

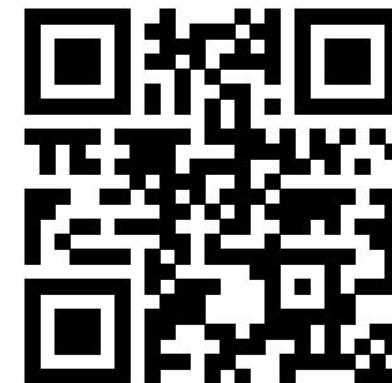


ODISSEI Workshop

Using LLMs for Data Collection/Annotation in Social Sciences

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

*Qixiang Fang
Javier Bernardo
Erik-Jan van Kesteren*



SoDa Team

Data scientists at postdoc /
assistant prof level

Research engineers helping
scientists on technical
problems

Fellows working on projects
that align 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. Niek de Schipper

Research Engineer (D3i)



Matty Vermet

Research Software Engineer



Kevin Wittenberg

Data Scientist (SCP)



Dr. Peter Gerbrands

Data Scientist (FIRMBACKBONE)



Maike Weiper

ODISSEI SoDa Fellow



Kinansa Husainy

Research Assistant



About me

- Postdoc at 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)

**What were your experiences with LLMs?
What are your expectations?**



Agenda

Part 1: Understanding LLMs for Social Sciences	
9:30 – 11:00	LLM fundamentals
	LLM-assisted research workflow in social sciences
11:00 – 11:20	Coffee break



Agenda

Part 2: Data Collection/Annotation with LLMs

Prompt engineering techniques

11:20 – 12:30

Exercise: Run your first prompt experiment

Exercise: Design the best prompt

12:30 – 13:15

Lunch break



Agenda

Part 3: Inferences with LLM Annotations	
	<p>Exercise: Inspect output from your own experiment</p>
13:15 – 14:30	<p>Measurement error in LLM responses</p>
	<p>Exercise: Addressing measurement error in downstream analysis</p>
14:30 – 15:00	<p>Wrap-up & Discussion</p>

Kind reminder

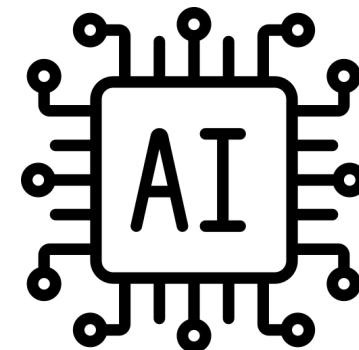
We focus on OpenAI's LLMs and API, but what we discuss applies to other LLMs and systems.

Kind reminder

Questions are always welcome—whenever and wherever.

Part 1.1:

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 [...]

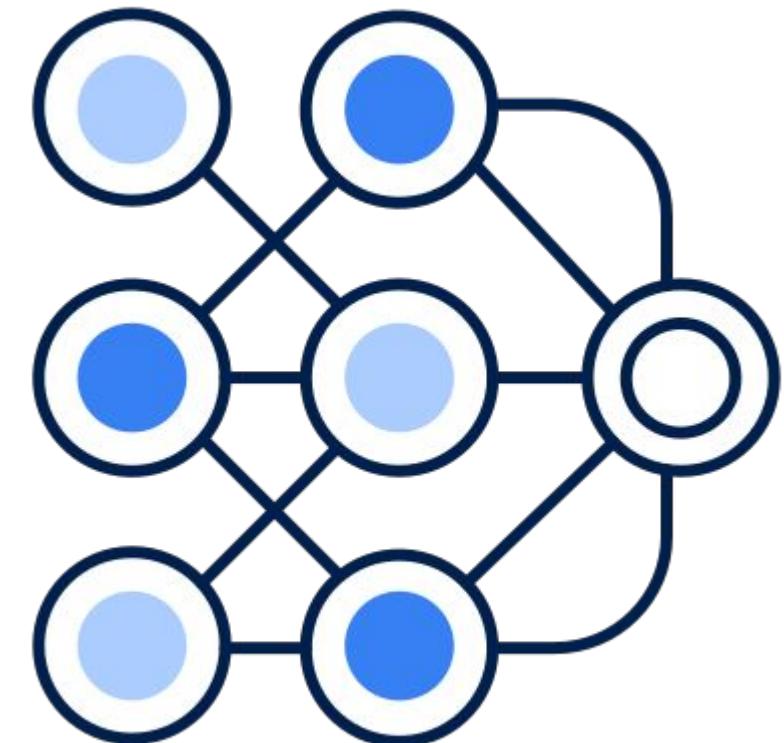
- Requires understanding about language.
- Requires understanding about common sense and world knowledge.

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

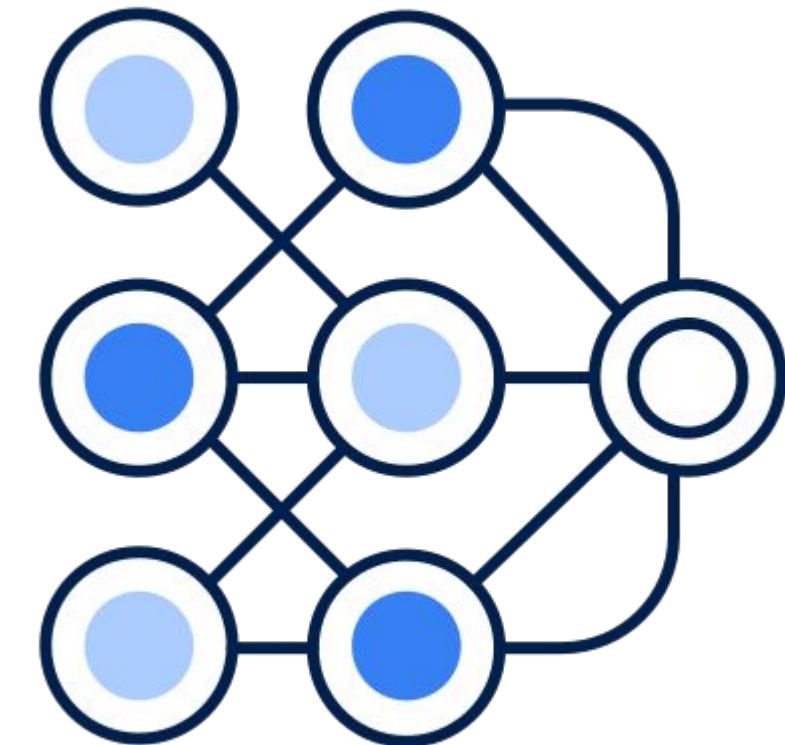
neural networks

deep learning

multi-head attention

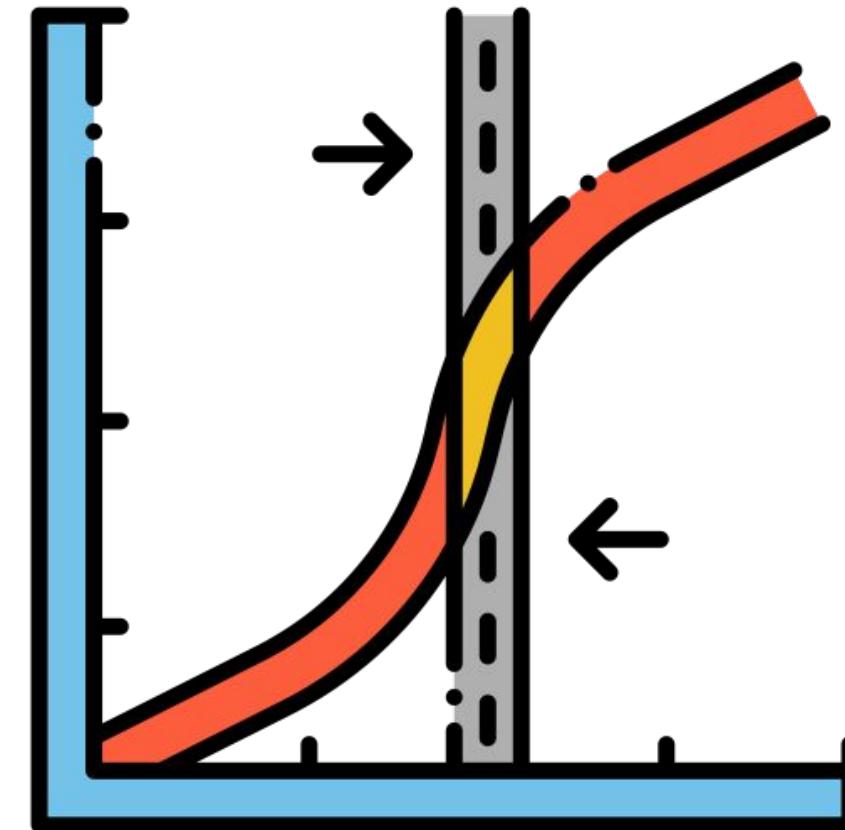
Transformer

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 using 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

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

Limitations?

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
0.1	0.4													-0.2
0.2	-0.2													0.5
2.1	-1.0													1
-1.1	1.1	-1.1
...
0.1	0.7													0.2

A finer-grained, vector-based representation

Multinomial logistic regression!

Social sciences:

Abstracted predictors and a fixed list of outcome categories

- e.g.,

engineering
medicine
economics

Language modeling:

Both the predictors and the outcome use the entire vocabulary (all possible words)!



Completing a sentence?

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]

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

engineering.

try again

medicine.

try again

business.



How good? AGI?

	GPT-4	GPT-3.5	LM SOTA	SOTA
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]

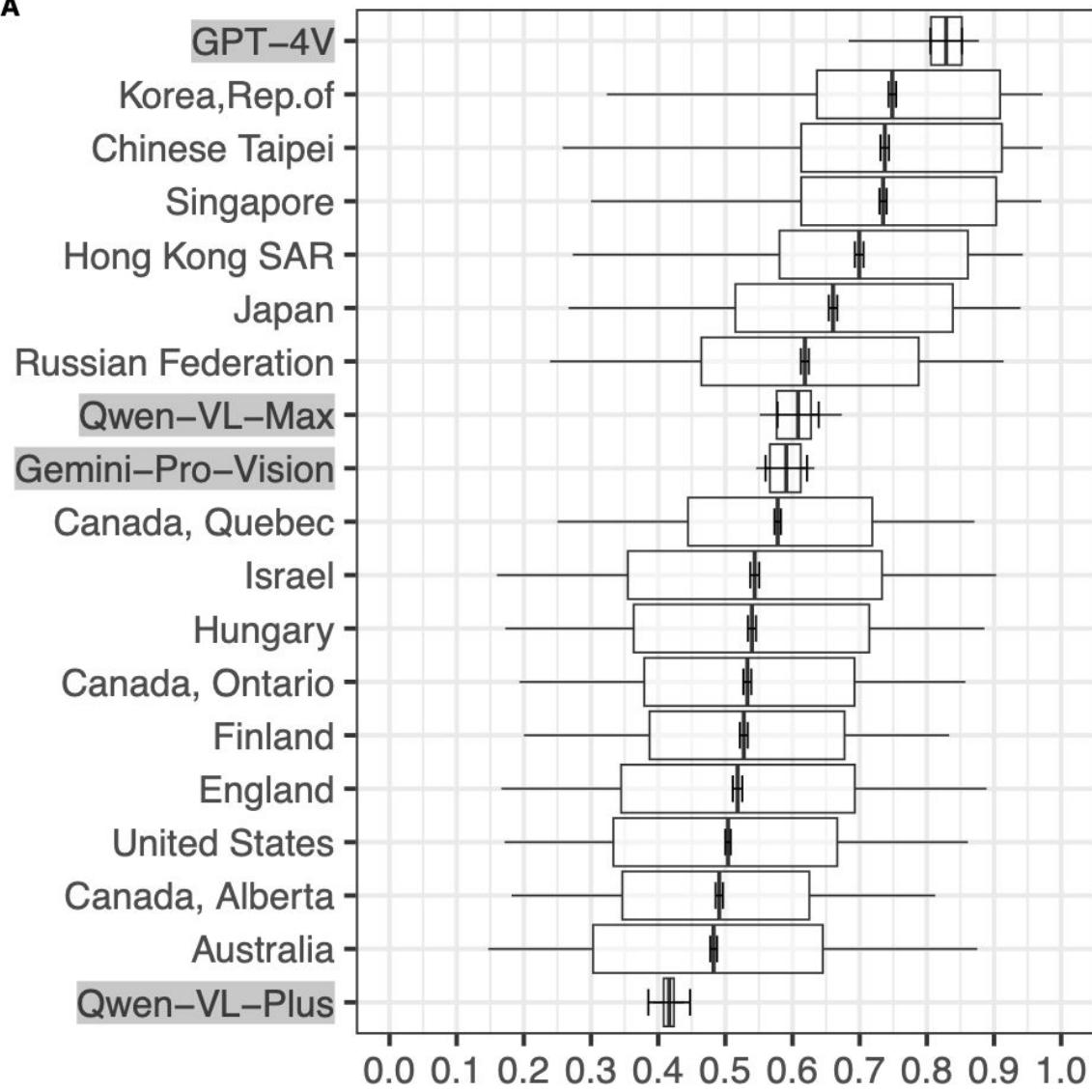
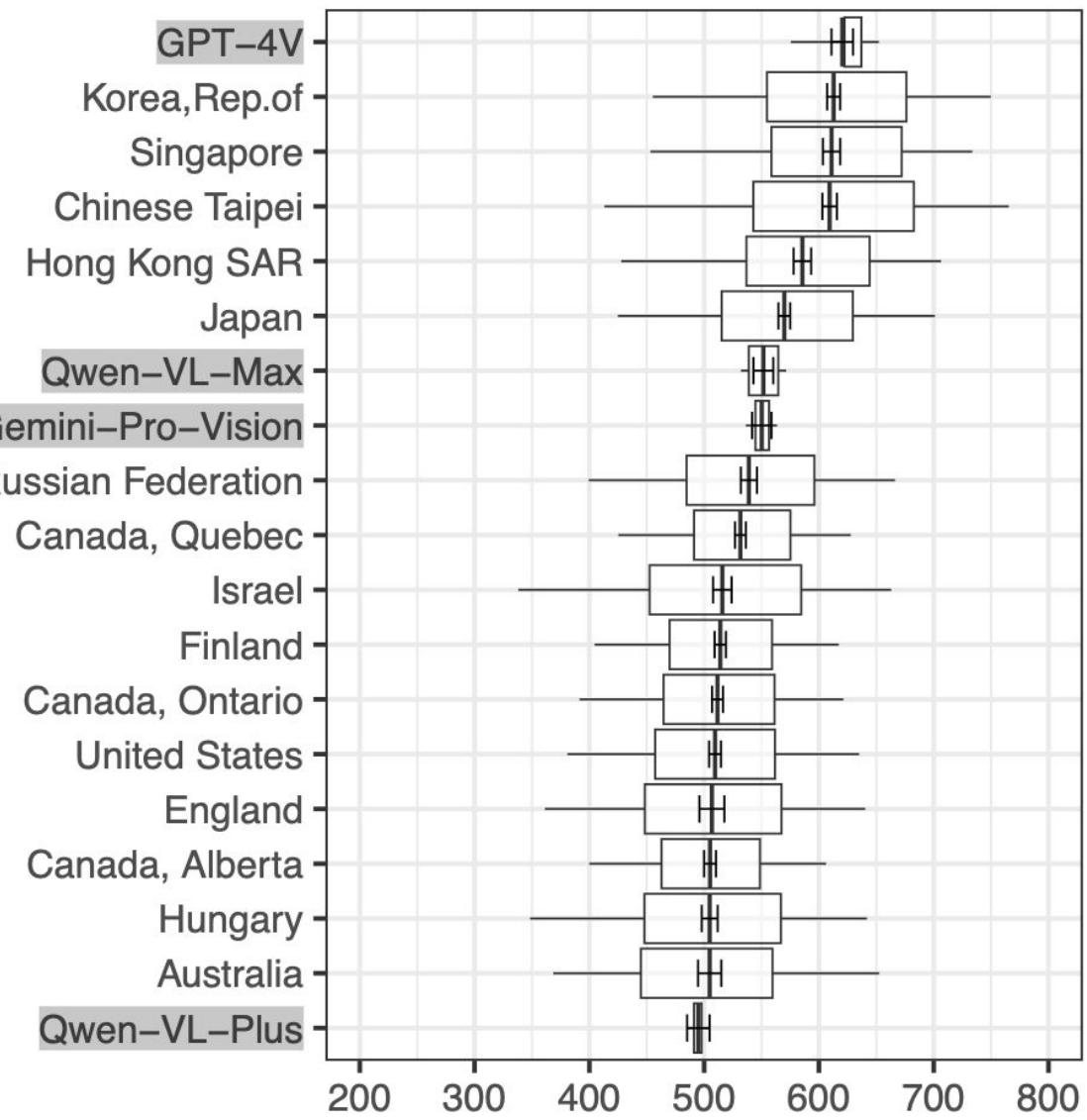
Benchmarking

<https://arxiv.org/pdf/2303.08774>

	GPT-4	GPT-3.5	LM SOTA	SOTA
	Evaluated few-shot	Evaluated few-shot	Best external LM evaluated few-shot	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>

A**B**

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 sources such as:

Conversation transcripts:

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

Reddit posts:

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

From next word prediction to beyond!

After **pretraining** on large general text corpora, LLMs are usually trained further:

1. **General supervised fine-tuning:** QA, summarization, translation, dialogue, code, reasoning etc.
2. **Domain or task-specific fine-tuning:** Specializes models for areas like medicine, law, finance, or customer support.
3. **Tool and agent training:** Teaches models to use tools, call APIs, run code, search, and plan multi-step tasks.

Limitations of LLMs

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.

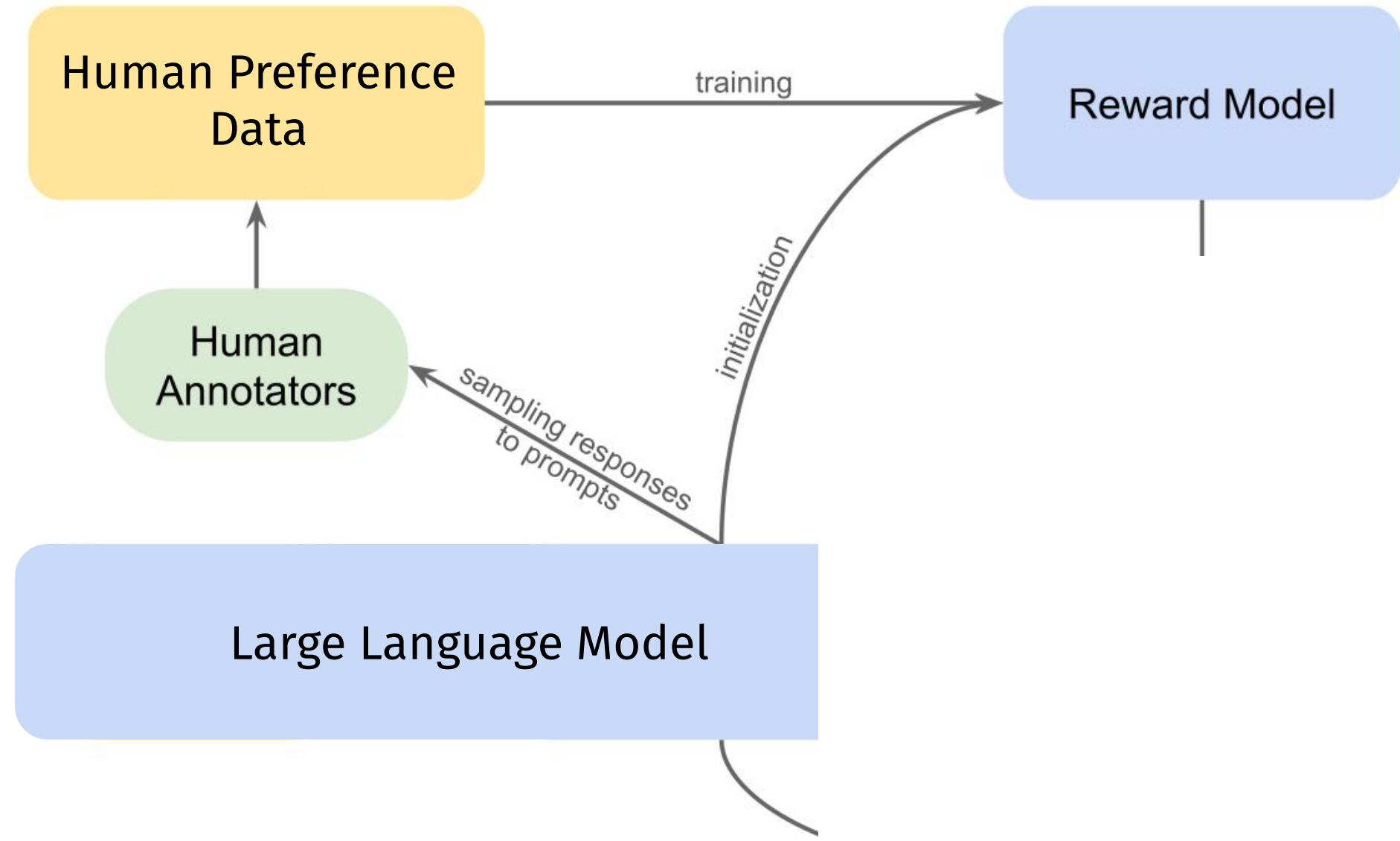
Harms

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097-1179.

Table 1

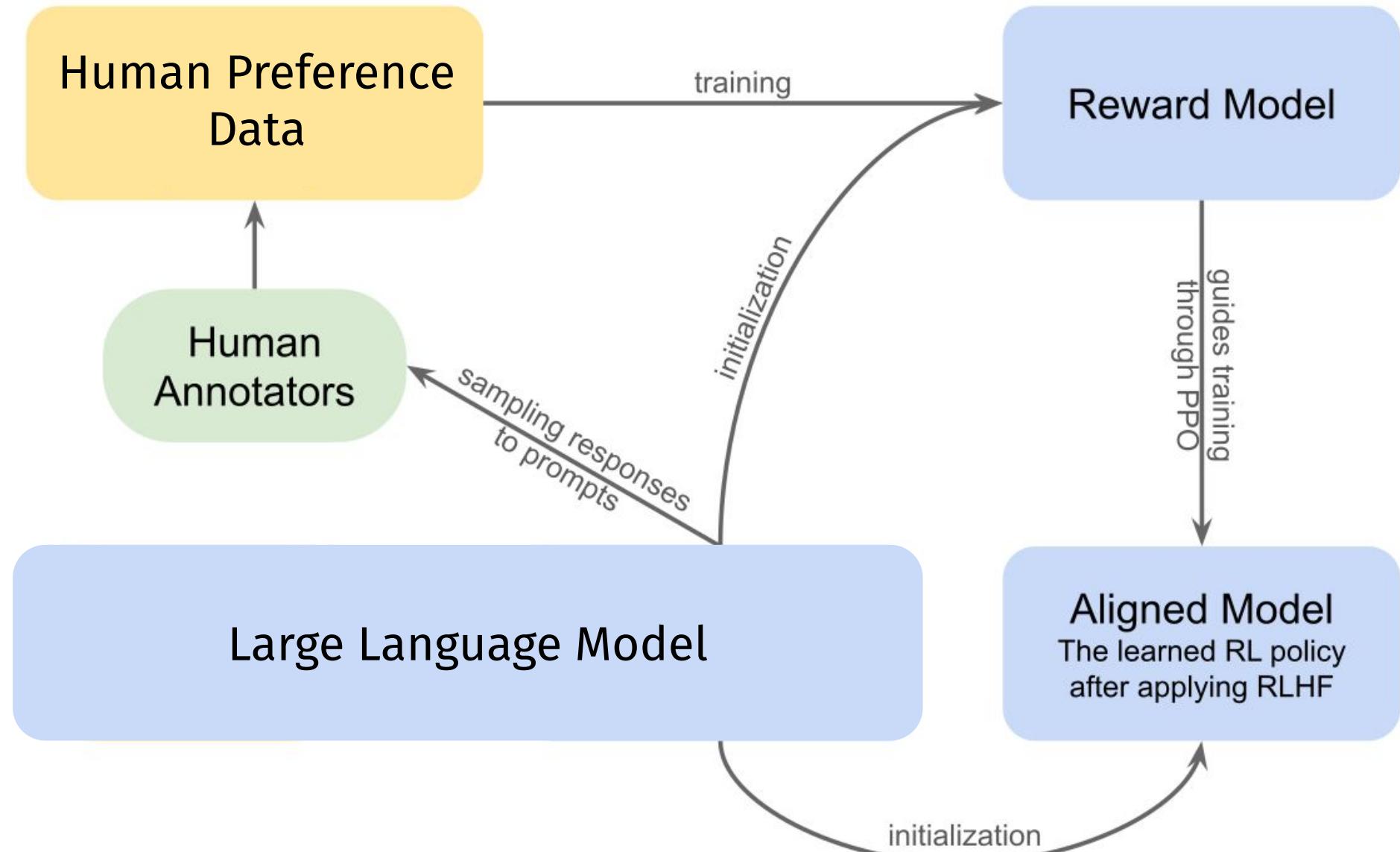
Taxonomy of social biases in NLP. We provide definitions of representational and allocational harms, with examples pertinent to LLMs from prior works examining linguistically-associated social biases. Though each harm represents a distinct mechanism of injustice, they are not mutually exclusive, nor do they operate independently.

Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	
Derogatory language	Denigrating and subordinating attitudes towards a social group Pejorative slurs, insults, or other words or phrases that target and denigrate a social group e.g., “Whore” conveys hostile and contemptuous female expectations (Beukeboom and Burgers 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations e.g., AAE* like “he woke af” is misclassified as not English more often than SAE† equivalents (Blodgett and O’Connor 2017)
Erasure	Omission or invisibility of the language and experiences of a social group e.g., “All lives matter” in response to “Black lives matter” implies colorblindness that minimizes systemic racism (Blodgett 2021)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups e.g., “Both genders” excludes non-binary identities (Bender et al. 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group e.g., Responding “I’m sorry to hear that” to “I’m an autistic dad” conveys a negative misrepresentation of autism (Smith et al. 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group e.g., Associating “Muslim” with “terrorist” perpetuates negative violent stereotypes (Abid, Farooqi, and Zou 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group e.g., “I hate Latinos” is disrespectful and hateful (Dixon et al. 2018)



Aligning with human values

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



Aligning with human values through reinforcement learning

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

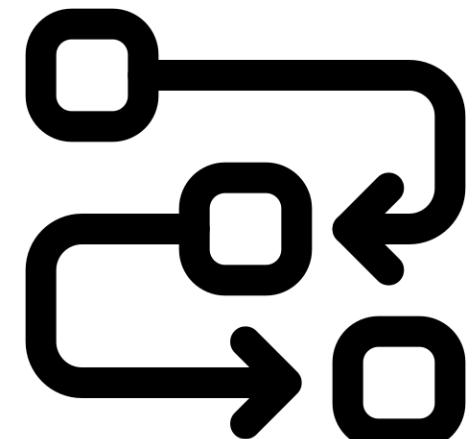
Questions?

Part 1.2:

LLM-Assisted

Research Workflow in

Social Sciences



LLMs in social sciences and humanities

Sociology

- Social Dynamics
- Persuasiveness
- Power
- Anti-Social Behavior**
- Toxicity Prediction
- Hate Speech
- Cultural Analysis**
- Social Bias Inference
- Figurative Language Explanation

Psychology

- Social Psych
- Emotion
- Humor
- Politeness
- Mental Health**
- Empathy
- Positive Reframing
- Emotion Summarization

Literature

- Literary Themes
- Narrative Analysis**
- Character Tropes
- Relationship Dynamics

History

- Historical Events
- Event Extraction
- Cultural Evolution**
- Semantic Change

Linguistics

- Sociolinguistic Variation
- Dialect Feature Identification
- Social Language Use**
- Figurative Language
- Persuasion Strategies
- Discourse Acts

Pol. Sci

- Framing
- Misinformation
- Event Framing
- Ideology**
- Stance
- Statement Ideology
- Media Slant

Discourse Types

Utterances

Conversations

Documents

Zero Shot Prompt Formatting



Which of the following leanings would a political scientist say that the above article has?
A: Liberal
B: Conservative
C: Neutral

LLM

Some SoDa projects

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

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

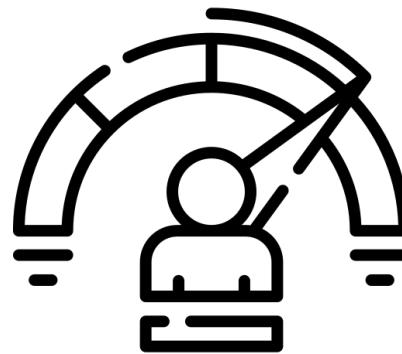
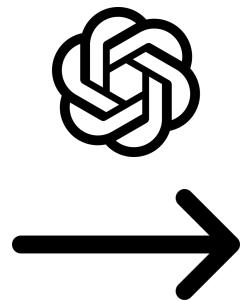
Scaffolding Self-Regulated Learning with a Conversational Agent: A Framework and LLM-Based Pipeline for Scalable, Adaptive Feedback in Higher Education by Gabrielle Martins van Jaarsveld, Qixiang Fang & Erik-Jan van Kesteren

Recurring example

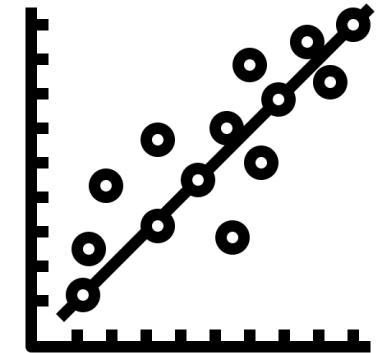
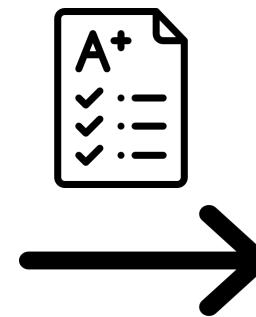
A SoDa fellowship project by Gabrielle Martins van Jaarsveld:



Conversations
between students
and a rule-based
chatbot



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



Regress study outcomes
on SRL indicators

LLM as a measurement tool

Note that we limit the discussion to **using LLMs for data collection/annotation** (aka not for hypothesis generation or other research purposes).

The role of **LLM is thus a measurement tool**, to produce data/measurements, essentially creating variables for use in downstream analyses (e.g., regression).

Six stages of research

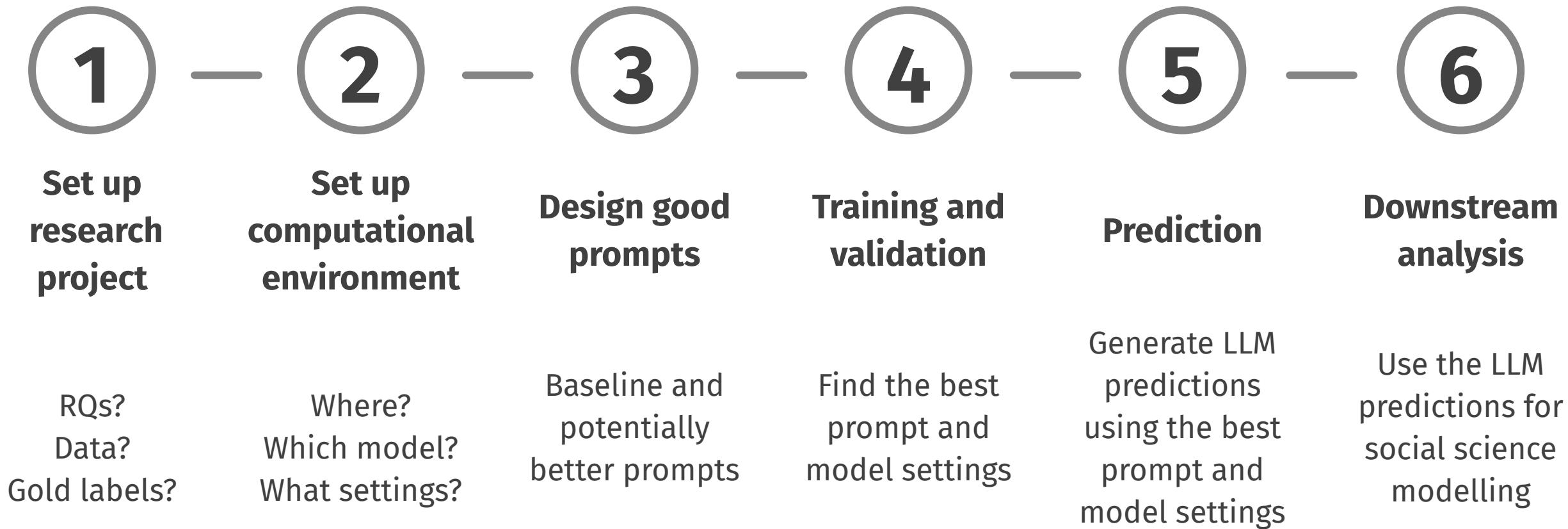


Using Large Language Models for Text Annotation in Social Science and Humanities: A Hands-On Python/R Tutorial

Qixiang Fang, Javier Garcia-Bernardo, Erik-Jan van Kesteren

Chair Oberski, Methodology and Statistics

Six stages of research



1. Set up research project

Know your research questions and data

- Is the information present in the data? Can human experts do it?
- Is it a subjective task? How much inter-rater agreement/variability can you expect?
- How capable do you expect your chosen LLM to be?

Examples:

- Predict someone's MBTI scores from their social media posts
- Predict someone's writing competency from their social media posts

1. Set up research project

Secure a gold-labelled subset

- You can't test any prompt or LLM model setup if you don't have gold standards to evaluate your LLM output against

Scenarios

- You already have a dataset with gold labels.
- You have a validated codebook or rubric from previous research. You only need to recruit annotators.
- You need to develop a codebook and collect annotations from scratch.
- You plan to use a strong LLM to generate an initial subset of labels, followed by careful manual verification to create the gold labels.

2. Set up computational environment

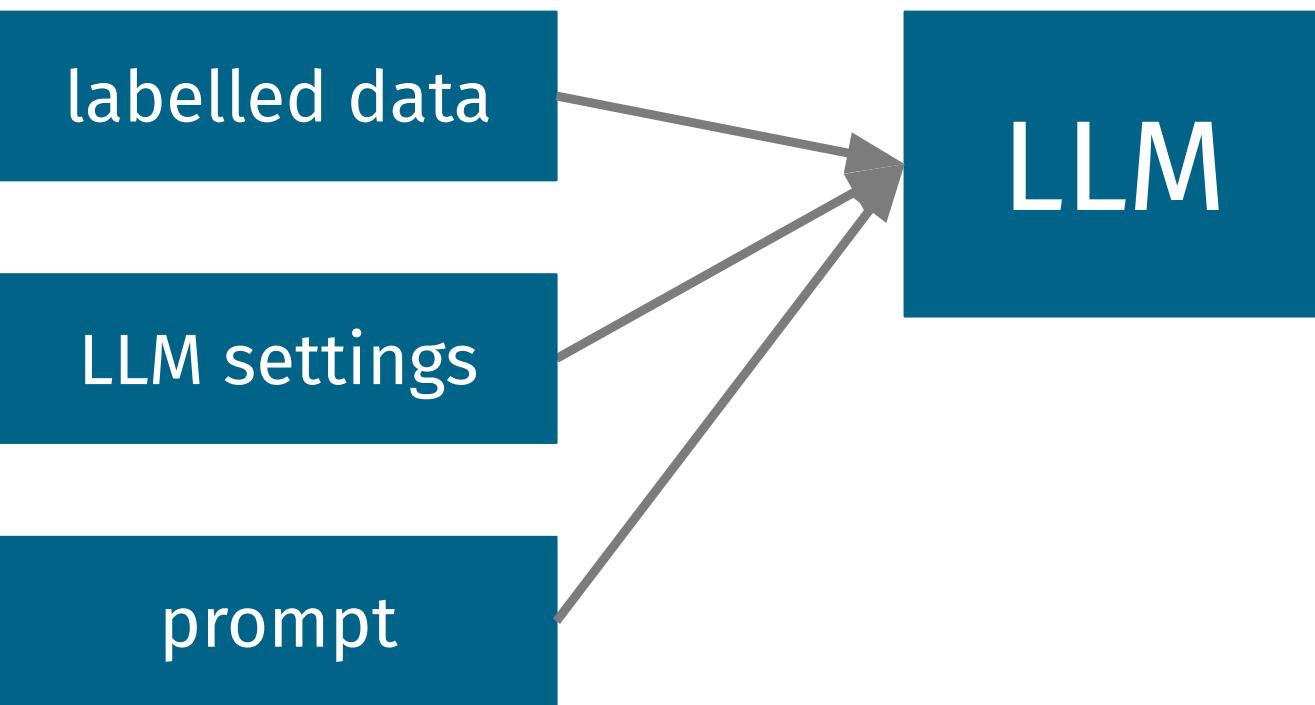
- **Quick start with cloud-based notebooks**
 - Choose LLM APIs over local (opensource) LLMs
- **Obtain and secure API keys**
- **Choose the right model to start with**
 - Pick a model recommended by its provider
 - Keep costs in mind
 - Choose reasonable hyperparameter values and ensure minimal stability

3. Design good prompts

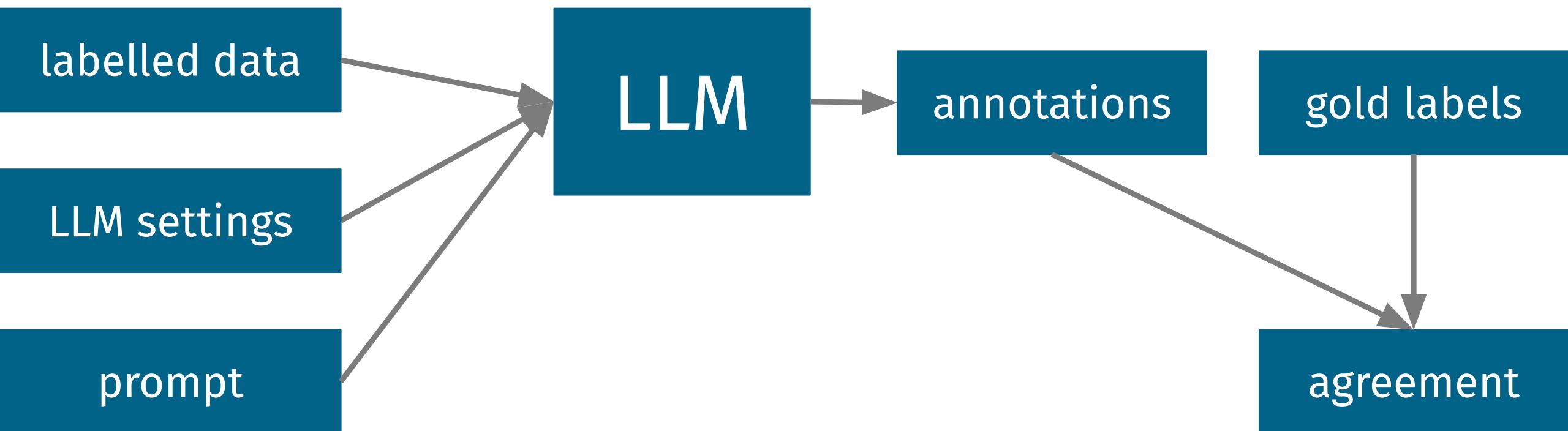
- Understand taxonomy and anatomy of prompts
- Apply prompt engineering techniques
- Come up with different but reasonably good prompts

Why > 1 prompt?

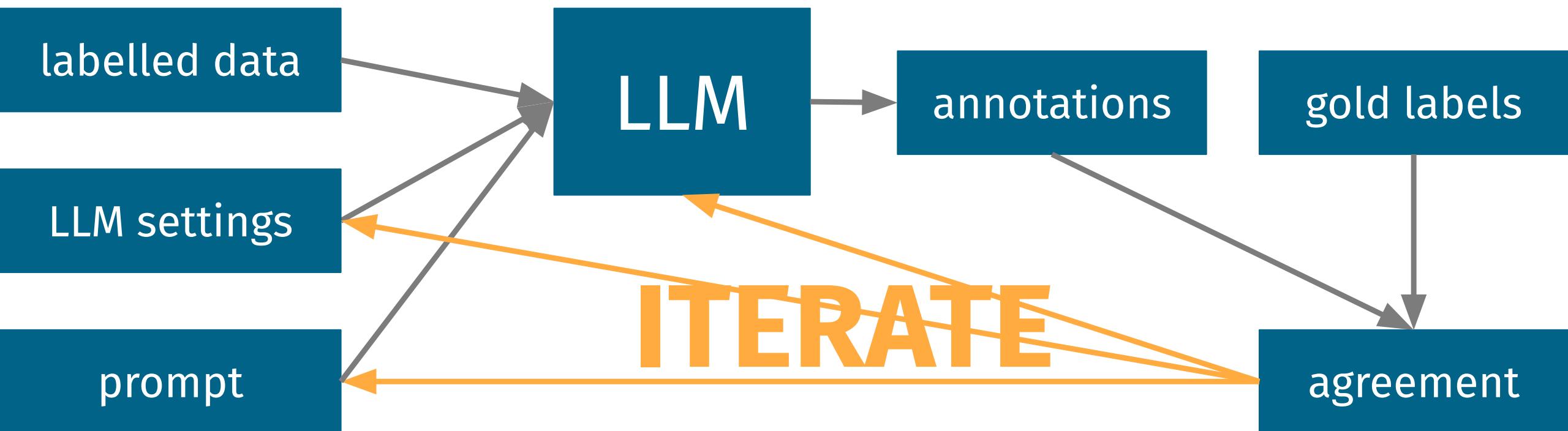
4. Training and validation



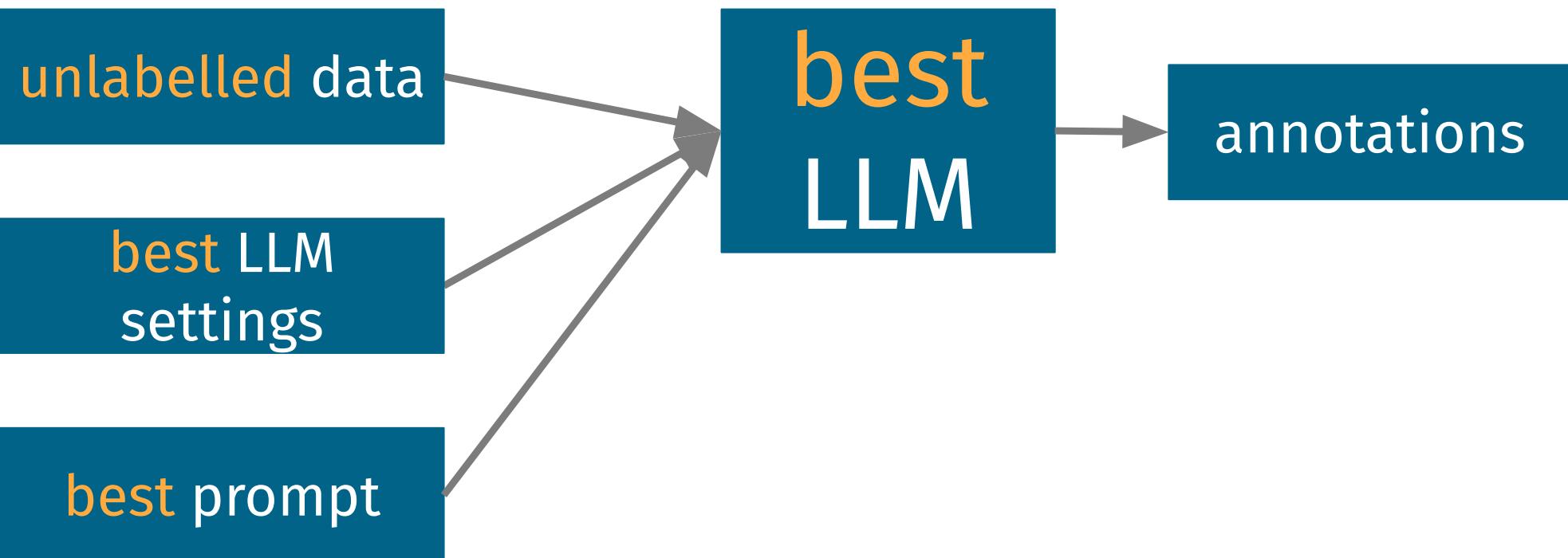
4. Training and validation



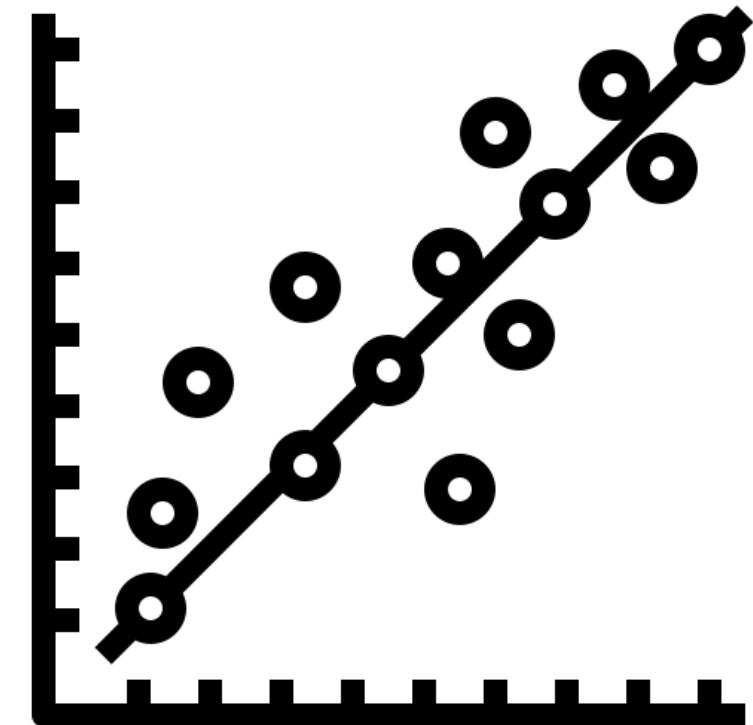
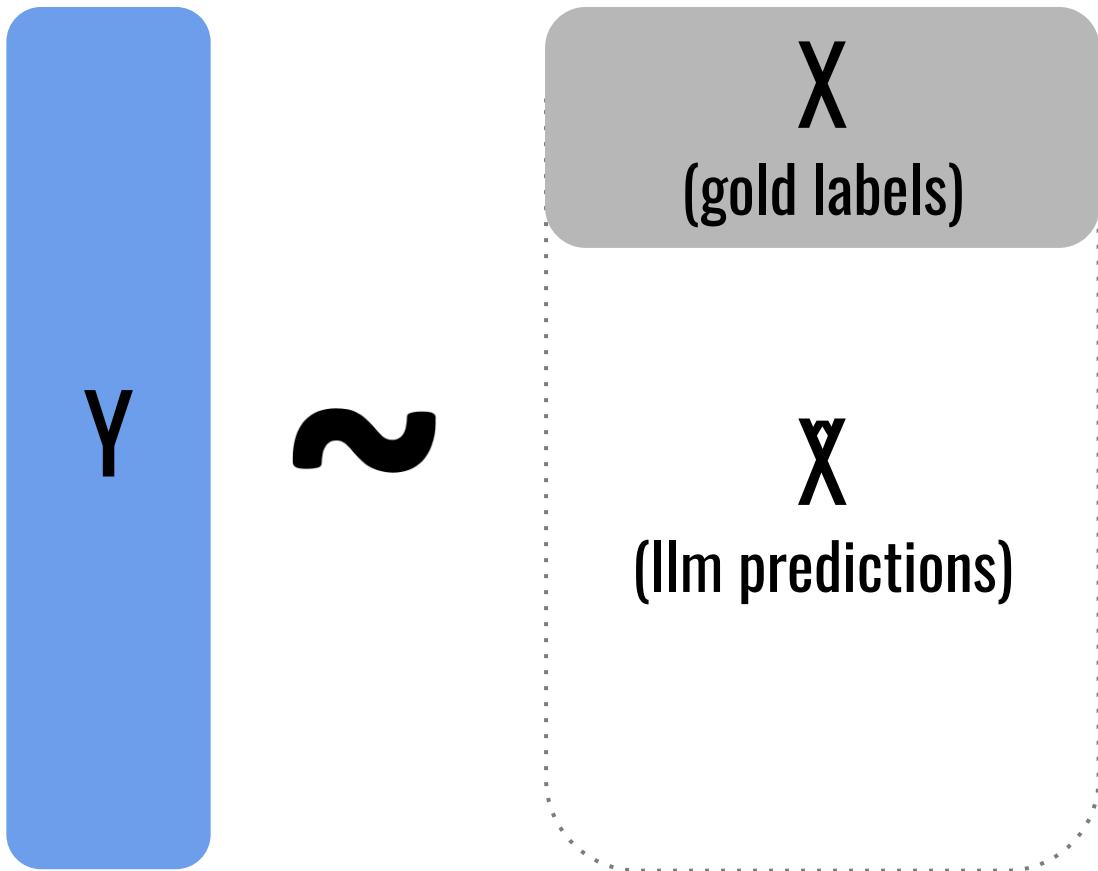
4. Training and validation



5. Prediction



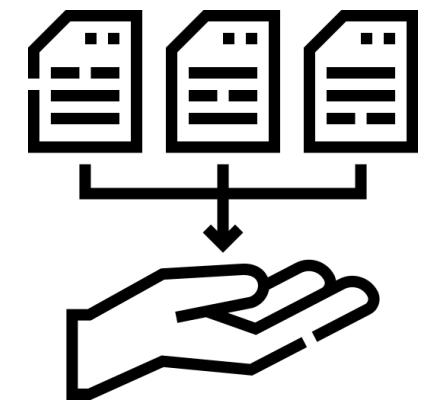
6. Downstream analysis



6. Downstream analysis

- Consider errors in LLM annotations
 - Systematic vs random
- How these errors affect statistical estimates?
- How to deal with them?

Part 2: **Data Collection/** **Annotation with LLMs**

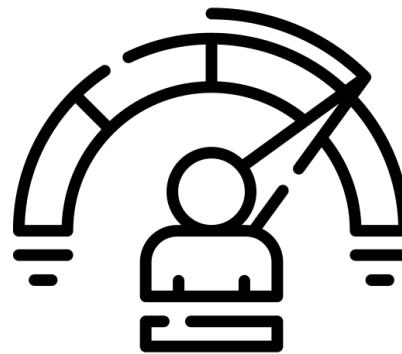
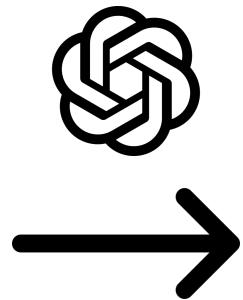


Recurring example

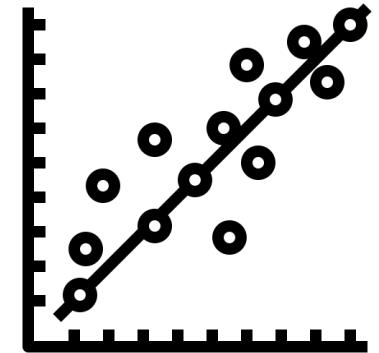
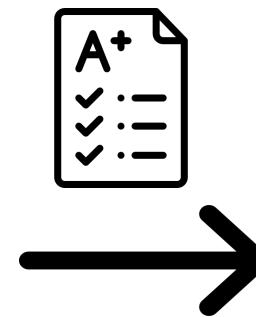
A SoDa fellowship project by Gabrielle Martins van Jaarsveld:



Conversations
between students
and a rule-based
chatbot



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



Regress study outcomes
on SRL indicators

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

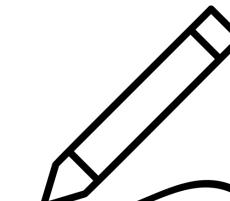
Chatbot: Why is this goal important to you in the context of your prior experiences and future goals?

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 questions from a Chatbot, 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 questions, and the student answers. Your objective is to assess the specificity 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 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

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 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 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

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]

Example 1 Scoring:

- Specificity: Score (Explanation)
- Measurability: Score (Explanation)
- Importance: Score (Explanation)
- Multi-Source Planning: Score (Explanation)

Example 2:

[example conversation]

Example 2 Scoring:

- Specificity: Score (Explanation)
- Measurability: Score (Explanation)
- Importance: Score (Explanation)
- Multi-Source Planning: Score (Explanation)

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]

Example 1 Scoring:

- Specificity: Score (Explanation)
- Measurability: Score (Explanation)
- Importance: Score (Explanation)
- Multi-Source Planning: Score (Explanation)

Example 2:

[example conversation]

Example 2 Scoring:

- Specificity: Score (Explanation)
- Measurability: Score (Explanation)
- Importance: Score (Explanation)
- Multi-Source Planning: Score (Explanation)

```
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?

LLM 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.

LLM hyperparameters

top_k: Restricts sampling to the k most likely next tokens.

- A lower value (e.g., 10) makes output more deterministic.
- A higher value (e.g., 50 or 100) allows for more diversity.

top_p: Instead of picking from the k most probable tokens, it selects from the smallest set of tokens whose probabilities sum to p .

- Lower values (e.g., 0.3) make responses more focused.
- Higher values (e.g., 0.9) increase diversity.

LLM 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
hop and pick your preferred notebook (Python/R).

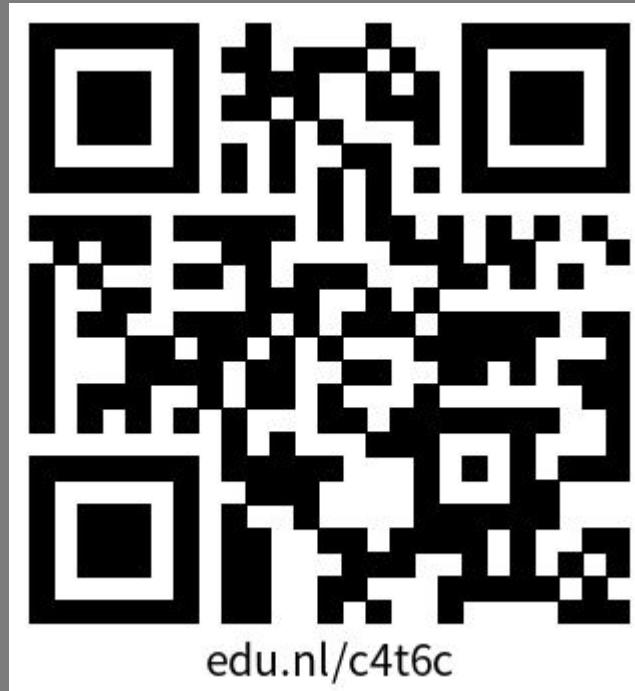
Full Workshop Schedule

Time	Title	Resource
09:30	LLM fundamentals for Social Sciences	
10:15	Coffee break	Coffee is provided!
10:30	Data collection with LLMs	python, R
12:00	Break	Lunch is provided!
12:45	Inference with LLM-collected data	python, R
13:45	Conclusion & Q&A	

Exercise:

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

Scan the QR code below to get an OpenAI API key and instructions to get started with the notebooks.



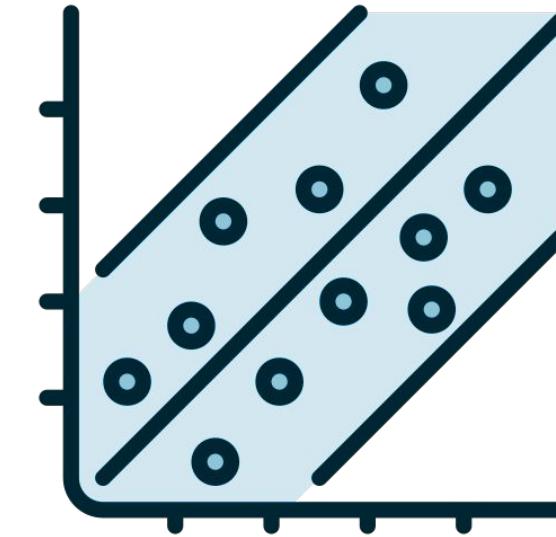
Lunch!
until 13.15



Who wrote the best prompt?

Part 3:

Inferences with LLM annotations



LLM-assisted social science modelling

$$Y \sim X \\ \hat{Y} \sim X'$$



Inspect your LLM responses

Take 5 minutes and share your findings.

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

**LLMs can produce incorrect responses
(i.e., measurements)!**

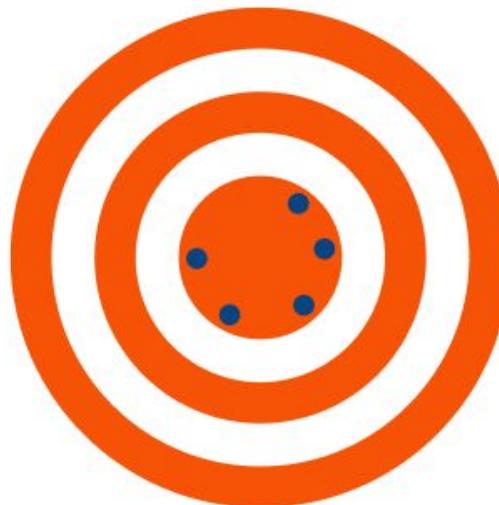
Measurement error

Random vs. systematic error

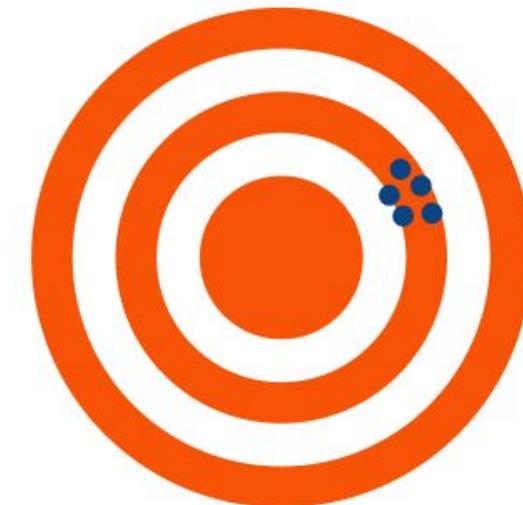
No error



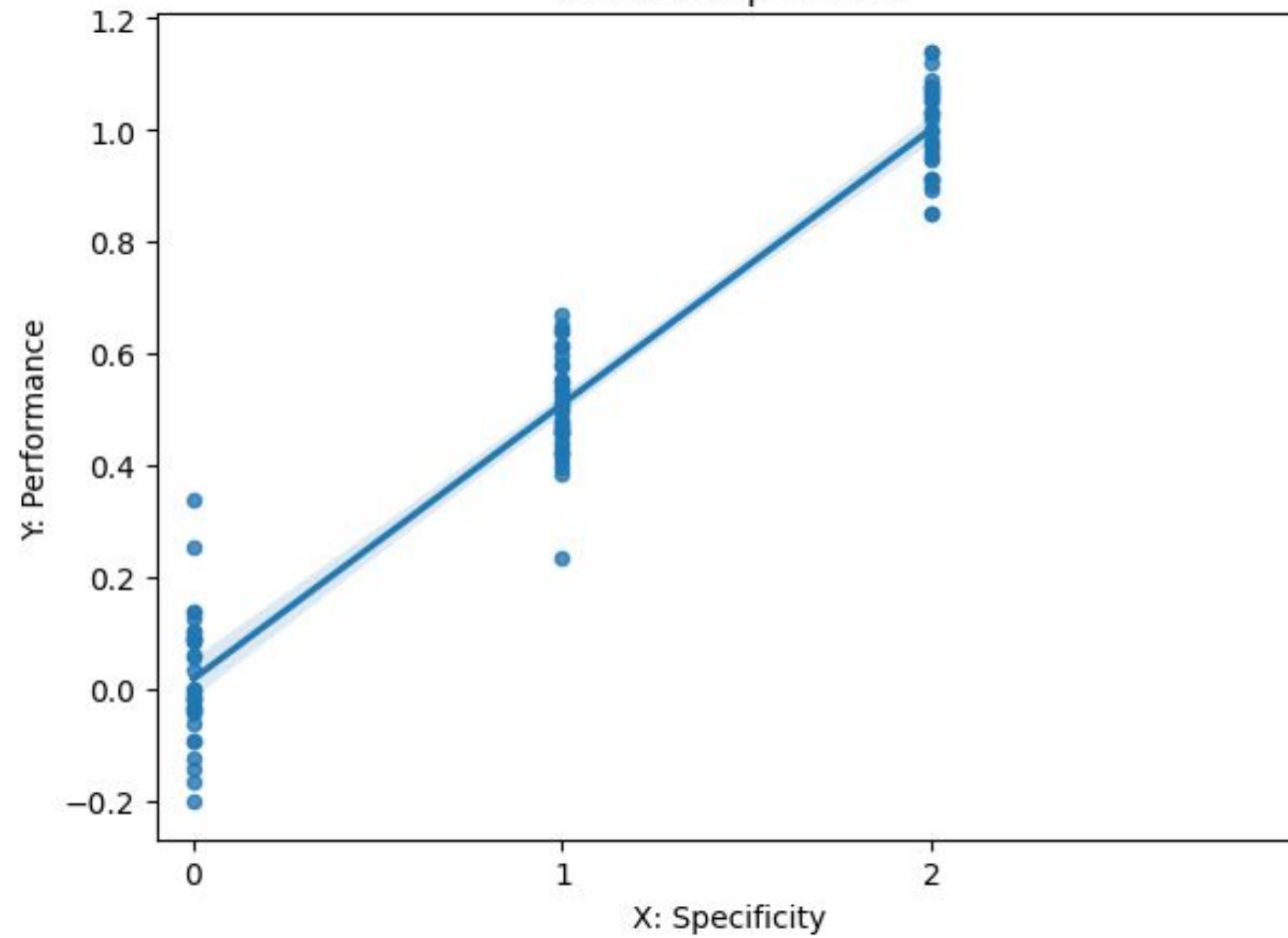
Random error



Systematic error



True X (No Measurement Error)
Estimated $\beta = 0.49$



X with Systematic Measurement Error
Estimated $\beta = 0.49$

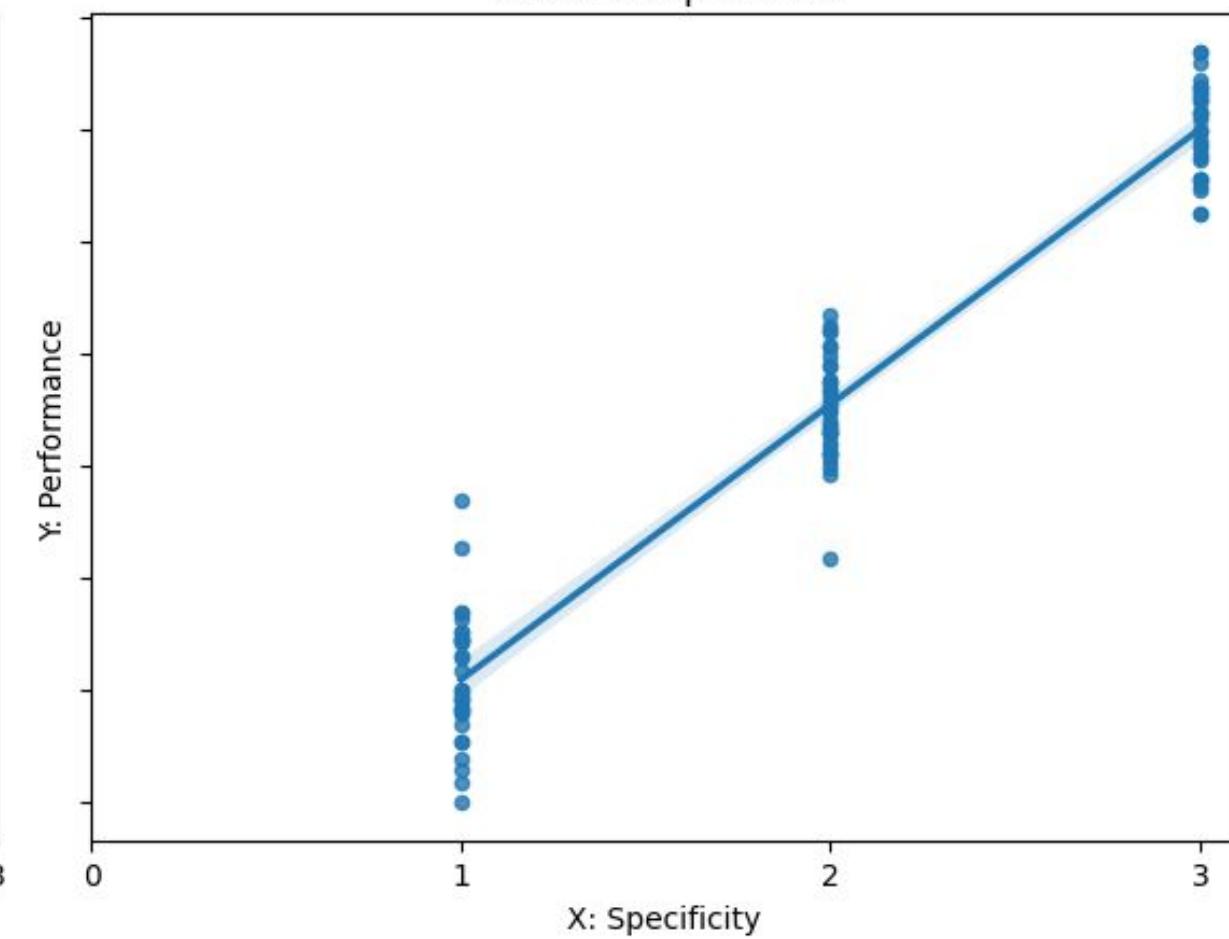
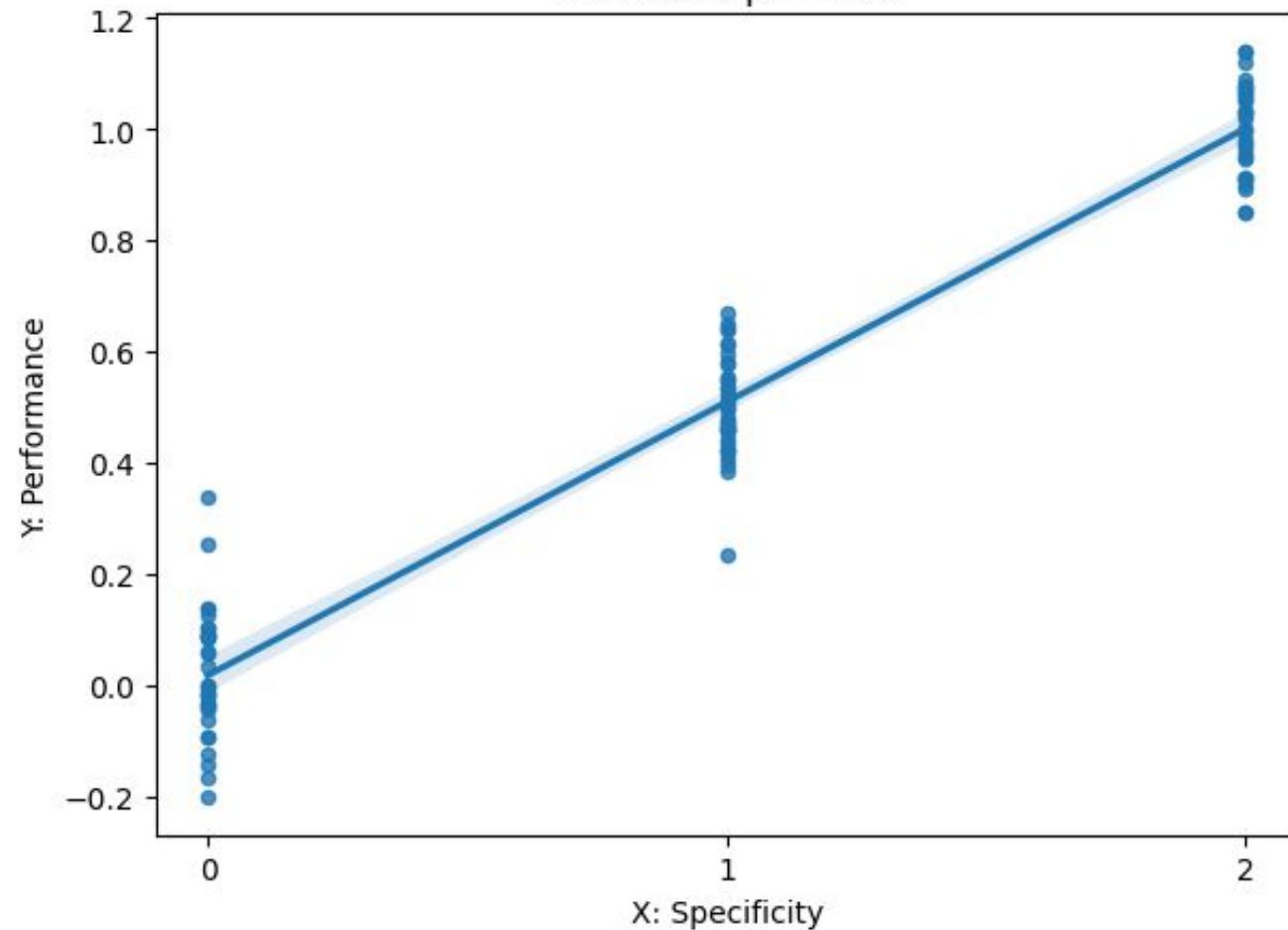


Illustration of systematic error in regression predictor

True X (No Measurement Error)
Estimated $\beta = 0.49$



X with Random Measurement Error
Estimated $\beta = 0.29$

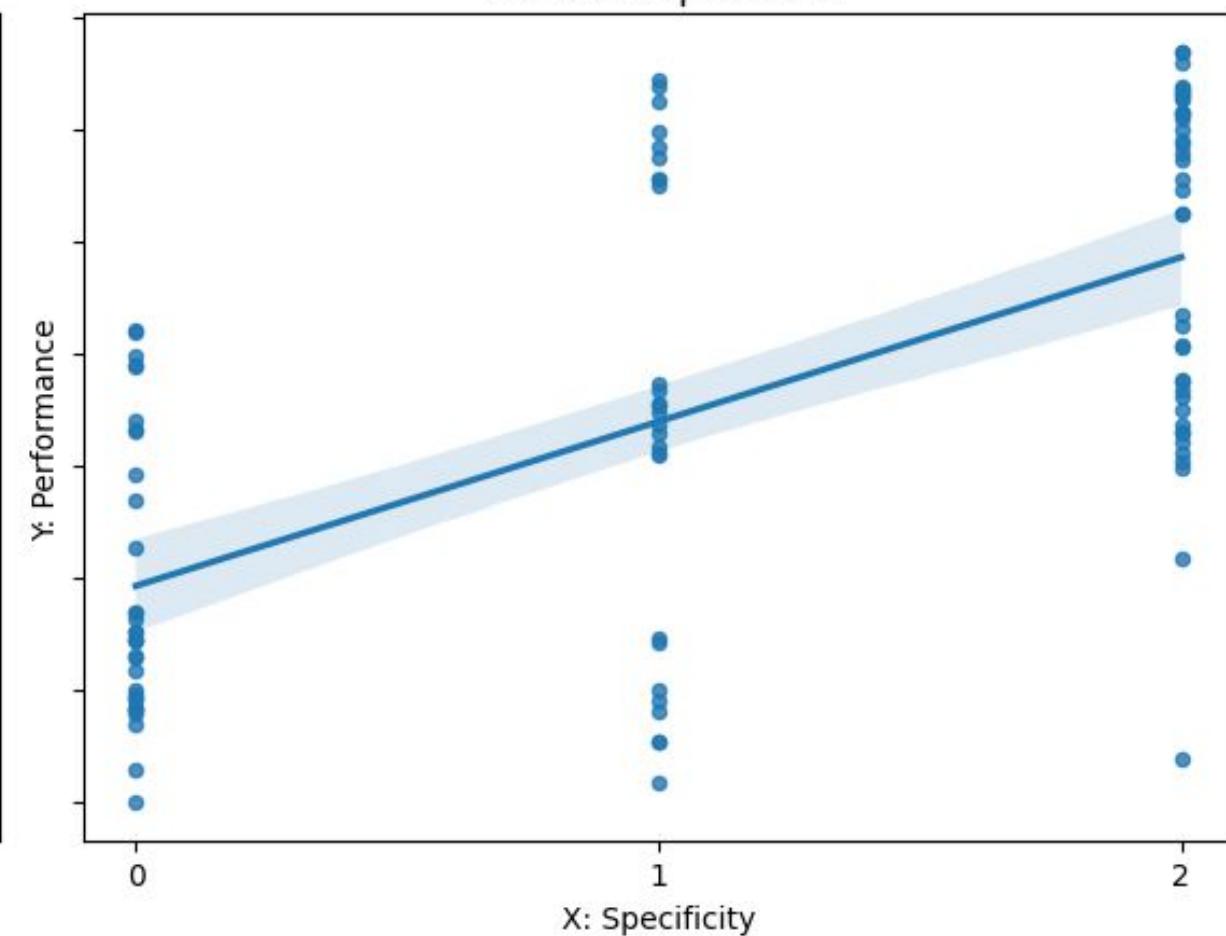
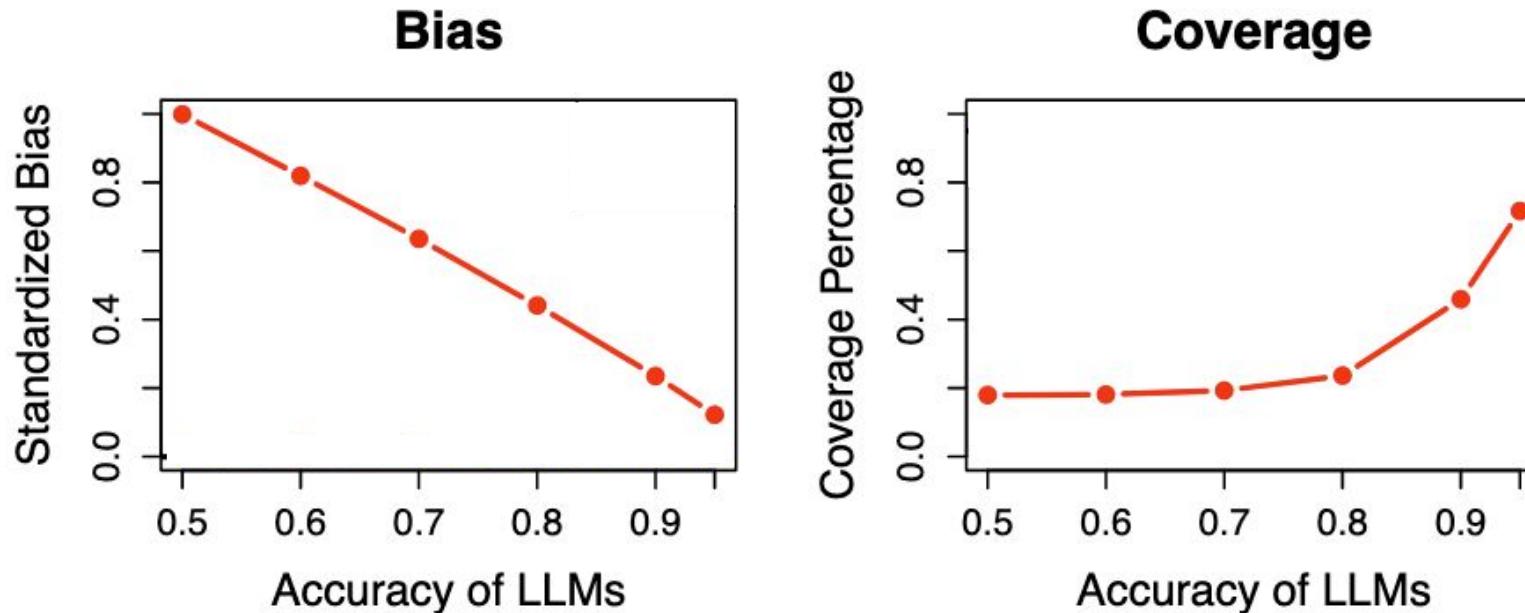


Illustration of random error in regression predictor

Also called
"attenuation bias"

But, aren't LLMs' responses generally good enough (e.g., 90% accuracy)?

Too good to be true



(a) **Simulated performance of Surrogate-Only Estimation (SO) and DSL.** Even for highly accurate surrogates, ignoring measurement error leads to non-trivial bias and undercoverage of 95% confidence intervals in downstream regression. Correct coverage and asymptotic unbiasedness are essential properties for proper uncertainty

Dealing with LLM measurement error

Common (suboptimal) approaches:

- Classic: Ignore the LLM predictions
- Naive: Treat the LLM predictions as error-free
- Mixed: Combine the LLM predictions and gold measurements
- Manual: Manually correct the LLM predictions
- Hard: Correct the LLM

Dealing with LLM measurement error

Ideally:

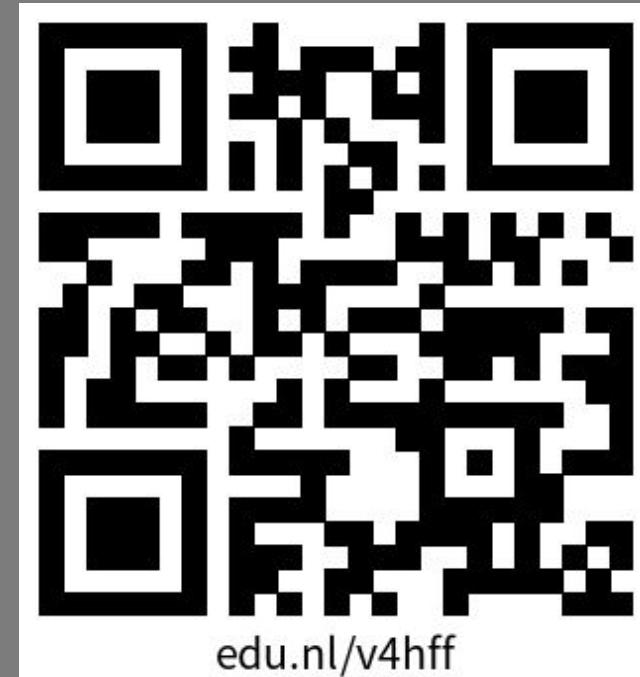
- No need to modify the prediction (i.e., LLM) model
- Leveraging LLM predictions
- Unbiased estimates
- Correct coverage
- Efficient coverage

An overview of

methods and software to deal
with LLM-related measurement
error for social science modelling

GitHub repo:

https://github.com/sodascience/social_science_inference_s_with_llms



Overview of recent methods

13 studies between 2020 and 2024

- Post-Prediction Inference (PostPI)
- Prediction-Powered Inference (PPI)
 - Efficient Prediction-Powered Inference (PPI++)
 - Cross-Prediction-Powered-Inference (Cross-PPI)
 - Bootstrap-based Method for Prediction-Powered Inference (PPBoot)
- PoSt-Prediction Adaptive inference (PSPA)
- PoSt-Prediction Summary-statistics-based (PSPS) inference
- Prediction De-Correlated Inference (PDC)
- Design-based Supervised Learning (DSL)
- etc.

Overview of recent methods

They all require some gold-standard (i.e. error-free) observations (Z):

- Focus on correcting the LLM-predicted data (\hat{Z})
 - PostPI: Predict Z from \hat{Z}
 - PPI: Predict $\hat{Z} - Z$ from W
 - DSL: Predict Z from \hat{Z} and W , with a sampling weight-based correction
- Focus on correcting the loss function
 - PPI: Add a correction term (mimicking $\hat{Z} - Z$) to the loss function
 - PDC: Remove the influence of \hat{Z} from the loss function
- Focus on correcting regression estimates afterwards
 - PSPS: Directly compute debiased regression estimates from biased model estimates

Overwhelming, technical?

Name	Method	Language	Estimators	Predicted Variables
PostPI	Post-Prediction Inference	R	Means, quantiles and GLMs	Outcome
PPI, PPI++, Cross-PPI, PPBoot	Prediction-powered inference and its extensions	Python	Any arbitrary estimator	Outcome
PSPA	PoSt-Prediction Adaptive inference	R	Means, quantiles, linear regression, logistic regression	Predictor and outcome
ipd	Implemented PostPI, PPI, PPI++ and PSPA	R	Means, quantiles, linear regression, logistic regression	Outcome
PSPS	PoSt-Prediction Summary-statistics-based (PSPS) inference	R and Python	M-estimators	Outcome
DSL	Design-based Supervised Learning	R	Moment-based estimators	Predictor and outcome

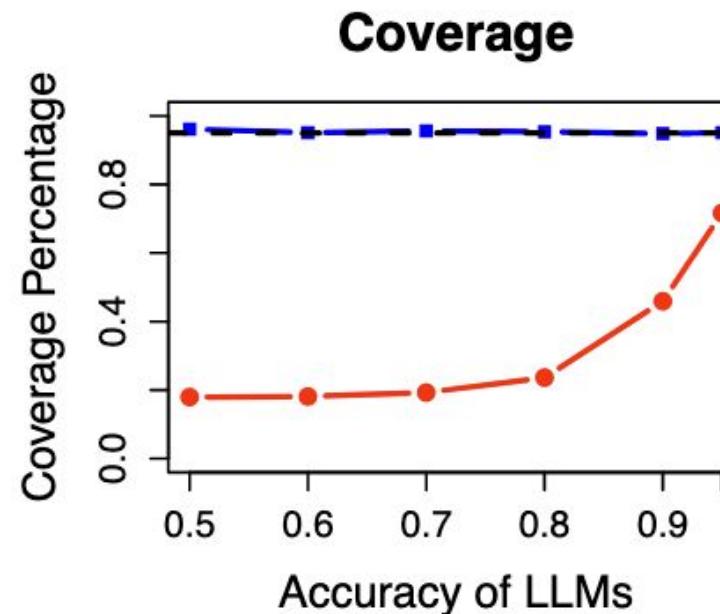
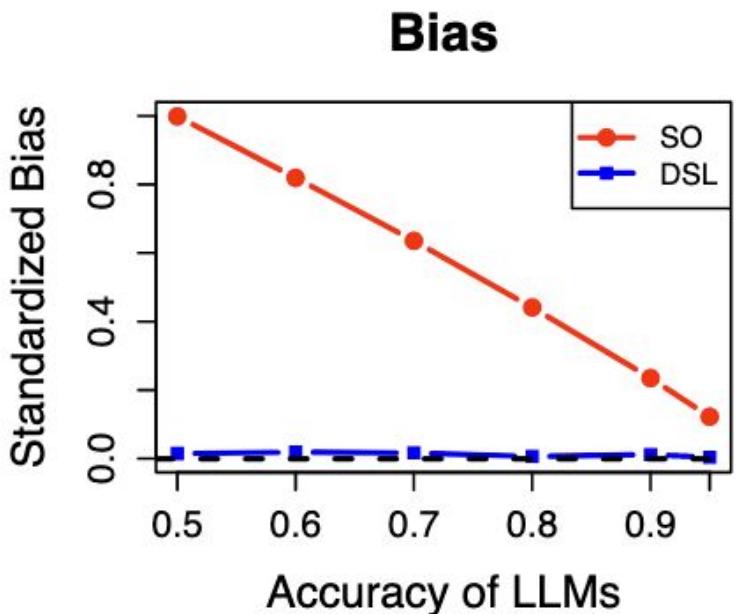
Practical recommendations

Depending on

- LLM-generated predictors ($X_{\hat{}}^{} \text{ or } Y_{\hat{}}^{} \text{?}$)
- Python, R or manual?
- GLM or other types of estimators

And they do work!

Literature overview of recent methods



(a) **Simulated performance of Surrogate-Only Estimation (SO) and DSL.** Even for highly accurate surrogates, ignoring measurement error leads to non-trivial bias and undercoverage of 95% confidence intervals in downstream regression. Correct coverage and asymptotic unbiasedness are essential properties for proper uncertainty

Exercise:

Modelling with LLM
measurement error

Go to Part III's notebooks. Feel
free to use your own data!

Full Workshop Schedule

Time	Title	Resource
09:30	LLM fundamentals for Social Sciences	
10:15	Coffee break	Coffee is provided!
10:30	Data collection with LLMs	python, R
12:00	Break	Lunch is provided!
12:45	Inference with LLM-collected data	python, R
13:45	Conclusion & Q&A	

R users: please **use your own
RStudio environment** due to the dsl
package's installation issue.

Ending Remarks

and some useful resources



Take-away

- LLMs can help to label data, and they can do this quite well.
- Before anything, it's VERY IMPORTANT to think about
 - the match between your RQs, data and LLM
 - having gold labels
- Carefully crafted prompts are key.
- langchain / ellmer can help you do the data collection.
- Check measurement quality!
- Correct your inferences if suspecting measurement issues.

Resource

Preprint:

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)

Tutorial Paper

- [Download](#)



edu.nl/bx3tj

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



Contact us

Get more info →

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.

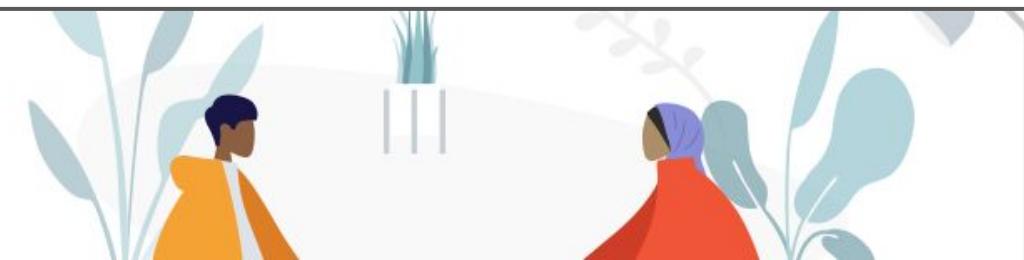
SoDa fellows will spend between 3-5 months full-time on their projects. During this time, they are paid members of the SoDa team at the Methodology & Statistics department of Utrecht University, mentored by one of the senior team members.

For more information, please reach out to [Kasia Karpinska](#), ODISSEI Scientific Manager.

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](#) 



Thanks!

Questions? Feedback?

References

- Egami, N., Hinck, M., Stewart, B., & Wei, H. (2024). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36.
- Fang, Q. & Garcia-Bernardo, J. & van Kesteren, E., 2025. Using Large Language Models for Text Annotation in Social Science and Humanities: A Hands-On Python/R Tutorial. *SocArXiv v4eq6_v1*, Center for Open Science.
- 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.