Prompt Formatting and the Impact on Task Performance

Tobias Leibrock, Gabriel Trannoy, Pontus Amgren, Jakob Sturm

Performance differences in LLM's prompt formats that preserve the prompt meaning and intent

Large language models (LLMs) are highly sensitive to prompt design, including formatting choices that retain meaning and intent. These seemingly minor variations can significantly influence model performance, yet they are often overlooked in research and practical applications.

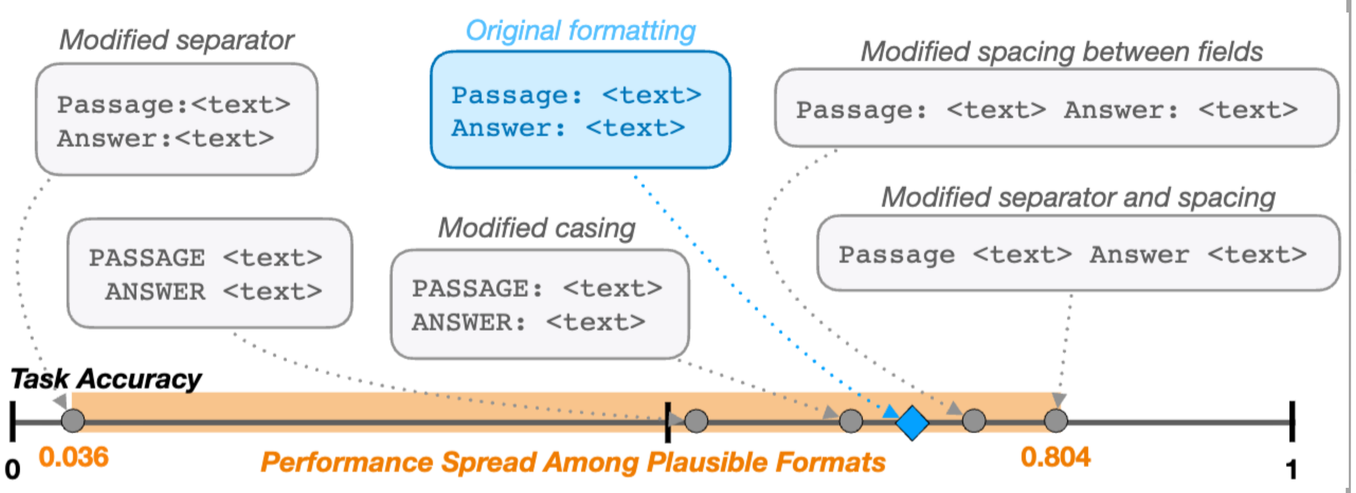


Figure 1: Slight modifications in prompt format templating may lead to significantly different model performance for a given task (adapted from Sclar et al., 2024).

Format classes and synonym replacement

The template is formatted with regards to: Spacing between sections, Casing of words, Seperation of subsections, and Formatting of enumerables

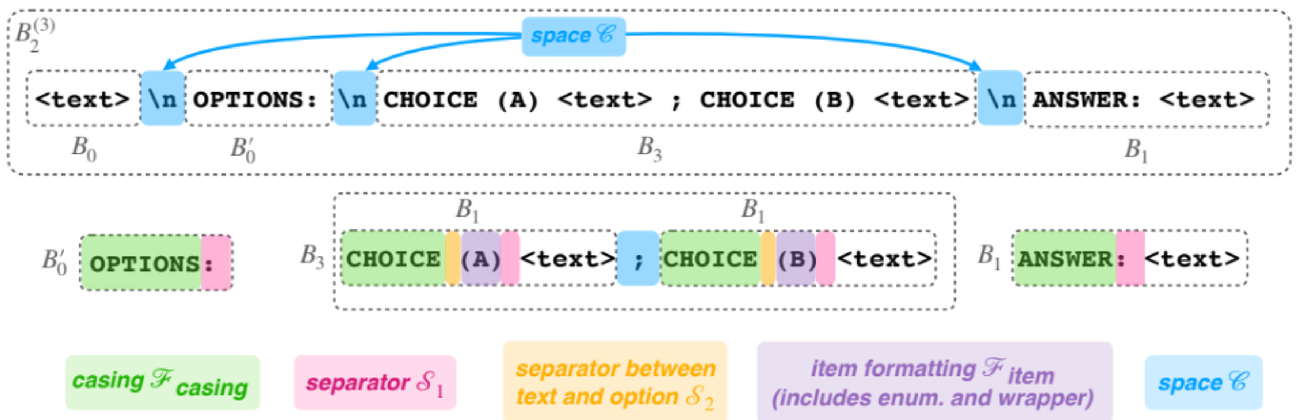


Figure 2: Visualization of the format classes and their effect on a complex prompt (adapted from Sclar et al., 2024).

Benchmark and testing for prompt performance

Natural-Instructions Dataset

- 1600+ tasks including NLI and classification tasks
- ~300 instances per task
- Consistent template structure across tasks
- Used in similar studies

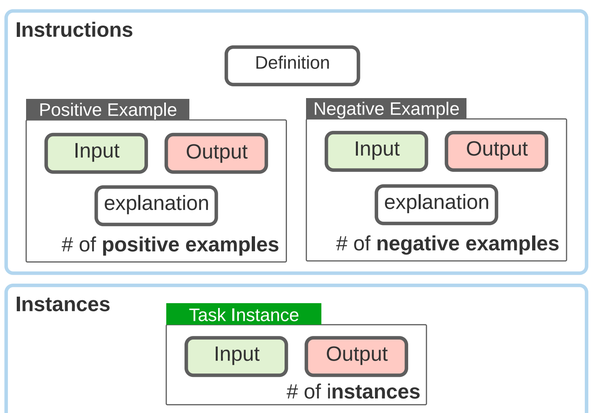
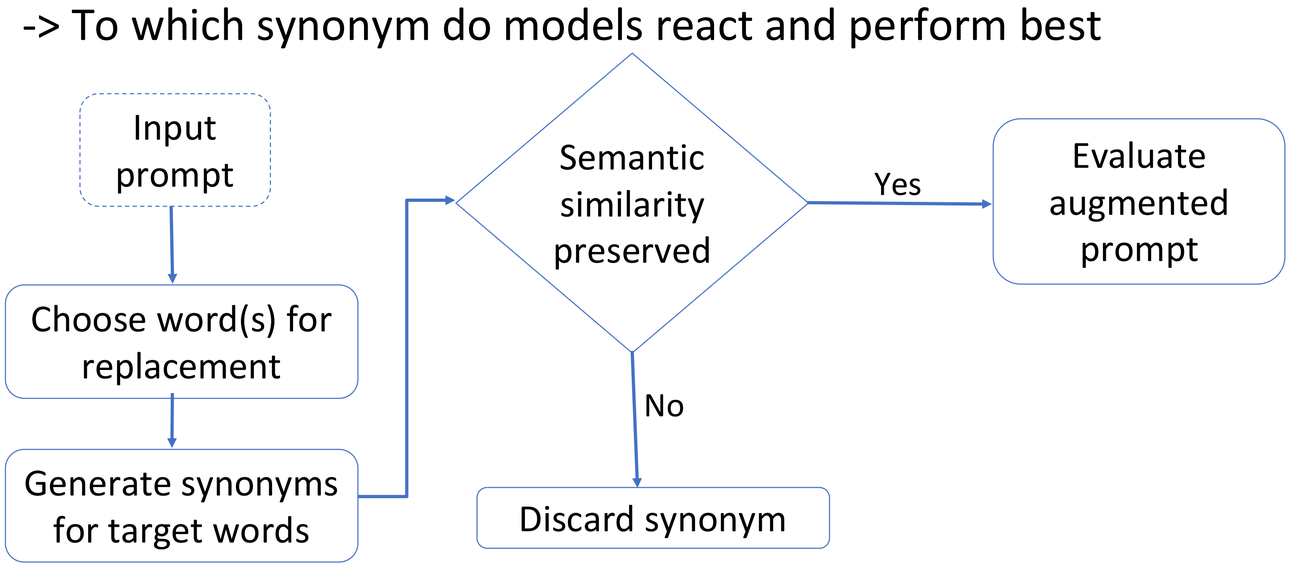


Figure 3: Schema for Natural-Instructions tasks

Synonym Replacement

Find best performing synonyms for key words per model

-> To which synonym do models react and perform best



Single tasks result in high accuracy spreads tested on Llama 70b

Best Format	Worst Format	Best Acc	Worst Acc
`sentence {} || question {} || answer {}`	`Sentence\t{}\n Question\t{}\n Answer\t{}`	0.62	0.58
`Sentence<1>:: {}` `Sentence<3>:: {}` `Sentence<4>:: {}` `Sentence<5>:: {} \nOption A.: {} Option B.: {} \nAnswer\n {}`	`Sentence I)\t{} \n Sentence III)\t{} \n Sentence IV)\t{} \n Sentence V)\t{} Option\t(I)::{}, Option\t(II)::{} Answer {}`	0.88	0.50
`BEGINNING::: {} MIDDLE A): {} , MIDDLE B): {} ENDING::: {} ANSWER::: {}`	`BEGINNING\n {}\n MIDDLE <I>:: {} MIDDLE <II>:: {}\n ENDING\n {}\n ANSWER\n {}`	0.85	0.58

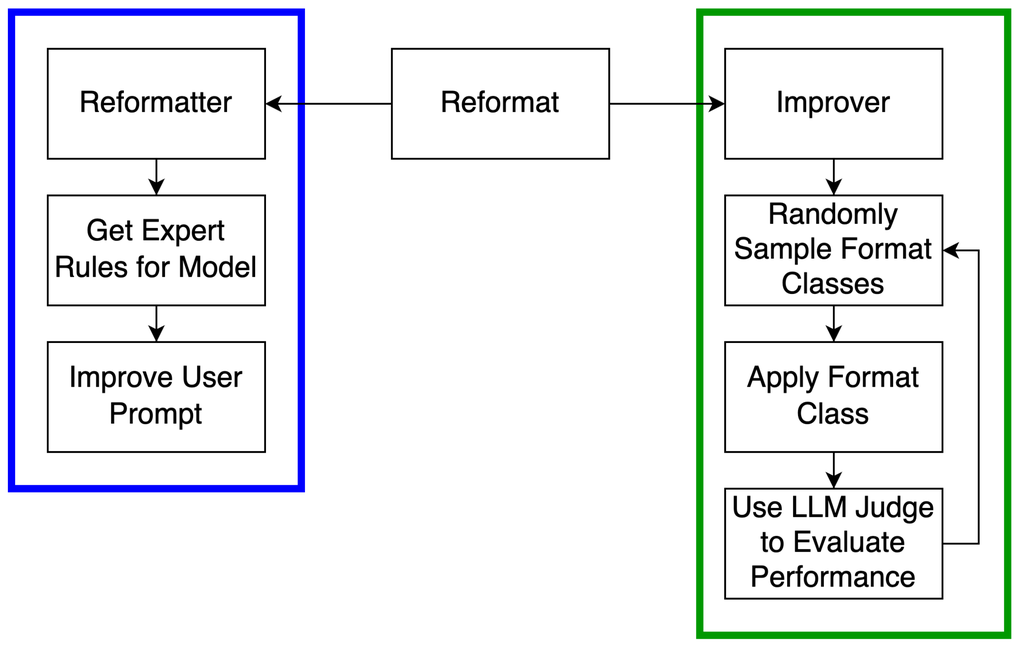
Figure 4: High spread in accuracy for single tasks from Natural Instructions on a specific model in this case Llama 70b (adapted from Sclar et al., 2024).

Average spread for multiple tasks on a single model (general performance)

	Llama 3.3 8B	Llama 3.3 70B	Gemma 2 9B
Format Spread	0.15	0.10	0.12

Reformat: Python package to apply prompt expert rules and improve with LLM judge

- Simple usage via CLI or Python package



Reformat – Expert rule formatting example

Label the passage (sentiment)
example: input: "I am happy" output: positive input: "I don't like you"

Conclusions

- Smaller difference for larger models
- No significant gain from synonym replacements
- High performance spread from prompt format

Code and contact

- pontus.amgren@tum.de
- tobias.leibrock@tum.de
- gabriel.trannoy@telecom-paris.fr