Large Language Models (LLMs) in Social Science Research

Session I: Introduction

Joshua Cova and Luuk Schmitz

Max Planck Institute for the Study of Societies

20-06-2024

Introduction

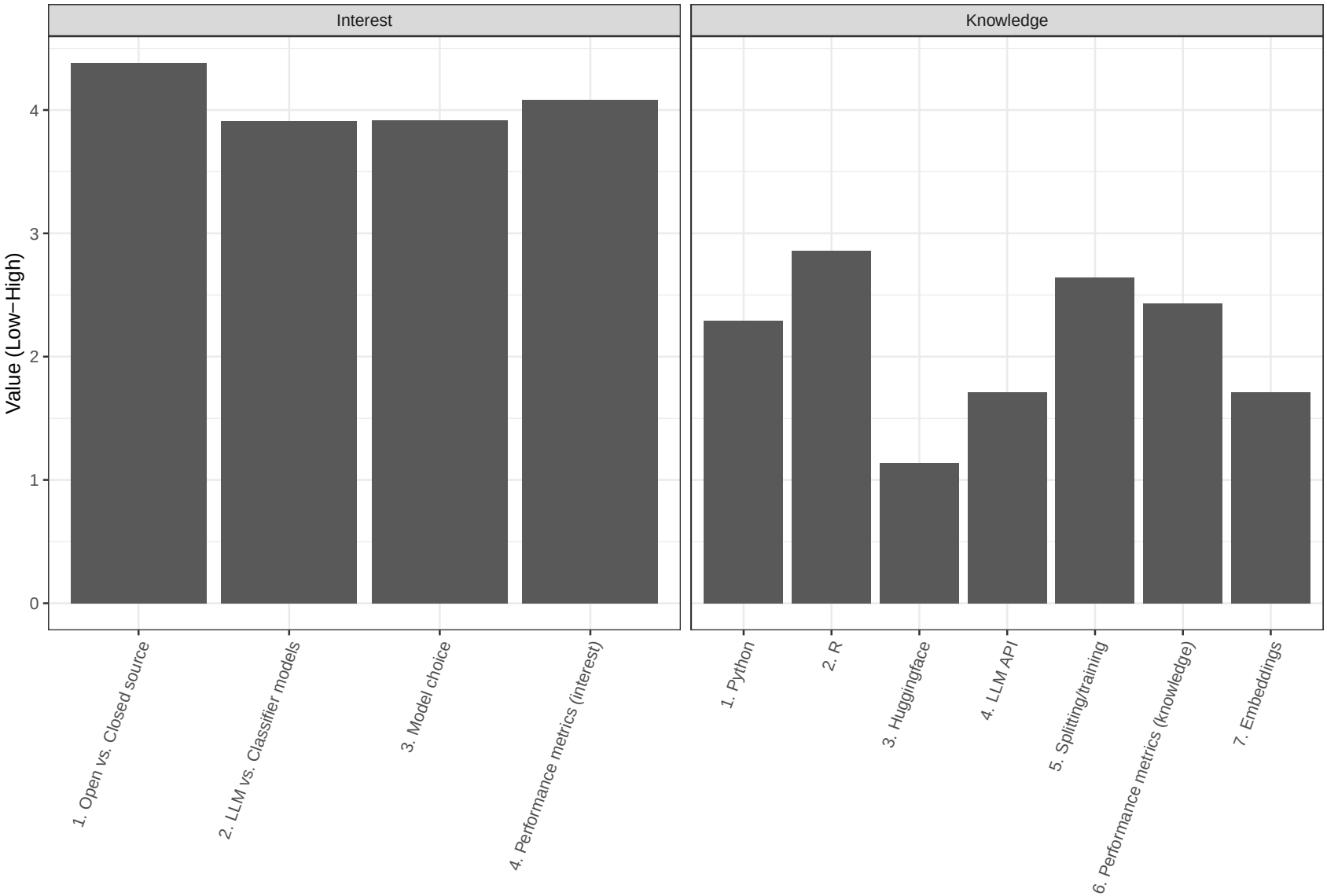
Introduction: Setting the stage

- Initial external workshop "Hands-On Text Coding with Large Language Models for Social Scientists"
- Perhaps a bit too advanced?
- Continue the conversation "in-house"
- Workshop series is oriented to **participants' interests and skill levels**
- Luuk and Joshua are interested users, not necessarily experts!

Introduction: Setting the stage

- Initial external workshop "Hands-On Text Coding with Large Language Models for Social Scientists"
- Perhaps a bit too advanced?
- Continue the conversation "in-house"
- Workshop series is oriented to **participants' interests and skill levels**
- Luuk and Joshua are interested users, not necessarily experts!
- **And what about you?**

Survey results



Prospective plan

Session 1

- The uses of LLMs in social science research (Luuk)
- Validation and performance metrics (Josh)
- Model selection (Luuk)

Session 2

- Presentation of working paper using LLMs (Luuk)
- Research ethics and closed vs. open-sourced models (Josh)

LLMs in social science research

The (R)evolution of LLMs

The current iteration of natural language-processing (NLP) machine learning models dates back to Vaswani et al. (2017). Their proposed model structure (transformer models) represented a giant leap in machine learning research. Compared to previous models, they are:

- Very efficient
- Awesome at detecting patterns
- Performance scales with size of training data

In the early days, development of *classifier* models and *generative* models ran in lockstep. As the commercial use cases of generative models became clear, development shifted towards the latter.

Classifier versus generative models

Without getting technical, this distinction is important. While both types of model share the same architecture and are great at predicting patterns, they are fundamentally different in their capabilities:

- Classifier: 'Understands' things and sorts them into categories.
- Generative: 'Understands' things and then creates new versions of them.

Moreover, generative models (e.g., OpenAI's GPT) are trained to be *instructed*. You can send a prompt to which they will respond.

Classifier models are much smaller than generative models in terms of training data and parameter count. But with high-quality training data, they are effective 'savants' that can perform a specific task (e.g., annotating a political corpus from the 1970s) just as well as a much larger generative model.

Example

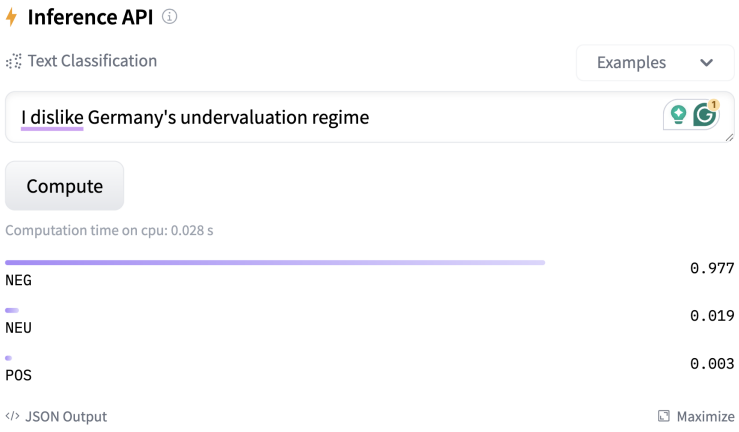


Figure 1: Classifier model

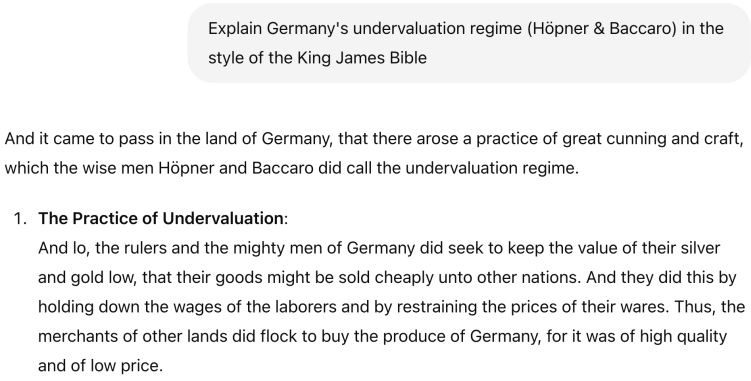


Figure 2: Generative model

The uses of large language models for the social sciences

LLMs have opened a wide spectrum of possibilities to scale and accelerate social scientific research. Based on a survey of the literature, we identify four use cases:

1. **Modeling human behavior computationally**
   - 1.1 *Testing & (potentially) running experiments*: Aher, Arriaga, and Kalai (2022); Dillion et al. (2023); Horton (2023)
   - 1.2 *Running surveys*: Tjuatja et al. (2023)
2. **Simulating social relationships**
   - 2.1 *Interactions between artificial agents*: Park et al. (2023); Wang et al. (2024)
   - 2.2 *Game theoretical simulations*: Akata et al. (2023)
3. **Interacting with human agents**
   - 3.1 *Chatbot for interviewing participants*: Chopra and Haaland (2023)
4. **Text annotation**
   - 4.1 *Zero-shot & few-shot text annotation*: Törnberg (2023); Gilardi, Alizadeh, and Kubli (2023); Leek, Bischl, and Freier (2024)
   - 4.2 *Synthetic data generation*: Laurer (2024)

Modeling human behavior

Generative AI: *Silicon parrot* or *Homo Sillicius*?

When presented with well-known behavioral experiments (e.g., *Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds*), do machine-based replies align with human judgement?

While some experiments are replicated, AI models also exhibit 'hyper-accuracy' distortions, based on our attempts to improve AI truthfulness and implement ethical guardrails. This offers some possible future inroads:

- Understanding the biases of AI systems (Aher, Arriaga, and Kalai 2022);
- AI models may act as 'effective proxies' to understand how sub-populations respond to treatments (Argyle et al. 2023);
- Potentially useful to help design future treatments in (survey) experiments (Horton 2023)

Simulating social relationships

LLMs may also be useful to simulate communities and societies. In much the same way, this reveals potential biases in AI systems and possibly sheds new light on theories of social organization.

> *In a large set of two players-two strategies games, we find that LLMs are particularly good at games where valuing their own self-interest pays off, like the iterated Prisoner's Dilemma family. However, they behave sub-optimally in games that require coordination (Akata et al. 2023).*

\

> *A society full of generative agents is marked by emergent social dynamics where new relationships are formed, information diffuses, and coordination arises across agents (Park et al. 2023)*

Interacting with human agents

What if we can use an AI-assisted approach to conduct semi-structured interviews? This is what Chopra and Haaland (2023) have tried to do.

Instead of a human interviewer, participants interact with an AI-assistant that is tasked to ask a predefined set of semi-structered questions.

Interestingly, a small majority of their respondents (n=395) indicates a preference for an AI-interviewer over a human interviewer.

The approach may be more scalable to yield high-quality interview data beyond what is typically possible.

Text annotation

Attempts to automate quantitative content analysis predate LLMs. Nevertheless, LLMs offer promising use cases to scale existing approaches and to increase conceptual complexity.

LLMs are particularly good at the following four things:

1. Labeling texts;
2. Identifying causality & modality;
3. Extracting metadata;
4. Generating synthetic data.

A good example: Leek, Bischl, and Freier (2024)

Side-note: LLMs can help streamline the research pipeline more generally

This workshop is about deploying LLMs for *research* purposes. Important to note is that their contribution doesn't stop here, but also applies to the entire research pipeline:

- Spitballing ideas
- Data prepration
- Summarizing texts (e.g., interview data)
- Creative writing
- Conducting preliminary analyses

See also: Korinek (2023)

Summary

The uses of LLMs are not confined to text annotation, and in many ways, it is the part where competitive alternative approaches still exist.

In many cases, they **do not** replace the human factor, but rather offer efficiency and scalability enhancements across the research pipeline.

The use of LLMs as research tools remains an 'academic Wild West' (Törnberg 2024, 2). There are, however, tools available to not enter the Wild West unarmed.

Validation

Validation

The use of LLMs is a very iterative process

- Thinking about prompt
- Splitting: Training/testing dataset
- Compare LLM result vs. human annotation
- Compare results across models
- Validation metrics
- Underlying tension between validity vs. reliability

We have been here before

Advance Access publication January 22, 2013

Political Analysis (2013) 21:267–297
doi:10.1093/pan/mps028

Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts

Justin Grimmer
Department of Political Science, Stanford University, Encina Hall West 616 Serra Street, Stanford, CA 94305
e-mail: jgrimmer@stanford.edu (corresponding author)

Brandon M. Stewart
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: bstewart@fas.harvard.edu

Edited by R. Michael Alvarez

We have been here before

- **Principle 1:** All Quantitative Models of Language Are Wrong—But Some Are Useful
- **Principle 2:** Quantitative Methods Augment Humans, Not Replace Them
- **Principle 3:** There Is No Globally Best Method for Automated Text Analysis
- **Principle 4:** Validate, Validate, Validate

Designing the right prompt is not simple and requires time

- LLM annotation in danger of becoming an academic 'Wild West' (Törnberg, 2024)

Designing the right prompt is not simple and requires time

- LLM annotation in danger of becoming an academic 'Wild West' (Törnberg, 2024)
- Draft a coding task
- Code a sample and compare your results ('gold standard') to LLM results
- Fine-tune your prompt by asking the LLM to motivate its decision
- Revise coding task
- Multilingual text might make it more difficult

Designing the right prompt is not simple and requires time

- LLM annotation in danger of becoming an academic 'Wild West' (Törnberg, 2024)
- Draft a coding task
- Code a sample and compare your results ('gold standard') to LLM results
- Fine-tune your prompt by asking the LLM to motivate its decision
- Revise coding task
- Multilingual text might make it more difficult
- *'If a man knows not to which port he sails, no wind is favorable'* (Seneca)

Example


Inflation
Days_lost
unemployment_rate
25
20
15
10
5
30000
20000
10000
0
5
4
3
2
1
1970
1974
1978
Germany
UK

Dataset

Small sample of 65 parliamentary interventions discussing inflation & trade unions (UK, 1979-1981), good distribution across categories (Conservative = 31, Labour = 29, Others = 5) \

Think about the unit of analysis (paragraph vs. sentence(s)) - iterative process

Prompt example

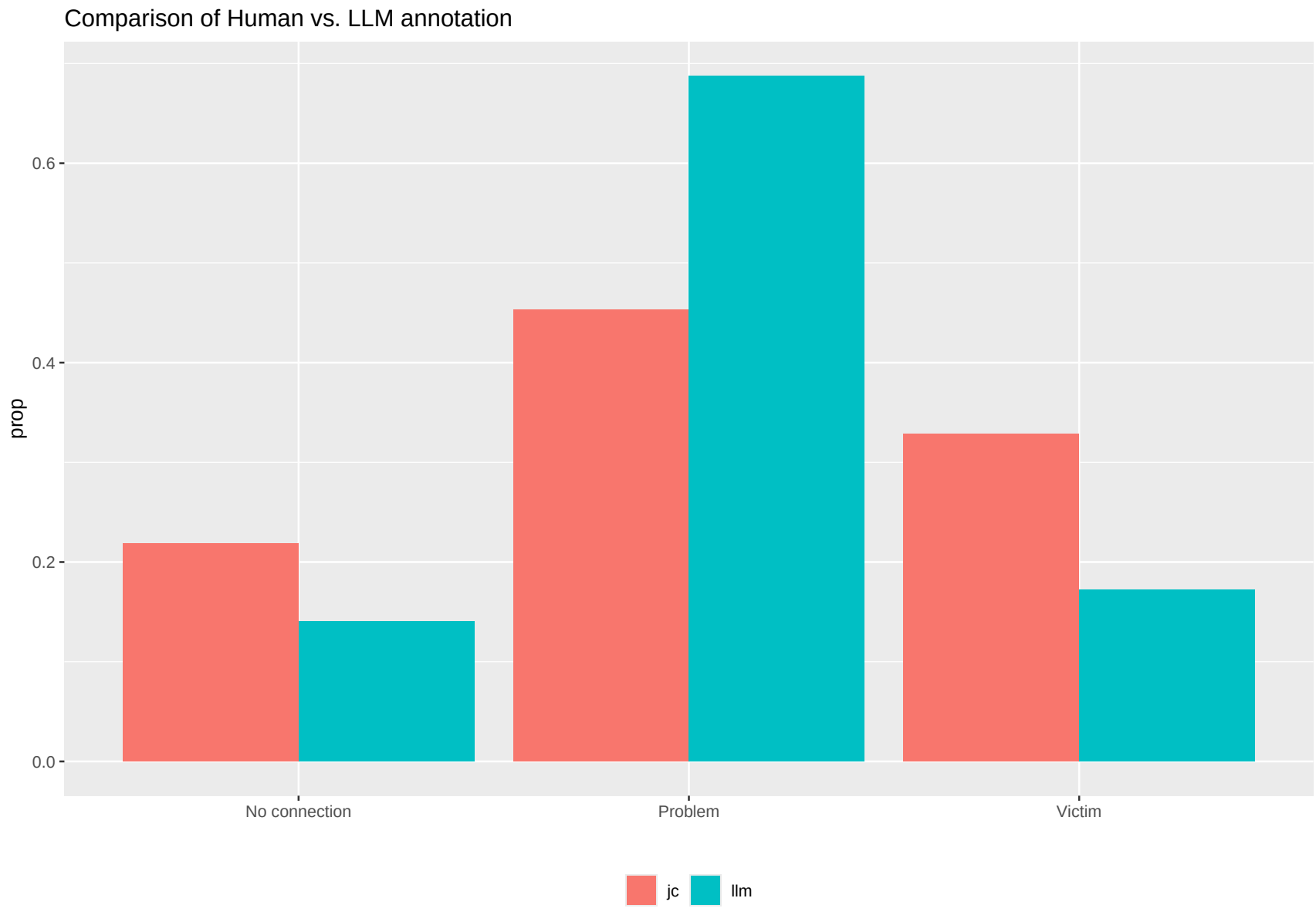
```
instructions = """You are a political science researcher.
            "You are tasked with classifying excerpts from a series of historical British parliamentary transcripts,
            dealing with the way in which trade unions are connected to wage push inflation.

            "Wage push inflation is defined as an overall rise in the cost of goods and services that results from a rise in wages.
            "Your task is to classify the statements by categorizing whether trade unions are seen as contributors to inflation (Problem)
              or as victims of inflation (Victim).

            "Example of a Problem statement: trade unions got everything they wanted without delivering anything of lasting benefit in return."
            "Example of a Victim statement: inflation has nothing to do with trade union monopoly. It has much to do with poor wage
            levels, particularly in engineering.

            "The output should be one and only one of these categories: 'Problem' 'Victim', 'No connection between inflation and trade unions'.
            Please motivate your answer with 1-2 sentences."""
```

New


Comparison of Human vs. LLM annotation
prop
0.6
0.4
0.2
0.0
No connection
Problem
Victim
jc
llm

Result: Where we agree

Statement 6: [. . .] The trade union movement has rightly indicated its concern for the welfare of its members and the protection of its members' jobs. That is its right. It is what one would expect. The movement has pointed out that the grave outlook for inflation will lead to higher wage claims simply to maintain, let alone improve the living standards of working people. Either the Government have not taken this prospect into account or they have not been given sufficient consideration to the implications of what is likely to happen.

Result: Where we agree

Statement 6: [. . .] The trade union movement has rightly indicated its concern for the welfare of its members and the protection of its members' jobs. That is its right. It is what one would expect. The movement has pointed out that the grave outlook for inflation will lead to higher wage claims simply to maintain, let alone improve the living standards of working people. Either the Government have not taken this prospect into account or they have not been given sufficient consideration to the implications of what is likely to happen.

Response 6: Victim

Motivation: The statement highlights the trade union movement's concern for the welfare and job protection of its members in the face of inflation. It portrays trade unions as victims of inflation, as they are forced to make higher wage claims to maintain living standards due to the economic circumstances.

Result: Where we agree

Statement 16: However, it will not do for the hon. Gentleman to say that that was an unfortunate episode in the course of a successful campaign against inflation. The success of the previous Labour Government's campaign against inflation can be gauged by the fact that their incomes policy collapsed in total disaster last winter. It did so not as a result of overwhelming Conservative opposition but because it was utterly rejected by the trade union movement and by the Labour Party.

Result: Where we agree

Statement 16: However, it will not do for the hon. Gentleman to say that that was an unfortunate episode in the course of a successful campaign against inflation. The success of the previous Labour Government's campaign against inflation can be gauged by the fact that their incomes policy collapsed in total disaster last winter. It did so not as a result of overwhelming Conservative opposition but because it was utterly rejected by the trade union movement and by the Labour Party.

Response 16: Problem

The statement suggests that the trade union movement played a significant role in the collapse of the incomes policy, which was aimed at combating inflation. This implies that trade unions are seen as contributors to inflation by rejecting measures to control it.

Result: Where we disagree

Statement 51: Government have just cut the percentage of rate support grant from 59% to 56% and, therefore, they immediately imposed an across-the-board increase of 3 per cent. or 4 per cent. on all ratepayers. The Government are imposing not only rent but rate increases. By that means, they are stoking up inflation and, at the same time, expecting trade unions and working people to hold their wage demands to 4% thus ensuring a swingeing cut in their standard of living.

Result: Where we disagree

Statement 51: Government have just cut the percentage of rate support grant from 59% to 56% and, therefore, they immediately imposed an across-the-board increase of 3 per cent. or 4 per cent. on all ratepayers. The Government are imposing not only rent but rate increases. By that means, they are stoking up inflation and, at the same time, expecting trade unions and working people to hold their wage demands to 4% thus ensuring a swingeing cut in their standard of living.

Response 51: Problem

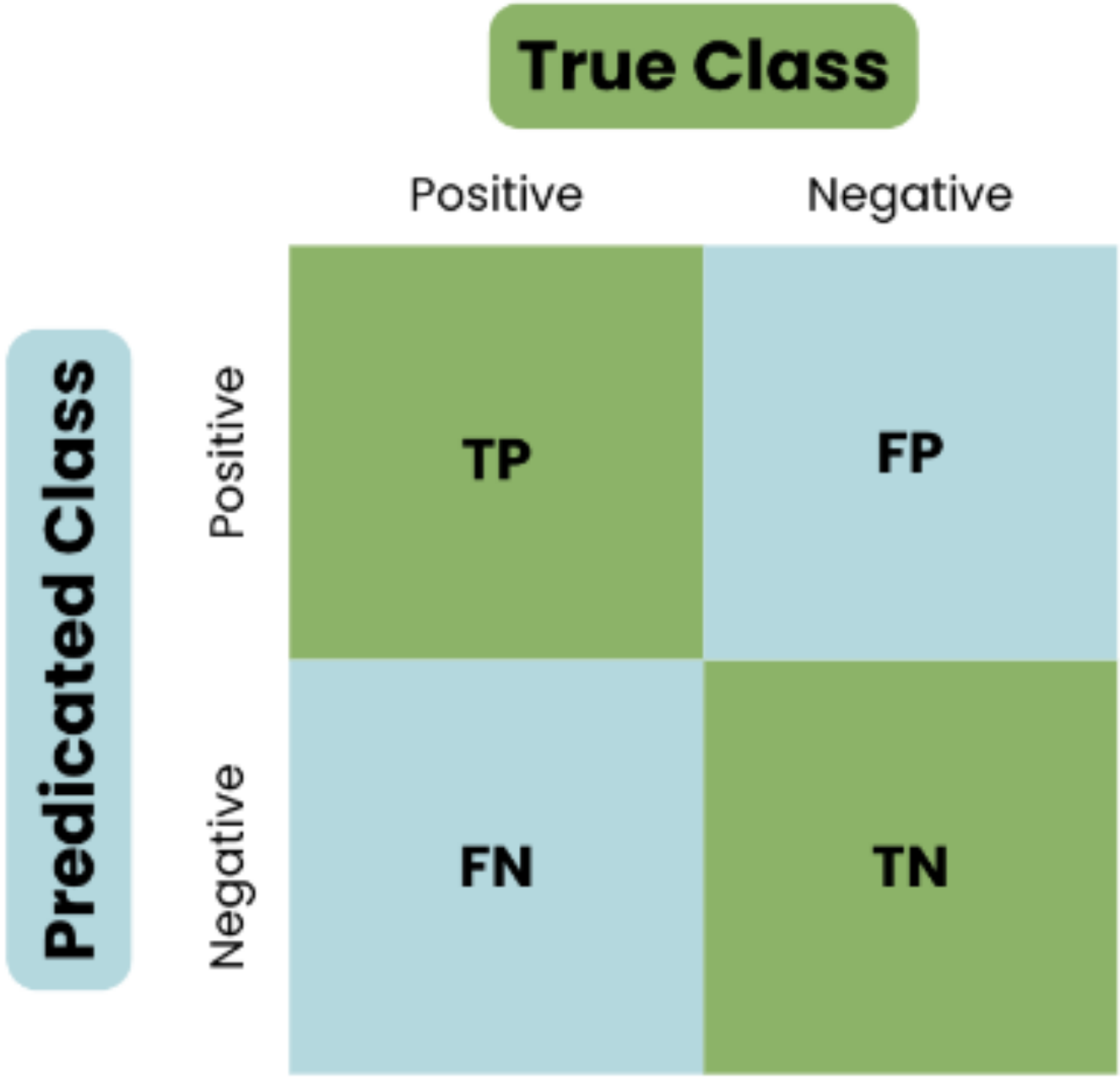
The statement suggests that the government's actions, including imposing rate increases and expecting trade unions to limit wage demands, are seen as contributing to inflation by putting pressure on wages."

Performance metrics

Confusion matrix

- We need to find a way to systematically study how well the LLM classification has performed
- Confusion matrix devised as a way to gauge classification performance of a model/algorithm
- Important to compare the results across different models
  - Positive: the label/category of interest
  - Negative: not the label/category of interest

Confusion matrix



Accuracy

$$Accuracy = \frac{Total\ \ correct\ \ predictions}{Total\ \ predictions}$$

Accuracy

$$Accuracy = \frac{Total\ \ correct\ \ predictions}{Total\ \ predictions}$$

- Bias: Imagine an annotation task where 99% of observations are negative and 1% positive.
- Annotator predicts all observations to belong to negative category (100%)
- Achieves 99% accuracy
- Problem of unequal representation in the dataset (class imbalance)

Accuracy

Table 1: Classification performance for sample of UK parliamentary texts

Dataset	Accuracy	Precision	Recall
Problem	0.73		
Victim	0.81		
No connection	0.92		

Precision

$$Precision = \frac{Correct\ \ positive\ \ predictions}{Total\ \ positive\ \ predictions} = \frac{TP}{TP + FP}$$

Precision

$$Precision = \frac{Correct\ positive\ predictions}{Total\ positive\ predictions} = \frac{TP}{TP + FP}$$

- Minimize mistakes in guessing positive labels
- Imagine an annotation task where 99% of observations are negative and 1% positive.
- We annotate everything as negative
- Achieves 0% precision
- However leaves out negative labels

Precision

Table 2: Classification performance for sample of UK parliamentary texts

Dataset	Accuracy	Precision	Recall
Problem	0.73	0.96	
Victim	0.81	0.47	
No connection	0.92	0.64	

Recall

$$Recall = \frac{Correct\ \ Positive\ guesses}{All\ \ positive\ \ labels} = \frac{TP}{TP + FN}$$

- How many positive predictions out of all the true positives that exist (also the misclassified ones)
- But if you label everything as positive, you will get no false positive

Recall

$$Recall = \frac{Correct\ \ Positive\ guesses}{All\ \ positive\ \ labels} = \frac{TP}{TP + FN}$$

- How many positive predictions out of all the true positives that exist (also the misclassified ones)
- But if you label everything as positive, you will get no false positive

Recall

Table 3: Classification performance for sample of UK parliamentary texts

Dataset	Accuracy	Precision	Recall
Problem	0.73	0.63	0.96
Victim	0.81	0.91	0.47
No connection	0.92	1.00	0.64

Inter-coder agreement metrics

- It is important that you are not the only person coding the text. It is important to include also other coders to increase validity and assess the consistency of annotations (*inter-coder reliability*)
- Also involve others in thinking about the 'right' prompt
- Difference between expert vs. crowd-sourced coders (Benoit et al. 2016)

Other important considerations

- Detailed Reporting: Document the validation process, including methodologies used, metrics calculated, and feedback gathered. This helps in understanding the validation results and communicating them to stakeholders.
- Transparent Communication: Clearly communicate the strengths and limitations of the LLM's annotations to users and stakeholders.

Model selection

Model selection

Given the tools in our hands to evaluate a model's performance, should we simply select the most performant model?

Let's consider Törnberg's six principles

1. Reproducibility
2. Ethics & legality
3. Transparency
4. Culture and Language
5. Scalability
6. Complexity

Another principle to consider: cost!

The teacher-student model

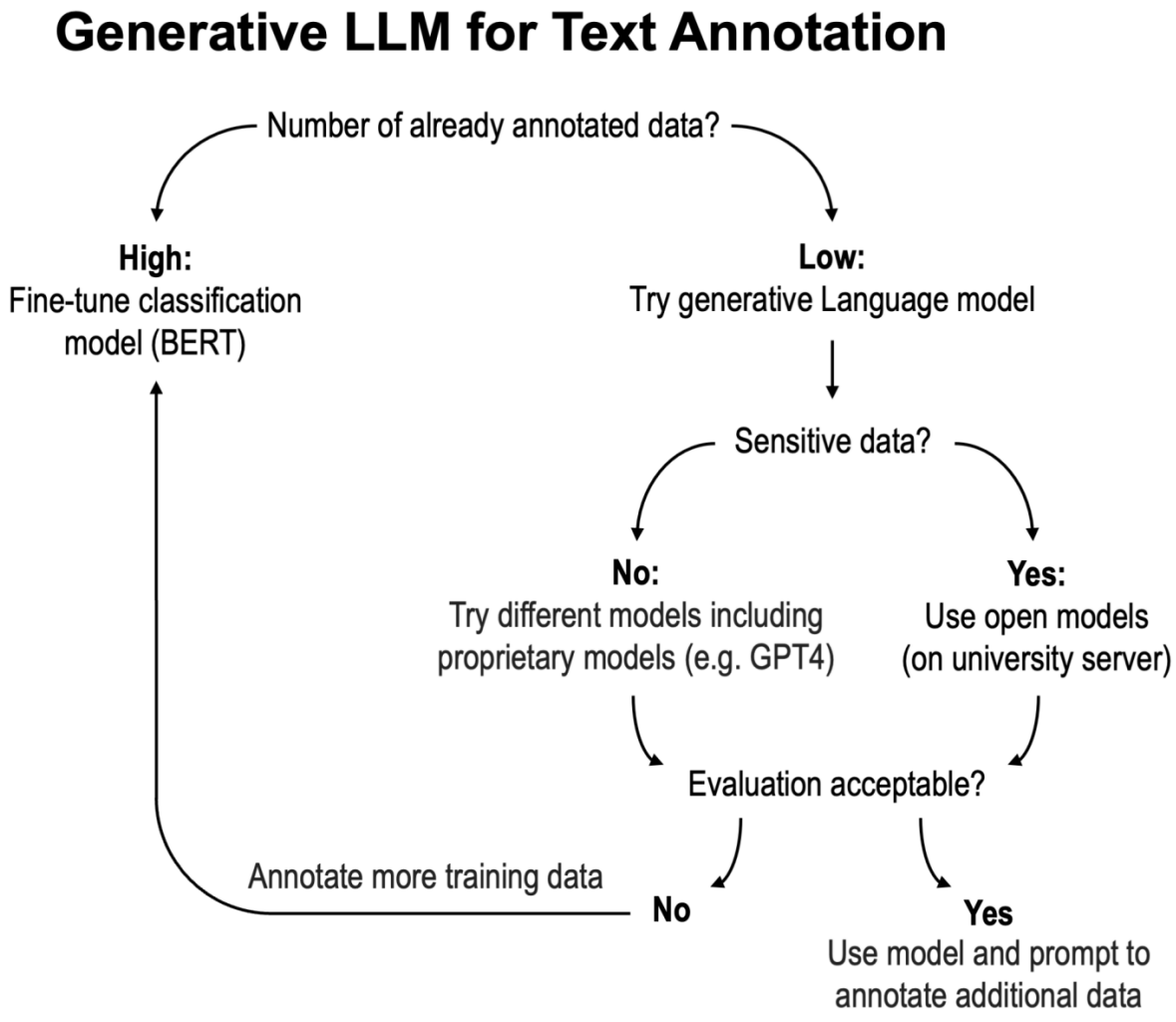


Figure 3: Decision tree for the use of generative LLM for text annotation

Drawing from my own experience I

I have tried to follow the Weber and Reichardt (2024) decision tree, but did not manage to get a classifier model to perform adequately.

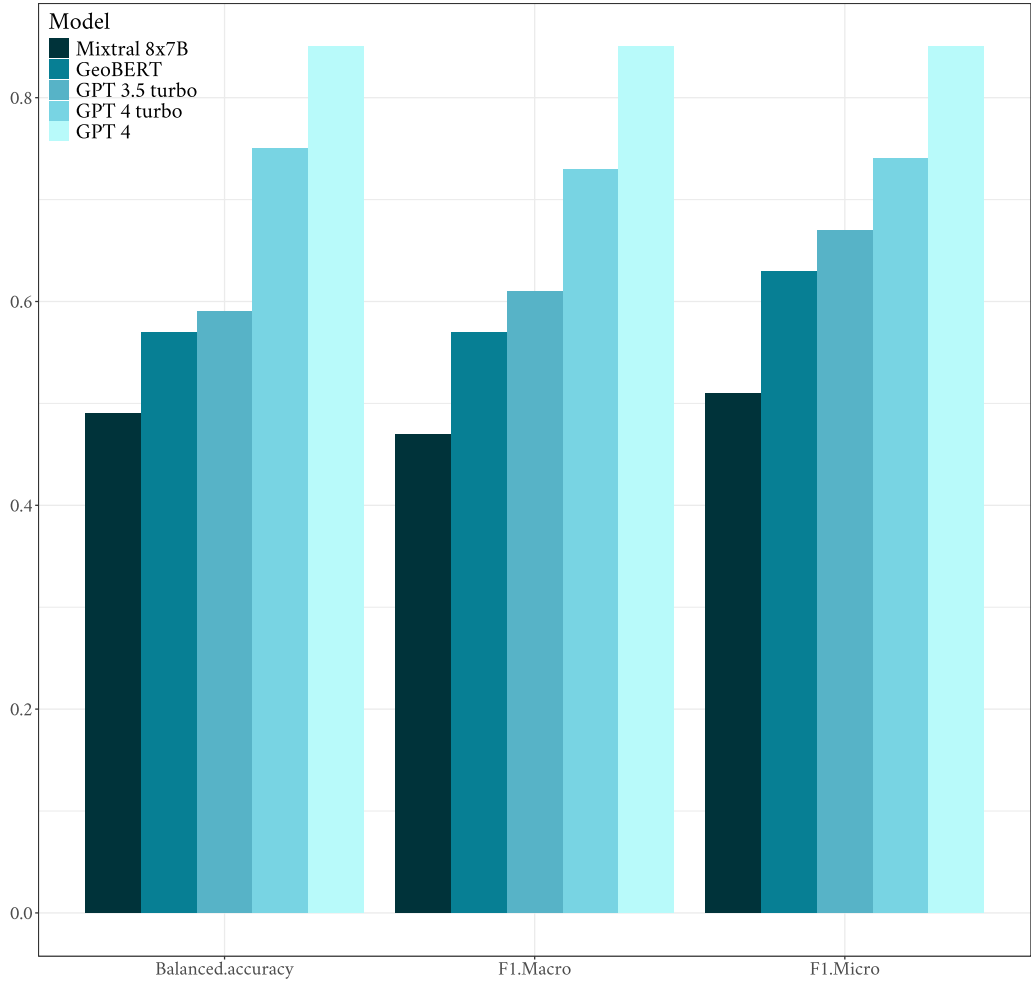


Figure 4: Model performance comparison

Drawing from my own experience II

The syntax for prompting LLMs is very similar, making cross-evaluation of LLMs feasible. Some nitpicking of specific models:

- Google Gemini: the syntax is a bit cumbersome, the web interface and myriad of different platforms to interact with are a bag of hurt
- Claude Opus: very similar to OpenAI and easy to set up. Not extremely cheap, but might be a cost-effective alternative for some tasks
- Mixtral & LLama: open-source LLMs, slightly more complex to set up, and problem of computing power

Drawing from my own experience III

At the moment, there seems to be a trade-off between performance and open science considerations.

- For simpler and less conceptually sophisticated tasks, classifier models remain the way to go.
- If classifier models do not perform well enough, the decision to continue improving them or opt for an LLM involves a careful trade-off between cost and time

Conclusion

LLMs offer large promises for social scientific research, but come with important pitfalls. Three principles to consider:

- Avoid the rule of cool: we should use tools when they offer substantive value, and not only if they are new and cool
- Exhaust alternatives before settling on an LLM as a primary classification tool
- Whether to use an LLM or a classifier model is often not an either-or question

References

Aher, Gati, Rosa I. Arriaga, and Adam Tauman Kalai. 2022. "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies." arXiv. https://doi.org/10.48550/ARXIV.2208.10264.

Akata, Elif, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. "Playing Repeated Games with Large Language Models." arXiv. https://doi.org/10.48550/ARXIV.2305.16867.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–51. https://doi.org/10.1017/pan.2023.2.

Chopra, Felix, and Ingar Haaland. 2023. "Conducting Qualitative Interviews with AI." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4583756.