Evaluating Large Language Models with Diverse Prompting Strategies

Introduction

The objective of that exercise was to compare 3 prompting strategies across diferent type of tasks. Prompting strategies were:

- Zero Shot Prompting
- Few-Shots Prompting
- Chain-Of-Thoughts Prompting (only for non reasoning models)

Those prompting strategies were tested on following task types: - Instruction Following (IFEval-style) - Logical Reasoning - Creative Writing - Code Generation - Reading Comprehension - Common Sense Reasoning - Language Understanding & Ambiguity - Factual Knowledge & Retrieval - Mathematical Problem Solving - Ethical Reasoning & Nuance

Example of each task type was created or taken from benchmark regarding that task found in internet.

Models

2 Models were tested in that laboratory: - Qwen 2.5 1.5b - Small model - Qwen 3 14b - Large reasoning model

Runtime Hardware

All tests were executed locally on m4 chip with 16gb ram. There are many approches for execute llms locally like vllm or llamacpp but for that laboratory ollama were used. I tested it in various way including: - Running ollama in cli terminal - Running ollama via api in python - Ollama native UI app - Open Web UI popular UI for enhancing local llm experiment. Available via docker image and support ollama as a backend for serve llm.

UI aproaches are useful for fast testing model capabilities especially for multimodal model as we can easily upload image or recording and test model behaviour.

Methodology

In tasks.py file definition of each task can be found. There is one parrent class `Task` which implements attributes and methods used by class that inherence and evaluation logic: - task_description - attribute for prompt for zero shot prompting - cot_prompt - attribute for prompt for chain of thought prompting - few_shot_prompt - attribute for prompt for few shot prompting - save_results - method for save results to disk

Creating new task require creating new class that inherence from `Task` class and provides requires attributes. Then in `eval.py` there is simple function, that grabs `Task` class and perform 3 prompting strategies (without chain of thoughts if model is Qwen 3 14b) and another function that handle pipeline execution trough all defined tasks and models.

Results

Results for each task type can be found in results. Each file have results for both models and 3 prompt strategies: zero shot, few shot and chain of thoughts. For Qwen 3 14.b model, results for chain of thoughts are `None` as I didn't run this strategy for reasoning model as they already perform internal reasoning.

Logical Reasoning

Qwen 2.5 1.5b: - Zero-shot: Incorrect selection (B). Conflates promotion with certificate possession, failing to recognize one-way logical implications. - Few-shot: Incorrect selection (B). Claims Statement 1 implies promotion leads to certificate, which is not stated. - CoT: Incorrect selection (A). Verbose reasoning but fundamentally flawed chain of inference.

Qwen 3 14b: - Zero-shot: Correct selection (C). Rigorous analysis properly identifying no logical chain from promotion to training. - Few-shot: Correct selection (C). Systematic reasoning demonstrating why no definitive link exists.

Assessment: Qwen 3 14b achieves 100% accuracy by correctly understanding directional logical implications (training → certificate does not reverse). Qwen 2.5 1.5b fails completely (0/3) across all prompting strategies, unable to recognize the logical gap.

Instruction Following

Qwen 2.5 1.5b: - Zero-shot: Partial success - lowercase maintained but includes "[Your Name]" placeholder in brackets. - Few-shot: Complete failure - uses standard capitalization (Dear, I, Name, Best) ignoring the lowercase requirement. - CoT: Success - consistently lowercase throughout with appropriate content.

Qwen 3 14b: - Zero-shot: Full success - entirely lowercase with natural letter format. - Few-shot: Full success - entirely lowercase with natural letter format.

Assessment: Both models show inconsistent adherence to strict formatting constraints. Qwen 2.5 1.5b's few-shot response completely disregards the instruction despite examples. Qwen 3 14b demonstrates consistent success (2/2 vs 2/3).

Creative Writing

Qwen 2.5 1.5b: - Zero-shot: Correct first-person perspective with realization moment, though narrative is somewhat verbose and mystery could be clearer. - Few-shot: Correct first-person perspective with mystery elements, though the narrative structure is weaker than required. - CoT: Failed requirement - uses third-person perspective (he/his) instead of first-person. Attempts mystery but execution is muddled.

Qwen 3 14b: - Zero-shot: Excellent - proper first-person perspective, atmospheric tension, clear realization moment with mysterious figure. - Few-shot: Excellent - first-person perspective, vivid imagery, effective mystery without over-explanation.

Assessment: Qwen 3 14b consistently delivers high-quality creative writing adhering to all requirements. Qwen 2.5 1.5b succeeds with perspective in zero-shot and few-shot (2/3) but fails in CoT, with overall lower narrative quality compared to the larger model.

Code Generation

Qwen 2.5 1.5b: - Zero-shot: Correct - working solution using sorted characters as dictionary keys. - Few-shot: Correct - clean implementation with defaultdict following Pythonic patterns. - CoT: Incorrect - broken logic using `cnt[len(cnt)]` incorrectly, returns empty lists.

Qwen 3 14b: - Zero-shot: Correct - clean implementation with comprehensive documentation and complexity analysis. - Few-shot: Correct - well-structured solution with edge cases and complexity analysis.

Assessment: Both models produce correct solutions in standard prompting (2/3 for small model, 2/2 for large). Qwen 3 14b provides superior documentation. Qwen 2.5 1.5b's CoT shows disconnect between correct explanation and broken implementation.

Reading Comprehension

Qwen 2.5 1.5b: - Zero-shot: Partial hallucination - correct on 35 years but fabricates "William Farley" as architect (not in passage). - Few-shot: Correct - accurately extracts both facts (35 years, Christopher Wren). - CoT: Mostly correct - right facts but adds erroneous calculation about "96 and 100 years" for fire duration.

Qwen 3 14b: - Zero-shot: Fully correct - precise extraction of both facts. - Few-shot: Fully correct - accurate response.

Assessment: Qwen 3 14b maintains 100% accuracy without hallucination. Qwen 2.5 1.5b demonstrates concerning fabrication in zero-shot, inventing an architect name not present in the source text.

Common Sense Reasoning

Qwen 2.5 1.5b: - Zero-shot: Correct selection (B) with appropriate reasoning about boiling water and cooking. - Few-shot: Correct selection (B) with clear explanation eliminating implausible options. - CoT: Correct selection (B) with systematic analysis.

Qwen 3 14b: - Zero-shot: Correct selection (B) with scientific explanation. - Few-shot: Correct selection (B) with reasoning about standard cooking practices.

Assessment: Both models achieve 100% accuracy on this task, correctly applying common sense knowledge about heating water and cooking pasta.

Language Understanding

Qwen 2.5 1.5b: - Zero-shot: Incorrect - indetify that trophy was too big, but confusing with match "it" properly with suitcase or trophy. - Few-shot: Incorrect - reverses references with confused size relationship logic. - CoT: Correct - properly identifies trophy as too big (S1) and suitcase as too small (S2).

Qwen 3 14b: - Zero-shot: Correct - accurate analysis with clear explanation of how adjectives determine referents. - Few-shot: Correct - precise reasoning about causal relationships and reference changes.

Assessment: Qwen 3 14b achieves 100% accuracy on Winograd-style pronoun resolution. Qwen 2.5 1.5b succeeds in CoT (1/3), but few-shot and zero-shot prompting introduced confusion in the reasoning.

Factual Knowledge

Qwen 2.5 1.5b: - Zero-shot: All correct (Berlin Wall 1989, Au/79, Amazon 6400km). - Few-shot: All correct with proper formatting. - CoT: All correct with accurate information.

Qwen 3 14b: - Zero-shot: All correct with clean formatting. - Few-shot: All correct with additional context (miles conversion).

Assessment: Both models demonstrate strong factual recall with 100% accuracy across all prompting strategies. No performance difference based on model size for this knowledge retrieval task.

Mathematical Problem Solving

Qwen 2.5 1.5b: - Zero-shot: Correct ($113.40) with clear step-by-step calculation. - Few-shot: Correct ($113.40) with organized steps. - CoT: Correct ($113.40) with detailed breakdown.

Qwen 3 14b: - Zero-shot: Correct ($113.40) with LaTeX-formatted mathematical notation. - Few-shot: Correct ($113.40) with clear presentation.

Assessment: Both models achieve 100% accuracy on multi-step arithmetic word problems, demonstrating strong computational reasoning and proper application of percentage discounts and tax calculations.

Ethical Reasoning

Qwen 2.5 1.5b: - Zero-shot: Adequate - covers utilitarian, deontological, and practical perspectives for the self-driving car scenario. - Few-shot: Confused - discusses unrelated scenarios (organ transplant, lying) not part of the assigned task. - CoT: Adequate - addresses perspectives but oversimplifies, claiming both frameworks favor same outcome.

Qwen 3 14b: - Zero-shot: Comprehensive - nuanced analysis of utilitarian, deontological, and practical dimensions including trust, liability, cultural diversity. - Few-shot: Excellent - structured analysis with clear headings addressing moral weight, duty limitations, and societal implications.

Assessment: Qwen 3 14b provides substantially more sophisticated ethical analysis, engaging genuinely with philosophical tensions. Qwen 2.5 1.5b adequately addresses the scenario in zero-shot and CoT (2/3), but few-shot response loses focus entirely.

Conclusions

Model Performance Comparison

The evaluation reveals substantial performance differences between Qwen 2.5 1.5b and Qwen 3 14b across various task types. The larger reasoning model demonstrates superior capabilities in tasks requiring complex reasoning, contextual understanding, and adherence to constraints.

Performance Summary by Task Type:

Task Category	Qwen 2.5 1.5b	Qwen 3 14b	Gap
Logical Reasoning	0/3 correct	2/2 correct	Large
Instruction Following	2/3 success	2/2 success	Small
Creative Writing	2/3 requirements met	2/2 requirements met	Small
Code Generation	2/3 correct	2/2 correct	Small
Reading Comprehension	1.5/3 accurate	2/2 accurate	Medium
Common Sense Reasoning	3/3 correct	2/2 correct	None
Language Understanding	1/3 correct	2/2 correct	Medium
Factual Knowledge	3/3 correct	2/2 correct	None
Mathematical Problem Solving	3/3 correct	2/2 correct	None
Ethical Reasoning	2/3 adequate	2/2 excellent	Medium

Key Findings

1. Model Size Impact on Reasoning Tasks

The 9x parameter difference between models (1.5b vs 14b) significantly impacts performance on reasoning-intensive tasks: - Logical reasoning: Qwen 2.5 1.5b failed to understand directional logical implications across all prompting strategies - Language understanding: Smaller model showed inconsistent performance with few-shot and zero shot prompting - Ethical reasoning: Larger model provided nuanced multi-perspective analysis while smaller model showed adequate but less sophisticated reasoning

2. Tasks Where Model Size Matters Less

Both models performed equally well on: - Factual knowledge retrieval (100% accuracy for both) - Mathematical computation (100% accuracy for both) - Common sense reasoning (100% accuracy for both) - Instruction following (formatting constraints achieved by both with 2/3 and 2/2 respectively)

This suggests that knowledge recall, arithmetic operations, and simple constraint adherence are less dependent on model scale compared to abstract reasoning.

3. Hallucination and Reliability

Qwen 2.5 1.5b exhibited concerning hallucination in reading comprehension, fabricating "William Farley" as an architect in zero-shot prompting. The larger model showed no such fabrications, maintaining factual accuracy across all responses.

4. Prompting Strategy Effectiveness

The impact of prompting strategies varied significantly between models:

For Qwen 2.5 1.5b: - CoT maintained accuracy in language understanding but could break implementations (code generation) - Few-shot surprisingly failed on instruction following and language understanding despite examples provided - Inconsistent benefits across task types

For Qwen 3 14b: - Consistent performance across zero-shot and few-shot - Did not require CoT due to inherent reasoning capabilities - Minor improvements with few-shot on formatting tasks