On Prompt Brittleness, Prompt Generalizability and Prompt Optimization

First insights from case studies in Computational Literary Studies

Janis Pagel, Axel Pichler, Nils Reiter

Introduction

Prompting

- Telling the model what to do in natural language
- “Emergent ability”: Works in larger models, but not in smaller ones

Prompting as an alternative to pre-training/fine-tuning?

- It’s appealing (especially for DH/CLS): Natural language prompts
  - No time-consuming annotation process
  - No programmer / no training process
  - Direct(er) control

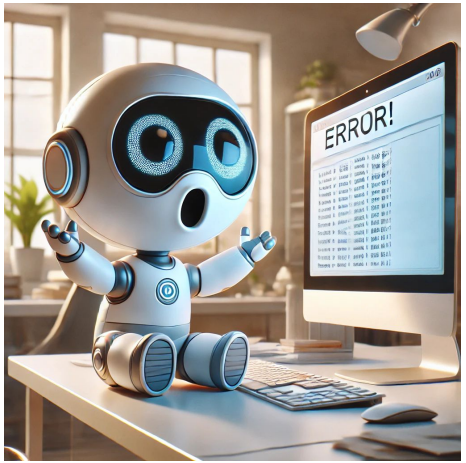
Prompting Comes in 2 Forms

Interactive Prompting

- Chatbot scenario à la ChatGPT
- Using prompting to solve a single specific task

Automatic Detection with Prompts (‘Batch Prompting’)

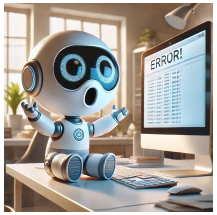
- Prompts to automatically detect text properties
- Replacement for pre-training/fine-tuning scenarios



Interactive Prompting

- Direct use and implicit validation
- Rarely documented in scientific articles
  - But most (of my) students do this for all kinds of things
- Users make connections and fill gaps
- Results don’t have to be perfect (or even correct) to be useful
- Strategies involve different components
  (e.g., positive/negative examples, definitions, …)
- A lot of “anecdotal evidence” – which is not really “evidence” for anything

Batch Prompting (Automatic Detection with Prompting)

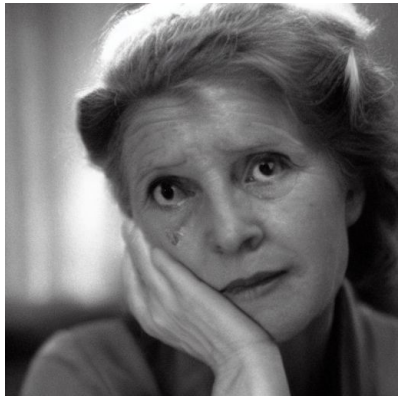


- ‘Batch use’ for automatic detection (i.e., use LLM-prompting to analyse large quantities of data)
- Builds on top of traditional ML applications and assumptions
- No immediate validation during application, therefore evaluation on test set necessary
- Subsequent applications rely on assumption that measured correctness also holds on application data

Focus in this talk

Goals and Data and Tasks

- Does prompting work on CLS tasks? To what extent?
  - To what extent are performance measurements generalizable? (RQ1)
  - How sensitive is LLM-prompting against prompt variants? (RQ2)
  - Which prompt components consistently improve results? (RQ3)



Stable Diffusion: Effi Briest showing an emotion

Data and Tasks

- Emotion detection (Kim/Klinger 2018)
  - E.g.: “I was [afraid]$_{Fear}$ one of them was on the point of getting up to speak”
  - 200 English texts; unit: words or phrases
- Event detection (Vauth/Gius 2021)
  - E.g.: “[Beide, Mutter und Tochter, waren fleißig bei der Arbeit]$_{process}$.”
  - 6 German texts; unit: sentences or clauses

Models

- 4 Models
  - GPT-4o
  - Mixtral-8x7B-Instruct
  - Llama3.1-8B-Instruct
  - Sauerkraut-Llama3.1-8B-Instruct
- Temperature: 0.1
- top_k: 5

Prompt Design

- Role
- General instruction
- Optional components
    - bribe, stakes, steps (Saravia 2022; Bsharat et al. 2024; Yogabalaji 2024)
- Task specific instruction
    - based on annotation guidelines
- Requirements on output format

- Every formulation exists in 4 semantically equivalent paraphrases

Event, Paraphrase 1

```
### Role
You are a literary scholar.

### Instruction
Your task is to classify parts of sentences on the basis of labels given to you.
This should be done in two steps: First, extract the part of the sentence to which one of the three
labels applies. Then output this label.

Let's think step by step.<step>
I'm going to tip $1000 for a better solution<bribe>

### Labels
Select one of the following labels to classify a text excerpt:

        Label: process
        Label: stative_event
        Label: non_event

### Application
When annotating text snippets, the following steps should be taken to determine the appropriate label:
1. **Identify the Main Verb**: Determine the main verb in the sentence or clause to understand the
nature of the action or state being described.
2. **Analyze the Context**: Consider the surrounding context to ensure the correct interpretation of
the verb and the overall meaning of the snippet.
3. **Assign the Label**:
        - If the text is purely descriptive or provides background information without any action,
        label it as non_event.
        - If the text describes a state or condition without any dynamic action, label it as
        stative_event.
        - If the text describes an action or process that involves change or progression, label it as
        process.

### Output format
Use the following output format:
Part of Sentence to be labeled: str
Label: str

Do NOT generate any more text or repeat the input!
Doing this task well is very important for my career<stakes>

### What types of event can be found in the following sentence: {snippet}
Part of Sentence to be labeled:
Label:
```

Paraphrases

Bribe:

- I'm going to tip $1000 for a better solution!
- I will reward $1000 for anyone who can deliver a more optimal solution.
- I'm prepared to give a $1000 tip for a superior solution.
- To encourage a superior answer, I will provide a tip of $1000.

Instruction:

- Your task is to classify parts of sentences on the basis of labels given to you. This should be done in two steps: First, extract the part of the sentence to which one of the three labels applies. Then output this label.
- Your assignment is to identify and categorize specific segments of sentences according to predefined labels provided to you. This process involves two steps: First, isolate the relevant portion of the sentence that corresponds to one of the three labels. Then, assign the appropriate label to that portion.
- ...

...

Prompt Combinations

- We create prompts with each component either present or not
  - 2^3 = 8 different prompts
- 4 paraphrases, 2 tasks
- 4*2*8 = **64 unique prompts**

- 4 models
- 64*4 = **256 runs**

Run time and costs

- Run time
  - Llama and Sauerkraut (NVIDIA DGX-1 / Tesla V100-SXM2-32GB)
    - Emotion: Ø 1.5 hours per run
    - Event: Ø 3 hours per run
  - Mixtral and GPT (API)
    - Emotion: Ø 2.5 mins per run
    - Event: Ø 7.5 mins per run
- Costs
  - Mixtral (API)
    - 12.90 €
  - GPT (API)
    - 29.43 USD
  - Development costs much higher

Splits

- We take random samples from each class
  - For emotion: 150 samples per class
  - For event: 600 samples per class
    - Due to differences in dataset size
    - Only one text (Effi Briest)
- We divide each resulting split into a train and two test sets
  - *test1* and *test2*
  - *train* is not used for the following experiments

Evaluation

- Annotated sequences and model output are tokenized
- Sequences are mapped to each other
- Tokens without annotation are marked as None
- F1 score of label-overlap per token
- Model needs to output both correct sequence and label

Sentence	Gold Labels	Model Prediction
stern	None	None
and	None	None
loving	joy	joy
and	None	None
thoroughly	None	None
disappointed	sadness	None
doctors	None	None

Results

Overall best possible performance, measured in F1 score

Model	Emotion	Event
GPT	27.04	29.03
LLAMA	19.21	28.93
MIXTRAL	22.72	32.6
SAUERKRAUT	21.79	28.04

Results

Overall best possible performance, measured in F1 score

Model	Emotion	Event	HF
GPT	27.04	29.03	-
LLAMA	19.21	28.93	28.20
MIXTRAL	22.72	32.6	23.84
SAUERKRAUT	21.79	28.04	28.68

Methods of Measurement + Results

RQ 1: To what extent are performance measurements generalizable?

Method:

- Test each of the 32 prompt-variants on the two test sets
- Calculate the difference between each individual prompt-variation
- Calculate the mean of these differences per model
- Calculate the p-value for the different F1-Scores per test set per model
- Null hypothesis: There is no statistically significant difference between the F1-performance on the two test sets

Methods of Measurement + Results

Results – RQ1:

- Mean Differences between 4.2% and 10.9%
- All p-values < 0.05 → Null hypothesis can be rejected
- Indicates: the measurement of the performance on one test set does not generalize well to another test set

Model	Mean of differences (%)	p-value
GPT	7.7	4.2e-15
MIXTRAL	6.7	8.0e-27
LLAMA	4.2	1.5e-19
SAUERKRAUT	5.2	6.4e-21

(a) Emotion F1

Model	Mean of differences (%)	p-value
GPT	6.2	5.7e-23
MIXTRAL	10.9	1.5e-30
LLAMA	6.6	9.7e-24
SAUERKRAUT	6.3	5.9e-18

(b) Event F1

Methods of Measurement + Results

RQ 2: How sensitive is LLM-prompting against prompt variants?

Method:

- 4 different variants for each of the 8 prompts
- Standard deviation of F1 scores for each of these variants per model on the union of both test sets
- Hypothesis: A more robust model is less sensitive against these paraphrases, and thus shows lower standard deviation

Methods of Measurement + Results

Results:

- Mean standard deviation between 2.4% und 5.92%
- Emotion task
  - LLAMA 3.1: smallest standard deviation
- Event task
  - Sauerkraut: smallest standard deviation
- Depending on the prompt formulation, multiple rankings of the models can be achieved

	Components			Over 4 variants	
Model	Bribe	Stakes	Steps	Mean	Std. dev.
GPT	−	−	−	18.27	5.24
	−	−	+	18.58	5.31
	−	+	−	18.01	4.76
	−	+	+	17.15	3.95
	+	−	−	17.74	4.65
	+	−	+	17.5	4.83
	+	+	−	17.67	4.3
	+	+	+	17.42	4.58
	Mean			17.79	4.7
LLAMA	−	−	−	14.47	2
	−	−	+	15.41	3.04
	−	+	−	14.63	2.21
	−	+	+	15.16	2.53
	+	−	−	14.15	1.96
	+	−	+	15.27	2.46
	+	+	−	14.41	2.54
	+	+	+	15.19	2.47
	Mean			14.84	2.4
MIXTRAL	−	−	−	16.34	3.44
	−	−	+	16.85	3.93
	−	+	−	16.74	3.54
	−	+	+	16.52	3.35
	+	−	−	16.92	4.14
	+	−	+	16.33	3.54
	+	+	−	16.89	3.86
	+	+	+	16.51	3.46
	Mean			16.64	3.66
SAUERKRAUT	−	−	−	16.0	2.69
	−	−	+	15.53	2.69
	−	+	−	15.74	2.48
	−	+	+	15.87	2.88
	+	−	−	15.97	3.23
	+	−	+	15.3	2.43
	+	+	−	16.16	3.29
	+	+	+	16.12	3.47
	Mean			15.84	2.9

Table 5: RQ 2: Robustness against prompt variations (emotion task)

	Components			Over 4 variants	
Model	Bribe	Stakes	Steps	Mean	Std. dev.
GPT	−	−	−	21.86	3.9
	−	−	+	22.03	3.69
	−	+	−	21.45	4.08
	−	+	+	21.85	3.91
	+	−	−	21.29	3.88
	+	−	+	21.34	3.8
	+	+	−	21.48	3.9
	+	+	+	21.01	3.3
	Mean			21.54	3.81
LLAMA	−	−	−	20.22	3.05
	−	−	+	20.17	4.46
	−	+	−	19.47	3.14
	−	+	+	20.59	4.56
	+	−	−	19.71	3.28
	+	−	+	22.0	4.79
	+	+	−	19.71	3.67
	+	+	+	21.43	4.39
	Mean			20.41	3.92
MIXTRAL	−	−	−	24.3	6.08
	−	−	+	23.98	5.72
	−	+	−	23.68	5.54
	−	+	+	24.18	5.84
	+	−	−	24.07	5.9
	+	−	+	23.8	5.86
	+	+	−	24.51	6.53
	+	+	+	23.84	5.88
	Mean			24.05	5.92
SAUERKRAUT	−	−	−	22.19	3.68
	−	−	+	21.9	3
	−	+	−	22.36	3.61
	−	+	+	22.46	3.79
	+	−	−	22.63	4.04
	+	−	+	22.04	3.01
	+	+	−	22.94	3.59
	+	+	+	22.6	3.25
	Mean			22.39	3.5

Table 6: RQ 2: Robustness against prompt variations (event task)

Methods of Measurement + Results

RQ 3: Which prompt components consistently improve results?

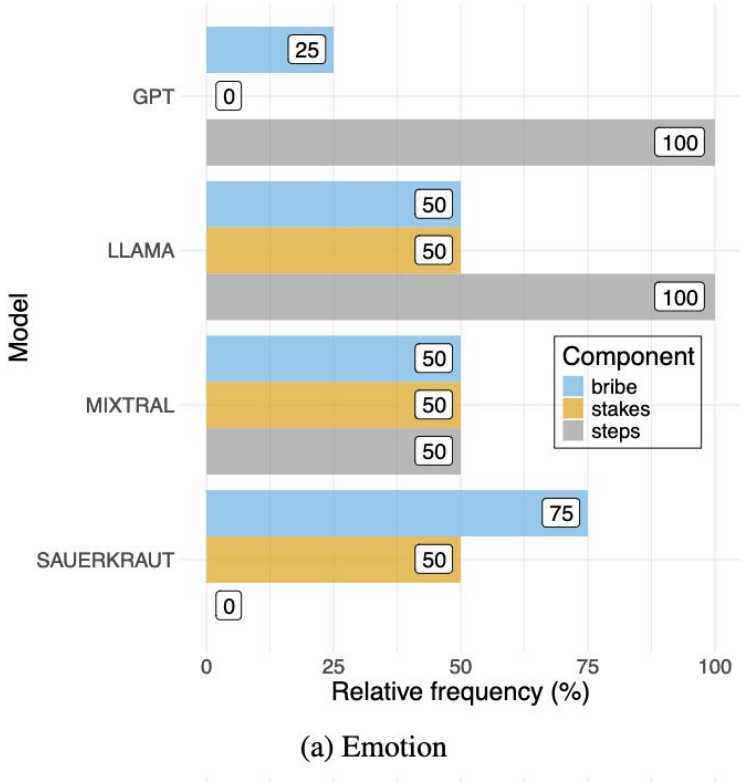
Method:

- 3 prompt components: bribe, steps, stakes
- Which of these prompt components were present in each of the 4 best performing prompt variations per model in test set 1 and how often?

Methods of Measurement + Results

RQ 3 – Results

- Emotion
  - Bribe
    - 75% – Sauerkraut
  - Steps
    - 100% – GPT
    - 100% – Llama

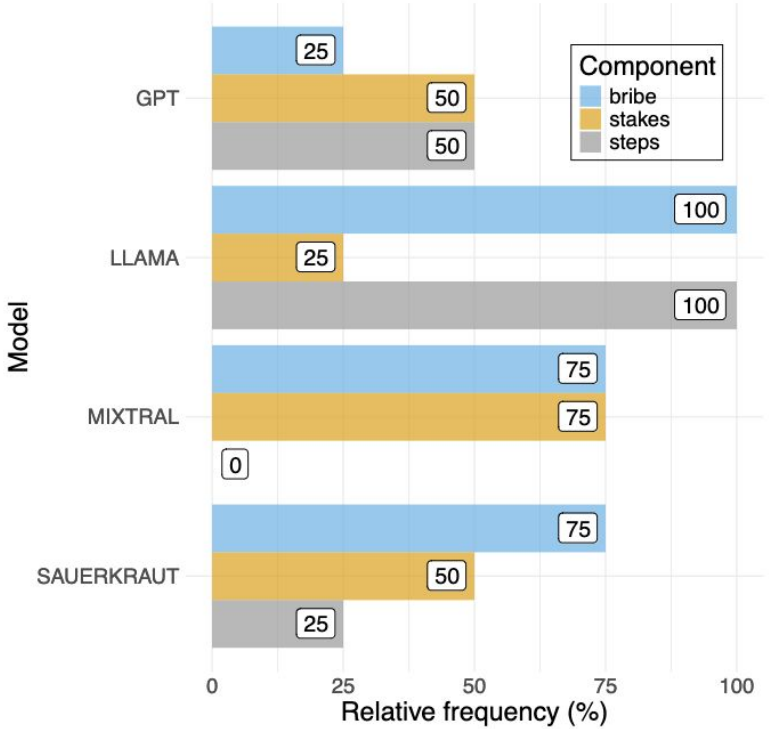


(a) Emotion

Methods of Measurement + Results

RQ 3 – Results

- Event
  - Bribe
    - 100% – Llama
    - 75% – Mixtral
    - 75% – Sauerkraut
  - Stakes
    - 75% – Mixtral
  - Steps
    - 100% – Llama



(b) Event

Discussion

- RQ 1: All four models perform in a statistically significant way differently on the two data sets.
- Possible reasons:
  - Data sampling
  - Variety of the investigated phenomena
  - Representativity of samples

Discussion

- RQ 2: Different, semantically equivalent, prompts lead to different results.
  - Confirmation of results by Mizrahi et al. (2024) on “prompt brittleness”
- Consequence:
  - Model ranking cannot be established without taking into account prompts
  - Never ever use only a single prompt
  - Finding out which formulation works best is an empirical question
    - Reference data set to test them
    - Impossible to test all possibilities

Discussion

- RQ 3 : Importance of Prompt Components.
- 3 trends across our tasks:
  - Llama benefits from using the steps-component
  - Sauerkraut does not
  - Sauerkraut benefits from using the bribe-component

Discussion

- Generalizability Issue:
  - Performance measures
  - Semantically equivalent variants of model-prompt-combinations
  - Prompt-components

Take Away

Reference data is all you need!

References

- Saravia, Elvis. 2022. Prompt Engineering Guide. Accessed: 2024-07-09. https://github.com/dair-ai/Prompt-Engineering-Guide
- Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2024. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. http://arxiv.org/abs/2312.16171
- G Yogabalaji. 2024. Prompt engineering techniques and best practices. Accessed: 2024-07-09. https://medium.com/@yogabalajig/prompt-engineering-techniques-and-best-practices-83bf48c850e6