Workshop on

Using Large Language Models for Data Collection and Modelling in Social Sciences

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

Qixiang Fang, Javier Bernardo & Erik-Jan van Kesteren
Utrecht University

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 →

SoDa Team

- **Data scientists** at postdoc / assistant prof level
- **Research engineers** with experience 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

Agenda

Part I: Understanding LLMs (9.30 - 10.15)

- LLM fundamentals
- LLMs and social sciences

Coffee break! (10.15 - 10.30)

Part II: Data collection with LLMs (10.30 - 12.00)

- Prompt engineering
- Exercise: Design your own prompt experiment

Lunch! (12.00 - 12.45)

Part III: Inferences with LLM-based data (12.45 - 14.00)

- Inspect output from your own experiment
- Measurement problems with LLM responses
- Addressing these problems

Who has worked with LLMs?

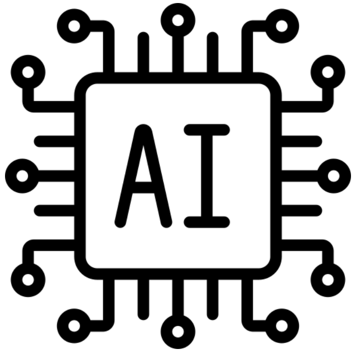
Kind reminders

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

Kind reminders

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

Part I:
Understanding LLMs



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 [...]

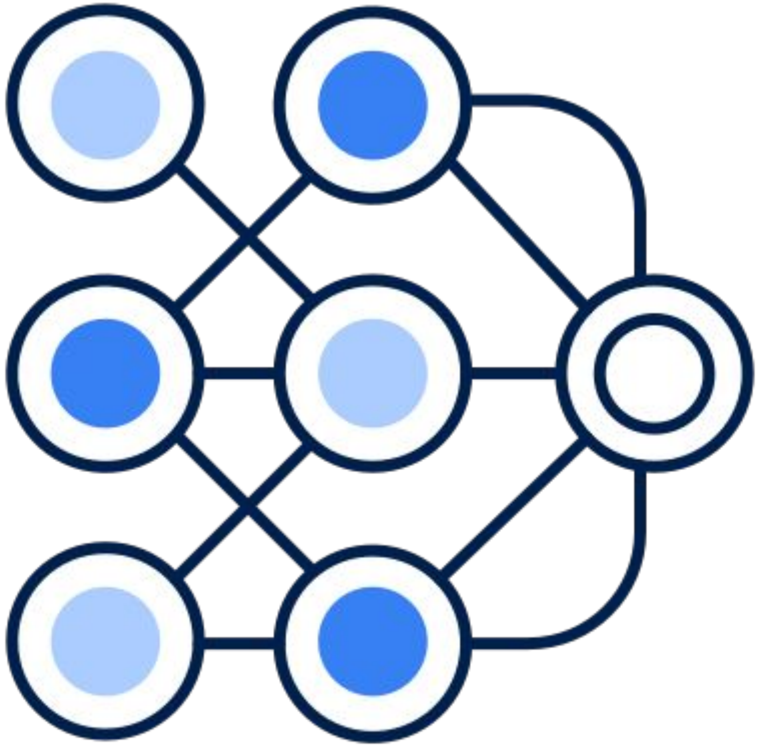
- Requires understanding about language.
- Requires understanding about common sense and world knowledge.

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

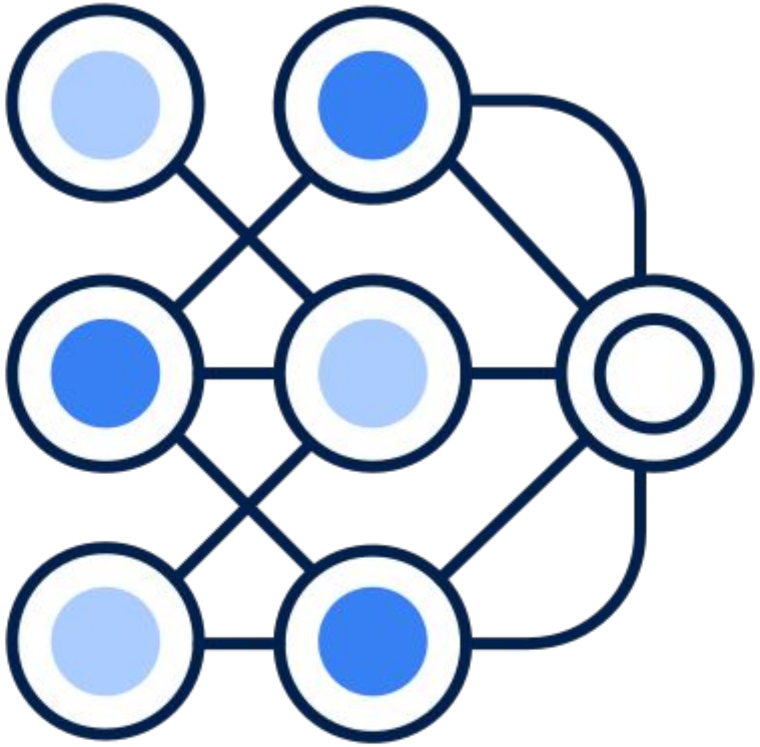
neural networks

deep learning

multi-head attention

Transformer

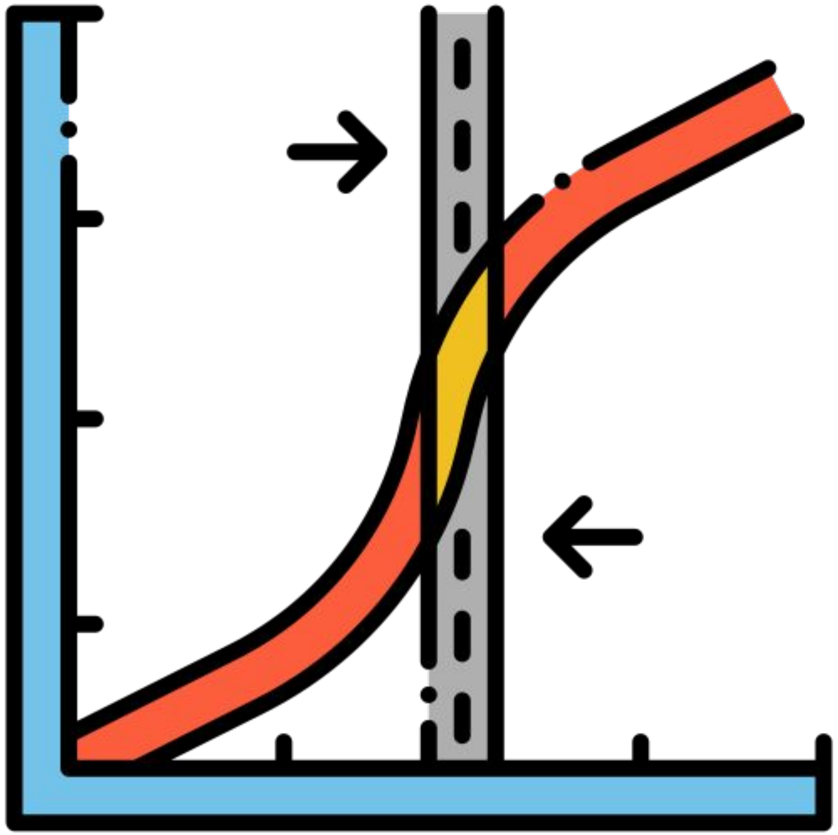
GPT



Modeling language by predicting it

It all comes down to a

logistic regression!



Binary logistic regression!

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

- Predictors: age, gender, nationality
- Outcome: business (yes/no)
- The model describes and has some understanding about this phenomenon (relationship between individual characteristics and academic choices)

gender	age	country	business
1	21	0	?
1	20	1	?
0	25	1	?
…	…	…	…

Social sciences (data)

Binary logistic regression of language

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

In the case of language modelling/prediction:

- Each word requires some numerical representation (just like the social science example).

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

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

this	the	21	19	y/o	male	fema le	stud ent	from	…	is	not	stud ying	learn ing	busi ness
1	0	1	0	1	1	0	1	1	…	1	0	1	0	?
0	1	0	1	1	0	1	1	1	…	1	1	0	1	?
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Language modelling (data)

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

this	the	21	19	y/o	male	fema le	stud ent	from	…	is	not	stud ying	learn ing	busi ness
0.1 0.2 2.1 -1.1 … 0.1	0.4 -0.2 -1.0 1.1 … 0.7	…	…	…	…	…	…	…	…	…	…	…	-0.2 0.5 1 -1.1 … 0.2	?

Language modelling (data)

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.

Multinomial logistic regression!

Social sciences:

- A fixed list of outcome categories
  - e.g.,

engineering
medicine
business

Language modeling:

- Use the entire vocabulary!

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	learn ing	busi ness
1	0	1	0	1	1	0	…	1	0	1	0	0	0	0
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

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	learn ing	busi ness
0	1	0	1	1	0	1	…	1	1	0	1	0	0	0
?	?	?	?	?	?	?	?	?	…	0	0	?	?	?

Iterating...

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

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

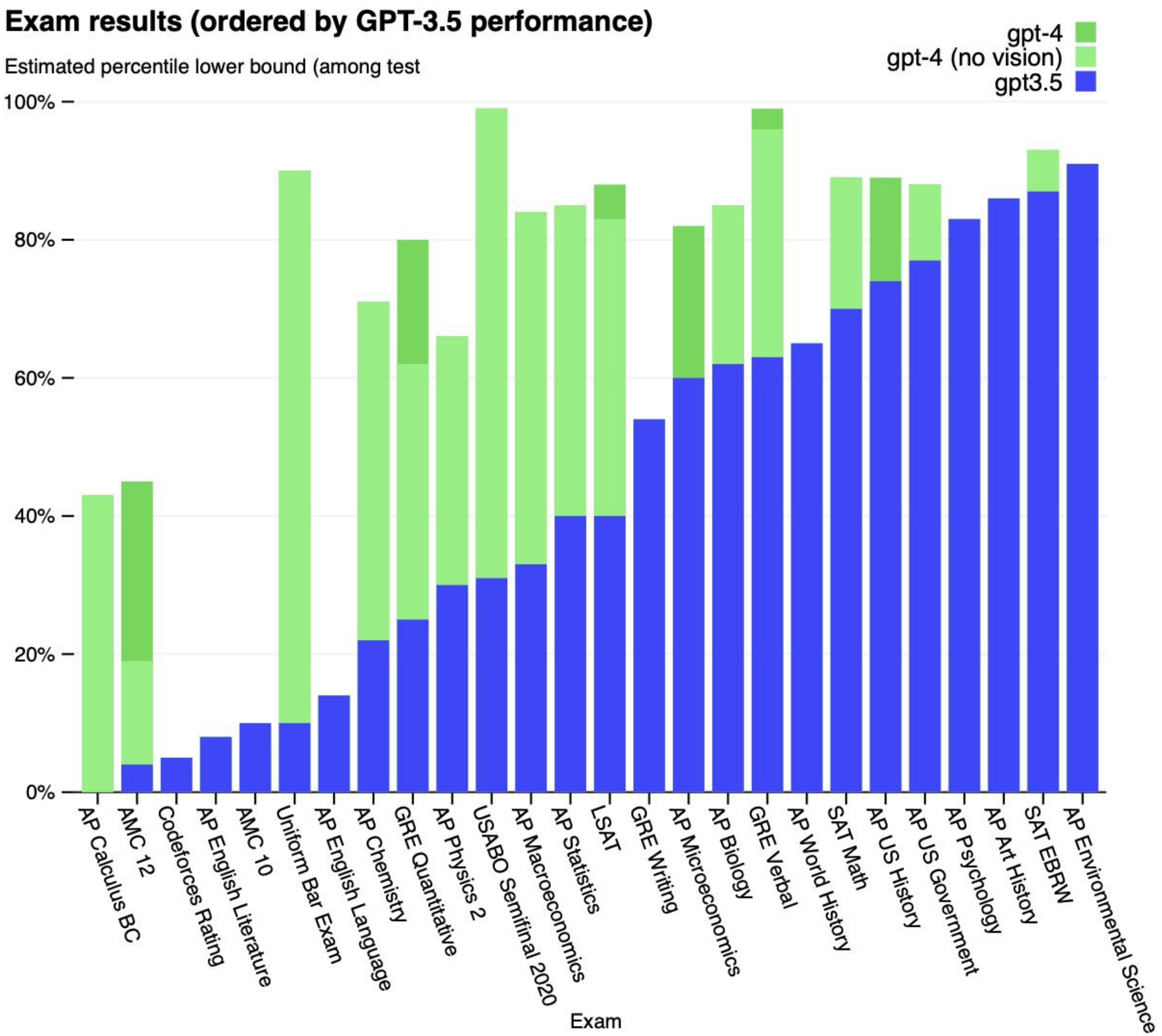


Figure 4. GPT performance on academic and professional exams. In each case, we simulate the

Benchmarking

https://arxiv.org/pdf/2303.08774

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]
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]

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 business student from Germany.

Reddit Q&A:

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

From next word prediction to beyond!

The model is further trained on task-specific prompt data.

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.

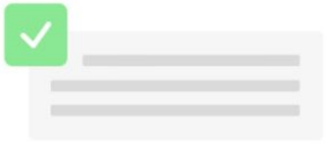
Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group e.g., "`Whore`" *conveys hostile and contemptuous female expectations* (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

Teach

We start by teaching our AI right from wrong, filtering harmful content and responding with empathy.

Test

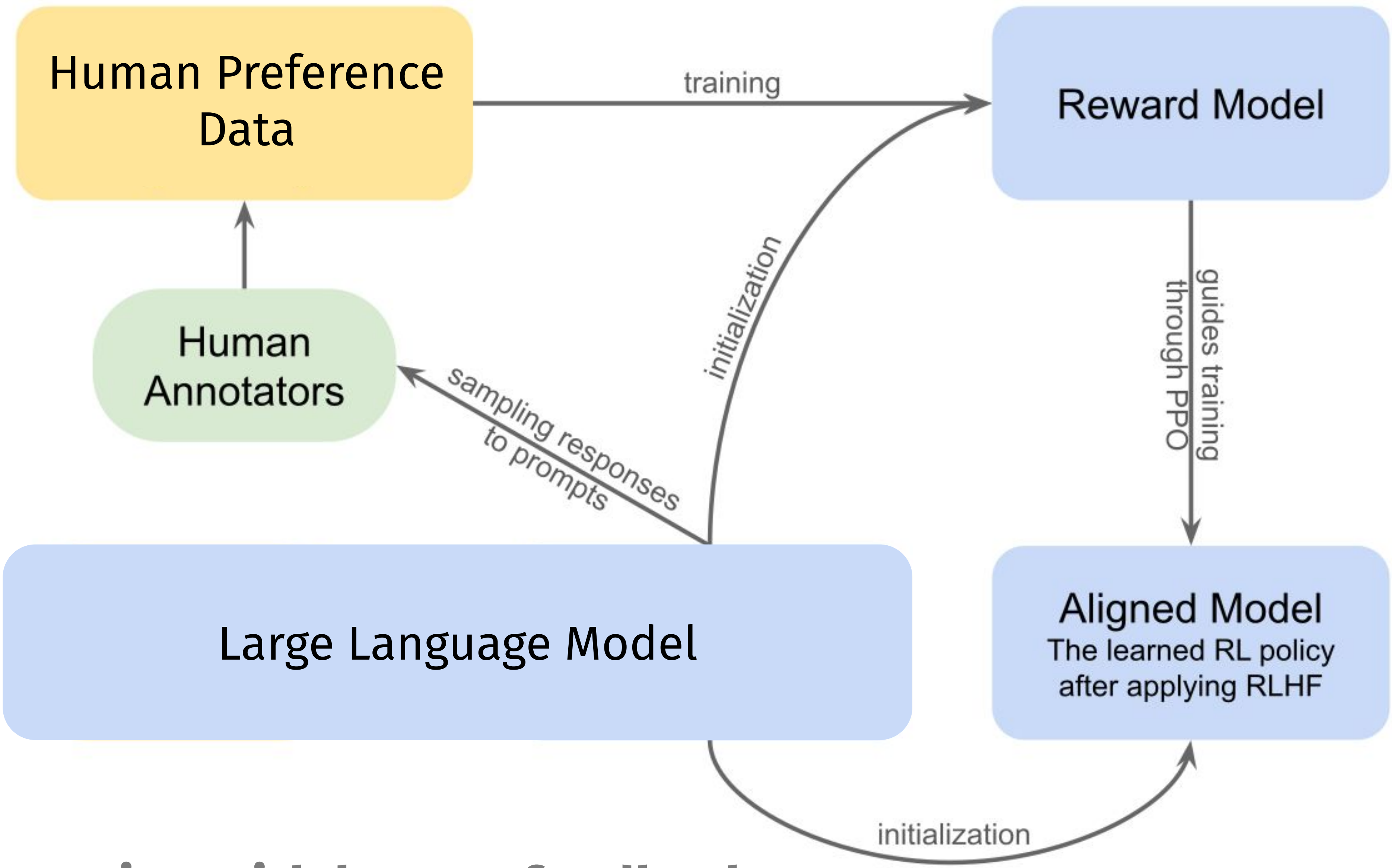


We conduct internal evaluations and work with experts to test real-world scenarios, enhancing our safeguards.

Share

We use real-world feedback to help make our AI safer and more helpful.

https://openai.com/safety/

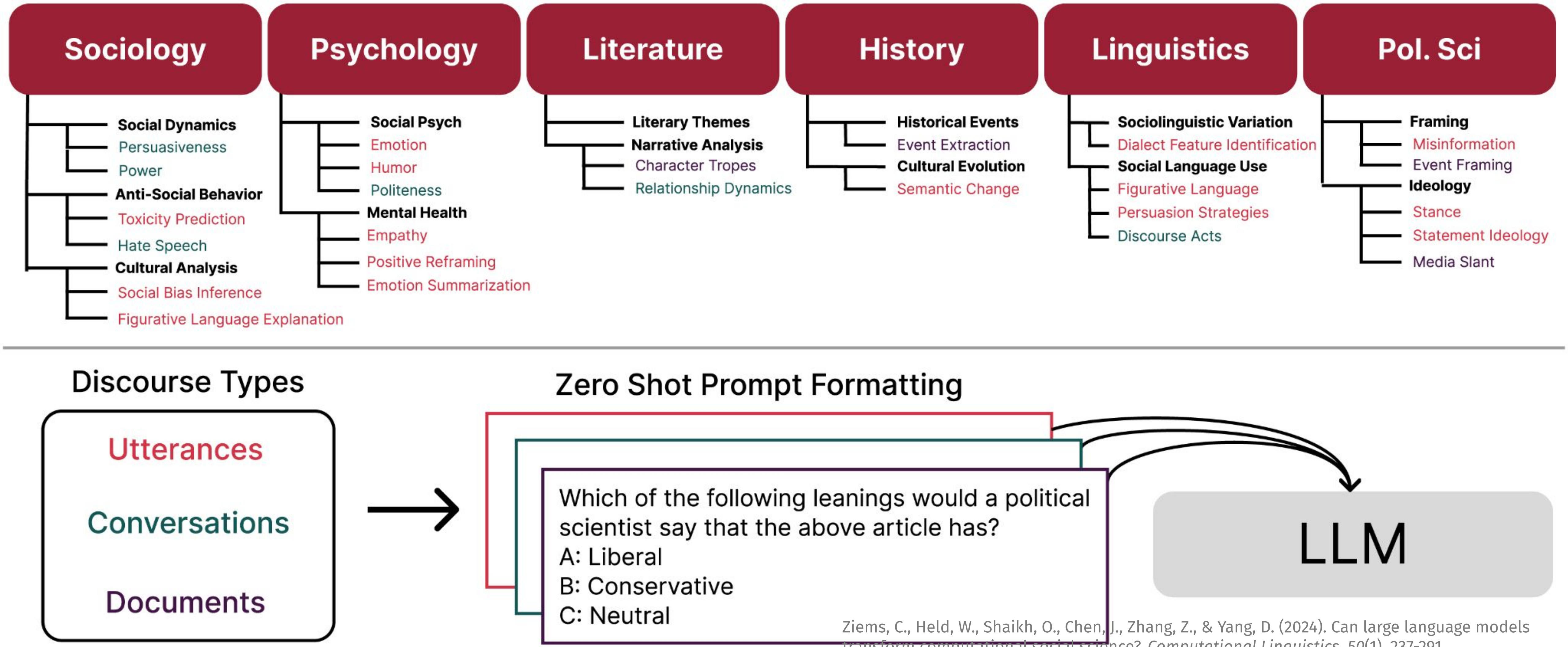


Reinforcement learning with human feedback

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

Questions?

LLMs in social sciences



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.

LLM in social sciences (more)

Psychometrics

- Pre-test the quality of potential test items

Opinion mining

- Estimate average opinions across different societies

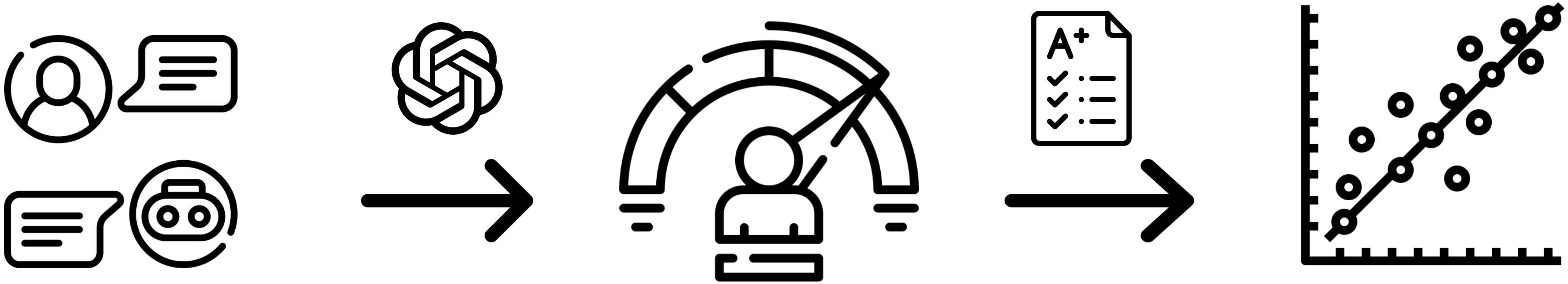
Automation of systematic reviews

- Ask for inclusion decision based on inclusion criteria and document

...

LLM in social sciences (SoDa)

A SoDa fellowship project by Gabrielle Martins van Jaarsveld:



Conversations between students and a learning chatbot

LLM-based measurements of indicators of self-regulated learning

Regress study outcomes on indicators of self-regulated learning

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

SRL Forethought Phase: Goal Setting & Planning

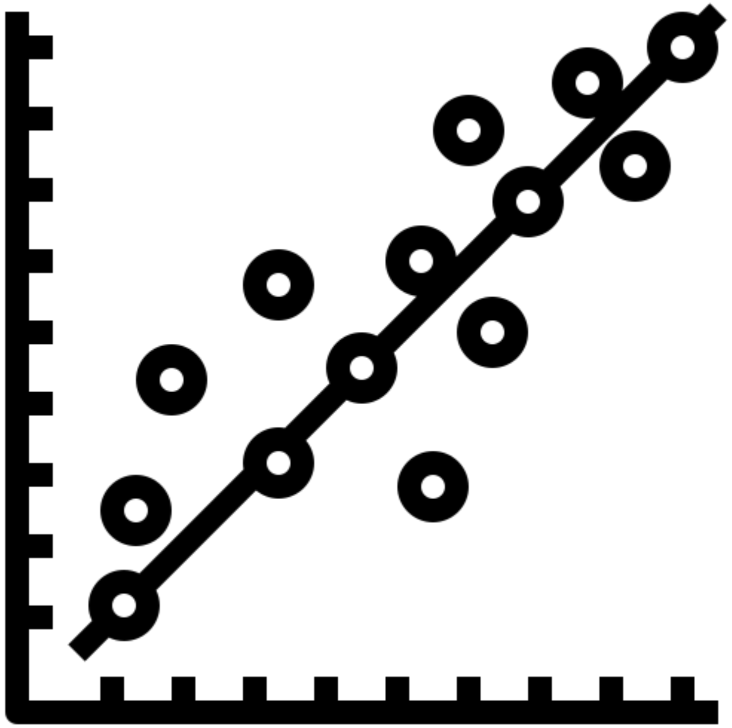
- Specificity (1/2)
- Measurability (1/2)
- Importance (2/2)
- Realistic multisource planning (1/2)

SRL Performance Phase: Monitoring

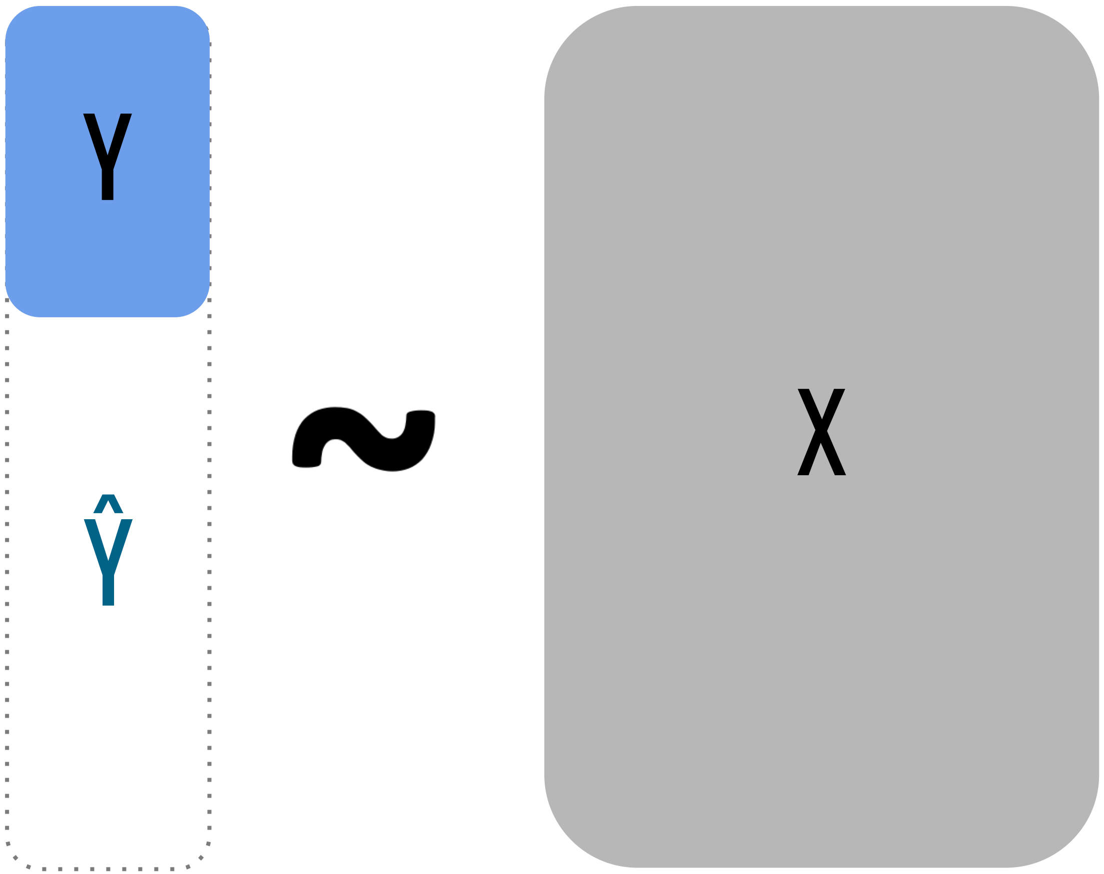
SRL Reflection Phase: Reflection & Adaptation

Traditional social science modelling

Y ~ X



LLM-assisted social science modelling



LLM-assisted social science modelling

Y ~ X

X'

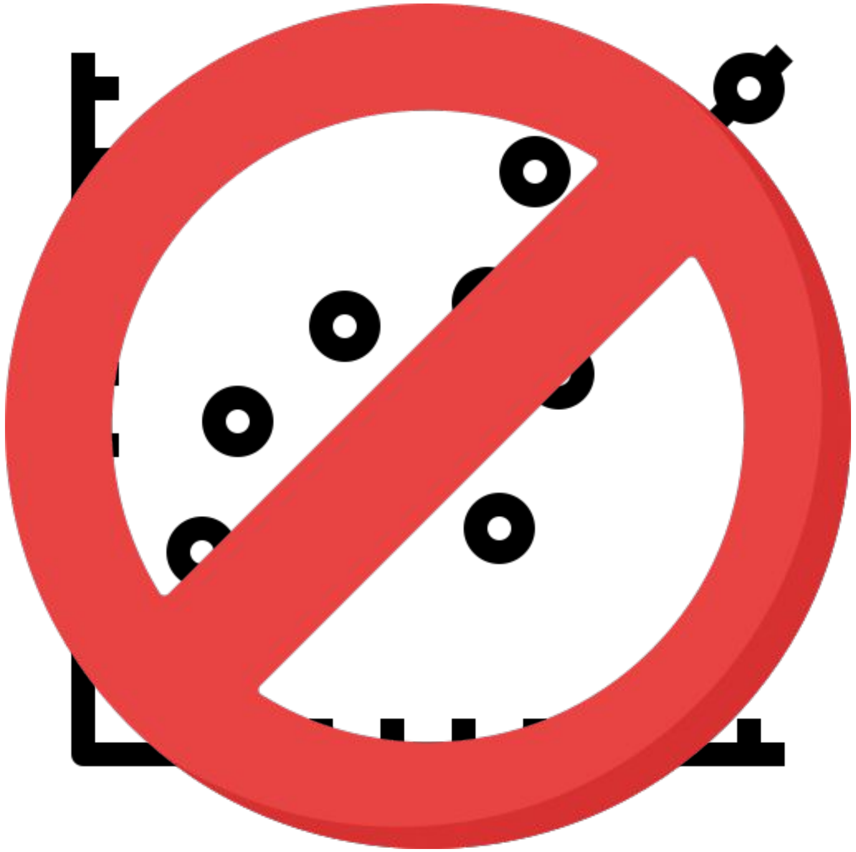
LLM-assisted social science modelling

$Y \sim X$

$\hat{Y} \quad X'$

LLM-assisted social science modelling

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

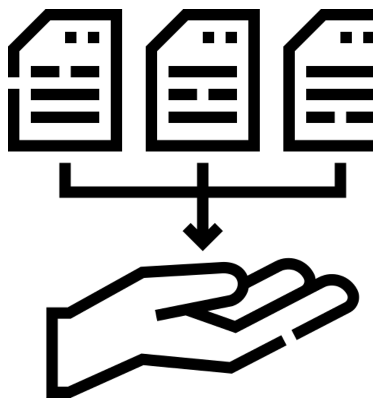


Questions?

Coffee break
until 10.30

Part II:

Data collection with LLMs



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:

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

A simple 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.

PROMPT: Set an academic goal for the upcoming week.
ANSWER: I would like to catch up on my geography reading
PROMPT: Add details to make your goal more specific.
ANSWER: 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.
PROMPT: How will you measure progress on and acheivement of your goal?
ANSWER: by the number of pages I write per day
PROMPT: Why is this goal important to you in the context of your prior experiences and future goals?
ANSWER: It is important to achieve because if I dont, I will fall behind and most likely wont be ready for the exam.
PROMPT: Create a step-by-step plan for achieving this goal in the coming week.
ANSWER: 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)
- User prompt (specific)

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

PROMPT: Set an academic goal for the upcoming week.
ANSWER: I would like to catch up on my geography reading
PROMPT: Add details to make your goal more specific.
ANSWER: 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.
PROMPT: How will you measure progress on and acheivement of your goal?
ANSWER: by the number of pages I write per day
PROMPT: Why is this goal important to you in the context of your prior experiences and future goals?
ANSWER: It is important to achieve because if I dont, I will fall behind and most likely wont be ready for the exam.
PROMPT: Create a step-by-step plan for achieving this goal in the coming week.
ANSWER: 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 & Specificity**
 - 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 & Specificity
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 & Specificity
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 & Specificity
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 & Specificity
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 & Specificity
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: If supported, setting a seed ensures reproducibility, generating the same response when used with the same prompt and parameters.

Prompt engineering 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.

Prompt engineering hyperparameters

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

More resources

OpenAI:
https://platform.openai.com/docs/guides/prompt-engineering

Prompt engineering guide: https://www.promptingguide.ai

Anthropic:
https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview

Exercise:

Design your own prompt experiment

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

Python: langchain package

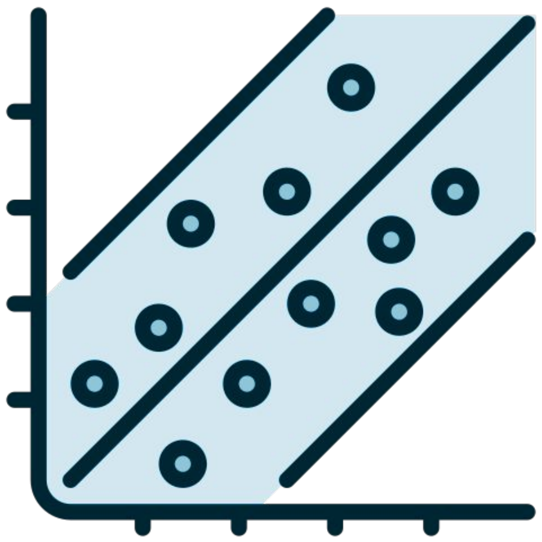
R: ellmer package

Need an OpenAI API key?

Lunch!

until 12.45

Part III:
Inferences with LLM-based data



LLM-assisted social science modelling

$Y \sim X$

$\hat{Y}$ 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., measurement)!

Measurement error

Systematic error (bias):

- Occurs consistently in the same direction (e.g., always overestimating or underestimating the true value).

- Caused by flaws in the measurement instrument, method, or external influences.
- Since it is predictable, it can often be corrected or adjusted for.

Measurement error

Systematic error (bias):

- Occurs consistently in the same direction (e.g., always overestimating or underestimating the true value).
- Caused by flaws in the measurement instrument, method, or external influences.
- Since it is predictable, it can often be corrected or adjusted for.

Random error (noise):

- Occurs unpredictably across measurements due to unpredictable factors like human variability, environmental changes, or instrument fluctuations.
- Leads to inconsistent results that scatter around the true value.
- While it cannot be eliminated completely, it can be reduced by averaging multiple measurements.

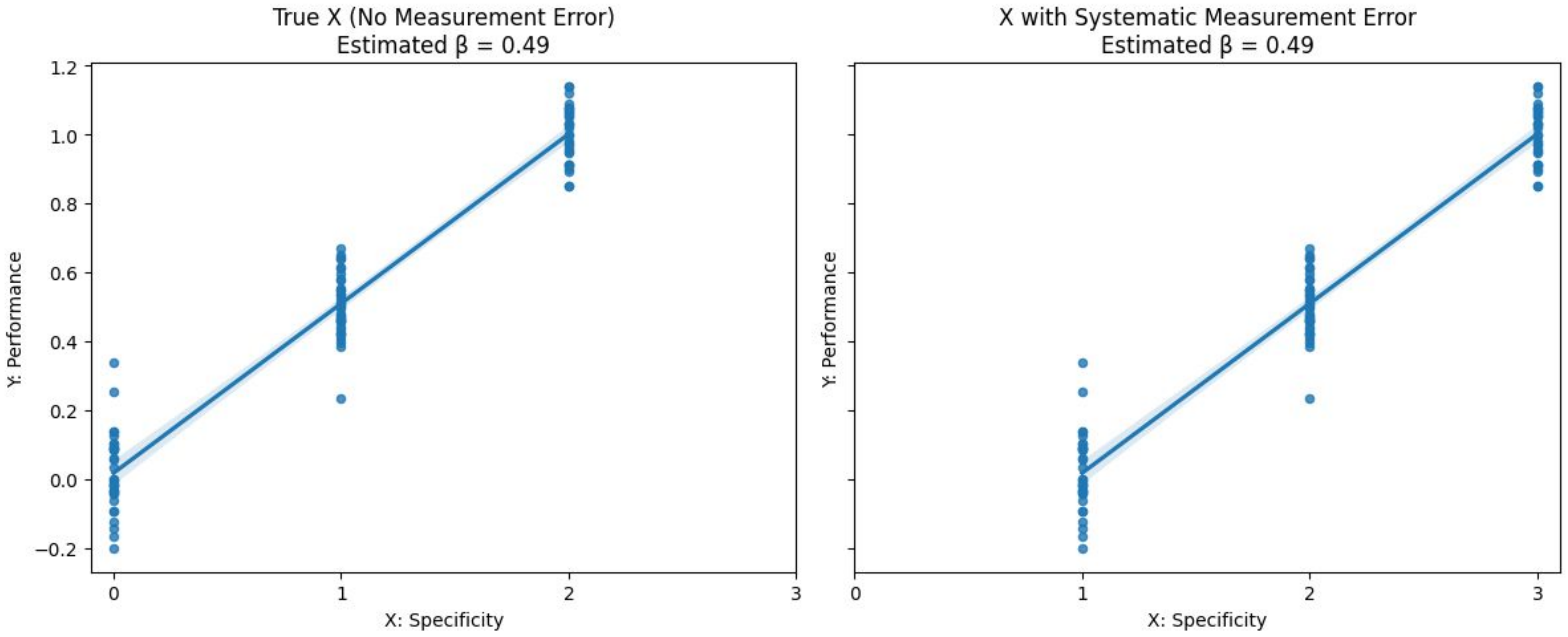


Illustration of systematic error in regression predictor

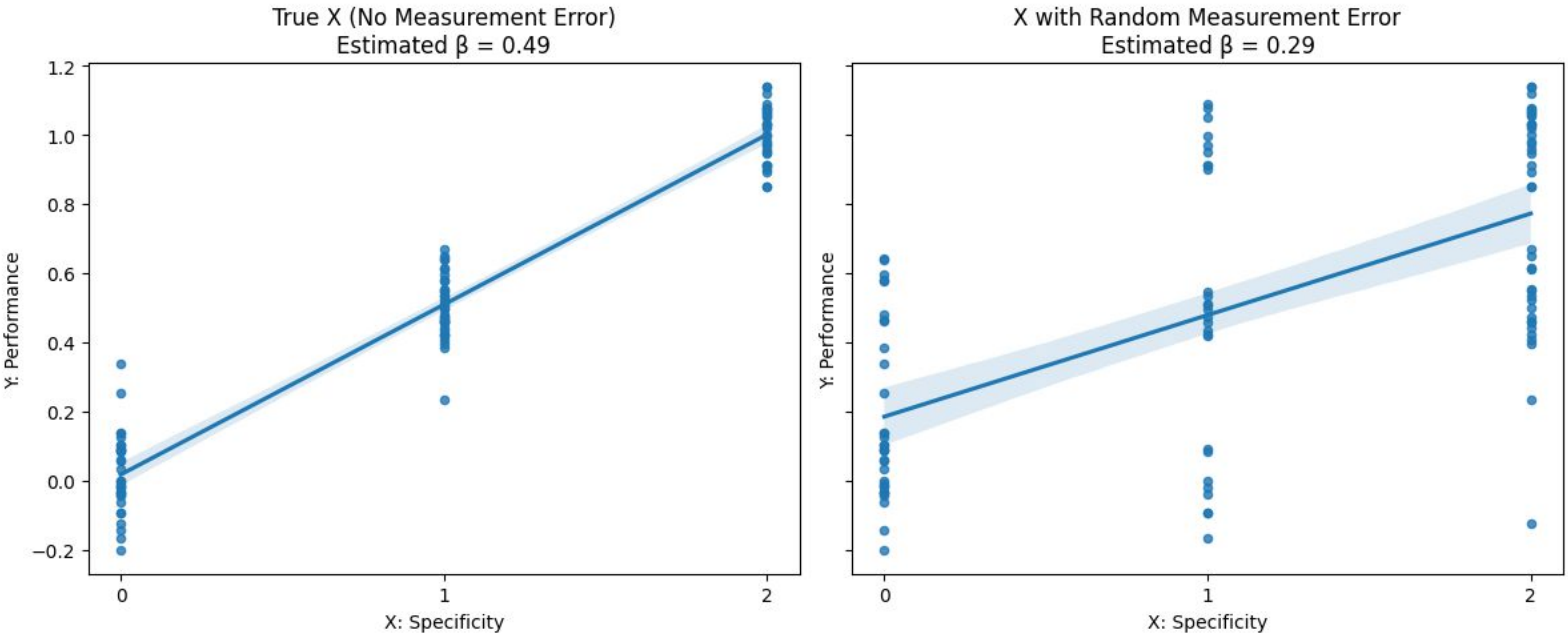
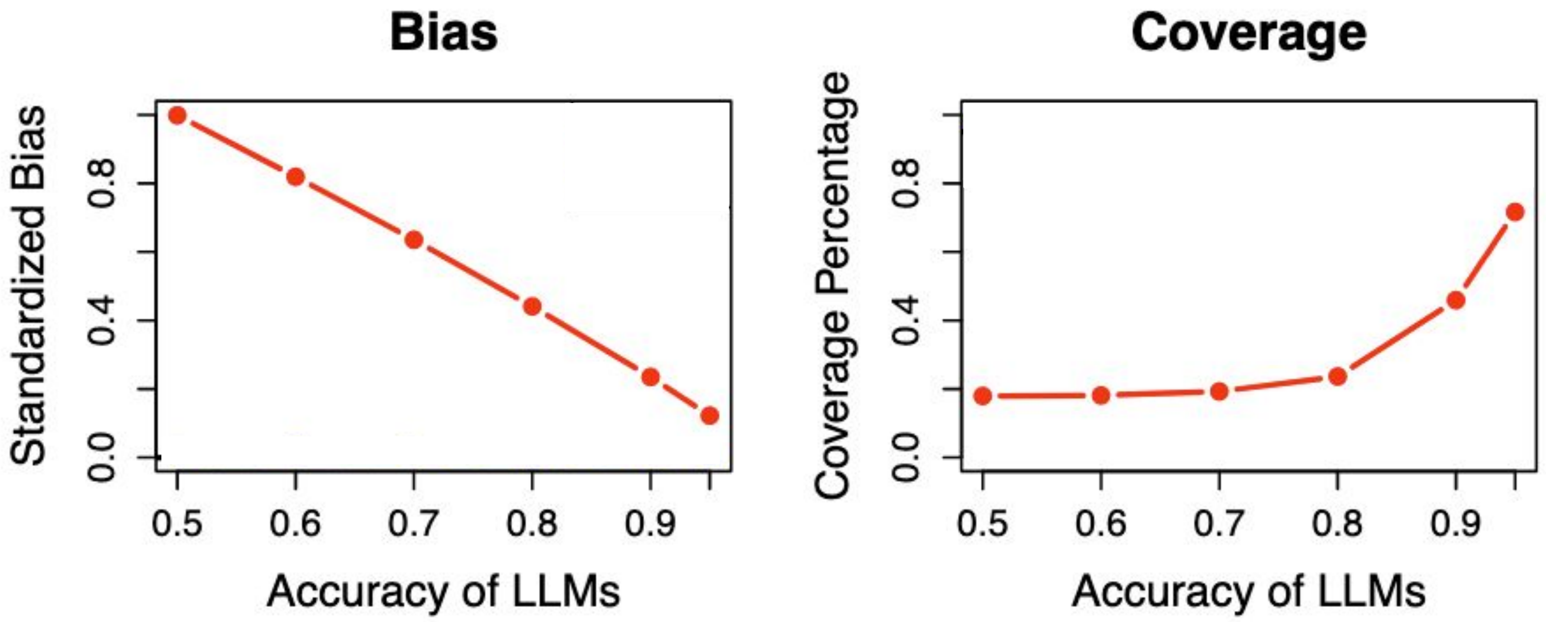


Illustration of random error in regression predictor

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

Egami et al. (2024)

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_inferences_with_llms

Literature 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.

Literature overview of recent methods

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

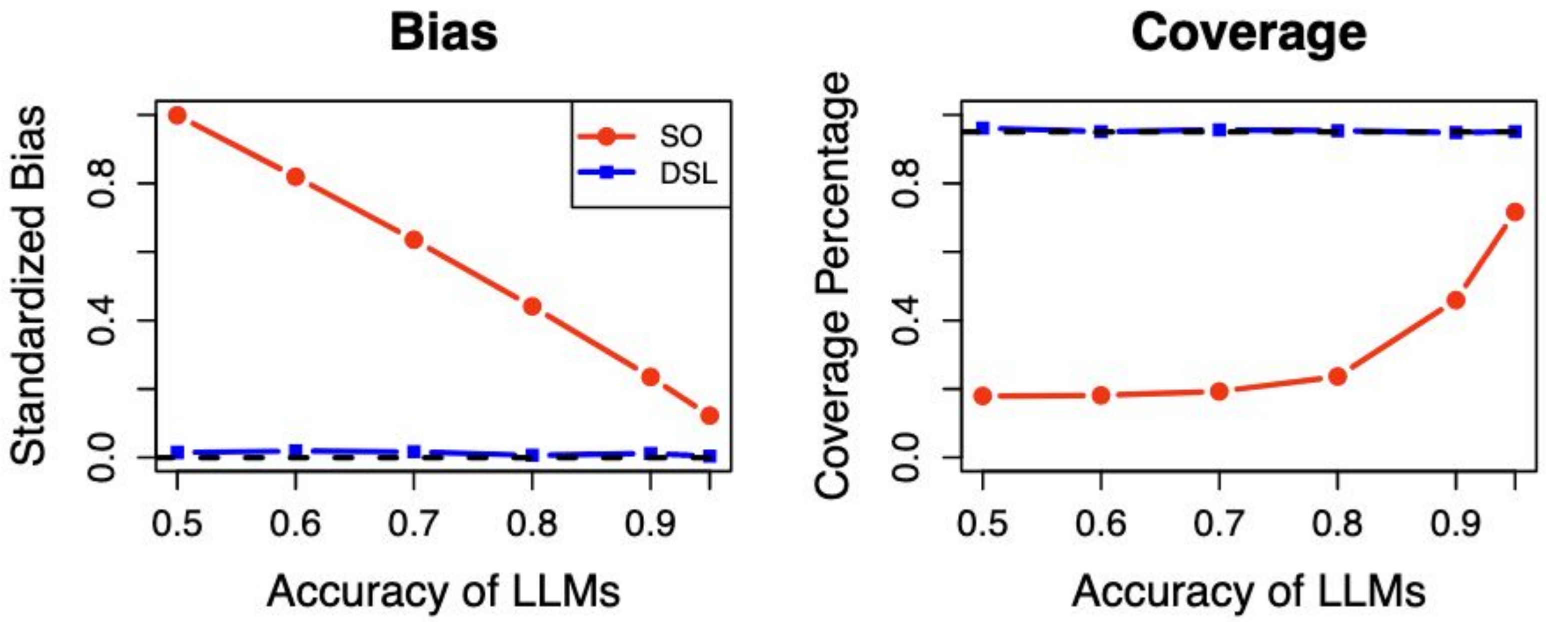
- Focus on correcting the LLM-predicted data ($\hat{Z}$)
  - PostPI: Predict Z from $\hat{Z}$
  - PPI: Predict $\hat{Z}$ - Z from W
  - DSL: Predict Z from $\hat{Z}$ and W, with a sampling weight-based correction
- Focus on correcting the loss function
  - PPI: Add a correction term (mimicking $\hat{Z}$ - 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

Literature overview of recent methods

Overwhelming, lacking comparisons, technical?

Luckily, they do work (depending on the scenario)!

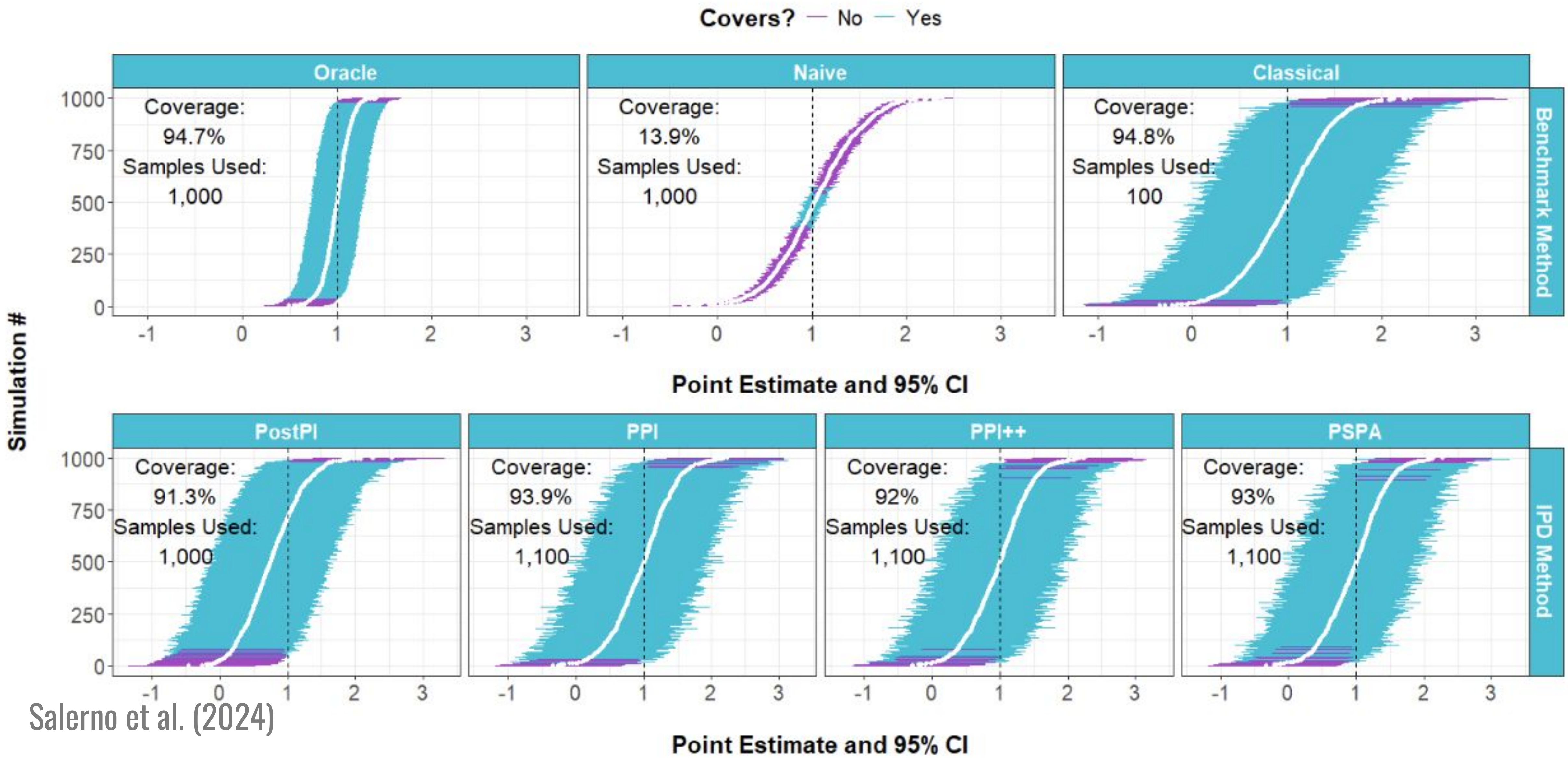
Literature overview of recent methods



(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

Egami et al. (2024)

Figure 1: Point estimate and corresponding 95% confidence intervals for four available IPD methods (`postpi`, `ppi`, `ppi_plusplus` and `pspa`; second row), as compared to three benchmark regressions (oracle, naive, and classical; first row) on 1,000 simulated linear regression datasets.



Salerno et al. (2024)

Practical recommendations

Depending on

- LLM-generated predictors (X_hat) or outcomes (Y_hat)?
- Python, R or manual?
- GLM or other types of estimators

Name	Method	Language	Estimators	Predicted Variables
PostPI	Post-Prediction Inference	R	Means, quantitles 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

Dealing with measurement error

Popular approaches in social sciences:

- Instrumental variable
- Attenuation correction
- Repeated measurements
- Structural equation modelling
- Bayesian

Exercise:

Modelling with LLM measurement error

Go to Part II's notebook. Try and feel free to use your own data!

Note that:

- Python: ppi_py package
- R: DSL package (use your local environment)

Conclusion

- LLMs can help to label data, and they can do this quite well.
- It's very important to create a good prompt.
- langchain / ellmer can help you do the data collection.
- Check measurement quality!
- Correct your inferences if suspecting measurement issues.

Bonus tip

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.

Bonus tip

Tutorial to be announced:

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

Questions? Feedback?

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.

Thanks!

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.

Salerno, S., Miao, J., Afiaz, A., Hoffman, K., Neufeld, A., Lu, Q., ... & Leek, J. T. (2024). ipd: An R Package for Conducting Inference on Predicted Data. *arXiv preprint* arXiv:2410.09665.

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.