# Using Large Language Models for Text-As-Data Studies in the Social Sciences

Ulrich Matter[1,2]

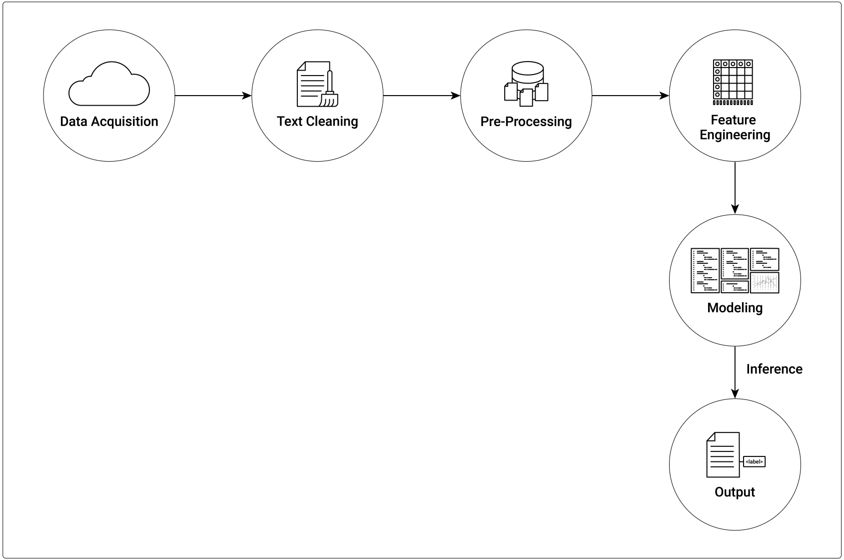
BPLIM Workshop on Empirical Research In the AI Era — December 17, 2024

[1]Bern University of Applied Sciences, [2]University of St.Gallen

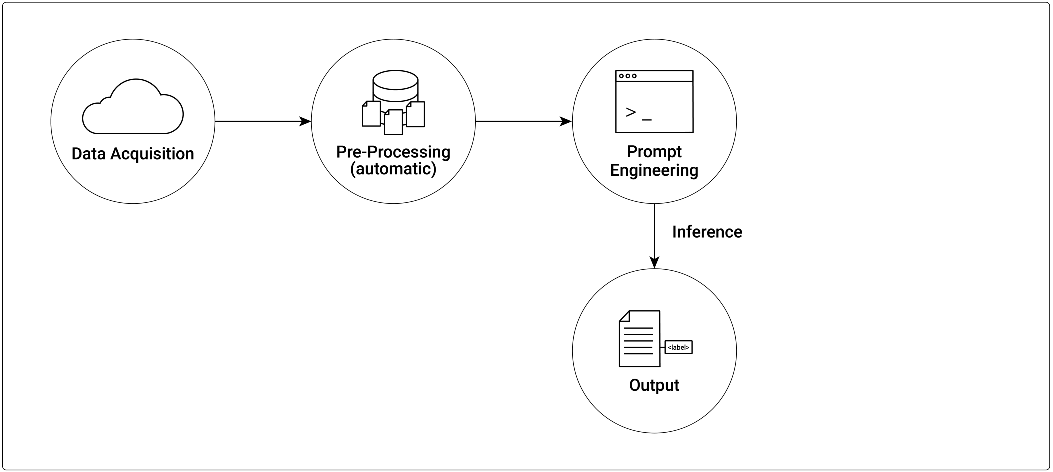
# Text as Data in Economics and Beyond

- **Text as Data (TaD) Approaches:** Rapidly increasing use in empirical social science research (Grimmer and Stewart, 2013; Gentzkow et al., 2019).
  - Employ machine learning algorithms for quantitative text analysis
- **Application Domains:**
  - Economics & Finance: (Tetlock, 2007; Gentzkow and Shapiro, 2010; Baker et al., 2016; Hansen et al., 2017)
  - Management Science: (Hoberg and Phillips, 2010; Guiso et al., 2015; Luca and Zervas, 2016; Netzer et al., 2012)
  - Political Science: (Laver et al., 2003; Quinn et al., 2010; Hopkins and King, 2010; Becker et al., 2017)
- **Challenges in Traditional TaD:**
  - High skill threshold (programming, ML expertise)
  - Labor-intensive (data preprocessing, model training)

# Motivation: Pre-LLM TaD



# Motivation: Potential TaD with LLMs



# Table of Contents

# LLMs vs Traditional NLP in TaD

# Attractiveness of LLMs for TaD: Two Examples

**Example 1: Hotel Reviews, (Zhang et al., 2023)**

> *"very poor value for money. £215 for a cramped room, which felt more like a bedsit, a pathetic shower, and a postage stamp view of the sky if you pressed your head up against the window and craned your neck. a complete rip-off. my complaint at check out was met with no more than empathy, which is no consolation at all. if you get room 107, prepare yourself. i will totally discount this hotel on future trips to london."*

Is this hotel review positive or negative?

# Attractiveness of LLMs for TaD: Two Examples

**Example 2: Economic News, (Barberá et al., 2021)**

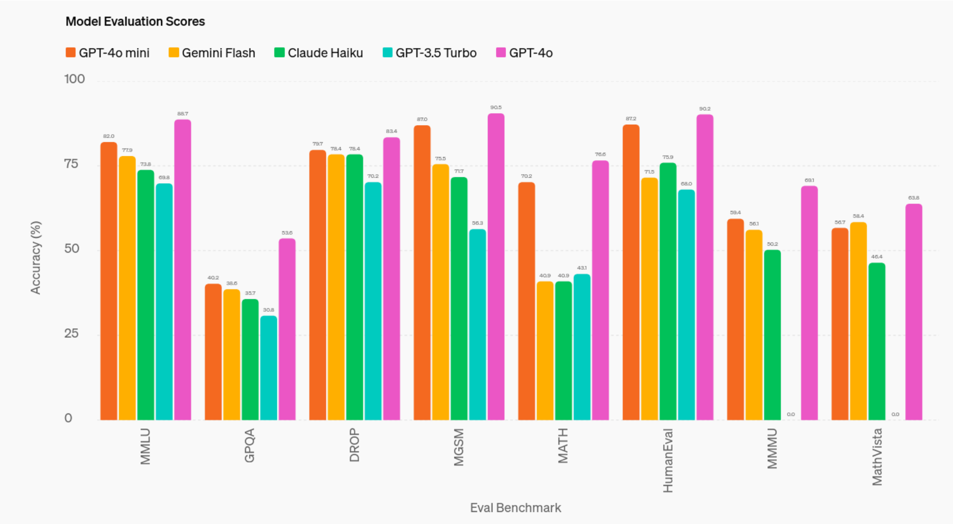
> *"Arthur E. de Cordova gesturing during the session yesterday at the New York Stock Exchange. He is one of the chief specialists trading I.B.M. shares. Trading in I.B.M. opened yesterday on the New York Stock Exchange at 11:20 A.M. The ticker tape showed that a block of S5.000 shares was sold at 8388 a share, followed by another 1,000 shares at the same price. The stock market rally, while powered by the heaviest trading in history, nevertheless slipped into lower gear yesterday and only managed to finish with a moderate gain..."*

Are these positive/optimistic economic news?

## Challenges

- Prompt Engineering:
  - How to counter hallucinations?
  - How to ensure stability/consistency
- Reference point?
  - Are humans (human labelers) still a reasonable "gold standard"?

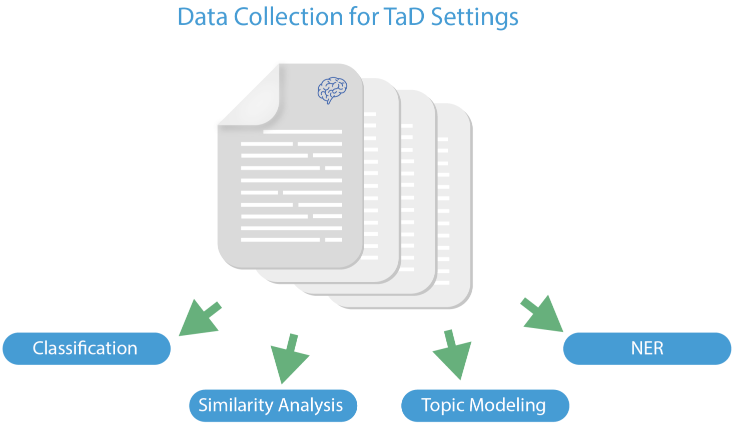
# Challenges


Model Evaluation Scores
GPT-4o mini
Gemini Flash
Claude Haiku
GPT-3.5 Turbo
GPT-4o
100
75
50
25
0
Accuracy (%)
82.0
77.9
73.8
69.8
88.7
40.2
38.6
35.7
30.8
53.6
79.7
78.4
78.4
70.2
83.4
87.0
75.5
71.7
56.3
90.5
70.2
40.9
40.9
43.1
76.6
87.2
71.5
75.9
68.0
90.2
59.4
56.1
50.2
0.0
69.1
56.7
58.4
46.4
0.0
63.8
MMLU
GPQA
DROP
MGSM
MATH
HumanEval
MMMU
MathVista
Eval Benchmark

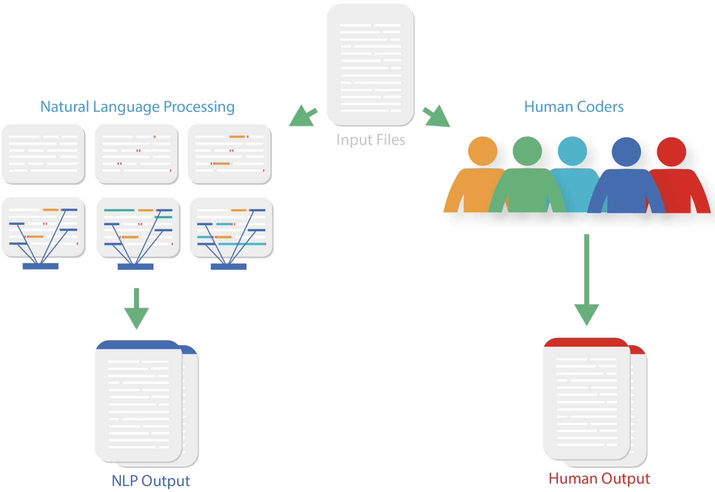
# Research Question, Aim

- **Research Questions:**
  - Can Large Language Models (LLMs) like GPT-4 (OpenAI, 2023; Zhang et al., 2023) improve TaD approaches?
  - Can LLM-based approaches be scaled consistently for standard NLP tasks (e.g., sentiment analysis)?
  - Do LLM-based methods outperform traditional methods and human coders in social science contexts?
- **Contributions:**
  - Introduce and evaluate an LLM prompting strategy for common TaD tasks.
  - Compare performance against conventional pipelines and expert human labeling.
  - Provide insights to enhance accessibility, accuracy, and efficiency of TaD in social sciences.

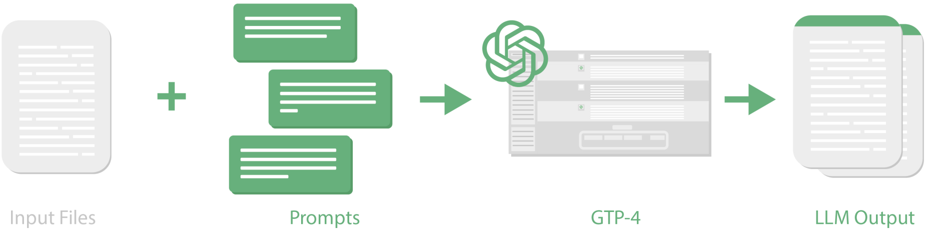
# Starting Point



# Get Two Reference Points: Traditional NLP and Humans

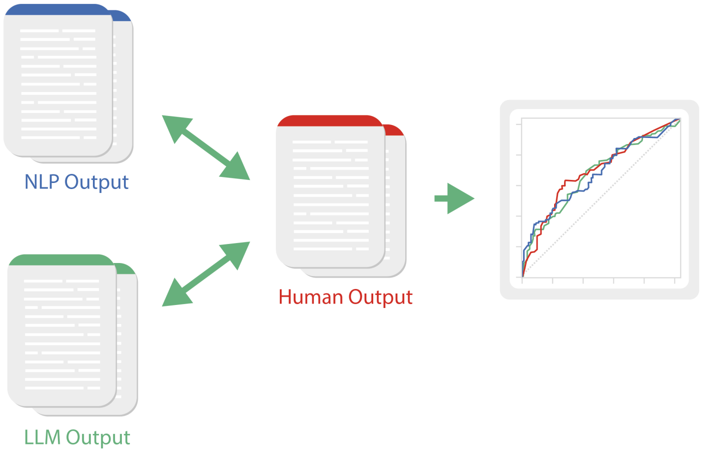


# GPT-based Approach



# Approach/Analysis Plan 1 (Alignment, Accuracy)

1.



# Approach/Analysis Plan 2 (Human Rating)

2.

MTurk Survey: Human Coding Recognition

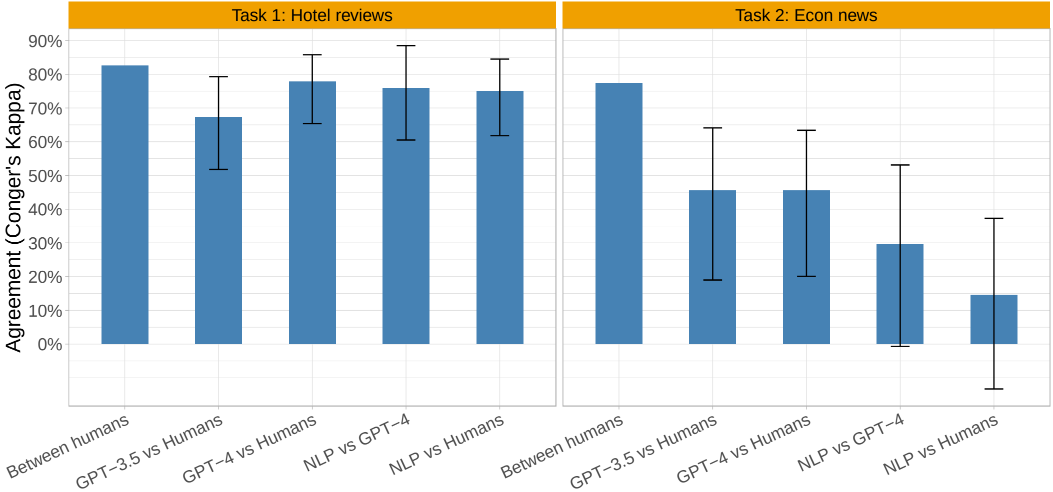
3.

MTurk Survey: Rating of Output

# Prompt Example: Hotel Reviews

```
You are a world-leading expert in sentiment analysis. Given a collection of hotel reviews, analyze each review and
categorize it as either positive or negative. For each review, consider factors such as guest satisfaction, service
quality, cleanliness, comfort, and overall experience. If a review reflects satisfaction, commendation, or positive
feedback on these aspects, categorize it as positive. Conversely, if a review expresses dissatisfaction, complaints, or
negative feedback, categorize it as negative. Provide a concise categorization for each review labeling it simply as
""Positive"" or ""Negative"". Use your language understanding capabilities to ensure the accuracy of the categorization.

For example, a review like ""the room was stylish and had a claw foot bath in the room with a separate bathroom with rain
forest shower."" would be considered positive.
A review like ""for the price paid i was left feeling let down."" would be considered negative.

Output:
 - Present your analysis by listing each review followed by its categorization (""positive"" or ""negative""). Ensure your
classifications are supported by the sentiment detected, showcasing your expertise in language interpretation. Make sure
that each analysis is separate and has separate bullet points or prefixes.

Important Facts to Remember:
- The task relies solely on your pre-existing knowledge and capability to discern and interpret subtle cues in the text.
- Your analysis should go beyond the surface level, considering both the explicit and implicit sentiments expressed.
- The goal is to mimic the nuanced understanding and categorization a human expert might provide, aiming for both accuracy
and depth in sentiment analysis.
- If you think that one review is balanced in sentiments (Neutral), do not create a new category for it such as ‘Neutral’.
Instead, categorize it based on the sentiment you think is more prevalent in the sentence.
- YOUR RESPONSE MUST BE ONLY THE PLAIN CATEGORY. EITHER REPLY WITH ""Positive"" OR ""Negative""! DO NOT ADD ANY ADDITIONAL
TEXT OR COMMENTS!

**************************************************
REVIEW (delimited by triple quotes) :
""""""
{text}
""""""
```

# Some Results (Alignment)



# Personal Learnings, Practical Tips, Open Questions

- Human Coders: Can we give them the same instructions as the model?
- Prompt engineering: Listen to developers/engineers
- Talk with linguists about prompt engineering
- Qualitative vs. quantitative evaluation
  - Qualitative: What is the model doing?
  - Relevance of setup for quantitative evaluation: Flexible serverless cloud-based infrastructure

# Qualitative Comparison of Prompts: Cloud Architecture

Monetary Policy Evaluation

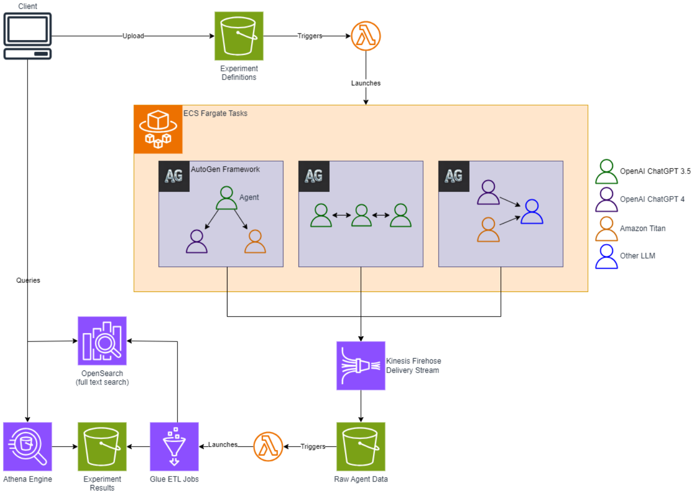
File Edit View Insert Format Data Tools Extensions Help

Menus 100% Comment only

A1 Prompt 1

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Prompt 1 | | | | ChatGPT Outputs | | | | | | | | |
| 2 | ID | Text | Original Category | Formatted Prompt | Response 1 | Category | Response 2 | Category | Response 3 | Category | | | |
| 3 | 157_1 | The action was taken | Dovish | You're a world-leadin<br><br>Instructions:<br>Read the Text: Carefu<br>Identify the Sentimen<br><br>Output Format: Each<br>Sentiment Category: '<br><br>Examples:<br>Text: "We are commit<br>Sentiment Category:<br>Rationale: The expres<br>Text: "The recent data<br>Sentiment Category:<br>Rationale: The phrase<br>Text: "The Committee<br>Sentiment Category:<br>Rationale: The decisi<br><br>Note: Do not include '<br>---<br>Text:<br>Sentiment Category: | Sentiment Category: | Dovish | Sentiment Category: | Dovish | Sentiment Category: | Dovish | | | |
| 4 | 161_2 | Such trends could fos | Hawkish | | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | | | |
| 5 | 52_0 | The Federal Open Ma | Neutral | | Sentiment Category: | Neutral | Sentiment Category: | Neutral | Sentiment Category: | Neutral | | | |
| 6 | 21_5 | Although continuing fa | Dovish | | Sentiment Category: | Dovish | Sentiment Category: | Dovish | Sentiment Category: | Dovish | | | |
| 7 | 78_7 | The Committee perce | Neutral | | Sentiment Category: | Neutral | Sentiment Category: | Neutral | Sentiment Category: | Neutral | | | |
| 8 | 115_6 | Nonetheless, the Con | Hawkish | | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | | | |
| 9 | 160_4 | As a consequence, th | Hawkish | | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | | | |
| 10 | 114_3 | Readings on core infl | Hawkish | | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | Sentiment Category: | Hawkish | | | |
| 11 | 60_0 | The Federal Open Ma | Dovish | | Sentiment Category: | Dovish | Sentiment Category: | Dovish | Sentiment Category: | Dovish | | | |
| 12 | 80_7 | The Committee judge | Dovish | | Sentiment Category: | Dovish | Sentiment Category: | Dovish | Sentiment Category: | Dovish | | | |

# Quantitative Comparison of Prompts

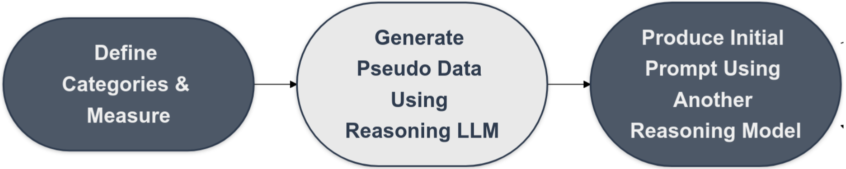


# Quantitative Comparison of Prompts

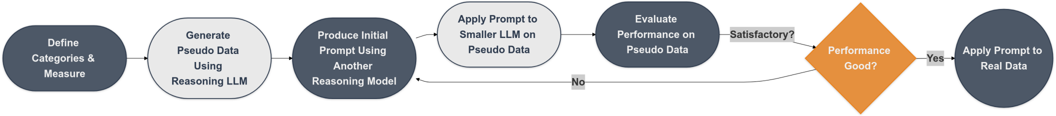
| | experiment | configname | promptname | inputname | prompttext | inputtext | chatprompt | chatresponse |
|---|---|---|---|---|---|---|---|---|
| 1 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_55 | You are a world-leading expert in sentiment analysis... | "great location for convention center. 5 min walk. l... | You are a world-leading expert in sentiment analysis... | Positive |
| 2 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_61 | You are a world-leading expert in sentiment analysis... | "no elevator, boiling eggs by yourself was a bit trick... | You are a world-leading expert in sentiment analysis... | Negative |
| 3 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_49 | You are a world-leading expert in sentiment analysis... | no iron | You are a world-leading expert in sentiment analysis... | Negative |
| 4 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_99 | You are a world-leading expert in sentiment analysis... | the carpark was full so we had to use a different on... | You are a world-leading expert in sentiment analysis... | Negative |
| 5 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_5 | You are a world-leading expert in sentiment analysis... | hotel staff are very friendly and try to help and assi... | You are a world-leading expert in sentiment analysis... | - Positive |
| 6 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_58 | You are a world-leading expert in sentiment analysis... | "nice room, and very relaxing view. the staff are ve... | You are a world-leading expert in sentiment analysis... | Positive |
| 7 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_14 | You are a world-leading expert in sentiment analysis... | pool and bbq area. we were there for australia day. ... | You are a world-leading expert in sentiment analysis... | Positive |
| 8 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_81 | You are a world-leading expert in sentiment analysis... | location and the room | You are a world-leading expert in sentiment analysis... | Positive |
| 9 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_39 | You are a world-leading expert in sentiment analysis... | "great location, friendly staff" | You are a world-leading expert in sentiment analysis... | - Positive |
| 10 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_6 | You are a world-leading expert in sentiment analysis... | "location is not the best, staff needs to improve thei... | You are a world-leading expert in sentiment analysis... | Negative |
| 11 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_7 | You are a world-leading expert in sentiment analysis... | location. fairly quiet considering its location. friend... | You are a world-leading expert in sentiment analysis... | Positive |
| 12 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_3 | You are a world-leading expert in sentiment analysis... | location was great and reception staff very helpful h... | You are a world-leading expert in sentiment analysis... | Negative |
| 13 | hotel-reviews-N100-4turbo | gpt-4-0125-preview_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_10 | You are a world-leading expert in sentiment analysis... | "privacy was an issue. old hotel with no insulation,... | You are a world-leading expert in sentiment analysis... | Negative |
| 14 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_89 | You are a world-leading expert in sentiment analysis... | rooms are pretty small | You are a world-leading expert in sentiment analysis... | - Negative |
| 15 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_85 | You are a world-leading expert in sentiment analysis... | very quiet and little to do | You are a world-leading expert in sentiment analysis... | - Negative |
| 16 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_54 | You are a world-leading expert in sentiment analysis... | breakfast is okay | You are a world-leading expert in sentiment analysis... | - Negative |
| 17 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_13 | You are a world-leading expert in sentiment analysis... | the room was a bit small near the door and bathroom | You are a world-leading expert in sentiment analysis... | - Negative |
| 18 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_23 | You are a world-leading expert in sentiment analysis... | a later departure time | You are a world-leading expert in sentiment analysis... | - Negative |
| 19 | hotel-reviews-N100 | 35turbo_temp02_topp01 | hotel_reviews_prompt | hotel_reviews_N100_text_65 | You are a world-leading expert in sentiment analysis... | "clean rooms with good aircon, very friendly helpful ... | You are a world-leading expert in sentiment analysis... | - Positive |

# Next Steps: Beyond Simple Prompt Engineering

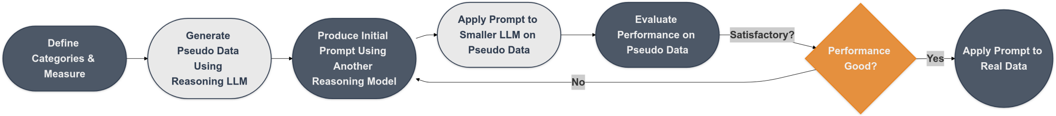
# Basic Idea


Define Categories & Measure
Generate Pseudo Data Using Reasoning LLM
Produce Initial Prompt Using Another Reasoning Model

# Basic Idea Continued



# Possible Extensions




- Automated Feedback Loop for LLM-based Prompt-Engineering
- Fine-tuning Based on Pseudo Data

# Taking Stock/“Conclusion”

# Taking Stock/ "Conclusion"

1. LLMs in TAD studies: A lot of potential.
    - Flexibility: little time for data preparation, no pre-training
    - Wide range of practical applications
    - Promising first applications
2. Challenges: Stability, consistency, validation.
    - Take prompt engineering seriously
    - Define proper benchmark
3. Tools: Quantitative and qualitative evaluation
    - Qualitative comparison of several prompts.
    - Quantitative: scalable/replicable setup, alignment, professional labellers

# Outlook/What's Next?

# Outlook/What's Next?

**Two perspectives on research in the era of AI agents:**

- Agent-based modelling is "back" in business! (economists might want to consider this)
- What happens to an economy if an increasing share of economic decisions are taken by AI agents driven by a handful of foundation models?
  - **Concerns:** supply-side inflation expectations? speed/extent of fluctuations?
  - **Suggestion:** study agents/multi-agent systems/agentic firms experimentally in the lab.

**Thanks for your attention!**

**Questions?**

umatter.github.io

# References

Baker, S. R., N. Bloom, and S. J. Davis (2016, July). Measuring Economic Policy Uncertainty*. *Quarterly Journal of Economics 131*(4), 1593–1636.

Barberá, P., A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler (2021). Automated text classification of news articles: A practical guide. *Political Analysis 29*(1), 19–42.

Becker, J., D. Brackbill, and D. Centola (2017). Network Dynamics of Social Influence in the Wisdom of Crowds. *Proceedings of the National Academy of Sciences 114*(26), E5070–E5076.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as Data. *Journal of Economic Literature 57*(3), 535–74.

Gentzkow, M. and J. M. Shapiro (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica 78*(1), 35–71.

Grimmer, J. and B. M. Stewart (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis 21*(3), 267–297.

Guiso, L., P. Sapienza, and L. Zingales (2015). The Value of Corporate Culture. *Journal of Financial Economics 117*(1), 60–76.

Hansen, S., M. McMahon, and A. Prat (2017, Oct. 31). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *Quarterly Journal of Economics 133*(2), 801–870.

Hoberg, G. and G. Phillips (2010, August). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *Review of Financial Studies 23*(10), 3773–3811.

Hopkins, D. J. and G. King (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science 54*(1), 229–247.

Laver, M., K. Benoit, and J. Garry (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review 97*(2), 311–331.

Luca, M. and G. Zervas (2016). Fake it till You Make it: Reputation, Competition, and Yelp Review Fraud. *Management Science 62*(12), 3412–3427.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science 31*(3), 521–543.

OpenAI (2023). GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science 54*(1), 209–228.

Tetlock, P. C. (2007, May 8). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance 62*(3), 1139–1168.

Zhang, L., X. Li, L. Chen, and L. Wang (2023). One Small Step for Generative AI, One Giant Leap for AGI: A Complete Survey on ChatGPT in AIGC Era. *arXiv* (2304.06488).

Zhang, Z., K. Yang, J. Z. Zhang, and R. W. Palmatier (2023). Uncovering synergy and dysergy in consumer reviews: A machine learning approach. *Management Science 69*(4), 2339–2360.