

# Language Models for Law and Social Science

## 11. LLMs for Research

# Outline

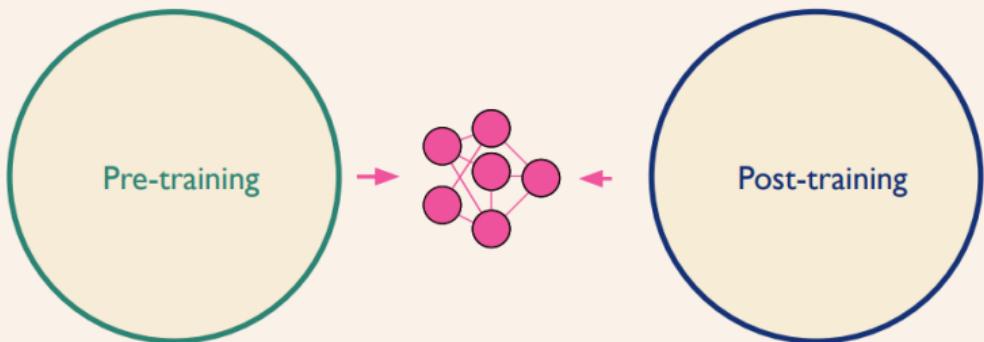
Recap: Recipe for AI

Application: Analyzing Proverbs in Industrial-Age England

NLP for Legal Research

AI and the Research Process

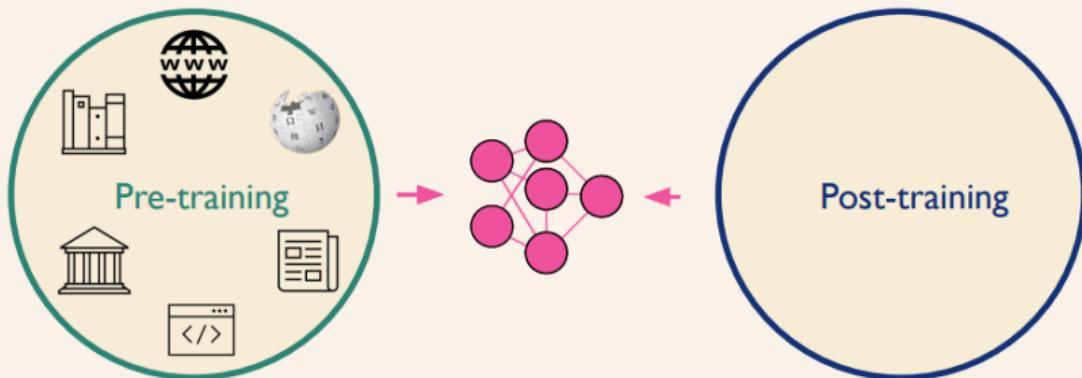
# Building a modern LLM



❖Ai2

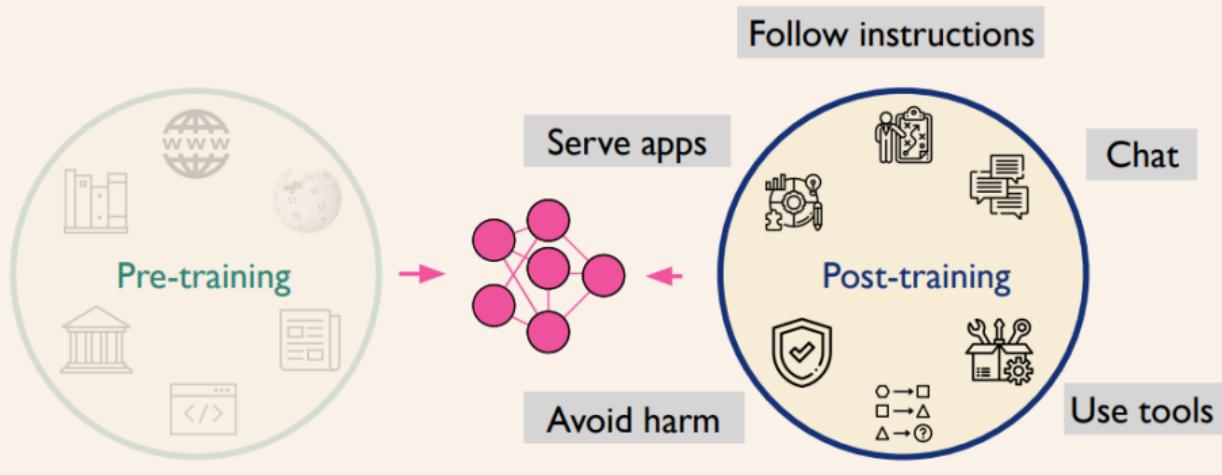
<https://rdi.berkeley.edu/adv-llm-agents/slides/OLMo-Tulu-Reasoning-Hanna.pdf>

# Building a modern LLM



Predict the next word in various contexts

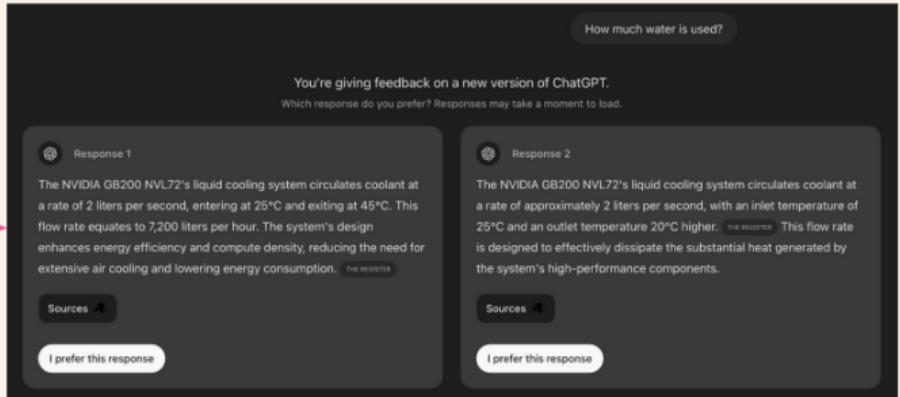
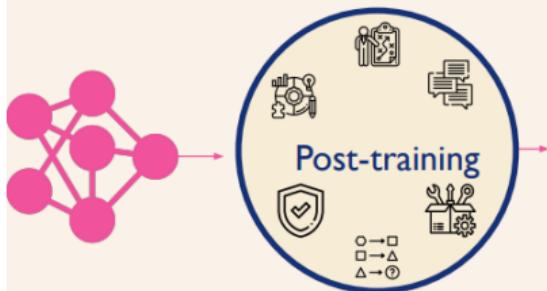
# Building a modern LLM



Predict the next word in various contexts

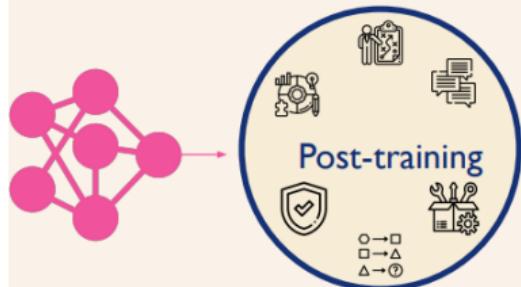
❖Ai2

# Post Training for Alignment with Human Preferences



[Photo by @FanaHOVA on X]

# Post Training for Tool Use /Agents



What are the voting numbers for Trump vs Harris?

As of November 7, 2024, with most votes counted, former President Donald Trump leads Vice President Kamala Harris in the popular vote by approximately 4.6 million votes. Trump has secured 72,641,564 votes, while Harris has received 67,957,895 votes. USA TODAY

In the Electoral College, Trump has obtained 312 electoral votes, surpassing the 270 required to win the presidency, while Harris has 226 electoral votes. REUTERS

These figures indicate that Trump has won both the popular vote and the Electoral College in the 2024 presidential election.

Sources 2

Visualize the numbers in pie chart.

2024 Presidential Election Popular Vote: Trump Vs H...

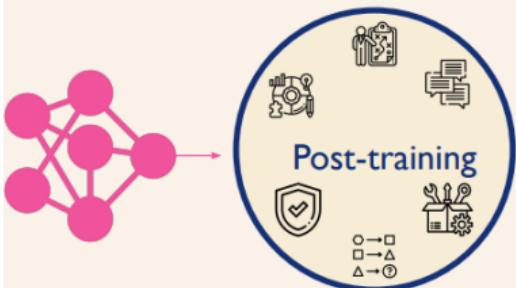
2024 Presidential Election Popular Vote: Trump vs Harris

A pie chart titled "2024 Presidential Election Popular Vote: Trump Vs H..." showing the results for "2024 Presidential Election Popular Vote: Trump vs Harris". The chart is divided into two equal halves. The left half is orange and labeled "Donald Trump 55.7%". The right half is orange and labeled "Kamala Harris 44.2%".

Searching

Code  
Execution

# Post Training for Reasoning



Find the smallest positive integer that can be expressed as the sum of two squares in two different ways.

To solve this problem, we need to find the smallest positive integer ( $n$ ) that can be expressed as the sum of two squares in two different ways. This means we need  $(n)$  such that there exist distinct pairs  $((a, b))$  and  $((c, d))$  where:

$$[ n = a^2 + b^2 = c^2 + d^2 ]$$

We will use Python to iterate through possible values of  $(n)$  and check if  $(n)$  can be expressed as the sum of two squares in two distinct ways.

Let's write a Python function to solve this problem step-by-step.

```
def find_smallest_sum_of_two_squares():
    def is_sum_of_two_squares(n):
        ways = set()
        for a in range(int(n ** 0.5) + 1):
            b_squared = n - a * a
            if b_squared >= 0:
                b = int(b_squared ** 0.5)
                if b * b == b_squared:
                    ways.add((a, b))
                    ways.add((b, a))
        return ways

    n = 1
    while True:
        ways = is_sum_of_two_squares(n)
        if len(ways) >= 2:
            return n, ways
        n += 1
```

```
smallest_n, ways = find_smallest_sum_of_two_squares()
print(smallest_n, ways)
```

```
(50, {(5, 5), (1, 7), (7, 1)})
```

The smallest positive integer that can be expressed as the sum of two squares in two different ways is (50).

The ways to express 50 as the sum of two squares are: [  $50 = 1^2 + 7^2$  ] [  $50 = 5^2 + 5^2$  ]

Thus, the answer is (boxed{50}).

## Some interesting questions for economists

- ▶ How should human preference data be collected?
- ▶ How should the preference data be processed or weighted?
- ▶ What can economists say about the specification of the utility function?
  - ▶ e.g., should it be transitive?

## Some interesting questions for economists

- ▶ How should human preference data be collected?
- ▶ How should the preference data be processed or weighted?
- ▶ What can economists say about the specification of the utility function?
  - ▶ e.g., should it be transitive?
- ▶ The utility function is only based on user feedback. But that might not be optimal from a welfare perspective – e.g., to teach students, the AI instructor shouldn't just give the answer.
  - ▶ can we use other reward signals?

## Some interesting questions for economists

- ▶ How should human preference data be collected?
- ▶ How should the preference data be processed or weighted?
- ▶ What can economists say about the specification of the utility function?
  - ▶ e.g., should it be transitive?
- ▶ The utility function is only based on user feedback. But that might not be optimal from a welfare perspective – e.g., to teach students, the AI instructor shouldn't just give the answer.
  - ▶ can we use other reward signals?
- ▶ What are the implications for the model's worldview from RLHF vs RLVR?
  - ▶ RLHF is a probabilistic approach, encourages interpolation.
  - ▶ RLVR is a discrete approach; discourages interpolation.

# AI assistants are going to be amazing

```
write a script to link judges_with_ai_summaries.csv and topics_summary_full_96.xlsx
✓ Planning step accepted

Let me check the existence and contents of these files to better understand their structure and format.

✓ ls -l judges_with_ai_summaries.csv topics_summary_full_96.xlsx >

Let me examine the structure of the CSV file to understand its columns and data format.

✓ head -n 5 judges_with_ai_summaries.csv >

Now let me examine the Excel file structure. I'll convert the first few rows to CSV format for better viewing in the terminal.

✓ python3 -c "import pandas as pd; df = pd.read_excel('topics_summary_full_96.xlsx'); print(df.head().to_csv())" >

Based on my examination of both files, I can now provide a clear task description for creating a script to link these files.

① A step-by-step plan to map judge summaries to topic model outputs and generate a combined dataset with topic associations. ▾

1 1. Import the necessary libraries (pandas, numpy, and a suitable NLP library such as scikit-learn or spaCy).
2 2. Read the judges_with_ai_summaries.csv file into a pandas DataFrame, ensuring memory efficiency by specifying appropriate
3 3. Read the topics_summary_full_96.xlsx file into another pandas DataFrame.
4 4. Preprocess text from both datasets (e.g., removing stop words, lowering case, tokenizing).
5 5. Represent the topic summaries and judge summaries using a vectorization method (e.g., TF-IDF or word embeddings).
6 6. Compute similarity scores between each judge summary and the topic representations (e.g., cosine similarity).
7 7. Identify the top topics for each judge based on similarity scores, and record them in a new column or data structure.
8 8. Merge the results back with the original judges DataFrame, preserving the judge IDs.
9 9. Create output columns for topic associations and their respective scores or ranks.
10 10. Write the final combined dataset to a CSV or Excel file for downstream analysis.
11
```

# Outline

Recap: Recipe for AI

Application: Analyzing Proverbs in Industrial-Age England

NLP for Legal Research

AI and the Research Process



## Cultural Change in the Industrial Revolution

Elliott Ash & Melanie Xue

# The Industrial Revolution

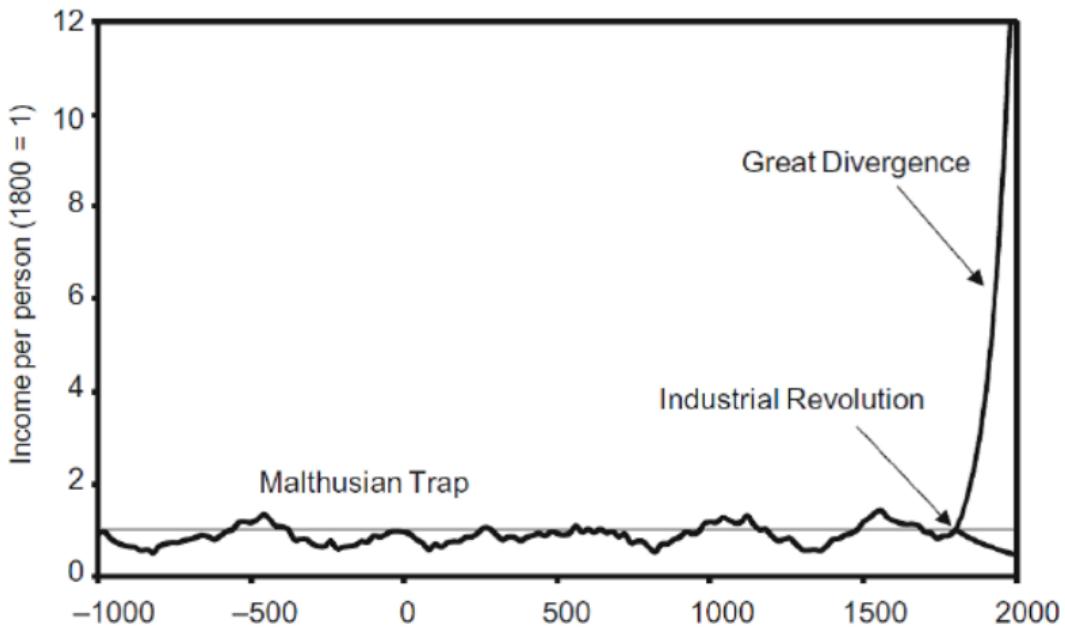


Figure: Income Per Person in the UK (Clark 2014, p. 219)

"The most fundamental transformation of human life in the history of the world recorded in written documents." (Hobsbawm, 1969).

# Proverbs and Culture

***“Hold on, wait for the grasshoppers.”***

***“Many get into a dispute well that cannot get out well.”***

- ▶ Proverbs tend to be metaphorical and cryptic, with unusual syntactic structure.
  - ▶ that makes them easier to memorize, and easier to transmit across generations while being resistant to information loss or manipulation (Villers, 2022).
  - ▶ also makes them difficult to understand just from reading them – you have to have some external knowledge.

# New Collection of Historical British Proverbs

- ▶ Lists of proverbs:
  - ▶ 5,975 proverbs from John Ray's *Compleat collection of English proverbs* (1617).
  - ▶ 20,773 proverbs from Robert Christy's *Proverbs, maxims and phrases of all ages* (1889).
  - ▶ 347 proverbs from Walter William Skeat's *Early English Proverbs, Chiefly of the Thirteenth and Fourteenth Centuries* (1910)
  - ▶ 10,316 proverbs from George Latimer Apperson's *English Proverbs and Proverbial Phrases: a Historical Dictionary* (1929).
  - ▶ 11,783 proverbs from Palmer Tilley's *A Dictionary of the Proverbs in England in the Sixteenth and Seventeenth Centuries* (1950).
- ▶ OCR'd the PDFs, regex extraction of proverbs, post-processing, checking, and cleaning.
- ▶ De-duplication / merging of close variants.
- ▶ Filtered out phrases that GPT-4o says are not proverbs.
- ▶ Final collection: 26,570 English proverbs.
- ▶ Main analysis: 12,831 proverbs that appeared at least once in our usage data (books and newspapers).

## Explaining Proverbs with AI

- ▶ Key challenge: we need not just the proverbs, but explanations that consistently lay out their social implications.
  - ▶ would require thousands of hours of work by human expert annotators.
  - ▶ standard NLP tools do not work with metaphor / implicit meaning in texts.

## Explaining Proverbs with AI

- ▶ Key challenge: we need not just the proverbs, but explanations that consistently lay out their social implications.
  - ▶ would require thousands of hours of work by human expert annotators.
  - ▶ standard NLP tools do not work with metaphor / implicit meaning in texts.
- ▶ What LLMs can do that makes this project possible:
  - ▶ draw on a giant multilingual text knowledge base
  - ▶ interpret metaphorical language
  - ▶ infer and interpolate meaning

What can we do now that we could not do before?

## What can we do now that we could not do before?

- ▶ 20 years ago:
  - ▶ annotate proverb explanations along some dimension, machine classify the rest.
  - ▶ learn LDA topics on the proverb explanations.

## What can we do now that we could not do before?

- ▶ 20 years ago:
  - ▶ annotate proverb explanations along some dimension, machine classify the rest.
  - ▶ learn LDA topics on the proverb explanations.
  - ▶ **But how do we get the explanations?**

## What can we do now that we could not do before?

- ▶ 20 years ago:
  - ▶ annotate proverb explanations along some dimension, machine classify the rest.
  - ▶ learn LDA topics on the proverb explanations.
  - ▶ **But how do we get the explanations?**
- ▶ 5 years ago:
  - ▶ use BERT to get more accurate machine classifications based on labels.
  - ▶ use BERTopic to get better, context-sensitive topics than LDA.
  - ▶ **But again: how do we get the explanations?**

## What can we do now that we could not do before?

- ▶ 20 years ago:
  - ▶ annotate proverb explanations along some dimension, machine classify the rest.
  - ▶ learn LDA topics on the proverb explanations.
  - ▶ **But how do we get the explanations?**
- ▶ 5 years ago:
  - ▶ use BERT to get more accurate machine classifications based on labels.
  - ▶ use BERTopic to get better, context-sensitive topics than LDA.
  - ▶ **But again: how do we get the explanations?**
- ▶ 2.5 years ago (RLHF models) and today (RLVR models):
  - ▶ **use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning.**
  - ▶ then apply the above tools (e.g. BERT) to classify/interpret the explanations.

## What about the future?

- ▶ Today (RLHF/RLVR LLMs):
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**
  - ▶ “give me a list of all English-language proverbs emphasizing the value of patience”

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**
  - ▶ “give me a list of all English-language proverbs emphasizing the value of patience”
  - ▶ “count the usage of these proverbs in a provided corpus (or one that is accessible online) by country and over time”.

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**
  - ▶ “give me a list of all English-language proverbs emphasizing the value of patience”
  - ▶ “count the usage of these proverbs in a provided corpus (or one that is accessible online) by country and over time”.
- ▶ **What LLMs cannot do, and won't be able to do:**

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**
  - ▶ “give me a list of all English-language proverbs emphasizing the value of patience”
  - ▶ “count the usage of these proverbs in a provided corpus (or one that is accessible online) by county and over time”.
- ▶ **What LLMs cannot do, and won't be able to do:**
  - ▶ “count the usage of these proverbs by county and over time, but I don't have the corpus”

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**
  - ▶ “give me a list of all English-language proverbs emphasizing the value of patience”
  - ▶ “count the usage of these proverbs in a provided corpus (or one that is accessible online) by county and over time”.
- ▶ **What LLMs cannot do, and won't be able to do:**
  - ▶ “count the usage of these proverbs by county and over time, but I don't have the corpus”
  - ▶ “Give me a measure of the counties in england over time by their cultural valuing of patience”

## What about the future?

- ▶ **Today (RLHF/RLVR LLMs):**
  - ▶ use gen AI (aligned LLM) to translate a given proverb into an explanation of its cultural meaning. then apply classifiers or topic models to interpret the explanations.
- ▶ **What LLMs cant do yet, but might be possible:**
  - ▶ “give me a list of all English-language proverbs emphasizing the value of patience”
  - ▶ “count the usage of these proverbs in a provided corpus (or one that is accessible online) by county and over time”.
- ▶ **What LLMs cannot do, and won't be able to do:**
  - ▶ “count the usage of these proverbs by county and over time, but I don't have the corpus”
  - ▶ “Give me a measure of the counties in england over time by their cultural valuing of patience”
  - ▶ “What is the causal effect of industrialization on the valuing of patience?”

## GPT-4o System Prompt

You are a helpful research assistant with knowledge of the literary history of proverbs. Explain the meaning of this proverb in a few sentences of plain language. Interpret the proverb in terms of the implied beliefs, values, attitudes, norms, or ideology.

## GPT-4o System Prompt

You are a helpful research assistant with knowledge of the literary history of proverbs. Explain the meaning of this proverb in a few sentences of plain language. Interpret the proverb in terms of the implied beliefs, values, attitudes, norms, or ideology.

Structure your answer as follows: "This proverb means that... If a person uses this proverb, it implies that this person ... If this proverb is prevalent in a society, it implies that the society...".

# Prompting Matters

## **Old prompt:**

Explain the meaning of this proverb in a few sentences of plain language, including an assessment of its moral, behavioral, or ethical implications.

Begin your answer with: "This proverb means that...".

## **New Prompt:**

Explain the meaning of this proverb in a few sentences of plain language. Interpret the proverb in terms of the implied beliefs, values, attitudes, norms, or ideology.

Structure your answer as follows: "This proverb means that... If a person uses this proverb, it implies that this person ... If this proverb is prevalent in a society, it implies that the society...".

# Prompting Matters

**Proverb: What belongs to the master is forbidden to the slave.**

# Prompting Matters

## **Proverb: What belongs to the master is forbidden to the slave.**

Old Explanation: This proverb means that there are certain privileges or possessions that are reserved for those in positions of power or authority and are not accessible to those who are subjugated or have lower status. It highlights the inequality and restrictions placed on individuals based on their social or hierarchical standing. **The moral or ethical implication of this proverb is a critique of social injustice and the unfair distribution of resources and rights, emphasizing the disparity between the powerful and the powerless.**

New Explanation: This proverb means that what is owned or controlled by someone in a position of power (the master) is not accessible or allowed for those in a subordinate position (the slave). If a person uses this proverb, it implies that this person recognizes and accepts the strict boundaries and inequalities that exist between different social classes or roles. **If this proverb is prevalent in a society, it implies that the society has a clear hierarchy and reinforces the idea that those in power have exclusive rights over their possessions and privileges, while those beneath them must adhere to limitations and restrictions.**

# Validation of Explanations Against Dictionaries

Out of debt, out of danger.

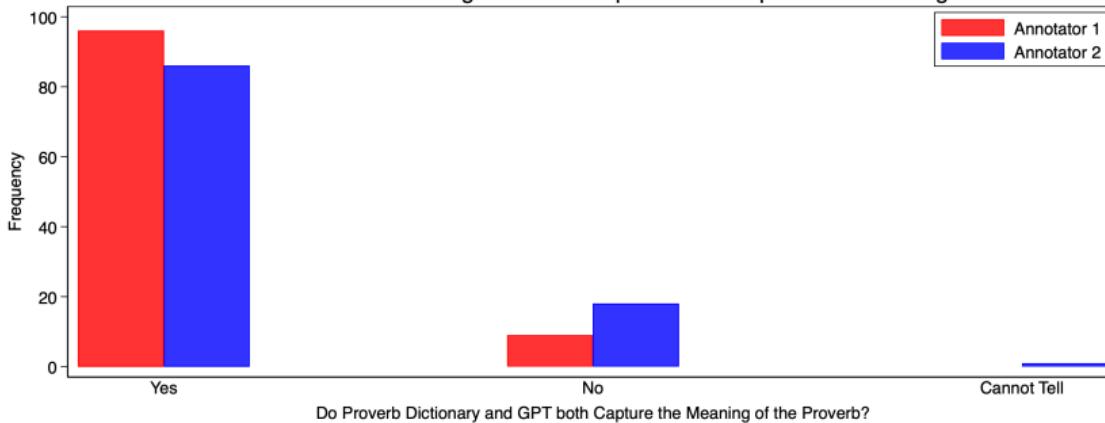


This proverb means that being free from debt removes the associated risks and dangers that come with owing money. Financial obligations can lead to stress, legal trouble, and other problems, so eliminating debt can result in a safer and more stable life. The ethical implication is that it promotes responsible financial behavior and cautions against the perils of borrowing beyond one's means.

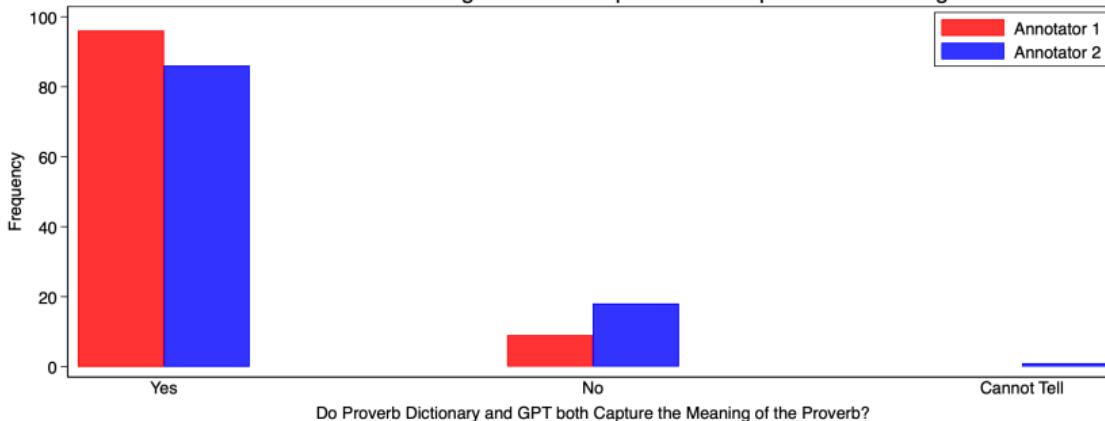
**out of debt, out of danger** Nobody who owes money can feel safe and secure: *I breathed a sigh of relief as I handed over the wad of bills—out of debt, out of danger.* The proverb was first recorded in 1639.

- ▶ Two blind annotators from UpWork viewed 150 proverbs accompanied by a proverb dictionary explanation and our explanation (with order randomized).
- ▶ They were asked:
  - ▶ Do both explanations capture the meaning of the proverb?
  - ▶ Which explanation do you prefer?

### Human Annotation of Proverb Meanings: Do both Explanations Capture the Meaning of the Proverb?

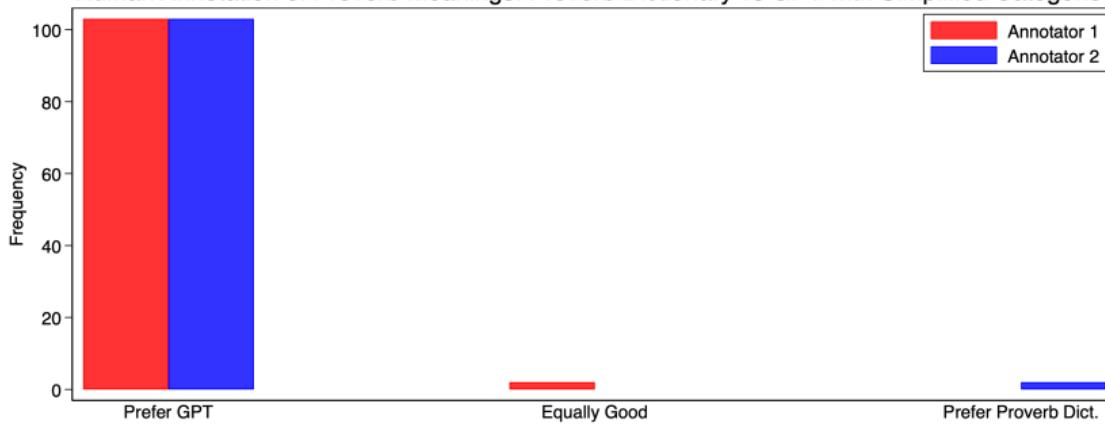


### Human Annotation of Proverb Meanings: Do both Explanations Capture the Meaning of the Proverb?



Do Proverb Dictionary and GPT both Capture the Meaning of the Proverb?

### Human Annotation of Proverb Meanings: Proverb Dictionary vs GPT with Simplified Categories



Preference between Proverb Dictionary and GPT Explanations

## Robustness: Other LLMs (e.g. Claude)

Proverb: Out of debt, out of danger.

This proverb means that being free from debt or financial obligations reduces one's exposure to potential risks or troubles. It suggests that debt can be a source of vulnerability or danger, and that by avoiding or eliminating debt, one can live a safer and more secure life.

This proverb has moral and ethical implications related to financial responsibility and prudence. It encourages people to live within their means, avoid excessive borrowing or overspending, and prioritize paying off any existing debts. The proverb implies that debt can be a burden or even a threat to one's well-being, and that financial freedom brings greater peace of mind and security.

- ▶ in sample of 20 hand-checked proverbs, 100% of generated meanings did not meaningfully disagree.
- ▶ in sample of 125 GPT-checked proverbs, 100% of generated meanings did not meaningfully disagree.

## Clustering Proverb Explanations With BERTopic

BERTopic dimension-reduces 12.8K proverbs down to 42 clusters based on the explanation texts.

## Clustering Proverb Explanations With BERTopic

BERTopic dimension-reduces 12.8K proverbs down to 42 clusters based on the explanation texts.

1. Encode proverb explanations to 384-dimensional embeddings.
  - ▶ MiniLM-L6-v2, a state-of-the-art encoder for sentences and paragraphs.
    - ▶ context-sensitive transformer architecture based on BERT.
    - ▶ learns embeddings to predict whether pairs of sentences are related or unrelated in giant pre-training corpus (1.1 billion sentence pairs).

# Clustering Proverb Explanations With BERTopic

BERTopic dimension-reduces 12.8K proverbs down to 42 clusters based on the explanation texts.

1. Encode proverb explanations to 384-dimensional embeddings.
  - ▶ MiniLM-L6-v2, a state-of-the-art encoder for sentences and paragraphs.
    - ▶ context-sensitive transformer architecture based on BERT.
    - ▶ learns embeddings to predict whether pairs of sentences are related or unrelated in giant pre-training corpus (1.1 billion sentence pairs).
2. Dimensionality reduction to learn topics:
  - ▶ UMAP to uncover latent manifold in our corpus
  - ▶ HDBSCAN: clustering of topics from areas with high density

# Clustering Proverb Explanations With BERTopic

BERTopic dimension-reduces 12.8K proverbs down to 42 clusters based on the explanation texts.

1. Encode proverb explanations to 384-dimensional embeddings.
  - ▶ MiniLM-L6-v2, a state-of-the-art encoder for sentences and paragraphs.
    - ▶ context-sensitive transformer architecture based on BERT.
    - ▶ learns embeddings to predict whether pairs of sentences are related or unrelated in giant pre-training corpus (1.1 billion sentence pairs).
2. Dimensionality reduction to learn topics:
  - ▶ UMAP to uncover latent manifold in our corpus
  - ▶ HDBSCAN: clustering of topics from areas with high density
3. Label the topics with GPT-4o based on sample of representative explanations.

## Clustering Proverbs by Explanations Works Well

#### A. Patience

Hold on, wait for the grasshoppers.

Haste comes late in the end.

He who pours water hastily into a bottle spills more than goes in.

### The Roman conquered by delay.

Haste makes waste and waste makes want, and want makes strife between the good man and his wife.

Nothing is well ordered that is hasty or precipitate.

He tires betimes that spurs too fast betimes, with eager feeding food will choke the feeder.

## B. Managing Conflict

Many get into a dispute well that cannot get out well.

It is better to keep out of a quarrel than to make it up afterwards.

Dissensions like small streams at first begun, scarce seen they rise but gather as they run.

It is madness for a sheep to treat of peace with a wolf.

Wranglers never want words though they may matter.

Take things always by the smooth handle.

Happy is the man who keeps out of strife

## Clustering Topics: Intruder Detection Validation

- ▶ Intruder detection:
  - ▶ Pair an anchor proverb with one from the same cluster, and with one from a random cluster.
  - ▶ Prompt: Which explanation is closer to the anchor?
- ▶ Same-cluster proverb is picked 84% of the time (GPT-4o, 200 pairs).
  - ▶ With humans, 78%.

## Clustering Topics: Intruder Detection Validation

- ▶ Intruder detection:
  - ▶ Pair an anchor proverb with one from the same cluster, and with one from a random cluster.
  - ▶ Prompt: Which explanation is closer to the anchor?
- ▶ Same-cluster proverb is picked 84% of the time (GPT-4o, 200 pairs).
  - ▶ With humans, 78%.
- ▶ Frequent problems: proverbs giving opposite advice get clustered together; many clusters are too generic.

## Clustering Topics: Intruder Detection Validation

- ▶ Intruder detection:
  - ▶ Pair an anchor proverb with one from the same cluster, and with one from a random cluster.
  - ▶ Prompt: Which explanation is closer to the anchor?
- ▶ Same-cluster proverb is picked 84% of the time (GPT-4o, 200 pairs).
  - ▶ With humans, 78%.
- ▶ Frequent problems: proverbs giving opposite advice get clustered together; many clusters are too generic.
- ▶ What we are probably going to do:
  - ▶ 1) use smaller clusters to reduce genericity
  - ▶ 2) ask GPT to split clusters if they contain proverbs giving opposite advice
  - ▶ 3) ask GPT to merge closely related clusters

# Comparison: Clustering on Proverb Text, Rather than Explanation

***"All that glitters is not gold."***

## Clustering on Metaphorical Explanation

- Never judge by appearances.
- He is like a silvered pin.
- The still sow eats up all the draught.

## Clustering on Literal Text

- Old women's gold is not ugly.
- Where gold avails, argument fails.
- Little discourse is gold, too much is dirt.

### ► Intruder detection comparison (GPT-4o, 200 pairs each):

- clusters built from explanations: Same-cluster proverb picked 84% of the time.
- clusters built from proverb texts: Same-cluster proverb picked 62% of the time.

## Additional Dimensions We Are Measuring with Targeted GPT Prompts

- ▶ practical/useful knowledge rather than abstract concepts and philosophy (e.g. Mokyr)
- ▶ McCloskey's bourgeois values: hard work/work ethic (Weber, 1905),
- ▶ patience/long-term orientation, honesty, rationality, curiosity, and learning (Clark, 2007).
- ▶ innovation/progress (Landes, 1969; Mokyr, 2016)
- ▶ individualism vs conformism (Macfarlane, 1978)

# Data Sources on Proverb Usage

- ▶ Pre-Industrial Usage:
  - ▶ EEBO (Early English Books Online)
  - ▶ ECCO (Eighteenth Century Collections Online)
  - ▶ 17 million pages from 146,000 books, 1475-1750
- ▶ Post-Industrial usage
  - ▶ British Newspaper Archive
  - ▶ 40 million pages from more than 1,000 newspaper outlets, 1750-1900

## Validation of proverb search hits

- ▶ Sample 1,000 proverbs with at least 5 hits in BNA.
- ▶ Sample 5 hits for each proverb.
- ▶ Check these samples (combination of algorithmic and human checking) to document <2% false positive rate – i.e., the proverb we were looking for is there over 98% of the time.

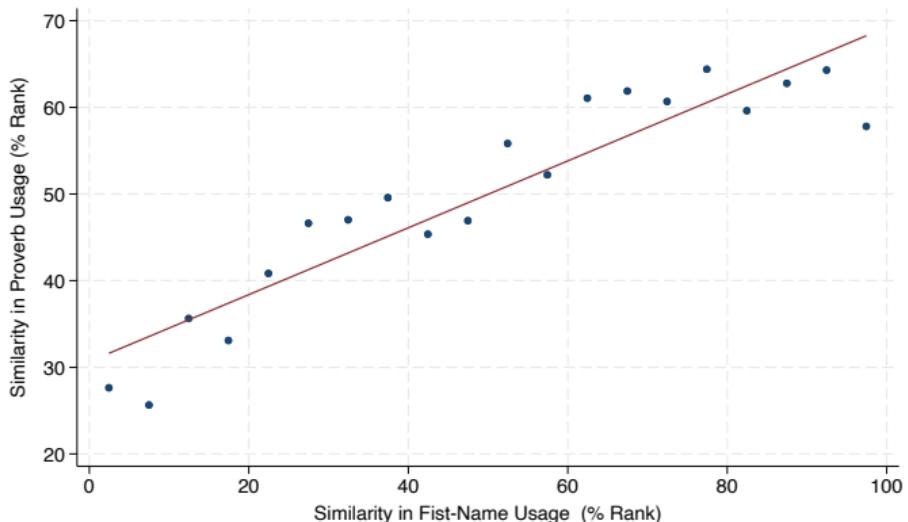
## Validation of proverb search hits

- ▶ Sample 1,000 proverbs with at least 5 hits in BNA.
- ▶ Sample 5 hits for each proverb.
- ▶ Check these samples (combination of algorithmic and human checking) to document <2% false positive rate – i.e., the proverb we were looking for is there over 98% of the time.
- ▶ GPT-4o prompt to evaluate in-context meaning:
  - ▶ proverb used as expected, intentionally, with the assigned meaning: 80.5%.
  - ▶ does not clearly intend the given meaning, but does not disagree or differ either: 17.1%.
  - ▶ proverb used intentionally with a different meaning: 2.4%.

## Proverb usage and first-name usage

- ▶ Proverb distance measure:
  - ▶ Compute vector with share of proverbs by topic for each county's newspapers, 1800-1900.
  - ▶ Compute cosine distance between each pair of counties.
- ▶ Name distance measure:
  - ▶ Compute vector of frequencies of first names by county in 1851 census.
  - ▶ Compute cosine distance between each pair of counties.

## Proverb usage captures cultural distance



Notes. Binscatter of county-to-county cosine similarity in proverb topic frequencies, 1800-1900 (vertical axis) against cosine similarity in first-name frequencies from 1851 census (horizontal axis). **Spearman correlation = 0.26.**

## Cultural Change in the Industrial Revolution

- ▶ Industrialization transformed England economically and culturally, especially starting in the late 1700s, with labor shifting out of agriculture and into manufacturing and services (Broadberry et al., 2015; Wallis et al., 2018).
- ▶ Thompson (1964): Anecdotal evidence on the rise of a “working class” identity as a separate group from employers, and especially changes in attitudes around time.
- ▶ **We assess how local proverb usage changed in localities experiencing industrialization.**

# Estimating Cultural Change After Industrialization

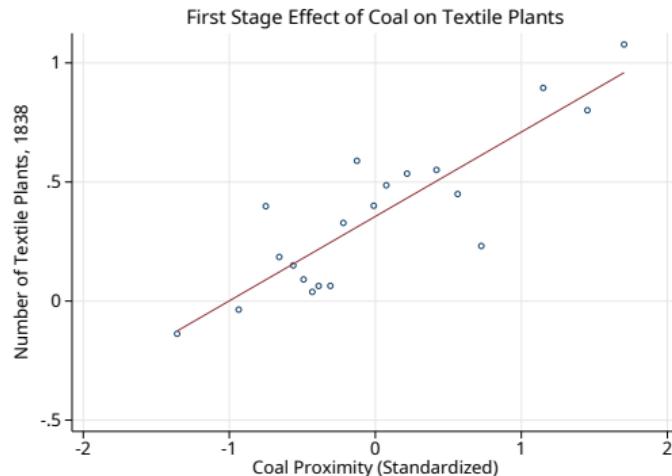
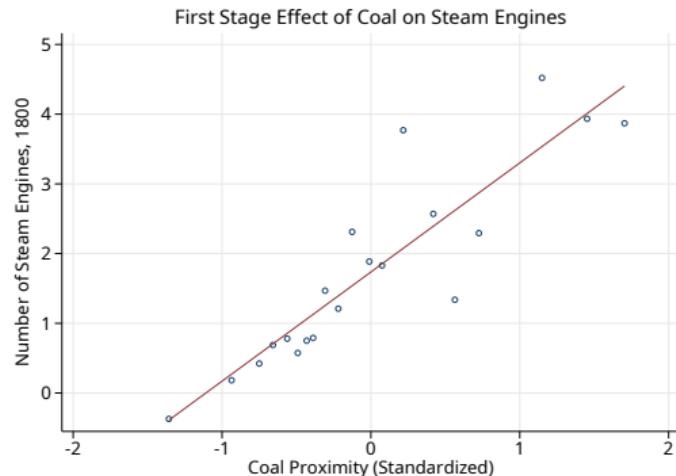
## Separate Regressions by Dimension or Cluster:

$$\begin{aligned}\ln(\#Proverbs_{v,i,t}) = & \theta_t(Industrialization_i \times Period_t) \\ & + \beta_{2,v} \cdot Geo\ Traits_i \times \lambda_t \\ & + \beta_{3,v} \ln(\#Overall\ Proverbs_{i,t}) \\ & + \beta_{4,v} \ln(\#Pages_{i,t}) + \gamma_i + \lambda_t + \varepsilon_{v,i,t}.\end{aligned}$$

- ▶  $Industrialization_i \times Period_t$ : Captures cultural changes before/after industrialization.
  - ▶  $Industrialization_i$  is measured using share of parishes having coal (reduced form).
- ▶ Control Variables:
  - ▶  $Geo\ Traits_i \times \lambda_t$ : Geographic traits interacted with time trends.
  - ▶  $\ln(\#Overall\ Proverbs_{i,t})$ : Log of the total number of proverbs in location  $i$  at time  $t$ .
  - ▶  $\ln(\#Pages_{i,t})$ : Log of the total number of pages in location  $i$  at time  $t$ .
  - ▶  $\gamma_i$ : Location fixed effects.
  - ▶  $\lambda_t$ : Time-period fixed effects.
  - ▶  $\varepsilon_{v,i,t}$ : Unobserved factors affecting proverb cluster prevalence, specific to location  $i$  at time  $t$ .

# First Stage Regression Results

## First Stage Effect of Coal Proximity



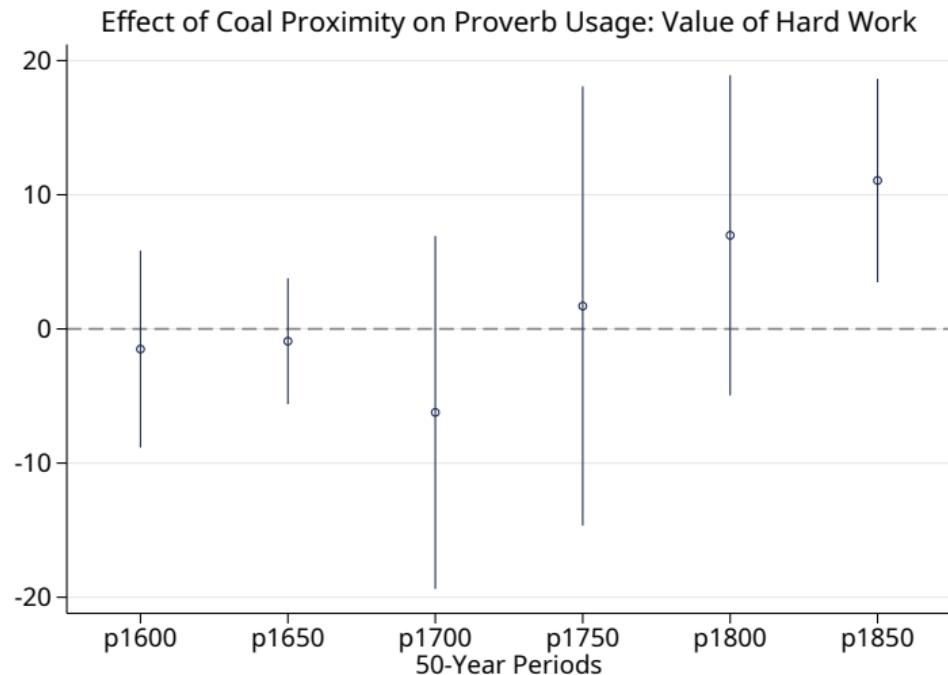
Notes. Binscatter of number of steam engines in 1800 (left panel) and number of textile plants in 1838 (right panel) by county against the share of parishes with coal in that county. Both relationships significant with  $p < .01$ . Kleibergen-Paap F-stat  $> 50$ .

## Proverbs About the Value of Hard Work (Clark, 2007; McCloskey, 2010)

**Topic 18: 'One should stay busy and active to avoid the negative consequences of idleness.'**

1. lazy sheep thinks its wool heavy
2. sloth is the devil's cushion
3. The sluggard's convenient season never comes.
4. to burn one's candle at both ends
5. always in his saddle, never on his way

# Effect of Industrializing on Valuing of Hard Work



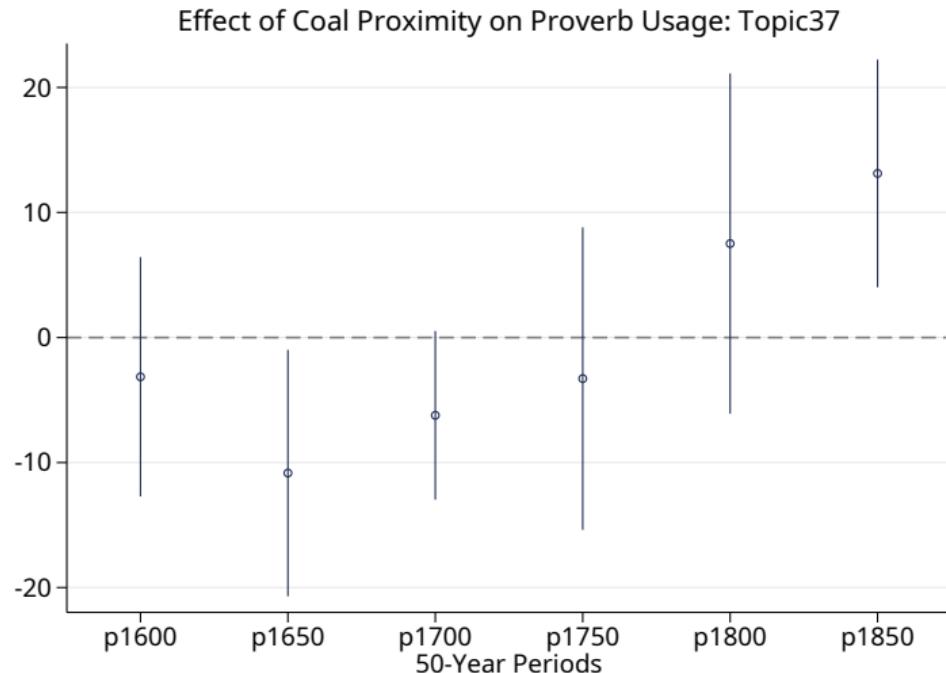
Notes. Regression of usage of proverbs valuing hard work by 50-year period on coal proximity for that county, interacted with period indicators. Regression include county and period FE's. Estimates are relative to the 1550-1599 period. 95% confidence intervals constructed with standard errors clustered by county.

## Proverbs About the Value of Patience (Thompson, 1967)

**Topic 37: 'You should prioritize patience and thoroughness over speed to avoid mistakes.'**

1. no haste but good speed
2. A hasty man never wants woe.
3. Make no more haste than good speed.
4. Hasty climbers have sudden falls.
5. Soft fire makes sweet malt.

# Effect of Industrializing on Valuing of Patience



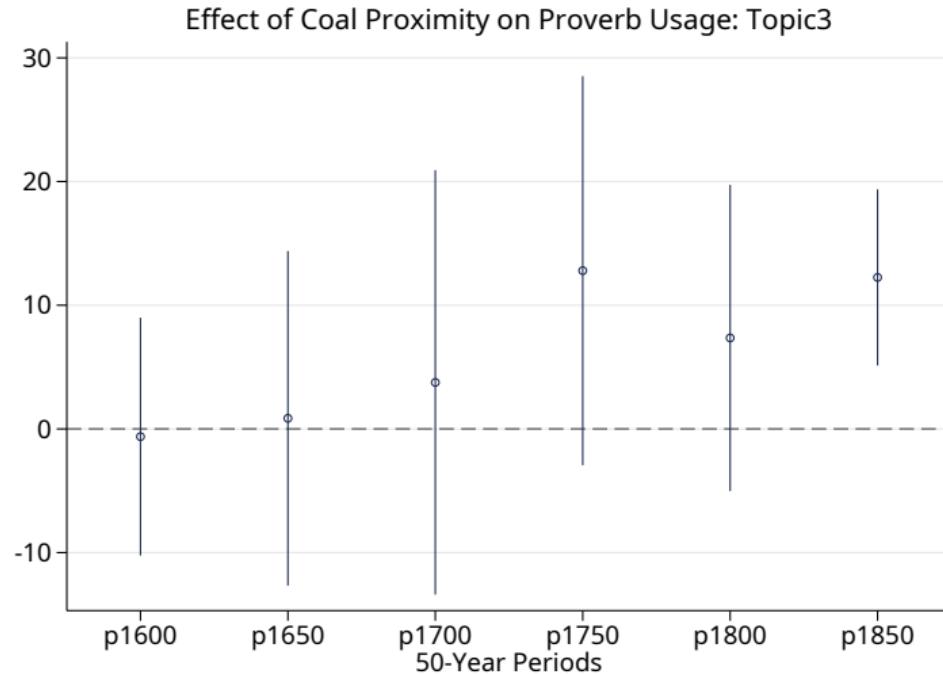
Notes. Regression of usage of proverbs valuing patience by 50-year period on coal proximity for that county, interacted with period indicators. Regression include county and period FE's. Estimates are relative to the 1550-1599 period. 95% confidence intervals constructed with standard errors clustered by county.

# Proverbs About the Value of Gender Norms

## **Topic 3: 'Traditional Gender Roles in Marriage: Women Should Fulfill Expected Roles'**

1. Young wenches make old wrenches.
2. An ill marriage is a spring of ill-fortune.
3. more belongs to marriage than four bare legs in a bed
4. a wise woman is the ornament of her house
5. He that marries a daw eats meikle dirt.

# Effect of Industrializing on Gender Norms



Notes. Regression of usage of proverbs valuing gender norms by 50-year period on coal proximity for that county, interacted with period indicators. Regression include county and period FE's. Estimates are relative to the 1550-1599 period. 95% confidence intervals constructed with standard errors clustered by county.

## Recap

- ▶ New data on proverbs.
- ▶ Use AI to extract the symbolic meaning and translate to the interpretable space.
- ▶ Use embeddings to recover the semantic connections and to do clustering of themes.
- ▶ Use human/AI validation to check validity of different steps of methods.
- ▶ Analyze cultural change.

## Questions on “Proverbs and Cultural Change”

1. What is the research question?
2. How is NLP used in this paper? Is it about measurement or prediction? Is it about supervised learning or unsupervised learning?
3. What is the goal of the data analysis? What are they trying to measure? What are they trying to predict or learn?
4. Is the right model being used? What other models could they have tried?
5. Did they provide validation that the model is delivering on the stated goals?
6. How were the model predictions/statistics used in a social-science analysis? What results seemed incomplete or non-robust?
7. Did they answer the research question? Highlight limitations and open questions.

# Outline

Recap: Recipe for AI

Application: Analyzing Proverbs in Industrial-Age England

NLP for Legal Research

AI and the Research Process

# Legal Texts

- ▶ Legislation
  - ▶ the statutes enacted by legislators, which are then added to a compiled code.
  - ▶ hierarchical structure, extensively cross-referenced.

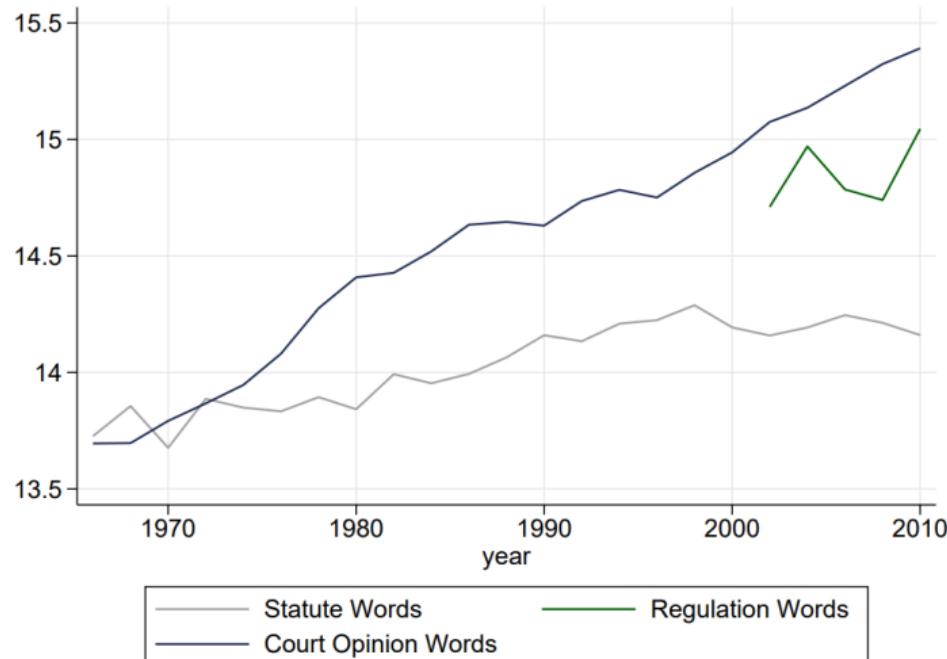
# Legal Texts

- ▶ Legislation
  - ▶ the statutes enacted by legislators, which are then added to a compiled code.
  - ▶ hierarchical structure, extensively cross-referenced.
- ▶ Regulations
  - ▶ the more specific rules to implement legislation, decided by more technocratic agencies.
  - ▶ e.g., tax agency should decide whether a gift counts as income.

# Legal Texts

- ▶ Legislation
  - ▶ the statutes enacted by legislators, which are then added to a compiled code.
  - ▶ hierarchical structure, extensively cross-referenced.
- ▶ Regulations
  - ▶ the more specific rules to implement legislation, decided by more technocratic agencies.
  - ▶ e.g., tax agency should decide whether a gift counts as income.
- ▶ Judicial opinions
  - ▶ when a dispute arises over the meaning of a statute or regulation, a judge decides.
  - ▶ judge will write an opinion, citing statutes and previous caselaw, explaining the interpretation.

# Legal Text Output in U.S. States (Ash, Morelli, and Vannoni 2022)



note log scale – per year we see:

- ▶ ~1.3M words in statutes
- ▶ ~3.3M words in regulations
- ▶ ~4.8M words in state court opinions

## Legal language is different from common language

1. legal documents tend to have more structure (e.g. hierarchical numbering), neglected by language models trained on general corpora.
2. legal language tends to be more precise → lawyers are rewarded for reducing ambiguity.
  - ▶ however:
    - ▶ definitions are often specified elsewhere in the document
    - ▶ extensive and pivotal references to other documents
    - ▶ and laws are often ambiguous anyway (next slide)

## Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

## Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

- ▶ ambiguity arises because legislators have bounded cognition and time.
- ▶ many legal questions are fact-based; sensitive to case specifics
- ▶ when provisions are contested, ambiguity might be used to overcome conflict.

## Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

- ▶ ambiguity arises because legislators have bounded cognition and time.
  - ▶ many legal questions are fact-based; sensitive to case specifics
  - ▶ when provisions are contested, ambiguity might be used to overcome conflict.
- helps explain why efforts to put law on a formal-logic basis, or to say “law is code”, have failed.

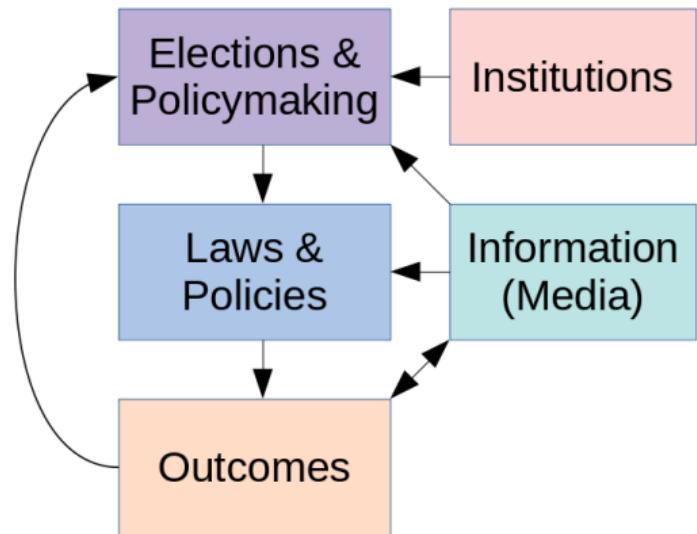
## Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

- ▶ ambiguity arises because legislators have bounded cognition and time.
  - ▶ many legal questions are fact-based; sensitive to case specifics
  - ▶ when provisions are contested, ambiguity might be used to overcome conflict.
- helps explain why efforts to put law on a formal-logic basis, or to say “law is code”, have failed.
- ▶ Which kind of model is better for these cases? IFT, RLHF, or RLVR?

**Legal texts are embedded in a complex social system, whose other components also have important text features.**

- ▶ Institutions
  - ▶ constitutions/charters/treaties
- ▶ Elections and policymaking
  - ▶ campaign ads, parliamentary debates, proposed bills
- ▶ Media
  - ▶ newspaper articles, TV transcripts, lobbying, academic research
- ▶ Laws and policies
  - ▶ legislation, regulation, judicial opinions
- ▶ Outcomes
  - ▶ contracts, culture



# Pragmatics

*When a diplomat says yes, he means 'perhaps';*

*When he says perhaps, he means 'no';*

*When he says no, he is not a diplomat.*

# Pragmatics

*When a diplomat says yes, he means 'perhaps';*

*When he says perhaps, he means 'no';*

*When he says no, he is not a diplomat.*

- ▶ language use depends on the social context.
  - ▶ e.g. social identity, relationships, setting, conversation history, shared knowledge...
  - ▶ e.g., how to use unreliable witness testimony?
- ▶ this is not that well explored in NLP.
  - ▶ standard RLHF/RLVR approaches based on user feedback are not well-designed to capture these points.

# Outline

Recap: Recipe for AI

Application: Analyzing Proverbs in Industrial-Age England

NLP for Legal Research

AI and the Research Process

# AI for Literature Review

<https://broneager.com/ai-literature-review-notebooklm>

The screenshot shows the NotebookLM interface. On the left, there's a sidebar with a list of sources under 'Sources' and a main content area titled 'Entrepreneurship Role Models'. The main content area displays a summary of research findings about stereotypical portrayals of women entrepreneurs as role models.

**Stereotypical portrayals of women entrepreneurs as role models.**

- Sources often depict a disconnect between the media's idealized portrayals and reality. For instance, one source points out that while entrepreneurship is seen as a path to success for many women, it is often framed through a male lens.
- This misrepresentation is further exacerbated by the lack of representation of women from disadvantaged backgrounds, such as ethnic minorities or low-income families, in these role models and envision themselves as successful entrepreneurs.
- Furthermore, the media frequently presents entrepreneurs as individuals who overcome career setbacks or balance work and family life. This narrative reinforces the stereotype that women are better suited for certain types of entrepreneurship than men.

**Practical Applications in Literature Reviews**

This chat feature can significantly streamline your literature review process:

1. Identifying Themes: Quickly spot common themes or contradictions across multiple papers.
2. Gap Analysis: Identify areas where research is lacking by asking about specific topics and noting where information is sparse.
3. Theoretical Framework: Explore how different papers approach theoretical concepts in your field.
4. Methodology Comparison: Compare and contrast methodologies used across various studies.

# “Prompt Engineering” for Research

## Prompting Science Report 1: Prompt Engineering is Complicated and Contingent

10 Pages

Posted: 5 May 2025

### Lennart Meincke

University of Pennsylvania; The Wharton School; WHU - Otto Beisheim School of Management

### Ethan R. Mollick

University of Pennsylvania - Wharton School

### Lilach Mollick

University of Pennsylvania - Wharton School

### Dan Shapiro

Glowforge, Inc; University of Pennsylvania - The Wharton School

Date Written: March 04, 2025

### Abstract

This is the first of a series of short reports that seek to help business, education, and policy leaders understand the technical details of working with AI through rigorous testing. In this report, we demonstrate two things:

- There is no single standard for measuring whether a Large Language Model (LLM) passes a benchmark, and that choosing a standard has a big impact on how well the LLM does on that benchmark. The standard you choose will depend on your goals for using an LLM in a particular case.

- It is hard to know in advance whether a particular prompting approach will help or harm the LLM's ability to answer any particular question. Specifically, we find that sometimes being polite to the LLM helps performance, and sometimes it lowers performance. We also find that constraining the AI's answers helps performance in some cases, though it may lower performance in other cases.

# Using LLM outputs in regressions

Using Large Language Model Annotations for the Social Sciences:

## A General Framework of Using Predicted Variables in Downstream Analyses\*

Naoki Egami<sup>†</sup> Musashi Hinck<sup>‡</sup> Brandon M. Stewart<sup>§</sup> Hanying Wei<sup>¶</sup>

First Version: May 13, 2024

This Version: November 17, 2024

### Abstract

Social scientists use automated annotation methods, such as supervised machine learning and, more recently, large language models (LLMs), that can predict labels and generate text-based variables. While such predicted text-based variables are often analyzed as if they were observed without errors, we show that ignoring prediction errors in the automated annotation step leads to substantial bias and invalid confidence intervals in downstream analyses, even if the accuracy of the automated annotations is high, e.g., above 90%. We propose a framework of *design-based supervised learning* (DSL) that can provide valid statistical estimates, even when predicted variables contain non-random prediction errors. DSL employs a doubly robust procedure to combine predicted labels and a smaller number of expert annotations. DSL allows scholars to apply advances in LLMs to social science research while maintaining statistical validity. We illustrate its general applicability using two applications where the outcome and independent variables are text-based.

## Inference for Regression with Variables Generated by AI or Machine Learning\*

Laura Battaglia<sup>†</sup> Timothy Christensen<sup>‡</sup>  
Stephen Hansen<sup>§</sup> Szymon Sacher<sup>¶</sup>

April 30, 2025

### Abstract

Researchers now routinely use AI or other machine learning methods to estimate latent variables of economic interest, then plug-in the estimates as covariates in a regression. We show both theoretically and empirically that naively treating AI/ML-generated variables as "data" leads to biased estimates and invalid inference. To restore valid inference, we propose two methods: (1) an explicit bias correction with bias-corrected confidence intervals, and (2) joint estimation of the regression parameters and latent variables. We illustrate these ideas through applications involving label imputation, dimensionality reduction, and index construction via classification and aggregation.

## Video Presentation

## Outline of the Exam

- ▶ Paper exam during our standard class period (one hour 45 minutes).
- ▶ You can bring one piece of A4 paper with notes, front and back.
- ▶ Content:
  - ▶ Knowledge questions from the slides and required readings, eg:
    - ▶ explaining / applying concepts
    - ▶ explaining / labeling diagrams
  - ▶ Apply skills learned from the response essay, eg:
    - ▶ identify research question
    - ▶ criticize a hypothetical research design
    - ▶ suggest improvements
- ▶ Preparation:
  - ▶ we will set up a Moodle forum with questions about the exam that the TA's will answer for the whole class.