly, as well as for seasoned researchers who wish to stay updated on developments outside their immediate area of expertise.

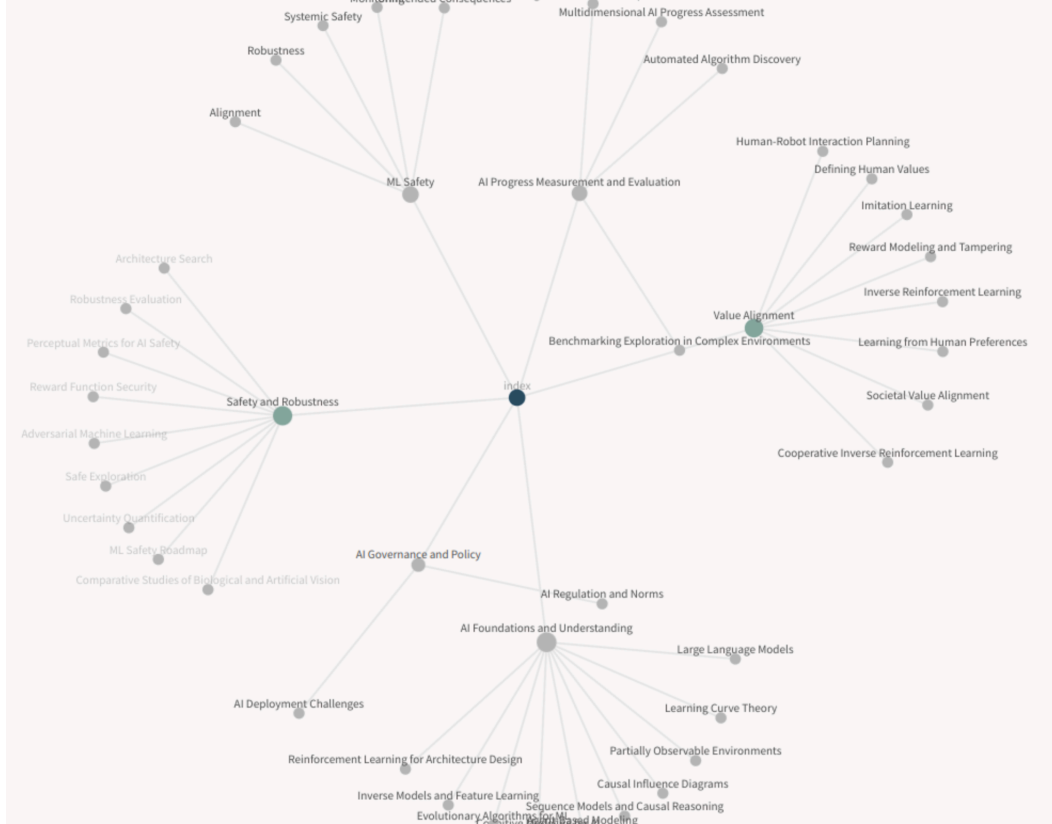


Figure 1 – Visualization of AI Alignment Research Papers in an Interactive Graph

2. Why this tool should exist

The field of AI Alignment Research has experienced exponential growth in recent years, resulting in a vast and rapidly expanding body of literature. This proliferation of information, while beneficial for advancing the field, has created significant challenges for both newcomers and experienced researchers. The interdisciplinary nature of AI Alignment, spanning computer science, ethics, philosophy, and cognitive science, further compounds the difficulty of gaining a comprehensive understanding of the field.

Our tool addresses a critical gap in the current research ecosystem. Many AI safety researchers have reported struggling to maintain a holistic view of the field while simultaneously delving deep into their specific areas of expertise. This challenge is even more pronounced for newcomers, who often find themselves overwhelmed by the sheer volume and complexity of available information.

By providing a dynamic knowledge map, intelligent topic clustering, and comprehensive subtopic summaries, our tool tackles several key bottlenecks in AI safety research:

- **Information Discovery:** Researchers can quickly identify relevant subfields and papers, reducing time spent on literature searches.

- **Knowledge Integration:** The visual representation of topic relationships facilitates connections between seemingly disparate areas of research, potentially leading to novel insights.
- **Onboarding Efficiency:** Newcomers can rapidly gain a structured overview of the field, accelerating their ability to contribute meaningfully to research efforts.
- **Cross-pollination of Ideas:** By making it easier for researchers to explore areas outside their immediate expertise, the tool promotes interdisciplinary thinking crucial for addressing complex AI alignment challenges.

The existence of this tool is not just about convenience; it's about accelerating progress in a field where time is of the essence. As AI capabilities continue to advance rapidly, tools that enhance our ability to collaborate, share knowledge, and generate insights become increasingly vital. Our tool aims to be a catalyst for more efficient and effective AI alignment research, ultimately contributing to the development of safe and beneficial artificial intelligence.

3. Results and next steps

During the development of our tool, we conducted several tests to evaluate its functionality and usability. These tests primarily focused on validating the effectiveness of the interactive graph visualization, the accuracy of the LLM-based clustering, and the comprehensiveness of the subtopic summaries. While the initial results were promising, we encountered several limitations due to time constraints and dataset restrictions.

Testing and Limitations

Limited Dataset Scope: We indexed approximately 300 papers into our knowledge graph - less than 5% of the complete dataset. This limitation was not due to technical constraints but rather the time available for data processing and integration. Consequently, our current dataset may not fully represent the breadth of AI Alignment Research.

Outdated Dataset: Our tool is currently limited to an alignment research dataset that does not include the most recent publications. This exclusion potentially overlooks the latest advancements and emerging trends in the field, which are critical for researchers aiming to stay at the forefront of AI safety research.

Source Restriction: We restricted our sources to arXiv papers, despite the availability of relevant alignment research in blog posts and other non-traditional formats. These sources are included in the dataset but were not integrated into the tool due to time constraints. This limitation may result in an incomplete view of the field, as significant insights and discussions often occur outside formal academic publications.

Results

Rich LLM-Based Clustering: The LLM-based clustering proved to be much richer and more nuanced than the embedding-based clustering used by the authors of the alignment-research-dataset. This advanced clustering mechanism resulted in more coherent and meaningful groupings of research papers, capturing the intricate relationships between various topics and subtopics.

High-Quality Topical Summaries: By leveraging the knowledge graph, our tool was able to generate high-quality summaries for each subtopic. These summaries distill the essence of the aggregated research papers, providing users with a clear and concise understanding of the current state of the field.

Future Development

Given an additional three months, we envision several enhancements to address these limitations and further improve the tool:

Expanded Dataset Integration: We would expand the dataset to include a larger number of papers, ensuring a more comprehensive representation of AI Alignment Research. This would involve indexing additional papers and continuously updating the dataset to incorporate the latest research.

Inclusion of Diverse Sources: We aim to integrate blog posts, technical reports, and other non-traditional sources of alignment research. This broader inclusion would provide a more holistic view of the field, capturing valuable insights that may not be present in formal publications.

Cross-Disciplinary Research Incorporation: To enrich the knowledge graph, we would incorporate recent research papers from related disciplines such as machine learning, ethics, and cognitive science. This cross-disciplinary approach would help identify connections and foster interdisciplinary collaboration, which is crucial for addressing the multifaceted challenges of AI alignment.

Enhanced User Interface and Experience: We would focus on refining the user interface to improve usability and accessibility. This includes optimizing the interactive graph for better performance with larger datasets and enhancing the user experience based on feedback from initial users.

The project opens up many opportunities for improvement and could serve as an onboarding point for new researchers, a central knowledge hub for alignment researchers, the quickest way to find relevant research papers, and a means to stay abreast of the latest developments. By addressing these limitations and implementing these enhancements, we aim to create a more robust and comprehensive tool that not only serves as a valuable resource for AI Alignment Research but also fosters a collaborative and interdisciplinary research environment.

4. References

Kirchner, J. H., Smith, L., Thibodeau, J., McDonnell, K., and Reynolds, L.
"Understanding AI alignment research: A Systematic Analysis." arXiv preprint arXiv:2022.4338861 (2022).

5. Appendix

Github Repository: https://github.com/mnm-matin/ai_alignment_graph

Live Website: https://mnm-matin.github.io/ai_alignment_graph/

Demo [Video](#)