ODISSEI Workshop

Using Large Language Models for Data Collection in Social Sciences

All materials at *https://is.gd/llm_data_workshop*

Qixiang Fang

Utrecht University

SoDa Team

Data scientists at postdoc / assistant prof level

Research engineers helping scientists on technical problems

Fellows working on projects that lines up with our goals

Dr. Erik-Jan van Kesteren
Data Scientist; Team Leader

Dr. Javier Garcia-Bernardo
Computational Scientist

Dr. Qixiang Fang
Data Scientist

PDEng. Parisa Zahedi
Research Software Engineer

Dr. Raoul Schram
Research Software Engineer

Dr. Peter Gerbrands
Data Scientist

Dr. Kristina Thompson
ODISSEI SoDa Fellow

Matty Vermet
Research Software Engineer

About Myself

- SoDa Team
  - Various NLP & LLM projects
  - Consultations
  - Make NLP and LLM more accessible
- PhD in Natural Language Processing (UU)
- MSc in Methodology & Statistics (UU)
- BAs in Psychology & Social Sciences (IUB)

Who has worked with LLMs?

Agenda

Time	Description
10:30 – 11:00	LLM Fundamentals
11:00 – 11:20	Prompt Engineering
11:20 – 11:45	*Exercise:* Design Your Own Prompt Experiment
11:45 – 12:00	Wrap-Up & Q&A

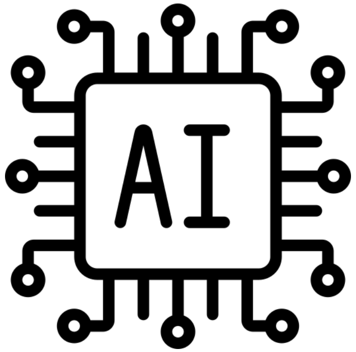
Kind reminder

We focus on OpenAI's LLMs and API, but what we discuss applies to other LLMs and systems.

Kind reminder

Ask questions whenever you want to.
If you don't follow, it's not your fault.

Part I:
LLM Fundamentals



Language and world understanding

This 21 y/o male student from Germany is studying [...]

Language and world understanding

This 21 y/o male student from Germany is studying [...]

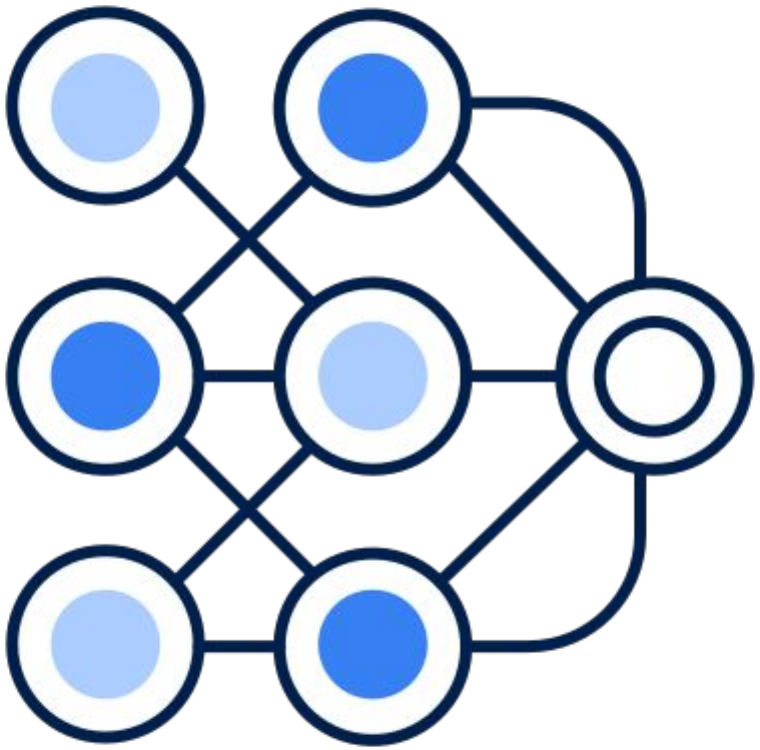
- Requires understanding about language.
- Requires understanding about common sense and world knowledge.

If a model can complete this sentence in a reasonable way, it demonstrates (some) knowledge and language understanding.

Modeling language by predicting it

The backbone of LLMs - **a language prediction model!**

Given some input text, you predict the next word(s).



Modeling language by predicting it

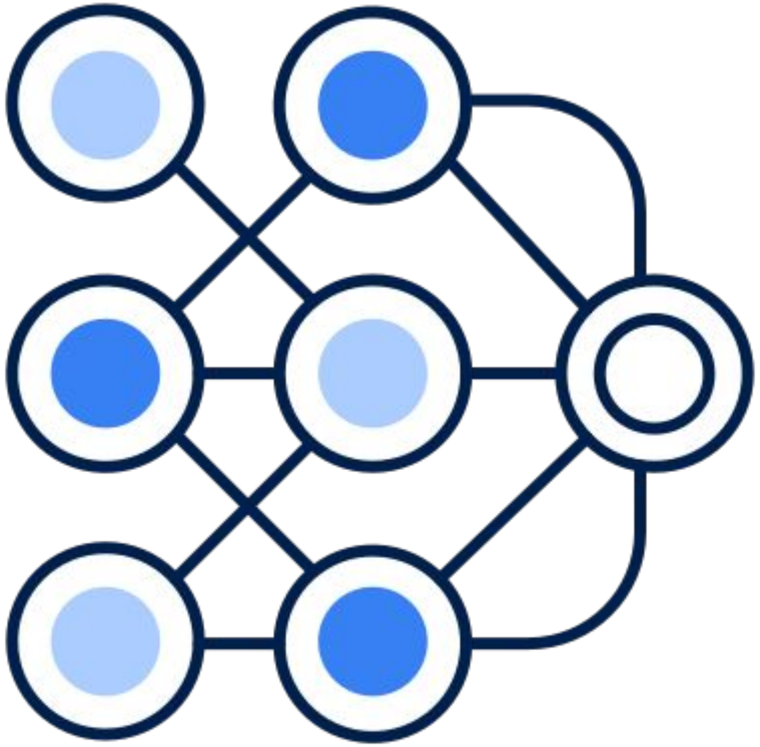
neural networks

deep learning

multi-head attention

Transformer

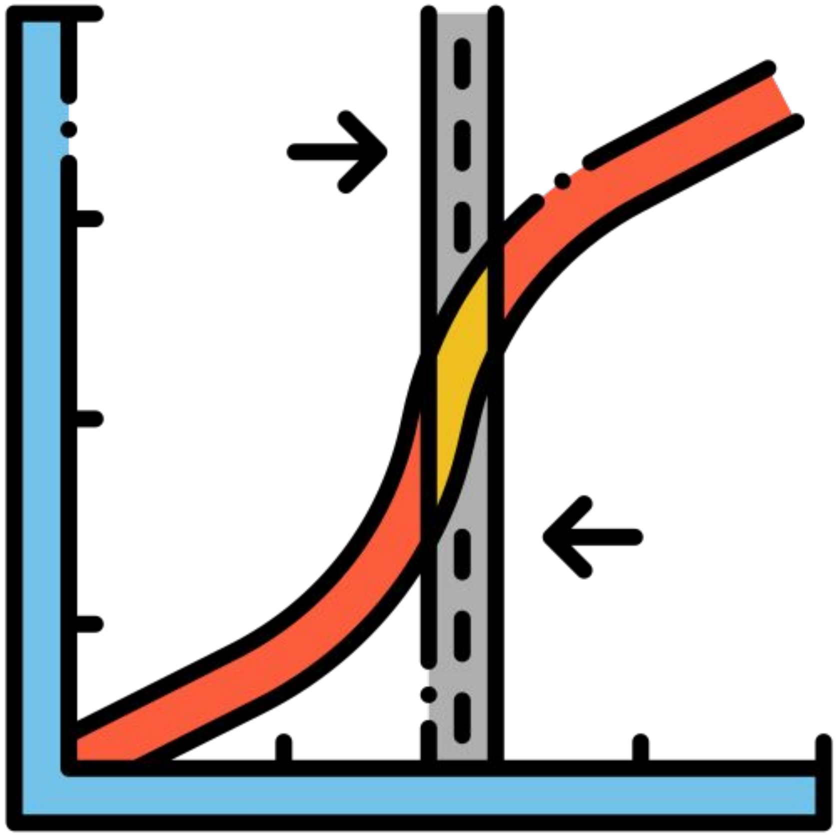
GPT



Modeling language by predicting it

It all comes down to a

logistic regression!



The social science way

This 21 y/o male student from Germany is studying [...]

- Predictors: age, sex, nationality
- Outcome: economics (yes/no)
- Model: binary logistic regression

Data with binary outcomes

sex	age	country	economics degree
male	21	Germany	yes
female	20	UK	no
female	25	US	yes
…	…	…	…

Data with multinomial outcomes

sex	age	country	degree
male	21	Germany	economics
female	20	UK	math
female	25	US	arts
…	…	…	…

with a multinomial logistic regression!

The LLM/NLP way!

This 21 y/o male student from Germany is studying [...]

In the case of language modelling/prediction:

- Each possible word is a predictor, requiring some numerical representation (just like the social science way but less abstraction).
- Each possible word is also a potential outcome.

This 21 y/o male student from Germany is studying [...]

this	the	21	19	y/o	male	fema le	…	is	not	stud ying	learn ing	engi neeri ng	medi cine	ecno mics
1	0	1	0	1	1	0	…	1	0	1	0	0	0	0
0	0	0	0	0	0	0	…	0	0	0	0	0	0	1

This 21 y/o male student from Germany is studying [...]

this	the	21	19	y/o	male	fema le	…	is	not	stud ying	learn ing	engi neeri ng	medi cine	ecno mics
1	0	1	0	1	1	0	…	1	0	1	0	0	0	0
0	0	0	0	0	0	0	…	0	0	0	0	0	0	1

The 19 y/o female student from UK is not learning [...]

this	the	21	19	y/o	male	fema le	…	is	not	stud ying	learn ing	engi neeri ng	medi cine	econ omic s
0	1	0	1	1	0	1	…	1	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Multinomial logistic regression!

Social sciences:

Abstracted predictors and a fixed list of outcome categories

- e.g.,

engineering
medicine
economics

Language modeling:

Use the entire vocabulary (all possible words)!

Iterating...

This 21 y/o male student from Germany is studying economics

- This [...]
- This 21 [...]
- This 21 y/o [...]
- This 21 y/o male [...]
- ...
- This 21 y/o male student from Germany is studying [...]
- This 21 y/o male student from Germany is studying economics [EOS]

ChatGPT 4o

complete the following sentence with one word: This 21 y/o male student from Germany is studying

engineering.

try again

medicine.

try again

business.

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	**95.3%** 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	**96.3%** 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	**87.5%** 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	**88.4** QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

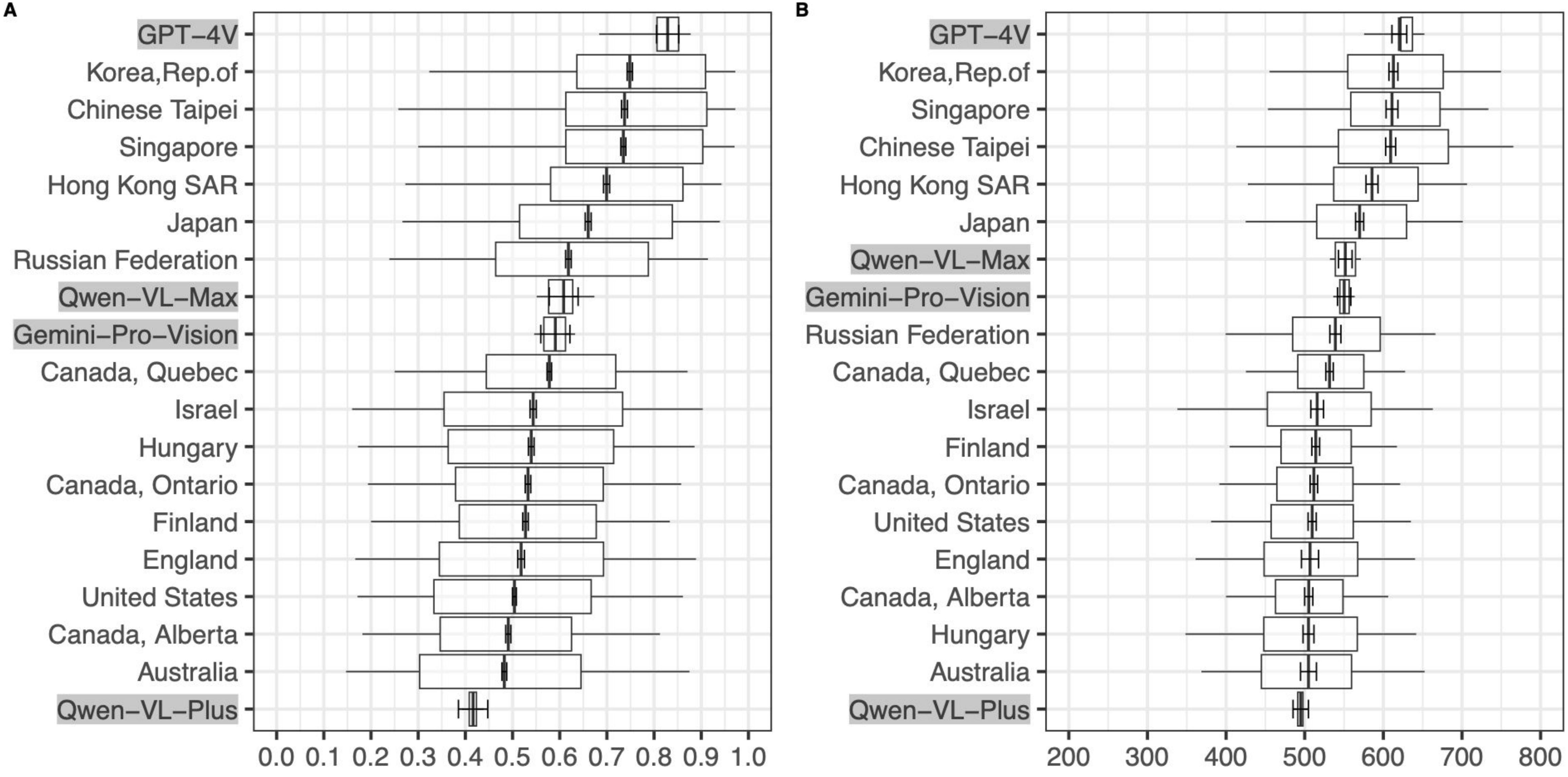
Benchmarking

https://arxiv.org/pdf/2303.08774

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]

Benchmarking

https://arxiv.org/pdf/2303.08774



Psychometrics-based benchmarking (8th grade math)

https://aclanthology.org/2025.gem-1.68.pdf

How come a language model understands so many tasks?

From next word prediction to beyond!

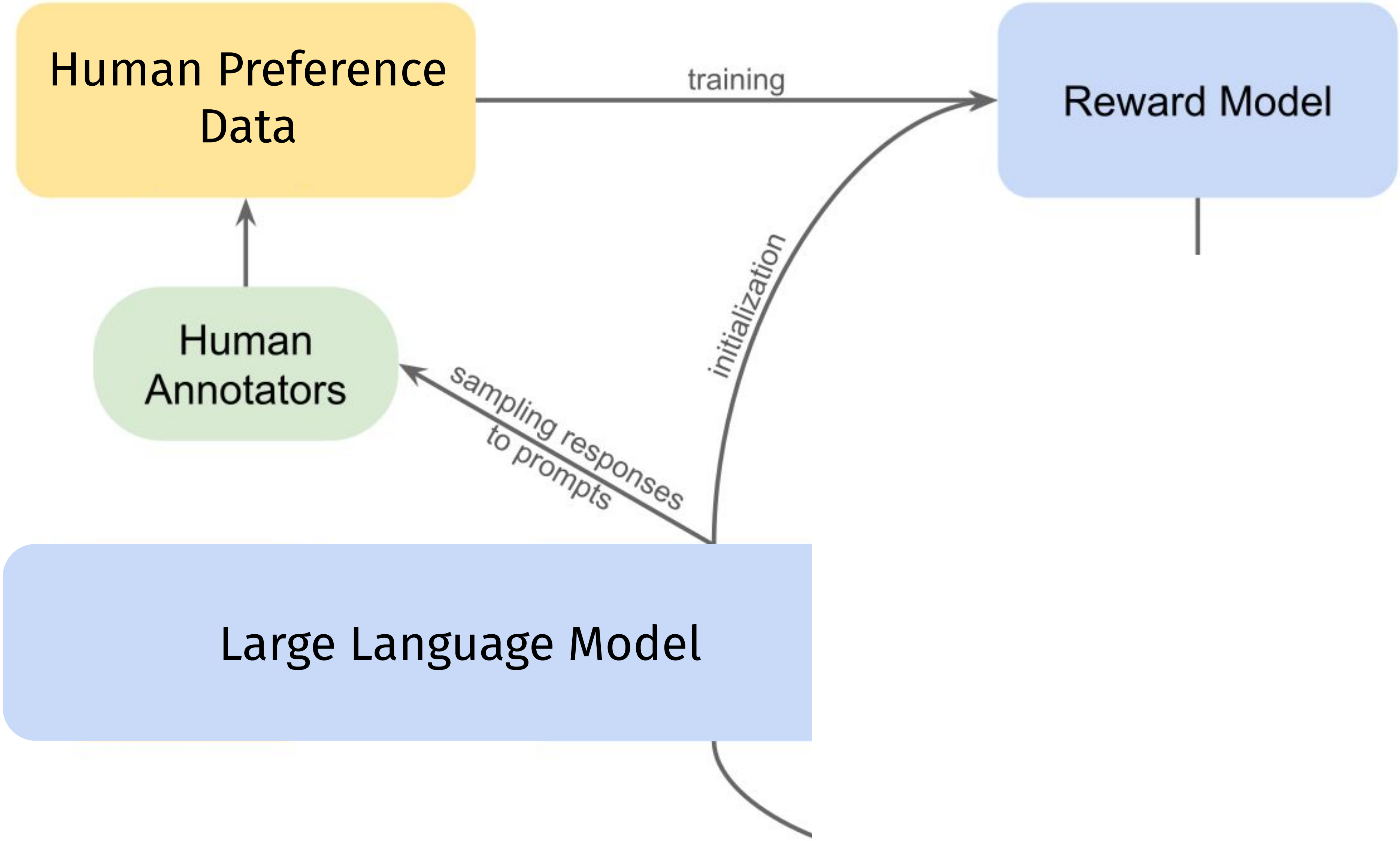
The model is trained on a variety of data such as:

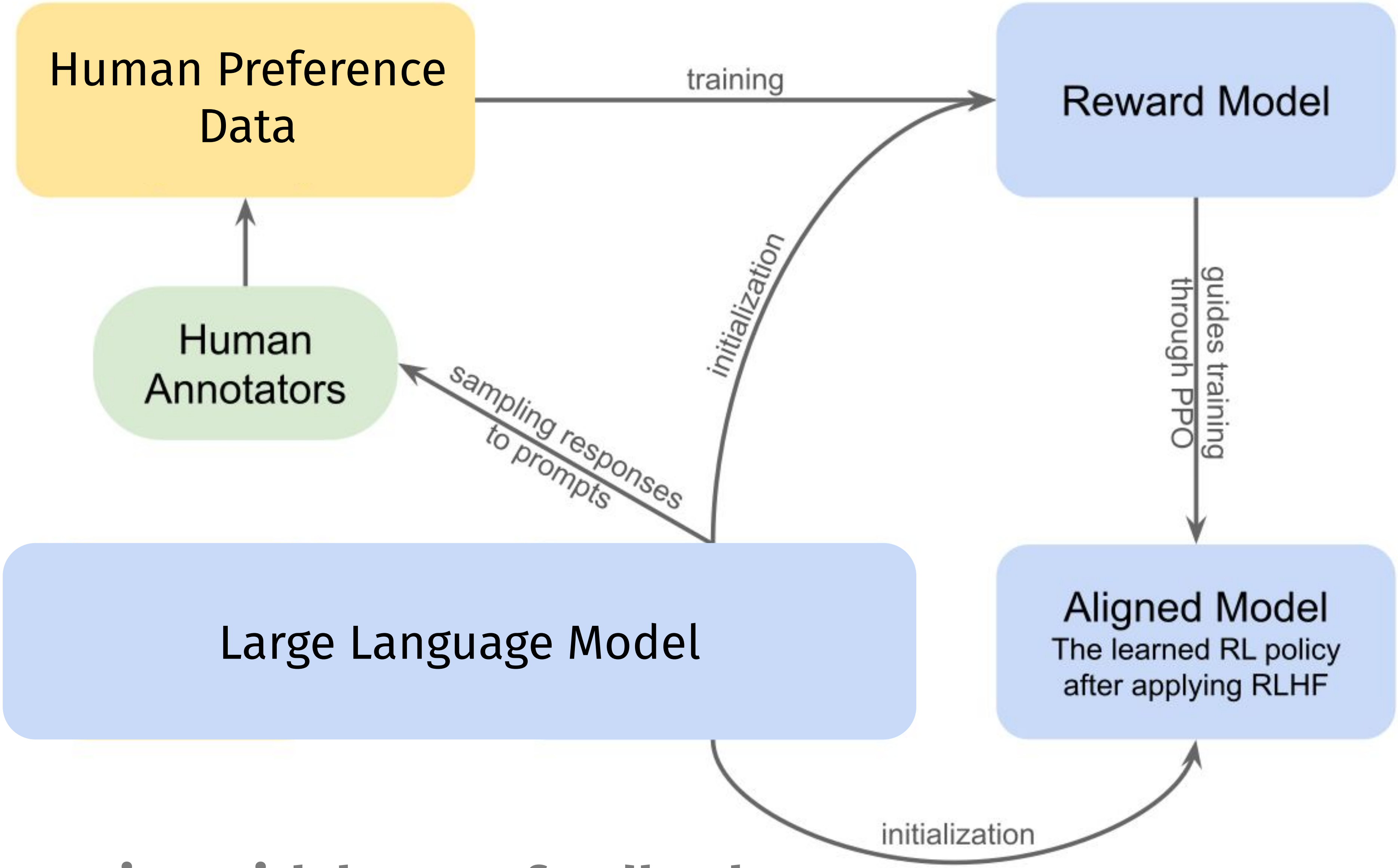
Conversation transcript:

- Interviewer: Introduce yourself.
- Interviewee: I'm a 21 y/o economics student from Germany.

Reddit Q&A:

- OP: "What would be a good university major for me? 21 y/o m from Germany."
- Anonymous user: "Economics!"


Human Preference Data
training
Reward Model
Human Annotators
initialization
sampling responses to prompts
Large Language Model



Reinforcement learning with human feedback

https://aitechfy.com/blog/how-does-chatgpt-work/

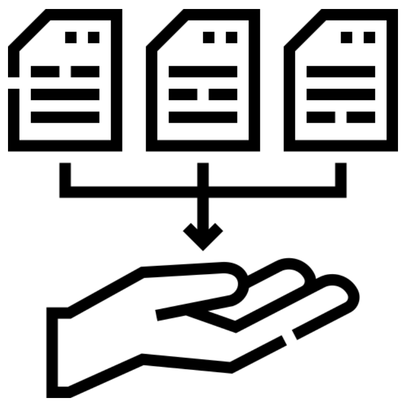
Questions?

Limitations

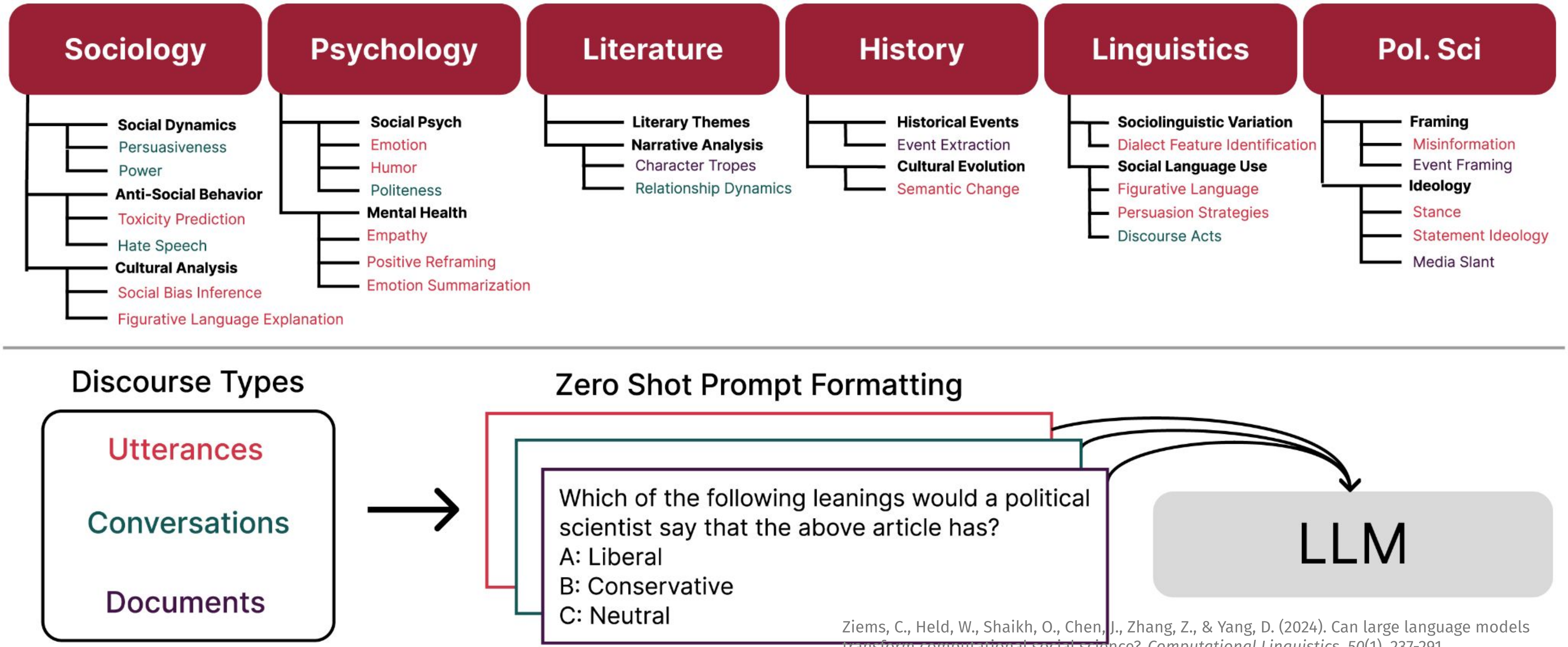
1. **Factual correctness:** prone to hallucinations and fabricated references.
2. **Context length limits:** struggle with very long documents.
3. **Temporal knowledge gaps:** outdated after training.
4. **Data and trainer bias:** reproduce social, cultural, or demographic biases from training data and human trainers.
5. **Poor reasoning consistency:** outputs can vary with small prompt changes or rewording.

Questions?

Part II:
Data collection with LLMs



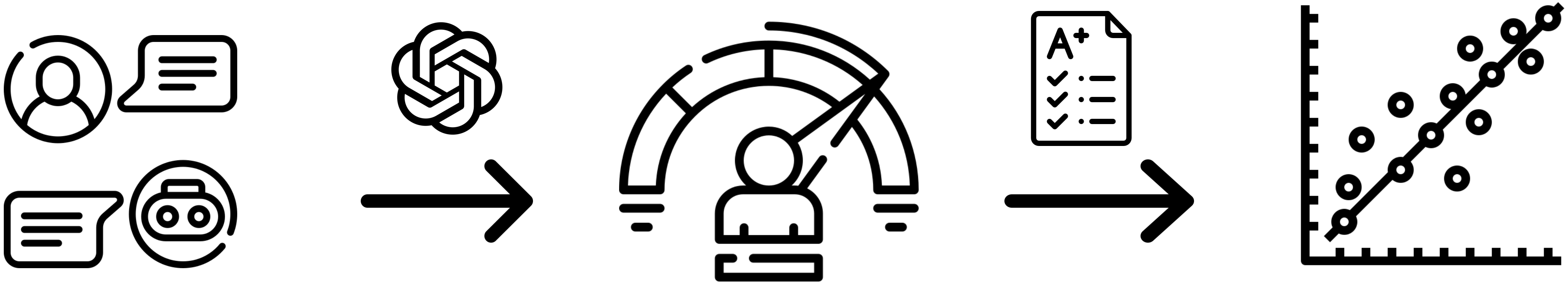
LLMs in social sciences and humanities



Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science?. *Computational Linguistics*, *50*(1), 237-291.

Recurring example

A SoDa fellowship project by Gabrielle Martins van Jaarsveld:

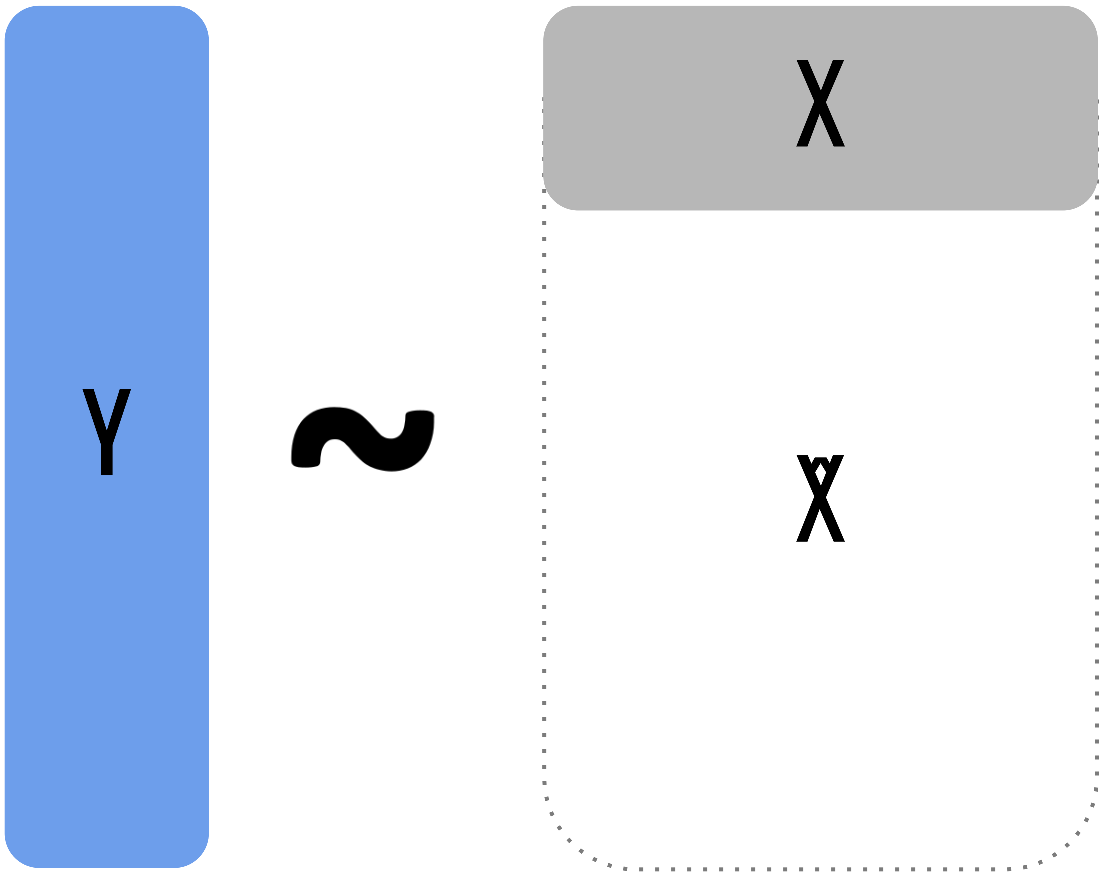


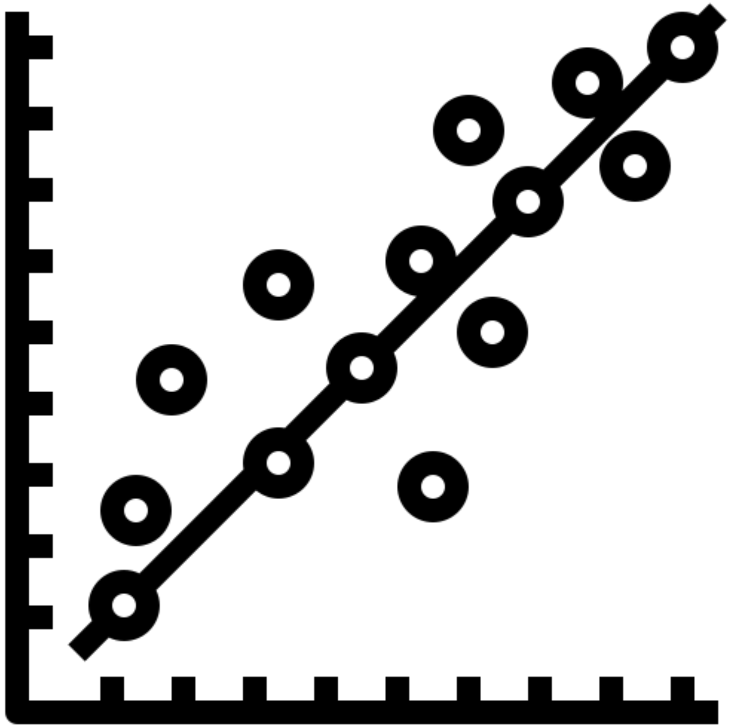
Conversations between students and a rule-based chatbot

LLM-based measurements of self-regulated learning (SRL) indicators

Regress study outcomes on SRL indicators

LLM-assisted regression modelling





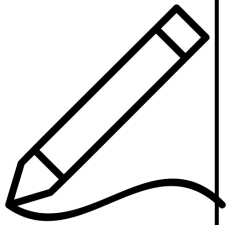
Chatbot: Set an academic goal for the upcoming week.

Chatbot: Add details to make your goal more specific.

Chatbot: How will you measure progress on and achievement of your goal?

Chatbot: Why is this goal important to you in the context of your prior experiences and future goals?

Chatbot: Create a step-by-step plan for achieving this goal in the coming week.



Chatbot: Set an academic goal for the upcoming week.

STUDENT: I would like to catch up on my geography reading

Chatbot: Add details to make your goal more specific.

STUDENT: I need to either read the book from last week and this week, or read my friends notes on the reading to take notes of my own so I dont fall behind.

Chatbot: How will you measure progress on and achievement of your goal?

STUDENT: by the number of pages I write per day

Chatbot: Why is this goal important to you in the context of your prior experiences and future goals?

STUDENT: It is important to achieve because if I dont, I will fall behind and most likely wont be ready for the exam.

Chatbot: Create a step-by-step plan for achieving this goal in the coming week.

STUDENT: 1. evaluate how much there is to do
2. get help from my friends
3. takes notes day by day

Goal Setting & Planning

- Specificity (0-2)
- Measurability (0-2)
- Importance (0-2)
- Realistic multisource planning (0-2)

Prompts

https://en.wikipedia.org/wiki/Prompt

- Prompt (natural language), instructions issued to a computer system (such as a text-to-image artificial intelligence) in the form of written or spoken language.

Elements:

- Context: external information or additional context that can steer the model to better responses
- Task: a specific task you want the model to perform
- Output indicator: the type or format of the output
- Input data: the input or question that we are interested in

A simple prompt

Context Task
Output indicator Input data

A university student was given a series of prompts, guiding them through the process of setting and elaborating on an academic goal for the coming week. You will be provided with the entire conversation including the prompts, and the student answers. Your objective is to assess the specificity of of the student's goal on a scale of 0 to 2 based on the entire conversation.

Chatbot: Set an academic goal for the upcoming week.
STUDENT: I would like to catch up on my geography reading
Chatbot: Add details to make your goal more specific.
STUDENT: I need to either read the book from last week and this week, or read my friends notes on the reading to take notes of my own so I dont fall behind.
Chatbot: How will you measure progress on and acheivement of your goal?
STUDENT: by the number of pages I write per day
Chatbot: Why is this goal important to you in the context of your prior experiences and future goals?
STUDENT: It is important to achieve because if I dont, I will fall behind and most likely wont be ready for the exam.
Chatbot: Create a step-by-step plan for achieving this goal in the coming week.
STUDENT: 1. evaluate how much there is to do
2. get help from my friends
3. takes notes day by day

Prompts

Main types:

- System prompt (overall, consistent)
- User prompt (specific, varying)

System prompt

A university student was given a series of prompts, guiding them through the process of setting and elaborating on an academic goal for the coming week. You will be provided with the entire conversation including the prompts, and the student answers. Your objective is to assess the specificity of of the student's goal on a scale of 0 to 2 based on the entire conversation.

User prompt

Chatbot: Set an academic goal for the upcoming week.
STUDENT: I would like to catch up on my geography reading
Chatbot: Add details to make your goal more specific.
STUDENT: I need to either read the book from last week and this week, or read my friends notes on the reading to take notes of my own so I dont fall behind.
Chatbot: How will you measure progress on and acheivement of your goal?
STUDENT: by the number of pages I write per day
Chatbot: Why is this goal important to you in the context of your prior experiences and future goals?
STUDENT: It is important to achieve because if I dont, I will fall behind and most likely wont be ready for the exam.
Chatbot: Create a step-by-step plan for achieving this goal in the coming week.
STUDENT: 1. evaluate how much there is to do
2. get help from my friends
3. takes notes day by day

Prompt engineering

https://en.wikipedia.org/wiki/Prompt

Prompt engineering is the process of structuring or crafting an instruction in order to produce the best possible output from a generative artificial intelligence (AI) model.[1]

Prompt engineering techniques

1. **Clarity**
 - Be explicit about what you want.
 - Avoid ambiguity and vague wording.

1. Specificity - Goal must be specific rather than general. The context and details of the goal should be explicitly stated and described, and all terms are explained.

- Score of 0: Extremely broad, with no details about what this goal entails. States the goal using vague terms without providing any descriptions of what they mean. Or the goal is an abstract concept to improve or work towards, without any explanation of how this could be actionable or concrete.
- Score of 1: States an actionable or concrete goal and offers some descriptions of the terms used. However, there are still some vague terms which are not fully described.
- Score of 2: No vague terms which are not described. Clearly states the goal and uses clear descriptions to describe exactly what they want to achieve. OR gives a boundary descriptor which offers context to the other unexplained terms in the goal.

2. Measurability - [...]
3. Importance - [...]
4. Multi-source Planning - [...]

Prompt engineering techniques

1. Clarity
2. **Role-based prompting**

- Assign a persona to the AI to guide its response style.

At the beginning of the system prompt:

"You are an expert in educational assessment and goal evaluation, with specialized expertise in applying deductive coding schemes to score the quality and content of student goals. You have a deep understanding of scoring rubrics and are highly skilled at analysing goals for specific characteristics according to well-defined criteria."

Prompt engineering techniques

1. Clarity
2. Role-based prompting
3. **Step-by-step reasoning** ("Chain-of-Thought" Prompting)
   - Encourage the model to explain its reasoning in stages.

##INSTRUCTIONS##

1. Understand the scoring rubric:

- REVIEW the rubric provided for each category to understand the criteria for scores of 0, 1, and 2.
- IDENTIFY the key elements that distinguish a low score (0) from a high score (2) in each category.

2. Analyse the conversation in relation to each category:

- SPECIFICITY: ASSESS the extent to which the goal is specific rather than general. Are context and details of the goal explicitly described, and all terms explained? Is the goal concrete and attainable and not something abstract?
- MEASURABILITY: DETERMINE if goal is measurable, assessable, documentable, or observable. Is the outcome measurable, and is it possible to track progress while working on the goal?
- PERSONAL IMPORTANCE: DETERMINE if there is an explicit reason for the goal which outlines why this goal is important to achieve on the basis of previous experience or in the context of future goals.
- MULTI-SOURCE PLANNING: EXAMINE whether there are specific activities mentioned, and whether these activities directly relate to the goal. Is there a schedule included mentioning days or times of day for working on these activities and accomplishing the goal?

3. Assign a score for each category:

- For each category, ASSIGN a score of 0, 1, or 2 based on the rubric.
- Use the provided scored examples as a reference to ensure consistency with previous assessments.

4. Provide a detailed rationale for each score:

- EXPLAIN why you assigned each score by directly referencing aspects of the goal that meet or fall short of the rubric criteria.

5. Check for consistency:

- DOUBLE-CHECK that each score aligns with both the rubric criteria and the rationale provided.
- MAINTAIN OBJECTIVITY by strictly adhering to the rubric without introducing personal biases.

##EDGE CASE HANDLING##

- If a goal is ambiguous or unclear, SCORE it on the lower end.
- If a goal appears to partially meet the criteria for two different scores, SELECT the score that best reflects the majority of the goals characteristics for that category.

##WHAT NOT TO DO##

- Never apply personal opinion or assumptions outside the rubric criteria.
- never give a score without a detailed explanation, even if the scoring seems obvious.
- never modify or assume student intent score the goal exactly as written.
- never ignore the rubric or provided examples when scoring

Prompt engineering techniques

1. Clarity
2. Role-based prompting
3. Step-by-step reasoning ("Chain-of-Thought" Prompting)
4. **Few-shot prompting**

- Provide examples to help the model learn the desired format or reasoning style.

##EXAMPLE SCORING##

Example 1:
[example conversation mentioned here – removed for data privacy reasons]

Example 1 Scoring:
- Specificity: Score (Reason)
- Measurability: Score (Reason)
- Importance: Score (Reason)
- Multi-Source Planning: Score (Reason)

Example 2:
[example conversation mentioned here – removed for data privacy reasons]

Example 2 Scoring:
- Specificity: Score (Reason)
- Measurability: Score (Reason)
- Importance: Score (Reason)
- Multi-Source Planning: Score (Reason)

Prompt engineering techniques

1. Clarity
2. Role-based prompting
3. Step-by-step reasoning (Chain-of-Thought Prompting)
4. Few-shot prompting
5. **Output Structuring**
  - Request a specific output format (e.g., bullet points, tables, JSON).

##EXAMPLE SCORING##

Example 1:
[example conversation mentioned here – removed for data privacy reasons]

Example 1 Scoring:
- Specificity: Score (Reason)
- Measurability: Score (Reason)
- Importance: Score (Reason)
- Multi-Source Planning: Score (Reason)

Example 2:
[example conversation mentioned here – removed for data privacy reasons]

Example 2 Scoring:
- Specificity: Score (Reason)
- Measurability: Score (Reason)
- Importance: Score (Reason)
- Multi-Source Planning: Score (Reason)

```python
class Structured_Response(BaseModel):
    Specificity_Score: int
    Specificity_Explanation: str
    Measurability_Score: int
    Measurability_Explanation: str
    Importance_Score: int
    Importance_Explanation: str
    Planning_Score: int
    Planning_Explanation: str
```

Functionality: Structured output

Prompt engineering techniques

1. Clarity
2. Role-based prompting
3. Step-by-step reasoning (Chain-of-Thought Prompting)
4. Few-shot prompting
5. Output Structuring
6. **Self-consistency prompting**
   - Asking for multiple responses and selecting the majority, average or best one.

Automatic prompt generator

Free: https://originality.ai/blog/ai-prompt-generator

Paid: https://console.anthropic.com/dashboard

Questions?

Prompt engineering hyperparameters

Temperature: Controls the randomness/creativity of the output.

- Low values (e.g., 0.3) make the model more deterministic and repetitive.
- High values (e.g., 0.6 or higher) increase diversity and creativity but may reduce coherence.

Seed: Setting a seed ensures reproducibility, generating the same response when used with the same prompt and parameters.

Prompt engineering hyperparameters

max_tokens: Limits the maximum number of tokens generated in the response.

Exercise:

Design your own prompt experiment and try to improve your prompt!

Go to **https://is.gd/llm_data_workshop** and pick your preferred notebook (Python/R).

```
••• Enter API key for OpenAI:
```

edu.nl/c4t6c

Inspect your LLM responses

- Wrong labels?
- Wrong explanations?
- Messy output?
- Inconclusive?
- ...?

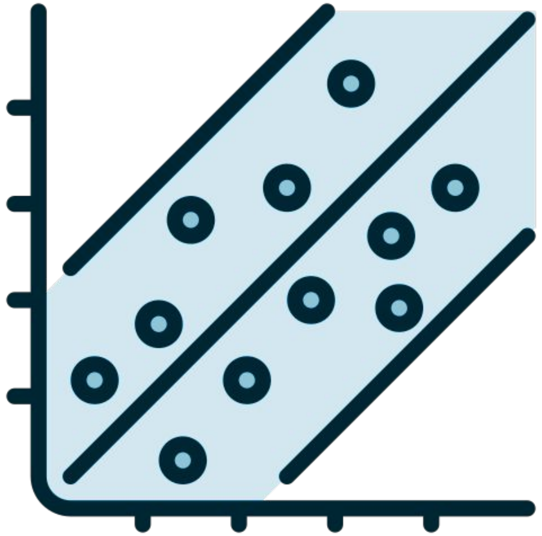
Who wrote the best prompt?

Ending Remarks:

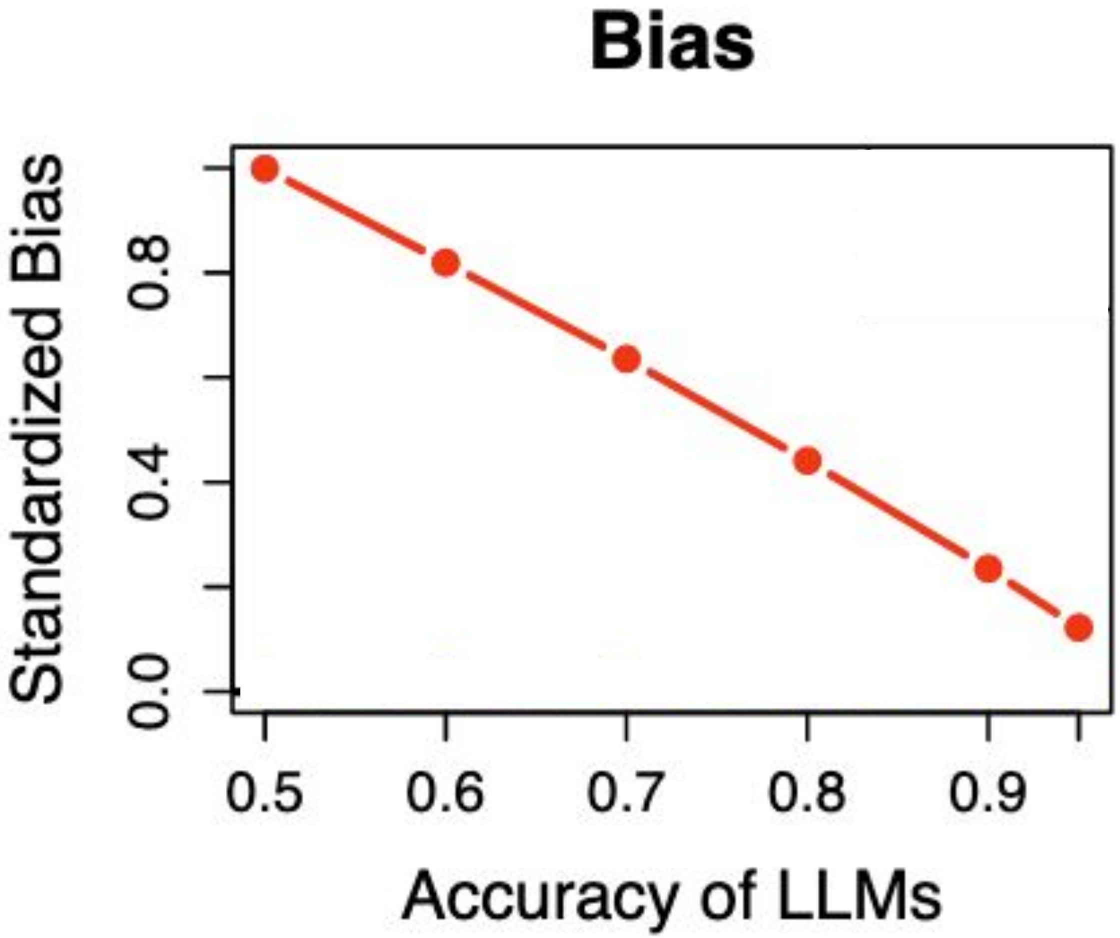
Next Full Workshop and Useful Resources

What comes after data collection?

Inferences with LLM-based data



Problem:



Egami et al. (2024)

What do we do?

Sign up for our next full workshop!

Next full workshop

- **Date**: 23 January 2026, from 9.30 to 15.00
- **Place**: Administration building (room “Van Lier & Eggink”), Utrecht University, Science Park
- **Content:** More depth + making inferences with LLM-generated data

Resource

Using Large Language Models for Text Annotation in Social Science and Humanities: A Hands-On Python/R Tutorial (Qixiang Fang, Javier Garcia Bernardo & Erik-Jan van Kesteren)

More insights on

Research Project Set-up

- Know Your Research Question and Data
- Secure a Gold-Labelled Subset

Evaluate Annotation Quality

- Compute Agreement between LLM and Gold Labels
- Identify Problematic Items and Examine Issues
- Fix Data Issues, Improve Prompt, Iteratively

Annotation Error in Downstream Analyses

- Systematic Error vs. Random Error in Annotations
- Effects on Downstream Analyses
- Methods to Account for Annotation Error

Taking Your Annotations to the Next Level

- Efficiency
- Performance
- Reproducibility
- Common Pitfalls

Resource

Tutorial:

The Best of Both Worlds: Saving Costs and Time When Using OpenAI's API - Combining OpenAI's Batch API and Structured Outputs

https://odissei-soda.nl/tutorials/llm_batch_structured_output/

Resource

OpenAI's researcher access program:
https://openai.com/form/researcher-access-program/

> We encourage applications from early stage researchers in countries supported by our API, and are especially interested in subsidizing work by researchers with limited financial and institutional resources. Researchers can apply for up to $1,000 of OpenAI API credits to support their work. Credits are valid for a period of 12 months and they can be applied towards any of our publicly available models.

We help social scientists with data intensive & computational research

Our goal is to enhance the evidence base and impact of social science by bringing the added value of new data sources and new data analysis techniques into social research in the Netherlands

part of ODISSEI

Contact us

Get more info →

Monthly Thursday SoDa Data Drop-In

If you have questions about your data or methods, join our monthly online SoDa Data Drop-In on the third Thursday of every month at 16:00. Add it to your calendar by clicking here, or just follow the link below.

Link to Teams meeting

ODISSEI SoDa Fellowship

ODISSEI SoDa Fellowship is a programme for early-career researchers in any domain of social sciences. During the appointment as a SoDa fellow, scientists work on data-related projects in social sciences.

SoDa fellows will spend between 3-5 months full-time on their projects. During this time, they are paid members of the SoDa team at the Methodology & Statistics department of Utrecht University, mentored by one of the senior team members.

For more information, please reach out to Kasia Karpinska, ODISSEI Scientific Manager.

Our posters

An AI approach to Investors Narratives Shaping Biodiversity As An Asset Class by Catalina Papar, Qixiang Fang & Helen Toxopeus

Scaffolding Self-Regulated Learning with a Conversational Agent: A Framework and LLM-Based Pipeline for Scalable, Adaptive Feedback in Higher Education by Gabrielle Martins van Jaarsveld, Qixiang Fang & Erik-Jan van Kesteren

Prompt Design Matters: Improving Cross-Cultural, Multilingual Text Classification Using LLMs in the Life Projects Dataset by Shiyu Dong, Pedro Miguel Silva Bastos, Qixiang Fang & Vinicius Coscioni

Thanks!

Questions? Feedback?