university of groningen

Quantitative Text Analysis – Essex Summer School

Text Annotation with Large Language Models

dr. Martijn Schoonvelde

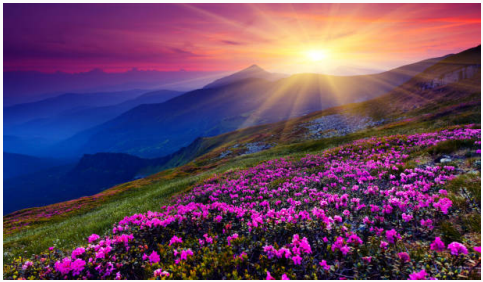
University of Groningen

Today's Class

- The promise (and pitfalls) of prompt-based LLMs
- Flash talk Ladislav
- Lab session: Try out prompt-based LLMs for text annotation
- AOB

Why text analysis matters in the social sciences

- Text is central to human communication – and to social science research
- We have seen applications in sociology, political science, communication studies, psychology, and beyond
- Digitalization has turned much of human communication into analyzable digital data
- $\Rightarrow$ The opportunities for text analysis are **greater than ever**



Source: getty images

The challenge of analyzing text

NLP and machine learning methods have advanced rapidly. Yet despite the potential, text analysis remains a **hard problem**

- Text is often ambiguous, contextual, emotional, or ironic
- Interpretation requires nuanced understanding and world knowledge
- Many of the traditional methods we have seen struggle with these complexities, especially for subtle or contextual meaning

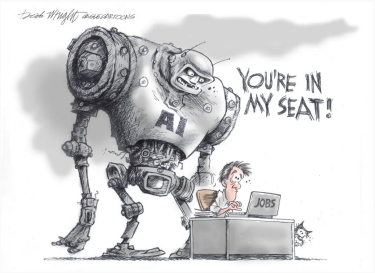
Source: tweetgen.com

Human annotation – still the gold standard?

- Human readers excel at context, emotion, and nuance
- But manual annotation has significant downsides:
  - Time-consuming and expensive
  - Limited in scale
  - Can suffer from bias, inconsistency, and limited reproducibility

The case for prompt-based LLMs

- Can we combine the **scale of machines** with the **understanding of humans**?
- Prompt-based LLMs offer a powerful new tool for text analysis
- Capable of performing tasks such as summarization, translation, classification, and code generation – often without task-specific training.
- Examples include **GPT-4**, **Claude**, **LLaMA**, **Mistral**, **DeepSeek**, etc.



However, "LLMs fit poorly into our existing frameworks for thinking about research methods: many of the lessons from machine learning are obsolete, and while LLMs can be eerily similar to working with a human coder, this model can be similarly misleading." (Törnberg, 2024)

Why are LLMs useful for annotation?

- LLMs encode vast language knowledge and can perform **zero-shot** or **few-shot** inference – greatly reduce manual annotation effort and cost.
- Useful in low-resource settings or multilingual contexts.
- Can adapt quickly to new tasks via prompt engineering

What is prompt engineering?

Definition: The practice of crafting effective natural language instructions to guide LLM behavior for tasks such as classification, summarization, or annotation.

- Eliminates the need for large labeled datasets
- Uses pre-trained knowledge in generative models

"It has become something of an academic Wild West" (Törnberg, 2024)

Prompt engineering vs. traditional sml

Aspect	Traditional sml	LLMs w/ Prompts
Training data	Large labeled set	Few or none (zero/few-shot)
Model training	Required	Pre-trained
Task design	Features, algorithms	Prompts and examples
Reproducibility	High (if open)	Depends on model + API

Best practices for LLM text annotations

Petter Törnberg (2023)

1. Choose a model (open-source preferred)
2. Define a systematic coding procedure
3. Build a **prompt codebook**
4. Validate with human comparison or test sets
5. Engineer prompt iteratively
6. Tune hyperparameters (e.g., temperature)
7. Address ethical and legal concerns

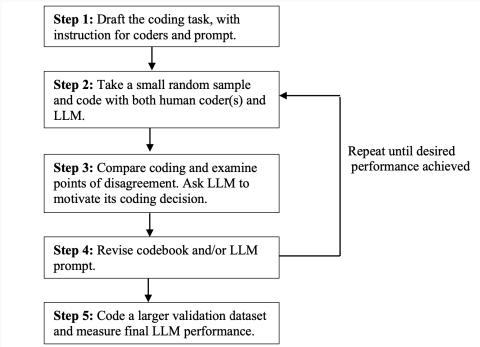
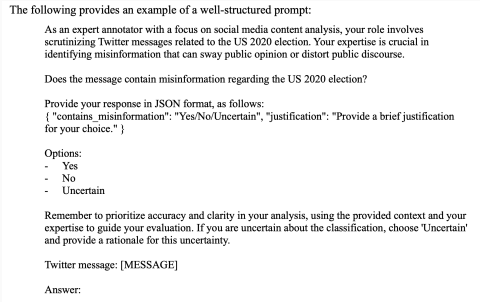


Figure 1: Example of a systematic coding procedure.

Source: Törnberg, 2024

Prompt engineering

- Be clear and specific in your instructions
- Include relevant context and constraints
- Choose **zero-shot**, **one-shot**, or **few-shot** mode
- Format output (e.g., JSON for structured data)
- Split complex tasks into steps (**chain-of-thought prompting**)
- Iterate and test


The following provides an example of a well-structured prompt:

As an expert annotator with a focus on social media content analysis, your role involves scrutinizing Twitter messages related to the US 2020 election. Your expertise is crucial in identifying misinformation that can sway public opinion or distort public discourse.

Does the message contain misinformation regarding the US 2020 election?

Provide your response in JSON format, as follows:
{ "contains_misinformation": "Yes/No/Uncertain", "justification": "Provide a brief justification for your choice." }

Options:
- Yes
- No
- Uncertain

Remember to prioritize accuracy and clarity in your analysis, using the provided context and your expertise to guide your evaluation. If you are uncertain about the classification, choose 'Uncertain' and provide a rationale for this uncertainty.

Twitter message: [MESSAGE]

Answer:

Source: Törnberg, 2024

Prompt example

```
Task description:
-
- You must annotate speeches made in the EU Council of Ministers during legislative negotiations. Does the speaker support the proposal under discussion? Did the intervention describe specific things that the speaker likes or dislikes? Are there aspects of the proposal that the speaker can support and other aspects that are problematic?
-
- You will evaluate each intervention on one dimension, relating to whether or not the speaker supports the proposal being discussed, returning a numerical score from the following 7-point scale in tabular format:
-
Support [-3= strongly against;
        -2= against;
        -1= somewhat against;
        0= neither support nor against;
        1= somewhat support;
        2= support;
        3= strongly support]
-
- You must return an annotation even if an expression of support is not explicit, and a one sentence explanation for the output provided.
-
- A high score nearing 3 represents full support, while a low score nearing -3 represents fully against. Use the full scale as appropriate.
-
######
-
Example output 1:
-
   Intervention_id | Support | explanation \n
   text_1          | -3      | The speaker explicitly states they cannot support the proposal. \n
-
######
-
Intervention text:
-
   {SPEECH_ID}     | {SPEECH}
-
######
```

Applications in political texts

Gilardi et al. (2023):

- Used GPT-3 to classify **topic**, **stance**, and **frames** in tweets
- Found comparable or superior accuracy vs. traditional methods

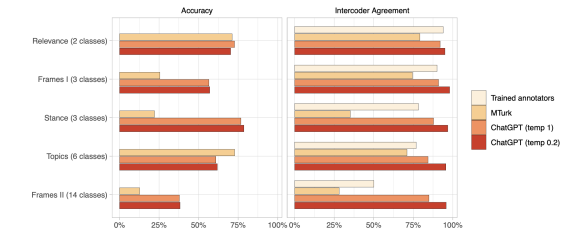


Figure 1: *ChatGPT zero-shot text annotation performance, compared to MTurk and trained annotators. ChatGPT's accuracy outperforms that of MTurk for four of the five tasks. ChatGPT's intercoder agreement outperforms that of both MTurk and trained annotators in all tasks. Accuracy means agreement with the trained annotators.*

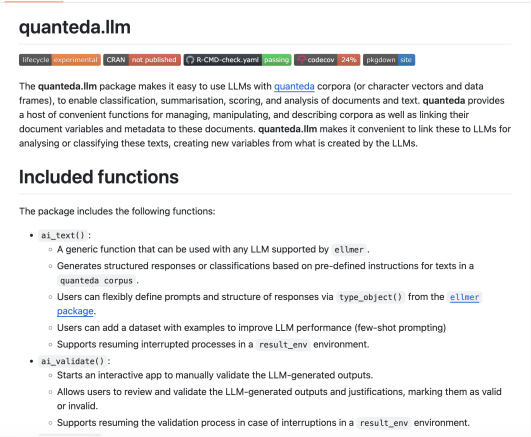
Source: Gilardi *et al.*, 2023

Ellmer

- Package developed by Wickham *et al.* (2025)
- Allows (among other things) for structured data extraction from a whole bunch of model providers
- Bit of a steep learning curve with regards to the syntax of developing structured prompts

quanteda.llm

- R package by Seraphine Maerz and Ken Benoit
- In development (not on CRAN yet) so expect quite some changes
- Integrates Ellmer functions with a Quanteda workflow



quanteda.llm

lifecycle experimental | CRAN not published | R-CMD-check.yaml passing | codecov 24% | pkgdown site

The **quanteda.llm** package makes it easy to use LLMs with quanteda corpora (or character vectors and data frames), to enable classification, summarisation, scoring, and analysis of documents and text. **quanteda** provides a host of convenient functions for managing, manipulating, and describing corpora as well as linking their document variables and metadata to these documents. **quanteda.llm** makes it convenient to link these to LLMs for analysing or classifying these texts, creating new variables from what is created by the LLMs.

Included functions

The package includes the following functions:

- `ai_text()`:
  - A generic function that can be used with any LLM supported by `ellmer`.
  - Generates structured responses or classifications based on pre-defined instructions for texts in a `quanteda corpus`.
  - Users can flexibly define prompts and structure of responses via `type_object()` from the `ellmer` package.
  - Users can add a dataset with examples to improve LLM performance (few-shot prompting)
  - Supports resuming interrupted processes in a `result_env` environment.
- `ai_validate()`:
  - Starts an interactive app to manually validate the LLM-generated outputs.
  - Allows users to review and validate the LLM-generated outputs and justifications, marking them as valid or invalid.
  - Supports resuming the validation process in case of interruptions in a `result_env` environment.

Rollama

- Developed and maintained by Johannes Gruber and Maximilian Weber
- Enables working with **local LLMs** from the Ollama repository
- Documentation and examples:: `https://jbgruber.github.io/rollama/`
- Preprint: `https://arxiv.org/pdf/2404.07654`

Ethical and practical considerations when using LLMs

- Biases: LLMs reflect training data (race, gender, politics)
- Environmental costs: High compute usage (Bender *et al.*, 2021)
- Reproducibility: Prefer open-source over closed APIs
- Copyright/privacy concerns: unclear legal landscape

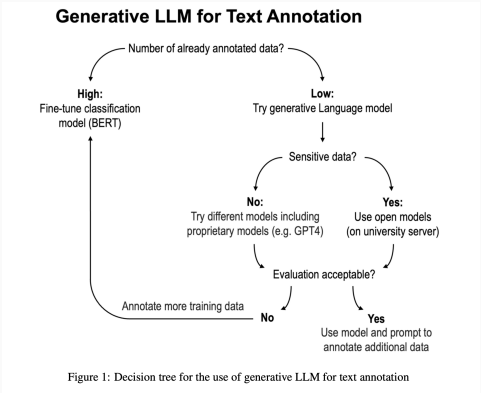


Figure 1: Decision tree for the use of generative LLM for text annotation

Source: Weber & Reichardt, 2023

Best practices when using prompt-based LLMs

- Clearly define task and expected format
- Choose appropriate LLM and setup (API vs local)
- Iterate on prompt engineering
- Validate results against benchmarks or human coders
- Document and share prompt/codebook for transparency

Policy negotiations in the Council of Ministers

- Paschal Donohoe sitting in ECOFIN
- The COM has proposed to harmonise corporate tax...

Paschal's aims

- **Strategic:** Convey and/or conceal underlying preferences (Schimmelfenning 2001; Cross 2012)
- **Informative:** Communicate positions; positives/negatives of a policy choice
- **Deliberative:** Justify views; Convince others; Build consensus (Risse 2000)

What you say and how you say it matters

- Paschal knows this
- He manipulates precision-vagueness in his intervention to navigate competing demands

Precision-vagueness in political communication

Growing literature in party competition that deals with vague political communication:

- "Vagueness denotes political statements that are **non-committal** in terms of the policy action to be taken or the outcome to be achieved" (Praprotnik & Ennser-Jedenastik, 2023)
- Vagueness concerns "the **variation in interpretation a single message may generate** amongst different recipients: the more varied the possible interpretations of a message, the vaguer it is" (Lefevere. 2023)

Precise political communications thought to have the opposite effect

- Clear message interpretation to those that observe it
- Policy commitments transparent
- When public, there is the potential for accountability later on

Precision-vagueness as a rhetorical strategy

To date focus has been on precision/vagueness of preferred policy position

- This ignores that one can be precise or vague about many other aspects of a policy decision

For any given policy position one can be precise or vague about...

- The position itself
  - A potential policy outcome that one can take a stance on
- The reasons for taking a particular stance on a position
  - Positives effects
  - Negative effects
  - Stakeholders affected

Our goal is to measure and then explain precision-vagueness for each of these aspects of Council debate interventions

What might explain variation in precision? Links to policy stance

Policy stance

- Degree to which one expresses support, opposition, or a neutral view towards a policy position

Hypotheses

- H1: Be more precise about position when you support or oppose a position
- H2a: Be more precise about the policy positives when you support a position
- H2b: Be more precise about the policy negatives when you oppose a position
- H3: Be more precise about policy stakeholders when you support or oppose a position

Empirical strategy

Step 1: The data - DICEU corpus (Wratil & Hobolt 2019)

Intervention data

- 1000+ transcribed ECOFIN Council interventions
- Broken down to 7915 sentences

Classification pipelines:

1. Human coding task
2. LLM coding task - Dynamic few-shot approach
3. Compare & refine prompt
4. Once performance strong enough, classify full corpus

Step 2: Human-coding

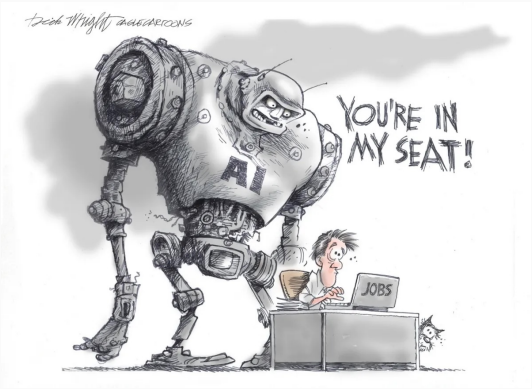
2x coders asked to identify the following in 500 sentences from ECOFIN interventions:

1. Presence / absence of concept
2. Evidence
3. Precision-vagueness of concept if present
4. Evidence

Assess intercoder reliability

- When acceptable, split sample into 250 examples for LLM; 250 examples for validation set

Step 3: Measuring precision with LLMs & dynamic few-shot learning



LLM experiment

- Same sample (250x sentences) and coding task
- Open source LLM: Llama3.1 405B
- **Dynamic few-shot prompting technique** w/ 250 validated examples
- 1x run @ temperature = 0
- Langchain for prompt construction and output validation

Results:

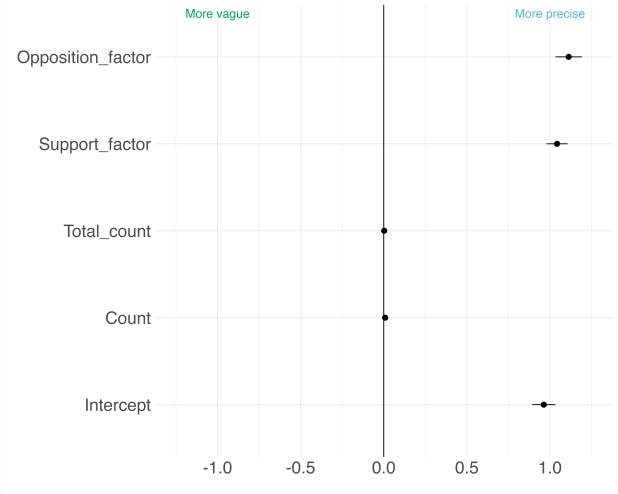
Classifier performance

How did our LLM classifier perform relative to our 'gold standard' test set?

Dimension	Measure	Cohen's κ	F1	% Agree	Precision	Recall
Position	Present	0.632	0.819	0.816	0.849	0.816
	Precision	0.426	0.554	0.596	0.630	0.596
Positives	Present	0.517	0.849	0.860	0.851	0.860
	Precision	0.425	0.798	0.824	0.789	0.824
Negatives	Present	0.585	0.906	0.912	0.905	0.912
	Precision	0.380	0.828	0.856	0.802	0.856
Stakeholders	Present	0.654	0.840	0.836	0.858	0.836
	Precision	0.585	0.795	0.784	0.817	0.784
Stance	Neutral	0.462	0.734	0.756	0.801	0.756
	Supp.-Opp.	0.489	0.719	0.748	0.806	0.748

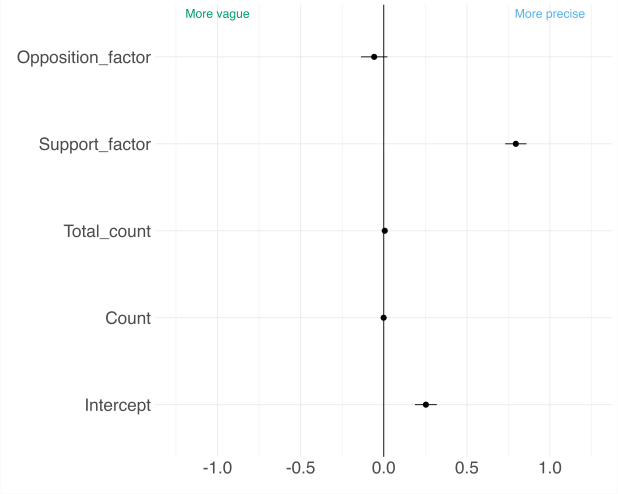
Table 1: Classifier performance across each concept and measure

Substantive results: H1



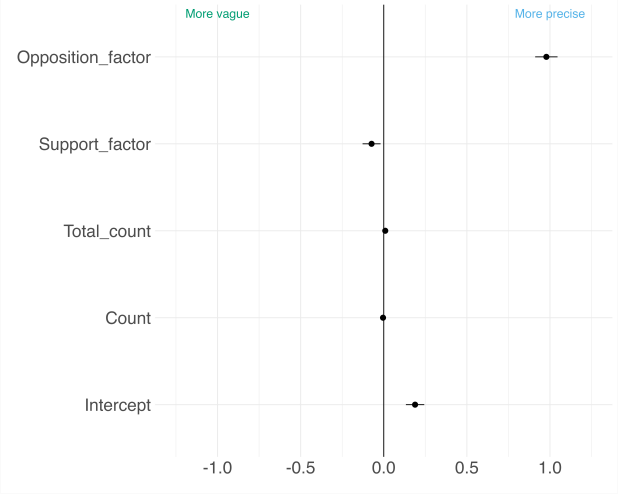
H1: Be precise about position when you support or oppose a position

Substantive results: H2a



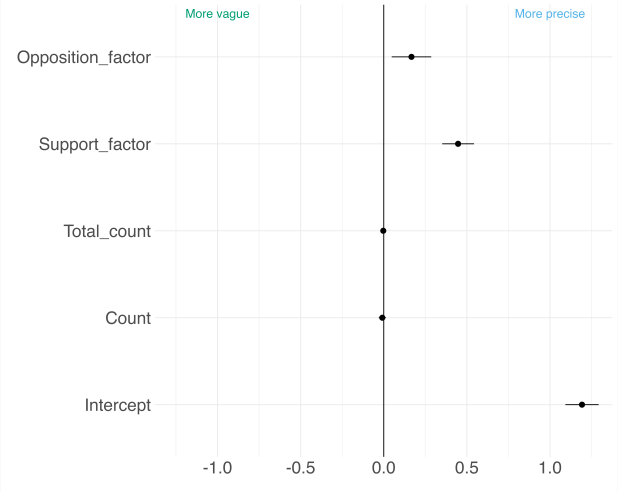
H2a: Be precise about the policy positives when you support a position

Substantive results: H2b



H2b: Be precise about the policy negatives when you oppose a position

Substantive results: H3



H3: Be precise about policy stakeholders when you support or oppose a position

Preliminary conclusions:

Overarching aims

1. Measure precision-vagueness in Council negotiations using LLMs
    - Position, policy effects (+/-) and stakeholders
2. Assess how they relate to policy stance

Results

- LLMs with dynamic few-shot learning approach can match human performance
    - Reasonably high performance metrics

Preliminary conclusions:

Overarching aims

1. Measure precision-vagueness in Council negotiations using LLMs
   - Position, policy effects (+/-) and stakeholders
2. Assess how they relate to policy stance

Results

- LLMs with dynamic few-shot learning approach can match human performance
  - Reasonably high performance metrics
- **Substantively:** Precision-vagueness varies as expected
  - Moves away from a neutral policy stance associated with more precise discussions of positions (H1) and stakeholders (H3)
  - Negotiators more precise about positives of policies they support (H2a) and more precise about the negatives of policies they are against (H2b)