
LLM Research Collaboration Recommender¹

David McSharry
Imperial College London

With
Jacques Thibodeau & Apart Research

Abstract

Good collaborators and mentors are very valuable in research, but finding them is often left up to chance.

An LLM could find people (eg on lesswrong) that are interested in similar research problems as you are, or have expertise in fields you are trying to understand. The LLM could then prompt you to check out their work, potentially leading to new research collaborations.

This could be especially useful for people who are new and therefore have weaker networks.

Keywords: Research tooling, AI safety

1. Tool overview

The tool uses a user researcher's profile to search Lesswrong (LW) for other researchers working on similar problems, have complimentary skills, or who have information that might be relevant to the user researcher.

This tool will take as input a profile of the researcher, for example their research interests, their experience, approaches they are excited about, and any other parts of the researcher's profile they think is relevant to finding potential collaborators. The tool will search LW by scraping information from tags relevant to the user researcher's research interests, find LW users with posts in these tags, and scrape their comments and posts. Finally, it will pass these posts and comments to a

¹ Research conducted at the [Research Augmentation Hackathon](#), 2024

system of LLMs that will output relevant researchers to the user researcher based on their posts and comments.

2. Why this tool should exist

High quality research collaborations are incredibly valuable, Jacques mentioned “John Wentworth said that he became 2-3x more productive once he started working every day with his research collaborator David Lorell”. However, finding collaborators that have aligned research interests and complementary skills is difficult partially due to the number of potential collaborators out there and the fact that what makes a collaboration work is different for everyone.

This tool exists to help a researcher find collaborators by searching forums for other researchers with aligned interests and complimentary skills, increasing the likelihood of finding high quality collaborators. This tool will likely initially be especially useful for junior researchers with less well developed networks.

3. Results and next steps

See below for a simple example of the tool in action.

AI Safety Friend Recommender

Select a tag:

deceptive-alignment



Enter your profile:

I am interested in what language models infer about the user during a conversation with the LLM user

Find Friends

Recommended Friends:

Based on the user's interest in understanding what language models infer about users during conversations, the following three forum users are the most relevant for potential collaboration:

1. Johannes Treutlein:

- His post on "inductive out-of-context reasoning" suggests an exploration of how language models (LLMs) can infer latent information from training data, which aligns closely with the user's interest in what LLMs infer about users in conversation.
- Treutlein's focus on the implications of LLMs learning potentially hazardous information is particularly relevant, as it touches on user perception and inference, which may inform a better understanding of user interactions.

2. Owain Evans:

- Evans's work on the "Situational Awareness Dataset" (SAD) investigates how LLMs recognize their context and self-knowledge, which directly relates to how language models can infer attributes about users based on their interactions.
- His insight into how LLMs may or may not accurately reflect their understanding of situations can provide important insights for your research on user-Language Model dynamics.

3. Eggsyntax:

- In their post related to demographic inference, Eggsyntax discusses how LLMs, like GPT-3.5, can infer demographic information about authors from text. This closely relates to your interest in the inferences made about users during a conversation.
- The insights gained on how models perceive and interpret author characteristics could provide valuable backgrounds in understanding how an LLM affects user experiences and what it infers in real-time interactions.

These three users showcase specific overlap in the research topics surrounding user inference, situational awareness in LLMs, and the broader implications of model understanding—all of which align well with your stated research interest.

Figure 1 – The tool successfully found researchers relevant to the user researcher (in particular eggsyntax who wrote a post very relevant to the profile). It also provides context and references to posts that could be useful for the user researcher to check out.

The user inputs a tag that is related to their research interest (to make searching LW more tractable for this MVP), and their profile (for demonstration we use a short research interest as a profile). The main limitation for this tool was LLM inference costs (the below demonstration was done with gpt-4o-mini but better LLMs will likely lead to better recommendations) and LW rate limits. These are minor limitations that could be overcome with some more work/funding.

Next steps would be using better LLMs, scraping more of LW for context while getting around rate limits, iterating on the tool with feedback from researchers (eg perhaps there are some other features of researchers that make collaborations fruitful that the system should weight more highly when making recommendations), and eventually making the tool work on other sites eg arxiv. The tool could eventually be made to work automatically by constructing a researcher's profile from their body of work.

The LLM recommender system is pretty simple in this iteration, but I have ideas about how to make the recommender system better (e.g. by generating a profile of each relevant user on LW. I started with this approach but found that given my time and cost constraints simply feeding many users posts and comments to a single LLM worked better).

This tool's functionality could also be extended to other similar use cases, such as generating a research profile for each relevant LW user, feeding profiles through a matchmaking algorithm (potentially with a human in the loop to ensure quality recommendations), and suggesting they make a connection, perhaps by pinging them via a LW DM.