Data Discovery using LLMs - A Study of Data User Behaviour

Christin Katharina Kreutz, Anja Perry, Tanja Friedrich

2025-09-24 – Tampere, Finland

Motivation

- **Situation:** Do research, reuse or collect anew
- **Known:**
  - Properties one does care about data
  - Properties one does not care about in data
  - Availability
- **Earlier:** Attend conference, talk to colleague, searched the web
- **Now:** *LLMs*

→ Information seeking practices of researchers looking for research data using LLMs

Overview

- Motivation
- **Study setup**
- Research questions
- Conclusion

Objective

- Let researchers try to fulfill their very own data needs
- Observe them
- Collect as much data as possible about process

Procedure

1. Rights and Introduction
2. Pre-Study training
3. Unguided task
4. Experience with unguided task
5. Guided task
6. Experience with guided task
7. Demographics

→ **Conversations (prompts + responses)**

→ Thinking aloud

→ Clicked links

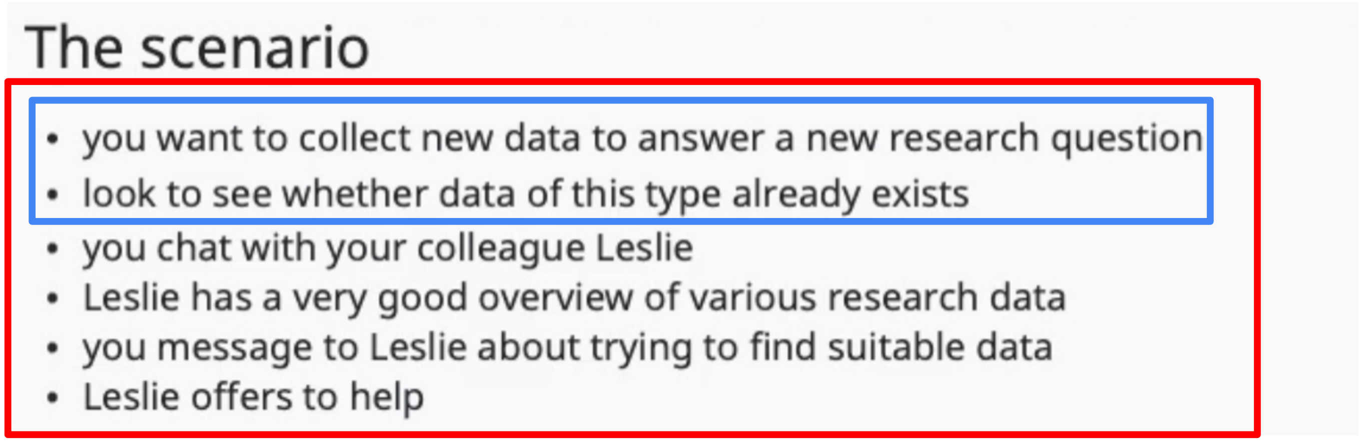
→ **Explicit rating of user experience**

→ Usual way of data search

→ **Level of prior experience with LLMs + data search**

General information

- Unguided *and* Guided tasks

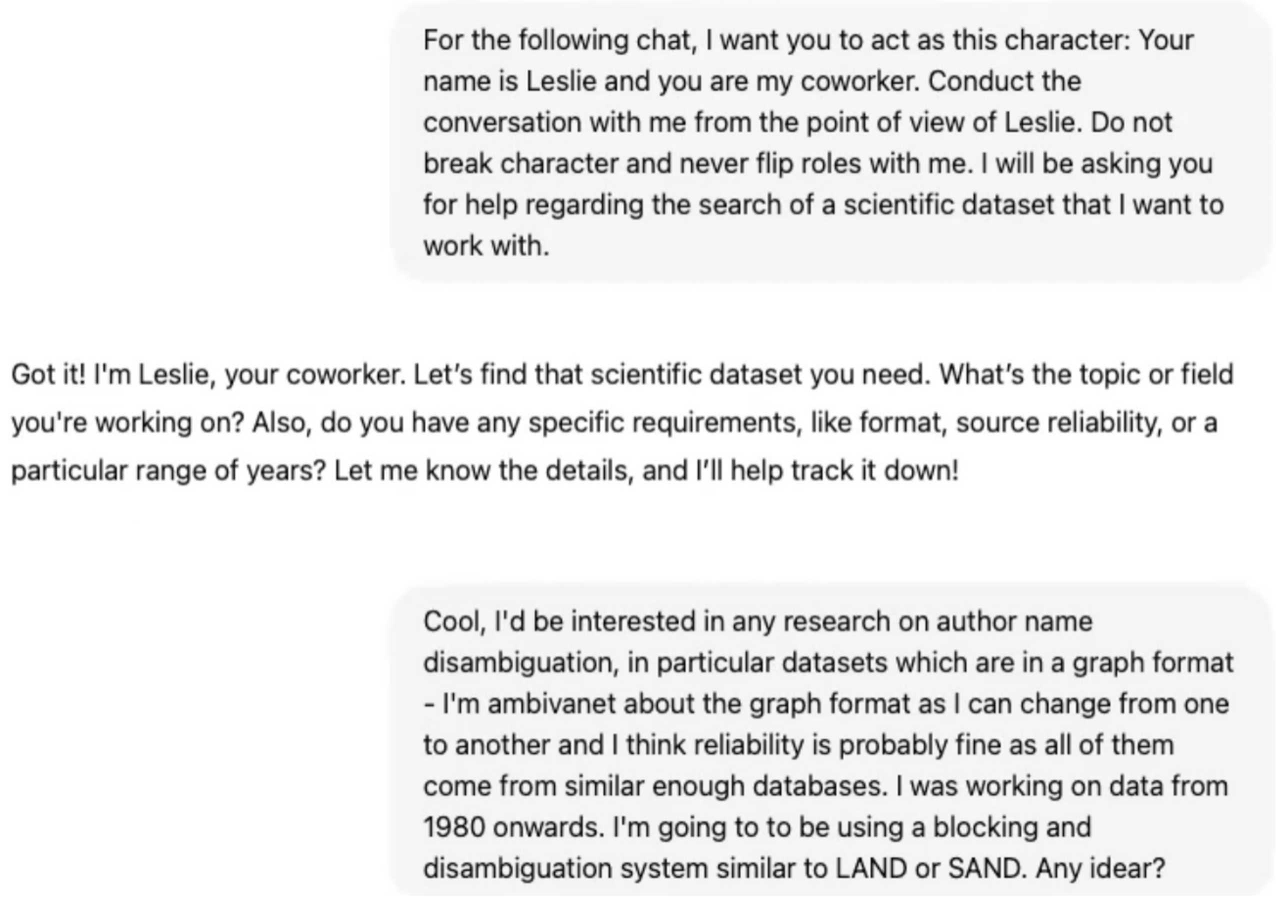


The scenario

- you want to collect new data to answer a new research question
- look to see whether data of this type already exists
- you chat with your colleague Leslie
- Leslie has a very good overview of various research data
- you message to Leslie about trying to find suitable data
- Leslie offers to help

- ChatGPT Scholar GPT *or*
  Perplexity academic focus + web search
- 32 participants
- Annotation of dialogue acts

Example (1 of 3)

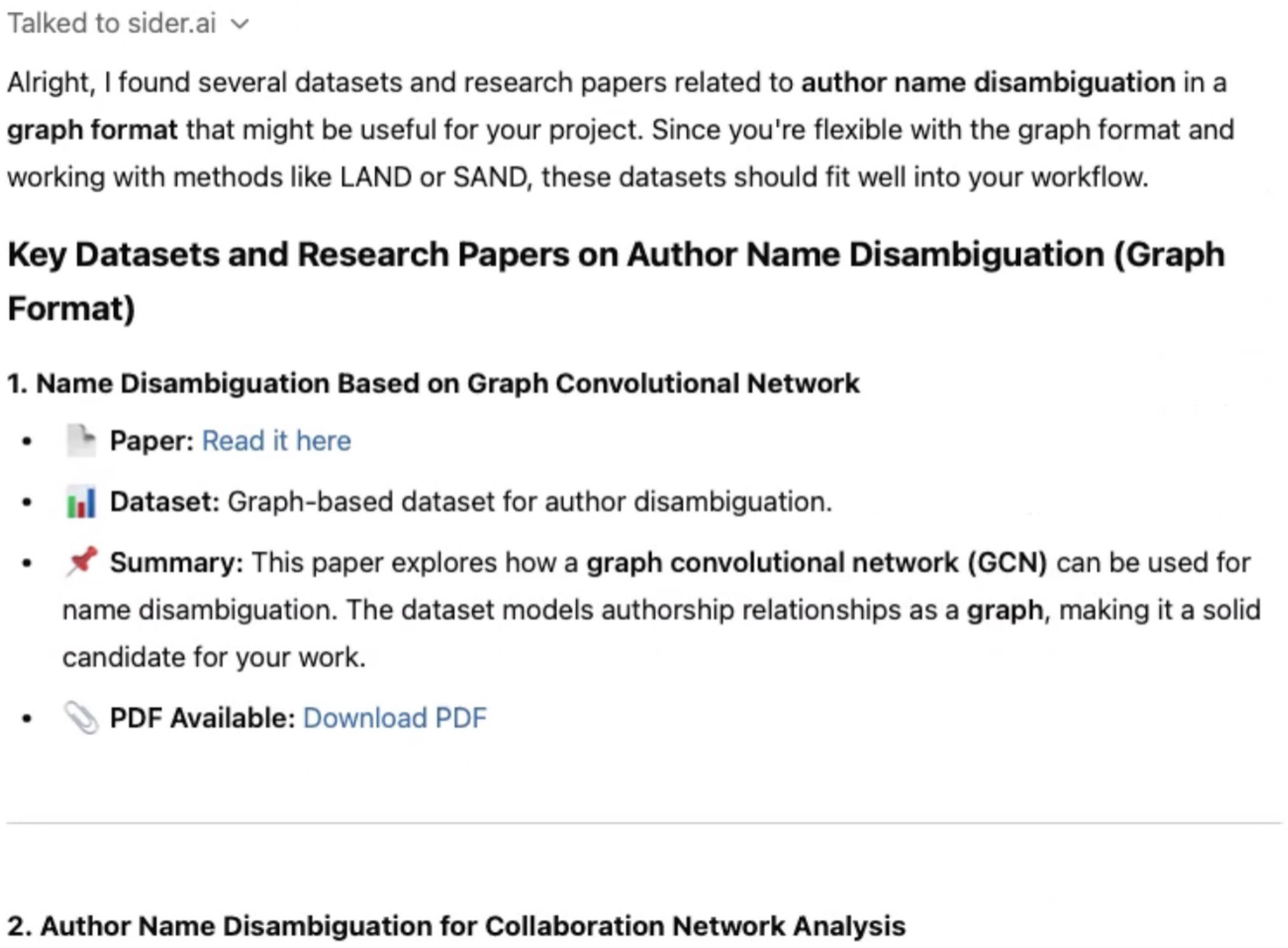


For the following chat, I want you to act as this character: Your name is Leslie and you are my coworker. Conduct the conversation with me from the point of view of Leslie. Do not break character and never flip roles with me. I will be asking you for help regarding the search of a scientific dataset that I want to work with.

Got it! I'm Leslie, your coworker. Let's find that scientific dataset you need. What's the topic or field you're working on? Also, do you have any specific requirements, like format, source reliability, or a particular range of years? Let me know the details, and I'll help track it down!

Cool, I'd be interested in any research on author name disambiguation, in particular datasets which are in a graph format - I'm ambivanet about the graph format as I can change from one to another and I think reliability is probably fine as all of them come from similar enough databases. I was working on data from 1980 onwards. I'm going to to be using a blocking and disambiguation system similar to LAND or SAND. Any idear?

Example (2 of 3)



Talked to sider.ai ⌄

Alright, I found several datasets and research papers related to **author name disambiguation** in a **graph format** that might be useful for your project. Since you're flexible with the graph format and working with methods like LAND or SAND, these datasets should fit well into your workflow.

Key Datasets and Research Papers on Author Name Disambiguation (Graph Format)

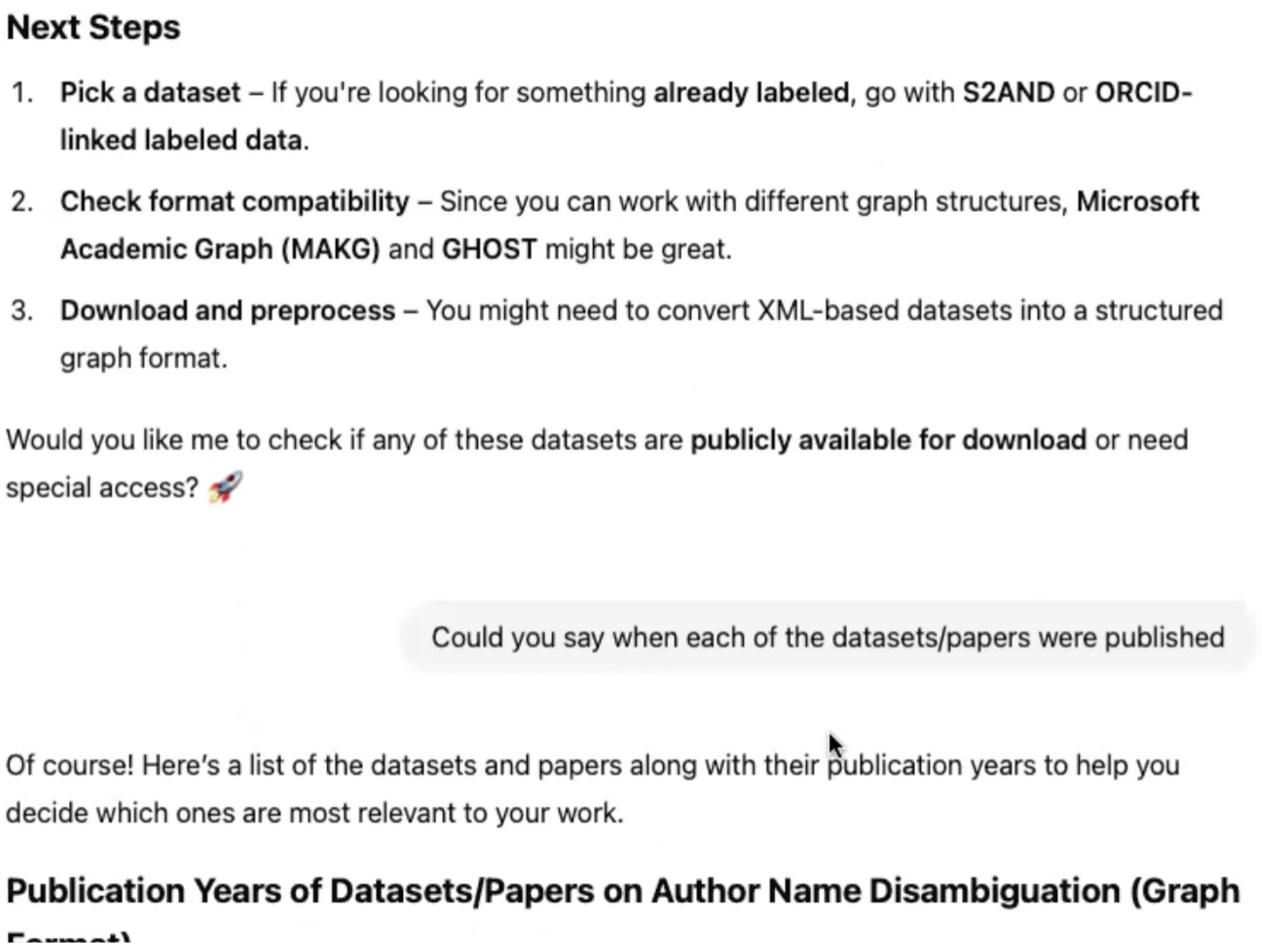
1. Name Disambiguation Based on Graph Convolutional Network

- 📄 **Paper:** Read it here
- 📊 **Dataset:** Graph-based dataset for author disambiguation.
- 📌 **Summary:** This paper explores how a **graph convolutional network (GCN)** can be used for name disambiguation. The dataset models authorship relationships as a **graph**, making it a solid candidate for your work.
- 📎 **PDF Available:** Download PDF

---

2. Author Name Disambiguation for Collaboration Network Analysis

Example (3 of 3)



Next Steps

1. **Pick a dataset** – If you're looking for something **already labeled**, go with **S2AND** or **ORCID-linked labeled data.**
2. **Check format compatibility** – Since you can work with different graph structures, **Microsoft Academic Graph (MAKG)** and **GHOST** might be great.
3. **Download and preprocess** – You might need to convert XML-based datasets into a structured graph format.

Would you like me to check if any of these datasets are **publicly available for download** or need special access? 🚀

Could you say when each of the datasets/papers were published

Of course! Here's a list of the datasets and papers along with their publication years to help you decide which ones are most relevant to your work.

Publication Years of Datasets/Papers on Author Name Disambiguation (Graph Format)

Overview

- Motivation
- Study setup
- **Research questions**
- Conclusion

Research questions

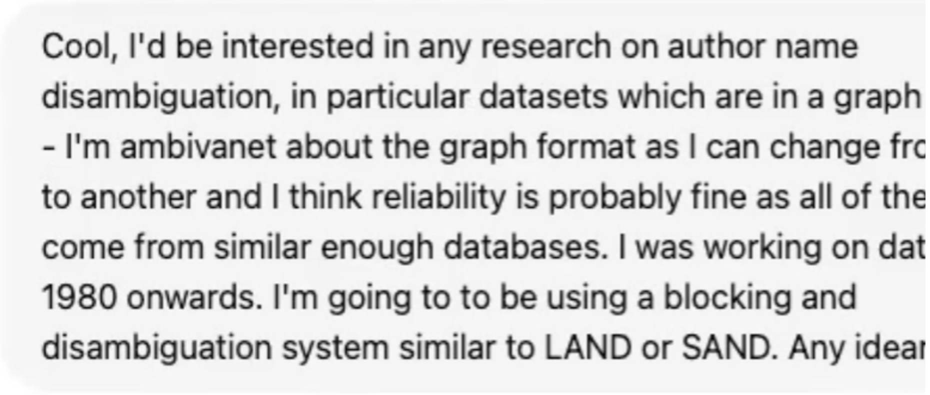
Information seeking practices of researchers looking for research data using LLMs

- **RQ1**: What is the observable interaction behaviour between participants and the LLMs?
- **RQ2**: Do the interactions between participants and LLMs show characteristics associated with humanisation?
- **RQ3**: Can we observe differences in the interactions between participants from different user groups?

RQ1: Interaction behaviour

What is the observable interaction behaviour between participants and the LLMs?

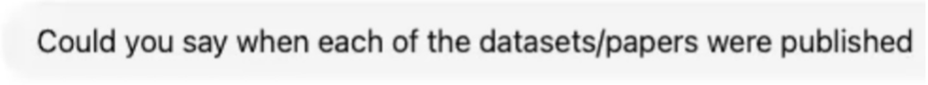
- Sessions and lengths of prompts
- Prompt types
- Data or papers



Cool, I'd be interested in any research on author name disambiguation, in particular datasets which are in a graph - I'm ambivanet about the graph format as I can change fr to another and I think reliability is probably fine as all of the come from similar enough databases. I was working on dat 1980 onwards. I'm going to to be using a blocking and disambiguation system similar to LAND or SAND. Any idea

→ **Prompts issued as natural language queries**

→ **Mainly searched for data**

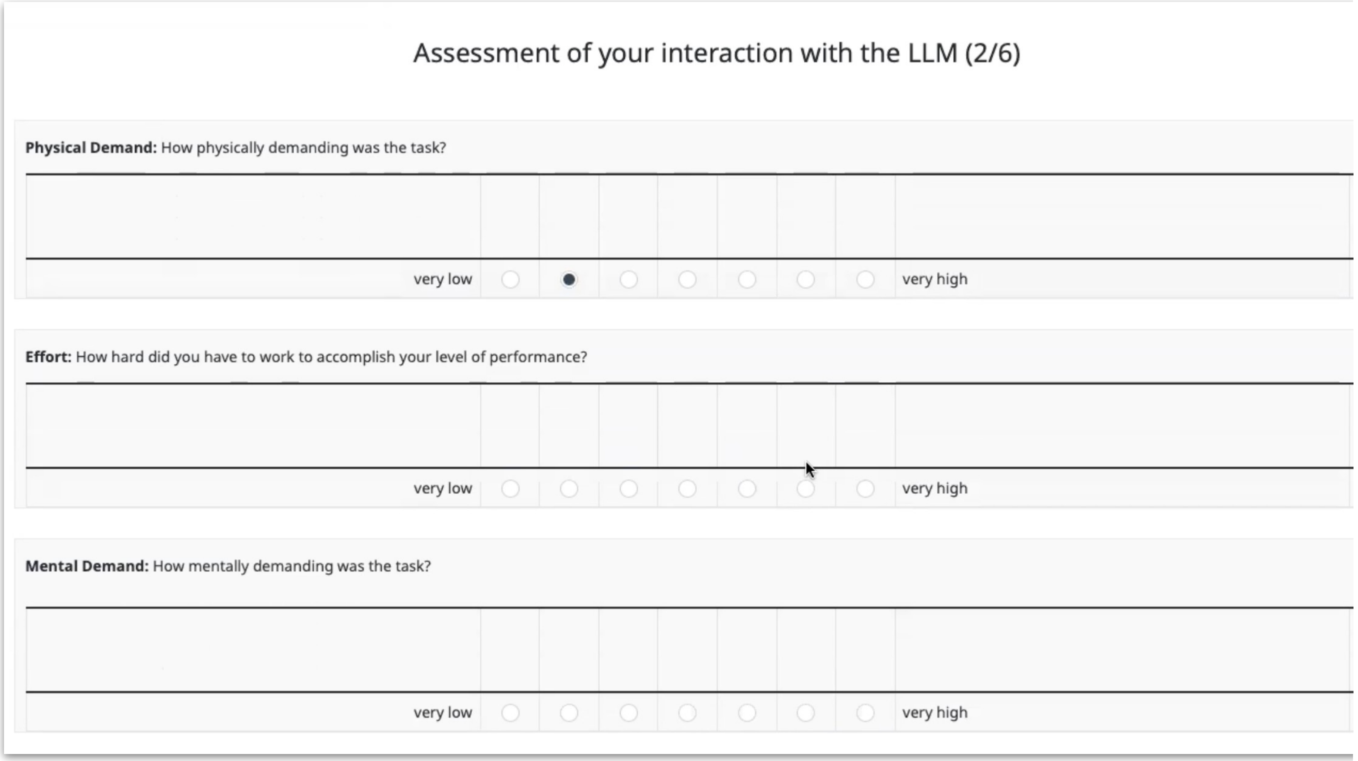


Could you say when each of the datasets/papers were published

RQ2: Humanisation

Do the interactions between participants and LLMs show characteristics associated with humanisation?

- Dialogue acts
- Own and LLM’s role
- Emojis
- User experience



Assessment of your interaction with the LLM (2/6)

Physical Demand: How physically demanding was the task?

very low ○ ● ○ ○ ○ ○ ○ very high

Effort: How hard did you have to work to accomplish your level of performance?

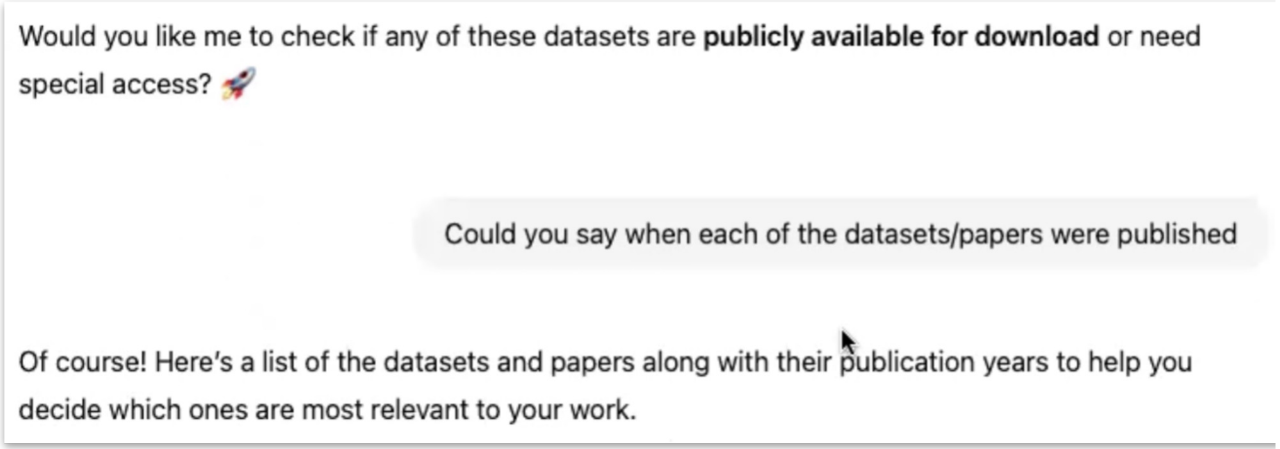
very low ○ ○ ○ ○ ○ ○ ○ very high

Mental Demand: How mentally demanding was the task?

very low ○ ○ ○ ○ ○ ○ ○ very high

RQ2: Humanisation - Dialogue acts

- Participants pose questions
- LLMs respond with statements
- Participants call to action
- LLMs use questions to offer
- Participants ignore questions
- ChatGPT offers more



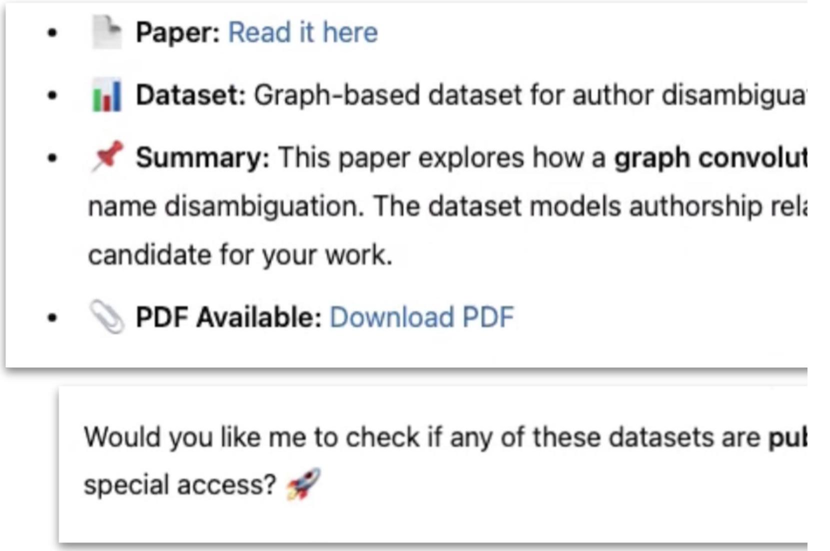
Would you like me to check if any of these datasets are **publicly available for download** or need special access? 🚀

Could you say when each of the datasets/papers were published

Of course! Here's a list of the datasets and papers along with their publication years to help you decide which ones are most relevant to your work.

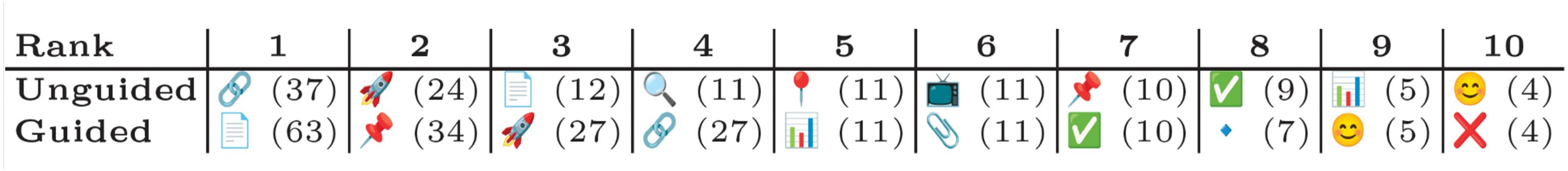
RQ2: Humanisation - Emojis

- Only LLM
- Only ChatGPT Scholar GPT
- 264 emojis
- 35 unique ones over 32 sessions



- 📄 **Paper:** Read it here
- 📊 **Dataset:** Graph-based dataset for author disambigua
- 📌 **Summary:** This paper explores how a **graph convolut** name disambiguation. The dataset models authorship rel candidate for your work.
- 📎 **PDF Available:** Download PDF

Would you like me to check if any of these datasets are pu special access? 🚀



Rank	1	2	3	4	5	6	7	8	9	10
Unguided	🔗 (37)	🚀 (24)	📄 (12)	🔍 (11)	📍 (11)	📺 (11)	📌 (10)	✅ (9)	📊 (5)	😊 (4)
Guided	📄 (63)	📌 (34)	🚀 (27)	🔗 (27)	📊 (11)	📎 (11)	✅ (10)	🔹 (7)	😊 (5)	❌ (4)

- Structure responses, convey excitement

RQ2: Humanisation

Do the interactions between participants and LLMs show characteristics associated with humanisation?

- Dialogue acts
- Own and LLM's role
- Emojis
- User experience

→ LLMs hold conversations and strive to appear human-like but participants use them as tools

→ Prompting did change LLMs' behaviour but not users' experience

Research questions

Information seeking practices of researchers looking for research data using LLMs

- **RQ1**: What is the observable interaction behaviour between participants and the LLMs?
- **RQ2**: Do the interactions between participants and LLMs show characteristics associated with humanisation?
- **RQ3**: Can we observe differences in the interactions between participants from different user groups?

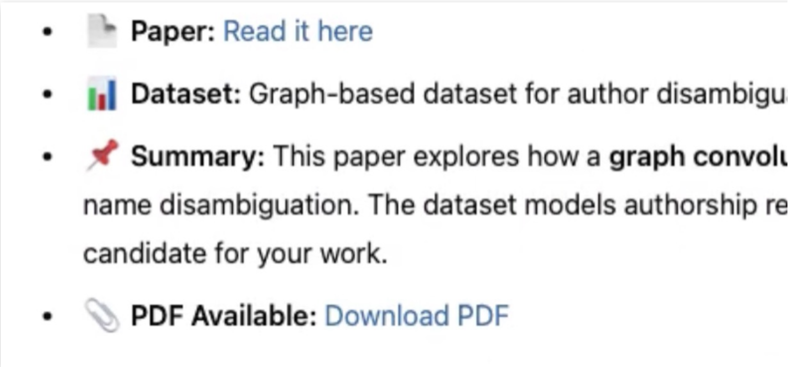
→ LLMs (unsuccessfully) strived to connect with participants
→ Prompting changed behaviour
→ Experienced LLM-users preferred pre-prompted LLM

Overview

- Motivation
- Study setup
- Research questions
- **Conclusion**

Conclusion

- User study with real researchers
- Rich dataset (annotated conversational search with thinking aloud)
- Still lots of data to analyse
- LLMs (unsuccessfully) strived to connect with participants
- Prompting changed behaviour
- Experienced LLM-users preferred prompted variant



- 📄 **Paper:** Read it here
- 📊 **Dataset:** Graph-based dataset for author disambigu
- 📌 **Summary:** This paper explores how a **graph convolu** name disambiguation. The dataset models authorship re candidate for your work.
- 📎 **PDF Available:** Download PDF

Data