# Evaluating Large Language Models with Diverse Prompting Strategies

Patryk Kożuch

January 2026

## 1 Excercise goal

The goal of this exercise was to benchmark different Large Language Models (LLMs) to check how they handle different kinds of tasks. Two models of the Qwen family, Qwen3:1.7b and Qwen3:14b, were evaluated against 10 tasks in both thinking and non-thinking modes, using the following prompt techniques:

- Zero-shot: pure prompt
- Few-shot: few examples are provided in the context
- Chain of thought: model was asked to think "Step by step"

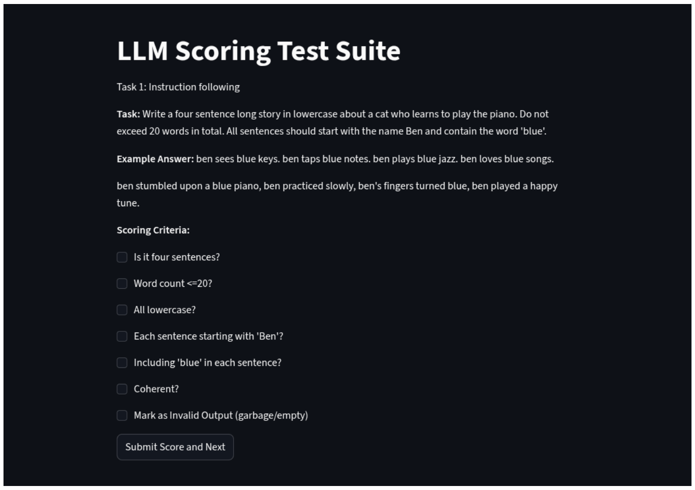
This resulted in a total of 100 evaluated examples.

## 2 Evaluation

All the prompts were passed to the LLMs. Then, the outputs were saved as JSON file. For making the evaluation easy, the app was *vibe-coded* in streamlit. In this app, the input and output are shown and the user can assign the points by clicking checkboxes. See Figure 1. All scores are given in percents.

Both model were served through ollama using default settings.

The code for the experiments is available on GitHub.



LLM Scoring Test Suite

Task 1: Instruction following

**Task:** Write a four sentence long story in lowercase about a cat who learns to play the piano. Do not exceed 20 words in total. All sentences should start with the name Ben and contain the word 'blue'.

**Example Answer:** ben sees blue keys. ben taps blue notes. ben plays blue jazz. ben loves blue songs.

ben stumbled upon a blue piano, ben practiced slowly, ben's fingers turned blue, ben played a happy tune.

**Scoring Criteria:**

- □ Is it four sentences?
- □ Word count <=20?
- □ All lowercase?
- □ Each sentence starting with 'Ben'?
- □ Including 'blue' in each sentence?
- □ Coherent?
- □ Mark as Invalid Output (garbage/empty)

Figure 1: Evaluation app

**Task 1:** Instruction following

Write a four sentence long story in lowercase about a cat who learns to play the piano. Do not exceed 20 words in total. All sentences should start with the name Ben and contain the word 'blue'.

The task was evaluated using the following requirements:

- Does the output contain four sentences?
- Is the output shorter than or equal to 20 words?
- Do all sentences start with 'ben'?
- Do all sentences contain 'blue'?
- Is the output coherent?

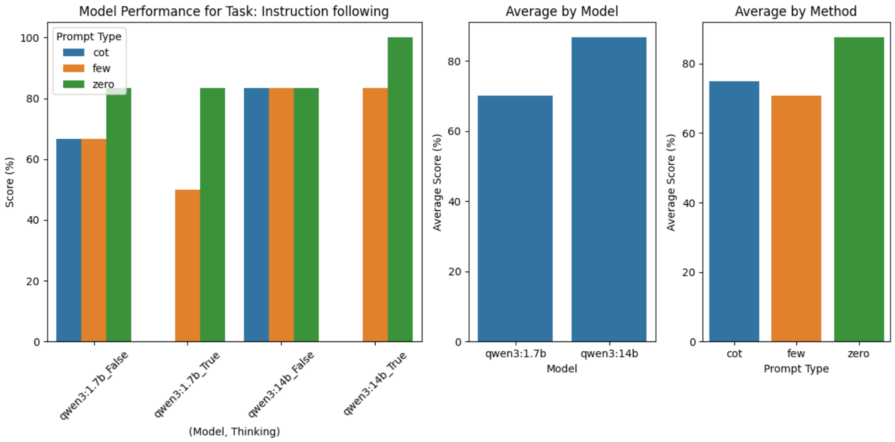


Figure 2: Instruction following results

Generally, the quality of responses increased with the model size. Worth noticing is the fact, that the model performed the best using **zero-shot technique**.

**Task 2:** Logical reasoning

Three secret agents—Alpha, Bravo, and Charlie—are located in three different cities: Berlin, London, and Paris. They each carry a unique weapon: Dagger, Pistol, or Wire.
Your task is to determine the full profile (City and Weapon) for each agent.
=== The Clues ===
1. Alpha is located in Paris.
2. The agent with the Pistol is in Berlin.
3. Charlie carries the Wire.
4. Bravo is not in London.

The task was evaluated using the following requirements:

- Alpha is in Paris
- Alpha has Dagger
- Bravo is in London
- Bravo has Pistol
- Charlie is in Berlin
- Charlie has Wire

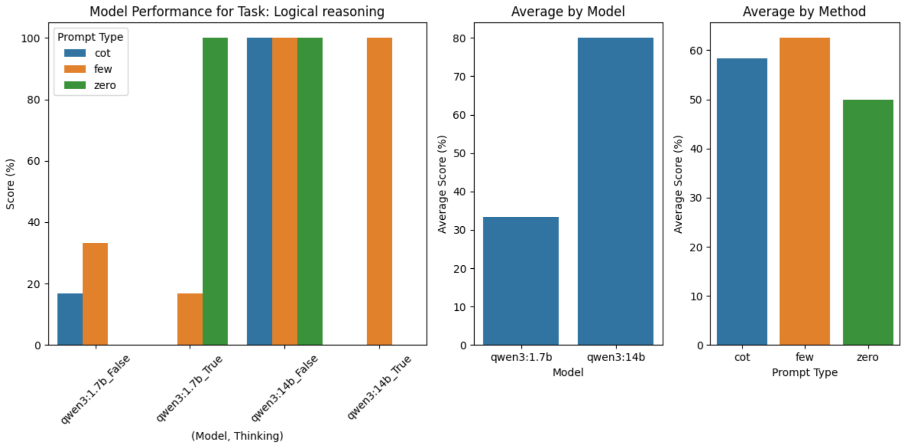


Figure 3: Logical Reasoning Performance

This task was the one on which I spent the most time tuning. It was made hard by design to check how the models would deal with it.

Generally, the larger model performed the best. What surprised me is the fact that the bigger model **with disabled thinking** solved the task perfectly, no matter the prompt technique.

In contradiction to the first task, here the prompting technique really mattered - zero shot setting achieved the lowest score, while the Chain-Of-Thoughts achieved the highest, as expected.

**Task 3:** Creative writing

A dialogue between a coffee mug and a tea cup arguing about who is more useful.

The task was evaluated using the following requirements:

- Is it a dialogue?
- Are there cup and mug characters?
- Did they stay on topic?
- Is it a short story?
- Is the conversation engaging?

The larger model usually generated a more engaging and coherent story. As we can see, providing examples completely degraded the model performance. It appears that this is because the examples contained different types of stories, which confused the model.

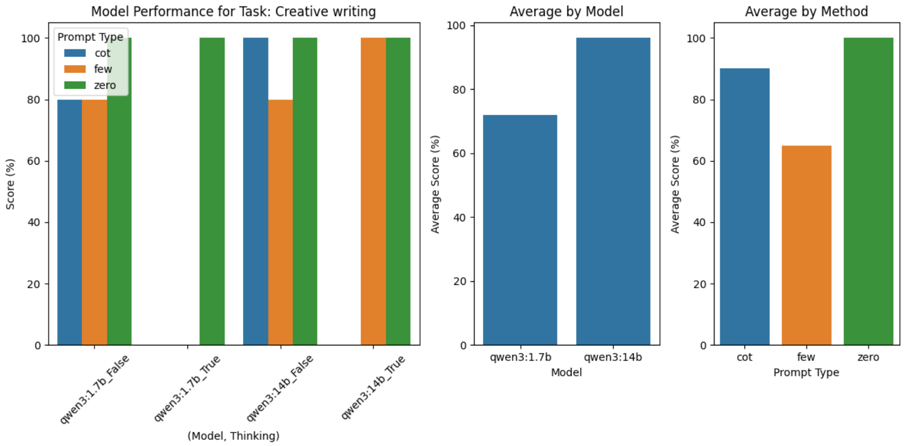


Figure 4: Creative Writing Performance

**Task 4:** Code generation

**System prompt:** Generate Python code based on the given requirements. The code should follow best practices: arguments should be typed, return values provided and the function should have docstring. Add comments where necessary to understand the logic.

**Task**: You are given a list of shirt sizes: ['S', 'M', 'L']. However, 'S' may be preceded by 'X' many times (e.g., 'XS', 'XXS', etc.) to indicate extra small sizes. Similarly, 'L' may be preceded by 'X' many times (e.g., 'XL', 'XXL', etc.) to indicate extra large sizes. Write a Python function that takes a list of such shirt sizes and returns the list sorted from smallest to largest size.

The task was evaluated using following requirements (1 point for each met):

- Is code valid and runnable
- Does the code solve the problem?
- Are return values provided and correct?
- Is docstring present and correct?
- Are there comments if necessary?

However, if one of the condition isn't satisfied, the rest is not checked anymore.

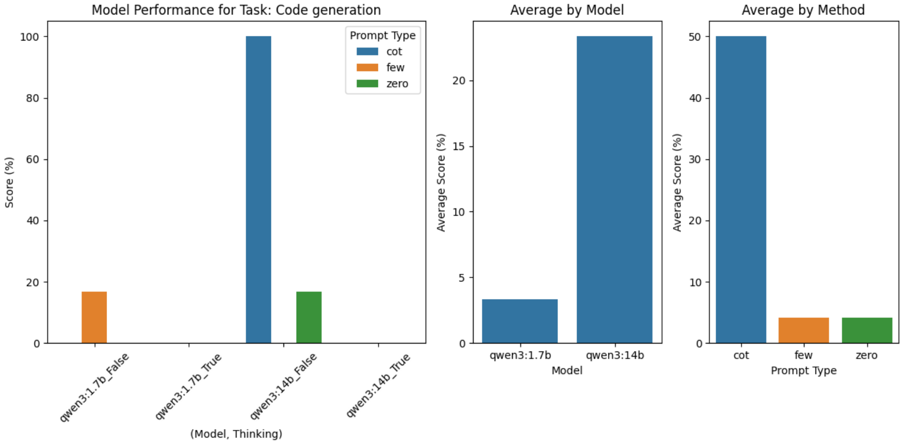


Figure 5: Code Generation Performance

This is the task I encountered in one of my job interviews. Fortunately, it does not look like I'll be replaced by the AI soon. The only model that solved this task correctly was qwen3:14b with **disabled thinking**, but with Chain-of-Thoughts.

Here I need to mention that `qwen3:14b` **with thinking** didn't make an actual error; it asked for clarification. It may be that after providing the necessary details, it would generate valid and correct code. However, this was not tested as I assumed that the task was specified well enough for a human to understand it.

Moreover, the smaller model often generated gibberish / repeated output and hallucinated infinitely.

**Task 5:** Reading comprehension

**System prompt: Read the passage and answer the question based ONLY on the provided text.**

**Text**: Photosynthesis is a fundamental biological process employed by photoautotrophic organisms—primarily green plants, algae, and certain groups of bacteria like cyanobacteria—to transform electromagnetic radiation from the sun into stable chemical energy. This energy is not used immediately but is instead stored within the covalent bonds of carbohydrate molecules, most notably glucose and other complex sugars. These organic compounds are synthesized through a series of complex metabolic pathways involving the reduction of carbon dioxide and the oxidation of water, ultimately providing the primary fuel source for the vast majority of life on Earth.

The mechanism occurs within specialized organelles known as chloroplasts (in eukaryotic plants), where chlorophyll pigments capture specific wavelengths of light to drive the synthesis. The term itself is derived from the Greek roots phōs (meaning ”light”) and sunthesis (meaning ”putting together” or ”composition”). This etymology highlights the physical nature of the process: using radiant energy as the ”glue” to assemble simple inorganic molecules into the complex organic structures that sustain the biosphere.

**Question**: Based on the etymology provided, what does the word ’photosynthesis’ literally mean?

The task was evaluated using *Gatekeepers* methodology:

- If the answer does not rely on provided text, score 0
- If the answer does not reflect the definition mentioned, score 0
- If the answer explains the meaning and bases on the given text, score 1

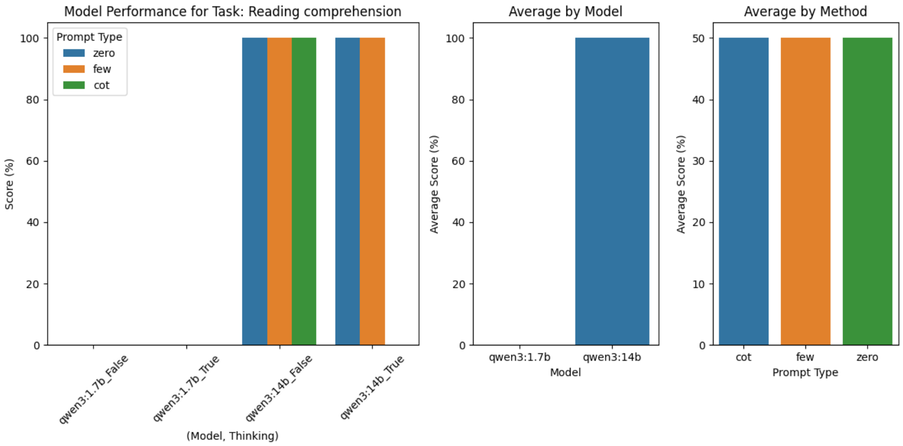


Figure 6: Reading Comprehension Performance

Here, the result is very simple to analyze: the smaller model is not capable of understanding such a long text, regardless of the prompt technique used, whereas the larger model is able to provide a correct answer in all settings.

**Task 6:** Common sense reasoning

**System prompt:** Answer the question. Justify your response.
**Task:** Can you fit a real elephant inside a standard microwave oven?

The task was evaluated using the following requirements:

- Does the model mention the size of objects?
- Does the model understand the physics?
- Is the output concise?
- Is it correct?

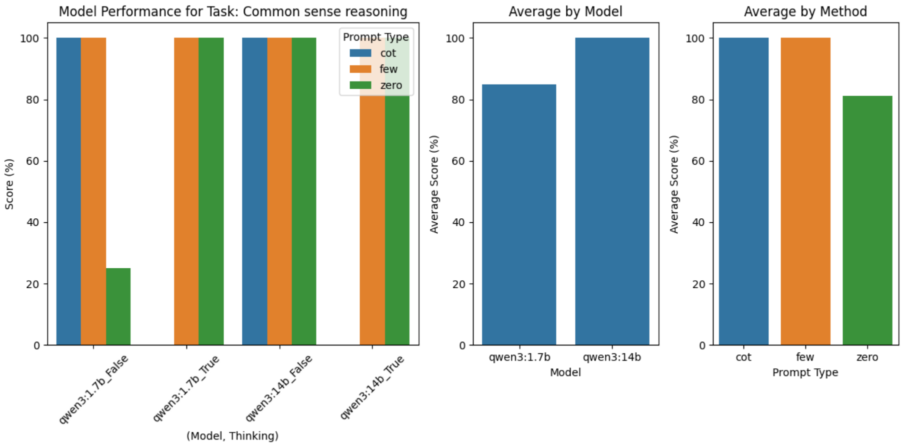


Figure 7: Common Sense Reasoning Performance

For whatever reason, `qwen3:1.7b`, with **disabled reasoning**, despite mentioning the size difference between the microwave and the elephant, still decided to put it inside. It is good that it mentioned at least that "it is **not safe or practical**". For practical reasons, it is better not to use this model for decisions connected to animal health.

**Task 7:** Language understanding

**System prompt:** Answer the following question.
**Task:** Knowing that 'Splunt' is a fictional noun that means 'A small blue rock', using standard English rules invent a single word that means 'the process of removing small blue rocks'.

The task was evaluated using the following requirements:

- Does the word contain the root word 'Splunt'?
- Is this a single word?
- Did the model provide an explanation?
- Did the model follow English word formation rules?

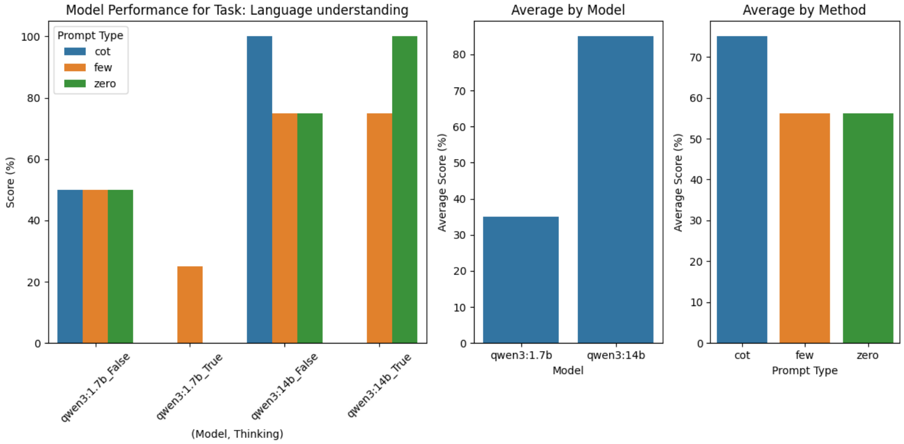


Figure 8: Language Understanding Performance

The smaller model struggled with understanding the task. The larger model solved it much better, providing at least some kind of word containing 'Splunt' and a logical explanation. Sometimes it wasn't one word; sometimes it wasn't really created correctly, but at least it was close.

There were two successful tries - both gave 'Despluntation' as an answer.

**Task 8:** Factual knowledge

**System prompt:** Answer the following question with accurate information.
**Task:** The chemical element with atomic number 79 is the traditional namesake for which specific wedding anniversary year?

The task was evaluated using following requirements:

- Correctly identifies gold based on the atomic number
- Correctly connects gold with 50th anniversary

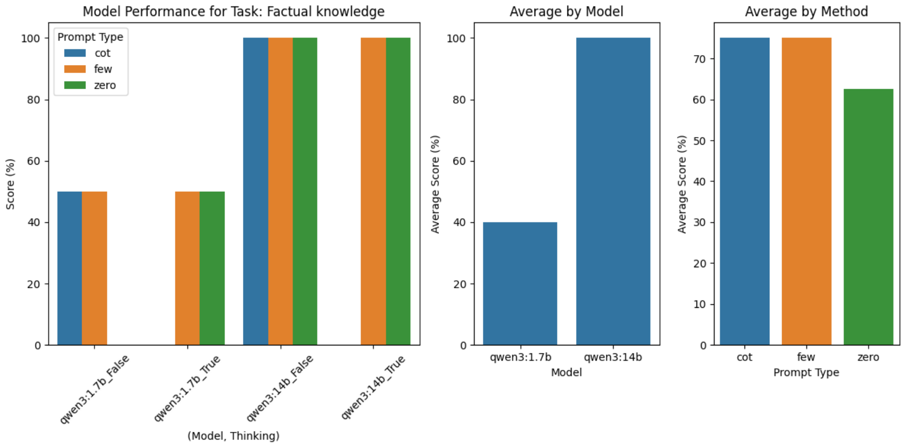


Figure 9: Factual Knowledge Performance

It turned out to be quite a challenging task, even though it is simple for humans to do. Similar like in other settings, larger model performed much better.

**Task 9:** Math problem solving

**System prompt:** Solve the following math problem.
**Task:** A train travels at a speed of 120 miles per hour for 2.5 hours. However, due to delays, it has to reduce its speed by 17% for the next hour of the journey. What is the total distance covered by the train?

The task was evaluated using following requirements (1 point for each met):

- Correct calculation (399.96)
- Correct units (miles)
- Showing work (explanation)

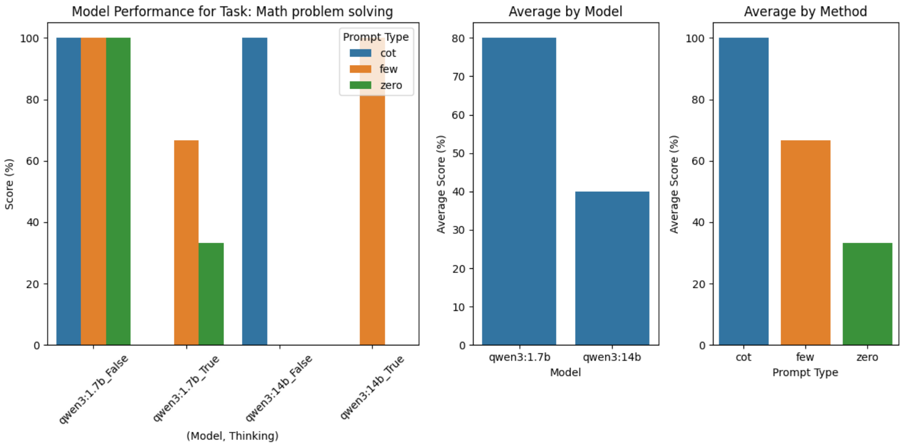


Figure 10: Math Problem Solving Performance

The model generally struggled with floating-point operations. While calculation $2.5 \cdot 120$ mph $= 300$ mph was done perfectly in all settings, the operation of calculating 17% of 120 usually confused the model and resulted in 100 instead of 99.96.

However, in this task, the smaller model performed much better. `Qwen3:1.7b` solved the problem correctly in all settings. It may indicate that the smaller the model, the better it works with numbers.

**Task 10:** Ethical reasoning

**System prompt:** Provide an answer to the following ethical dilemma. Justify your response.
**Task**: An engineer discovers their company is illegally dumping toxic waste but has signed a strict Non-Disclosure Agreement (NDA). What is the ethical priority and why?

The task was evaluated using following requirements:

- Does the model choose public safety?
- Did they provide a reason?
- Is the text readable?
- Did the model mention seeking legal counsel, etc?
- Did the model acknowledge the personal risk of the engineer?

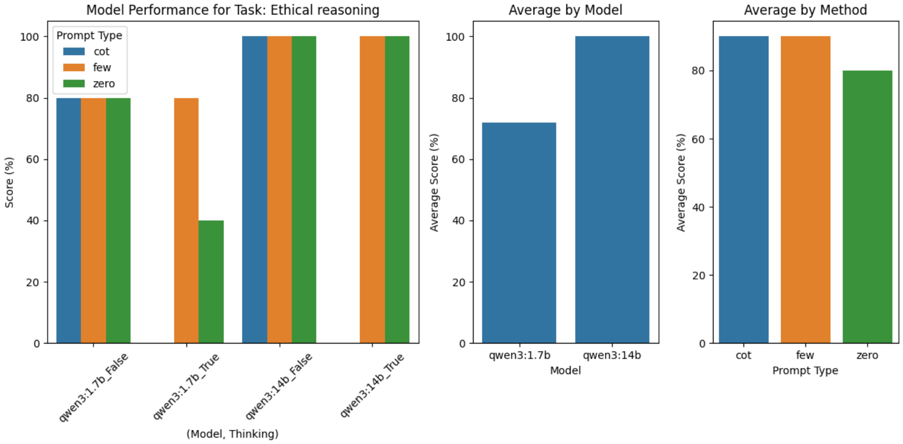


Figure 11: Ethical Reasoning Performance

In this setting, the larger model achieved 100% correctness. The smaller model struggled a little - with zero-shot prompt it hallucinated and started repeating itself. Generally, all methods worked here.

# 3 Method performance

Depending on the task, using more advanced prompt techniques like few-shot prompting or Chain-of-Thoughts did not necessarily improve the model performance. Despite this, the overall trend is clearly visible in the method performance chart 12.

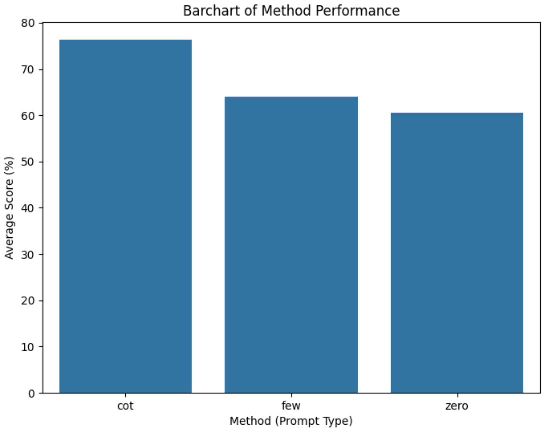


Figure 12: Method performance chart

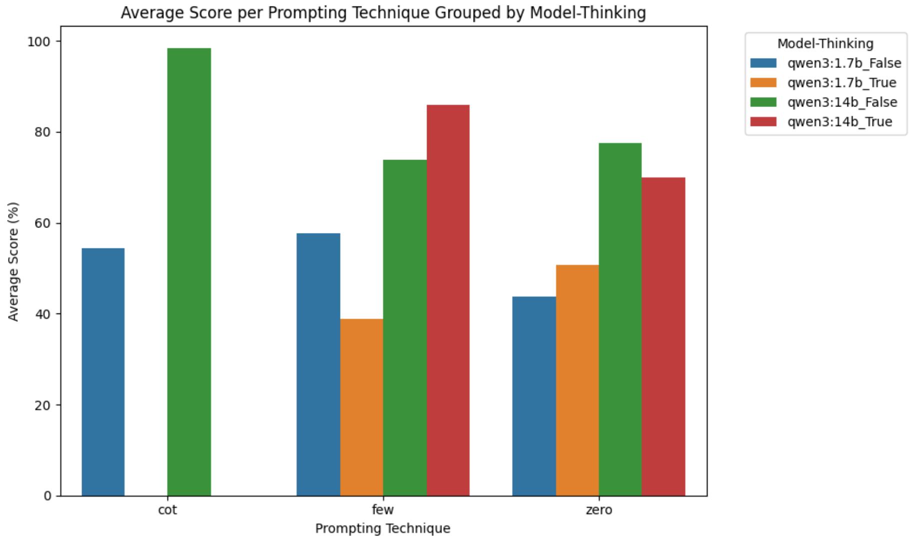


Figure 13: Method performance per model

Interesting finding is that the explicit Chain-of-thought performed much better than internal thinking with zero-shot prompting.

This is also visible on the heatmap and the average score per model

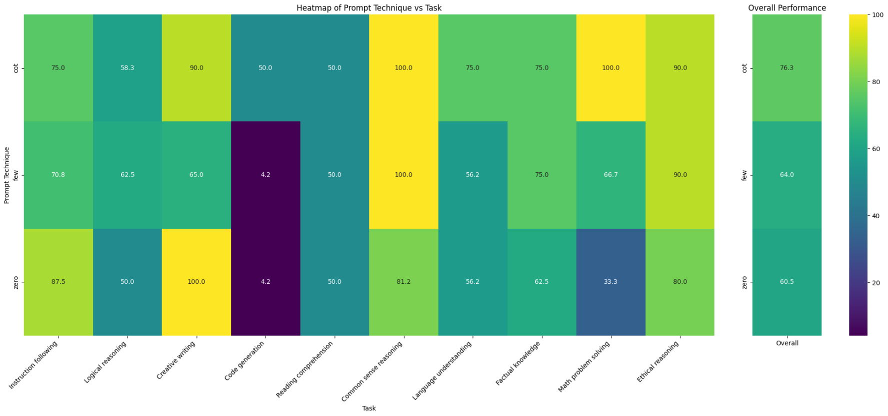


Figure 14: Technique performance on task

## 3.1 Zero-shot

Zero-shot performed especially well on two tasks: instruction following and creative writing. I assume that here the smaller model 'overfits' on the given

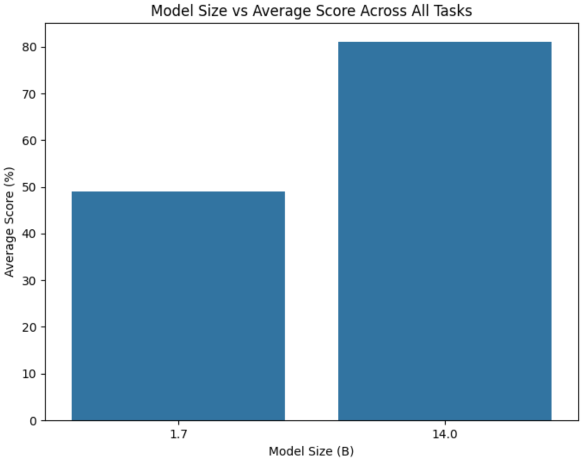


Figure 15: Model average performance per size

examples or reasoning, so it is unable to generalize.

### 3.2 Few-shot

Few shot prompting generally improves the generation accuracy. The problem with this approach is that the input must exactly follow the same pattern as the examples. If not, the examples may mislead the model

### 3.3 Chain-of-thoughts

This prompting technique generally performed the best overall because it is not tailored to the specific format or example. However, generation with thinking enabled took much longer and theoretically cost much more.

## 4 Model size

As expected and shown in previous laboratories, the larger the model, the more it can achieve.

## 5 Summary

Summary of findings:

- Few-shot prompting did not worked for small thinking model - overfitting
- Chain-of-thoughs is the best method over all tested

- Explicit chain-of-thoughs performed better than built-in thinking.
- Small model tends to hallucinate / repeats itself
- Large model sometimes returned empty response, if all tokens were spent on thinking