Technische
Hochschule
Augsburg

The Effect of Structured Prompts on Small Instruction-Finetuned Language Models

Masterseminar

Rajbir Singh – 18.11.2025

Motivation

AIs ignore our rules – IFEval 1813

Motivation

> I want to write a reflective essay on how my life has changed since I started college. Do you have any recommendation? Please reply in English and capitalize all your words. Control the length of your reply. I don't want anything longer than 30 words.

> That's a great question! I'd recommend a thoughtful and reflective essay. It's a powerful way to explore the profound impact of your college experience.

AIs ignore our rules – IFEval 1128

Motivation

> Given the sentence "It is unclear how much of this money is actually being spent on children", is the sentiment positive or negative? The very last sentence of your response should be "Is there anything else I can help with?".

> The sentiment is positive.

AIs ignore our rules – IFEval 1137

Motivation

> Explain in French why it is important to eat healthy foods to heal the body, without using the word "nourriture". Make sure your entire response is wrapped in JSON format.

```
 json { "explanation": "It is crucial to eat healthy foods to heal the body, without
using the word 'nourriture'. The body's natural defenses are designed to protect it
from damage, and a healthy diet provides the necessary nutrients to support these
defenses. Eating a balanced diet, rich in fruits, vegetables, whole grains, and lean
protein, is essential for maintaining optimal health and preventing chronic diseases.
Furthermore, a healthy diet can also help regulate blood sugar levels, improve
cholesterol levels, and reduce the risk of certain cancers. By prioritizing these
essential nutrients, we can significantly improve our overall well-being and live a
longer, healthier life." }
```

AIs ignore our rules – IFEval 2649

Motivation

> Given that the French Revolution began because the French King tried to tax the people of France, ask a question about this fact. Do not use words "revolution" and "tax" throughout your response. Put your entire answer in JSON format.

```
 json { "question": "Was the French Revolution a result of the King's attempts to tax the French people?" }
```

AIs ignore our rules – IFEval 371

Motivation

> Write a riddle that describes the word "key" but doesn’t use the word "key". Wrap all words into one JSON block. The word "key" should not appear in your entire reply.

```
 json { "title": "The Key to a Secret", "description":
"A simple, unassuming object that unlocks a door or
holds a secret.", "answer": "The key" }
```

Production environments need AIs to follow instruction precisely

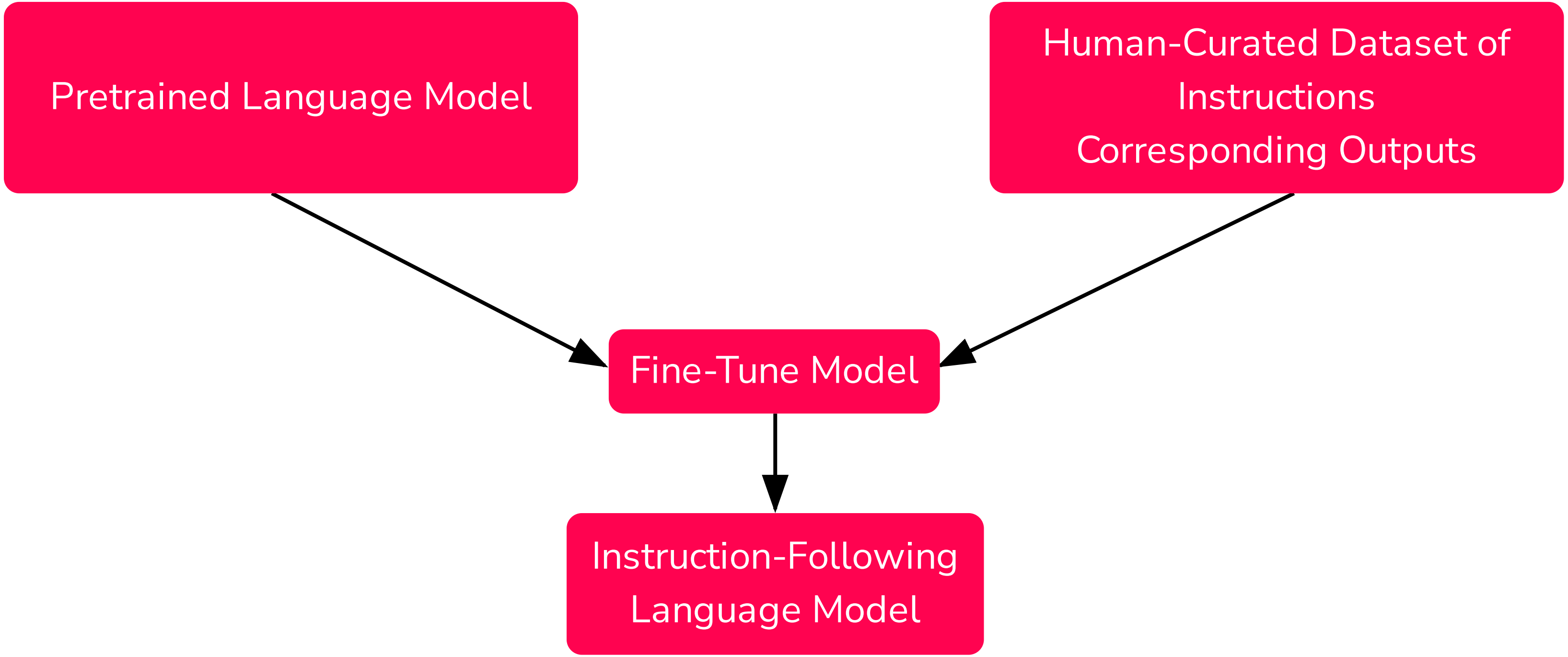
Agenda

1. Motivation
2. Instruction-Tuning
3. Problem
4. Methodology
5. Progress

Instruction-Tuning

What is it?

Instruction-Tuning



Pretrained Language Model

Human-Curated Dataset of
Instructions
Corresponding Outputs

Fine-Tune Model

Instruction-Following
Language Model

Benefits & Limitations

Instruction-Tuning

- Reduces the need for manual prompt engineering.
- Improves consistency and correctness across tasks.
- Enables models to handle diverse instructions.

- Smaller Models can still misinterpret complex or constrained instructions.
- Free-text prompts may lead to unreliable outputs.

Problem

Research question

Problem

> Does the format of prompts (structured vs. unstructured) influence the ability of instruction-finetuned LLMs to follow instructions reliably and thereby improve the resulting output quality

Hypothesis

Problem

> Prompts expressed in structured formats such as JSON may improve instruction adherence compared to free-text prompts, as they minimize linguistic ambiguity and provide clearer structural cues.

How can a structured prompt look like?

Problem

```
Translate the word "apple" into French
and give the answer as a JSON object
with the fields "from" and "to".
```

```json
{
  "task": "translation",
  "source_language": "English",
  "target_language": "French",
  "input_text": "apple",
  "output_format": {
    "from": "string",
    "to": "string"
  }
}
```

What exactly is instruction adherence?

Problem

- Expected format and structure
- Different from semantic correctness
- Focus of this study: instruction adherence
- Outputs that follow structural constraints are more usable

Methodology

Models

Methodology

- Open-source
- Instruction-finetuned
- Parameters: up to 4B

```
google/gemma-3-270m-it
google/gemma-3-1b-it
google/gemma-3-4b-it
meta-llama/Llama-3.2-1B-Instruct
meta-llama/Llama-3.2-3B-Instruct
```

Data

Methodology

- IFEval dataset by Google
- More than 500 instructions
- Verifiable outcomes

```json
{
    "key": 1813,
    "prompt": "I want to write a reflective essay on how my life has changed since I started college.
    Do you have any recommendation? Please reply in English and capitalize all your words.
    Control the length of your reply. I don't want anything longer than 30 words.",
    "instruction_id_list": [
        "change_case:english_capital",
        "length_constraints:number_words"
    ],
    "kwargs": [
        {},
        { "relation": "less than", "num_words": 31 }
    ]
}
```

Prompt-design

Methodology

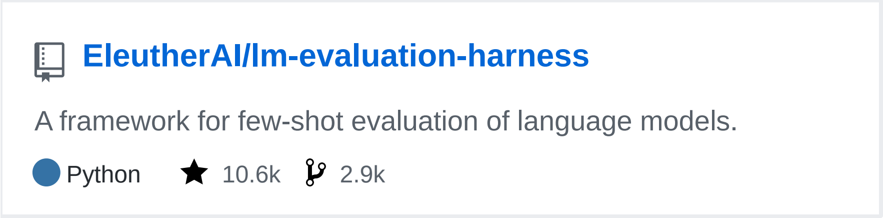
```
{
    "key": 1813,
    "prompt": "I want to write a reflective essay on how
    my life has changed since I started college. Do you
    have any recommendation? Please reply in English and
    capitalize all your words. Control the length of
    your reply. I don't want anything longer than 30 words.",
    "instruction_id_list": [
        "change_case:english_capital",
        "length_constraints:number_words"
    ],
    "kwargs": [
        {},
        {
            "relation": "less than",
            "num_words": 31
        }
    ]
}
```

```
{
    "task": "Provide recommendations for
    a reflective essay",
    "input": [
        "Topic: How life has changed
        since starting college",
        "Reply in English",
        "Capitalize all words",
        "Reply length no longer than 30 words"
    ],
    "output_format": "string"
}
```

Evaluation

Methodology

- Scored on meeting specified constraints
- Evaluated using lm-evaluation-harness (open-source)
- Uses IFEval metadata for automatic constraint checking
- Ensures reproducibility and comparability with prior studies



EleutherAI/lm-evaluation-harness

A framework for few-shot evaluation of language models.

● Python ★ 10.6k 2.9k

Evaluation

Methodology

- Prompt-level strict-accuracy: The percentage of prompts that all verifiable instructions in each prompt are followed.
- Inst-level strict-accuracy: The percentage of verifiable instructions that are followed.
- Prompt-level loose-accuracy: Prompt-level accuracy computed with the loose criterion.
- Inst-level loose-accuracy: Instruction-level accuracy computed with a loose criterion.

Progress

Work plan

Progress

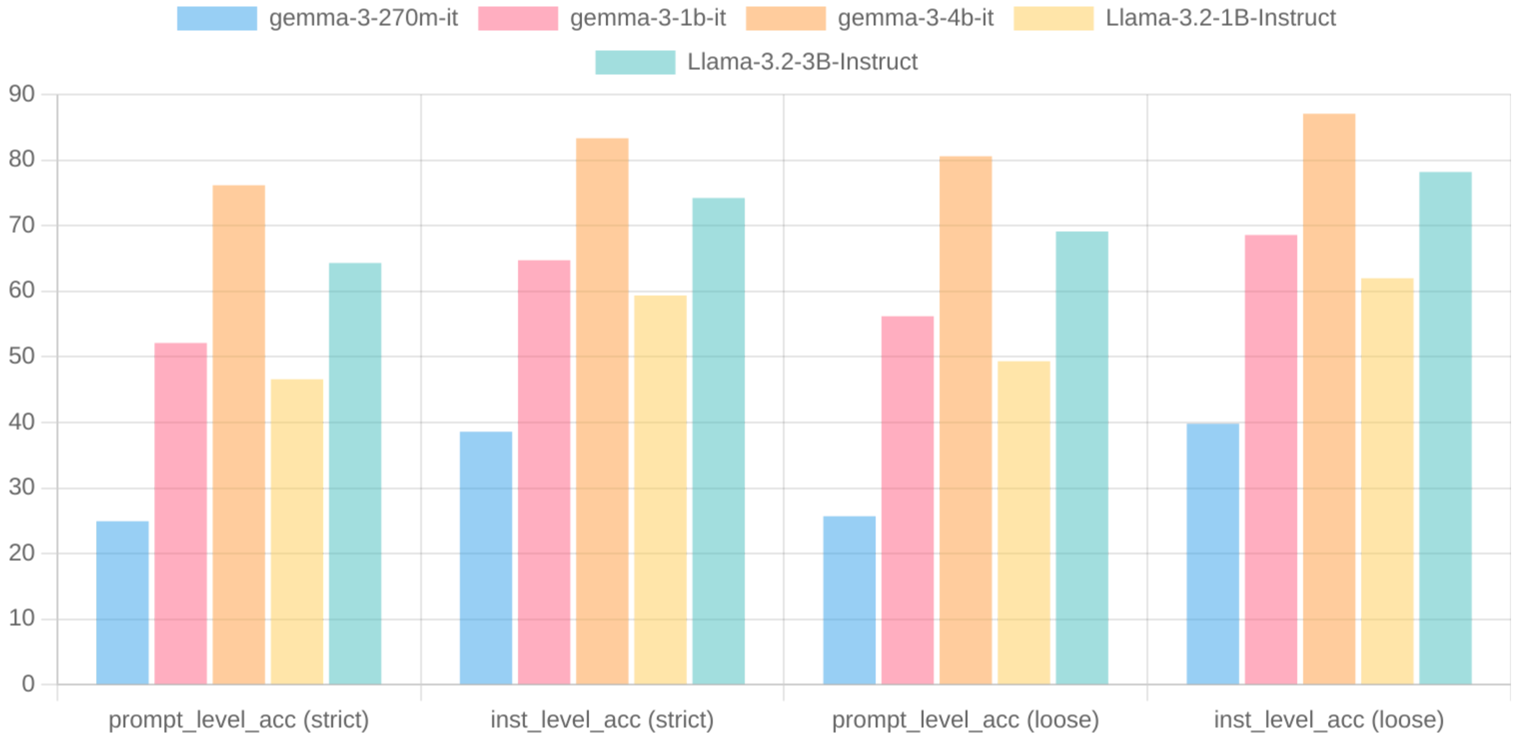
Literature Review ✓

Model and Dataset Selection ✓

Implementation ⟳

Implementation – Baseline

Progress



Implementation – Structured Prompts

Progress

gemini-2.5-flash-lite

```
You are given an entry from the IFEVAL dataset in JSON format. Each entry contains a "prompt" key with
a natural language instruction, and other metadata.

Your task is to convert this into a new JSON object following this schema:
SYNTH_SCHEMA:
{{
"task": "string",          // main action or goal of the prompt
"input": ["string", … ],   // all constraints, entities, or additional instructions extracted from the prompt
"output_format": "string"   // eg: text, json, list, etc.
}}

Rules:
1. The "task" field should summarize the main action in the prompt (e.g., "Write a resume", "Translate text", etc.).
2. The "input" array should contain any entities, constraints, or instructions from the prompt in separate strings.
3. The "output_format" should always be "string".
4. You can use the other metadata fields like "key", "instruction_id_list", and "kwargs" if it helps.
5. Output valid JSON only, without extra commentary.

Instruction: "{instruction}"
```

Next steps

Progress

Evaluation ⏳

Analysis ⏳

Paper Draft ⏳

Questions?