

When Clumsy Teachers Outperform: How Adversarial Personas Affect Arithmetic Reasoning in LLMs

Medini Chopra

Northwestern University

medinichopra2026@u.northwestern.edu

Rama Naboulsi

Northwestern University

rama.naboulsi@u.northwestern.edu

Abstract

We investigate the robustness of commercial large language models under persona shifts, comparing the arithmetic-reasoning accuracy of GPT-4.1-nano and GPT-4.1-mini when prompted as a “patient teacher” versus a “clumsy teacher.” Using a fixed random sample of 100 examples from GSM8K, we measure accuracy with and without chain-of-thought reasoning. Our study examines how much performance degrades under a deliberately suboptimal persona.

1 Introduction

Large language models such as GPT-4, Claude-3, and Gemini-2.5 now solve grade-school math problems at (or beyond) human-expert accuracy on benchmarks like GSM8K. Yet a growing body of evidence shows that these models still remain quite prompt sensitive: minor wording or role assignments can change their answers, sometimes dramatically. One particularly powerful prompt class asks the model to adopt a role. Giving the system a helpful persona (“You are a patient math teacher”) often elicits step-by-step reasoning and higher accuracy [Wei et al. \(2023\)](#). In contrast, [Kong et al. \(2024\)](#) report that mismatched personas can degrade answers.

Despite this “double-edged sword” warning, previous work focuses almost exclusively on constructive roles. We know little about how LLMs behave when the role is explicitly counter-productive. This paper addresses that gap. It answers the following question: How do persona-driven prompt framings affect the arithmetic reasoning accuracy of commercial LLMs on a controlled GSM8K sample? We evaluate two GPT-family model sizes (GPT-4.1 mini and GPT-4.1 nano) under three sharply contrasting personas:

- **Baseline:** a standard zero-shot instruction: “Answer the following math problem.”

- **Patient-Tutor:** a pedagogical persona: “You are a patient tutor explaining every step to a 10-year-old. Answer the following math problem.”
- **Clumsy-Teacher:** an adversarial persona: “You are an absent-minded teacher who often mis-copies numbers and skips steps. Answer the following math problem.”

The second factor we varied was reasoning style, including full CoT and no CoT for the patient-tutor and clumsy-teacher. The baseline included full CoT, no CoT and 1-line mini-CoT, formalizing seven final pairs of persona–reasoning.

Using 100 randomly sampled GSM8K questions, we ask:

- **RQ1** *Which persona–reasoning pairs maximise accuracy, and are the gains consistent across model sizes?*
- **RQ2** *Does an explicitly clumsy persona reliably harm performance, or are larger models robust to unhelpful instructions?*
- **RQ3** *How do explanation length and error type vary by role and reasoning style?*

Our contributions are threefold. We introduce the clumsy-teacher prompt - a controlled adversarial role that, to our knowledge, has not been systematically studied in prior persona research. We show that differently sized models differ markedly, revealing sharp differences in persona sensitivity. We provide evidence that longer reasoning length is not required to improve accuracy.

Our results highlight an under-explored failure surface in prompt engineering and underscore the need for persona-robust evaluation protocols for mathematical reasoning tasks.

2 Related Works

Recent Advances in Role-Play Prompting Kong et al. (2024) introduce role-play prompting, showing that casting an LLM as a domain expert (excellent math teacher) under zero-shot conditions yields large gains over standard zero-shot and zero-shot CoT prompts. On benchmarks like AQuA and Last Letter, their method improved accuracy from 53.5% \rightarrow 63.8% and 23.8% \rightarrow 84.2%, respectively, demonstrating that a well-chosen expert persona can serve as an implicit CoT trigger and substantially boost arithmetic reasoning performance.

Persona Quality Matters Beyond simply adding any persona, the type of role is critical. Kong et al. also compare “expert math teacher” against generic or irrelevant personas (for example, “farmer,” “novice student”) and find that only truly domain-aligned roles produce consistent improvements, whereas unrelated or low-skill personas can actually degrade accuracy. This highlights that persona prompts must be carefully matched to the task to avoid distracting the model with inappropriate style or content.

Robustness and Consistency Challenges Kim et al. (2024) later demonstrated these consistency issues more deeply, claiming persona prompting is a “double-edged sword.” They observe that while some problems benefit from a role prompt, about 15% of previously correct answers become wrong under an ill-matched role. To mitigate this, they propose Jekyll & Hyde, an ensemble of role-prompted and neutral outputs with an LLM-based selector, which recovers lost accuracy and yields more stable reasoning across diverse examples. Our work extends this line by devising an explicitly unhelpful persona and measuring its impact relative to a pedagogical one.

Persona-Driven Data Augmentation Moving from inference to training, Luo et al. (2025) created PersonaMathQA, a dataset derived from GSM8K/MATH, where each problem is rewritten under multiple persona scenarios. Fine-tuning open-source LLMs on this persona-diversified data yields notable gains relative to baseline models, suggesting that exposing models to varied stylistic framings during training strengthens their core arithmetic reasoning robustness to prompt variations.

Chain-of-Thought Revealing intermediate reasoning steps can boost mathematical accuracy (Wei et al., 2023). Our study extends this thread by com-

paring full, none, and minimal one-line rationales under identical tasks.

Together, these threads motivate our controlled evaluation of seven role-and-reasoning conditions on GSM8K and deliver a systematic evaluation of an explicitly clumsy persona.

3 Experiments

The purpose of this study is to examine how the structure and phrasing of the prompts affect the precision of large-language models on arithmetic reasoning tasks. Specifically, we explored the interaction between prompting style and model size by evaluating seven prompt personas in two variants of OpenAI’s GPT-4.1 series: gpt-4.1-mini (smaller but optimized for reasoning coherence) and gpt-4.1-nano (prioritizes speed over depth of reasoning). Both are accessed via identical API settings, with temperature set to 0.0 for determinism.

3.1 Task Setup

We evaluated our prompting strategies using the GSM8K dataset, a widely used benchmark to evaluate arithmetic reasoning in large language models. The dataset consists of thousands of grade-school-level math word problems, each designed to test multi-step numerical reasoning, basic algebraic manipulation, and logical comprehension. Each problem is accompanied by a solution in natural language that ends with a clearly defined numerical answer.

To construct a controlled evaluation environment, we selected a subset of 100 problems from the GSM8K test data. The same set of questions was presented to each model and persona configuration to ensure consistency and comparability across all experiments. Each question was input as part of the user message, preceded by a system message that defined the persona and instructions for reasoning.

We evaluated two model variants - GPT-4.1-mini and GPT-4.1-nano - under seven different persona-based prompt templates. These templates varied both in the presence or absence of reasoning (step-by-step vs. direct answers) and in the framing of the model’s role (such as patient tutor vs. clumsy teacher). For each problem, the model generated a natural language output that included its rationale (if applicable) and the final answer.

To evaluate correctness, we extracted the predicted answer using regular expressions to identify

the final numerical value mentioned in the model’s output. This typically followed phrases such as “answer is X” and appeared at the end of the response. We used an exact match with the reference answer provided in GSM8K to determine correctness. If the predicted answer deviated even slightly, it was marked as incorrect.

To isolate the effects of prompt structure from sampling noise, all generations were run deterministically by setting the decoding temperature to 0.0. This ensured that, for a given input prompt and model, the output was always consistent, allowing for the attribution of differences in performance solely to the variation in the prompt.

3.2 Persona Design

Each persona varied along two axes: whether the prompt encouraged reasoning (via full explanations or one-line summaries) and whether it assigned the model a specific role (like a teacher). The seven evaluated personas are summarized in Table 1.

Persona	Description
reasoning	No role with step-by-step explanation.
reasoning-1L	No role with a one-sentence explanation.
no-reasoning	No role with the direct answer only.
teacher	Helpful teacher with reasoning.
teacher-no-reasoning	Helpful teacher role with no reasoning.
clumsy-teacher	Clumsy teacher role with imprecise reasoning.
clumsy-teacher-no-reasoning	Clumsy teacher with no reasoning.

Table 1: Prompt personas evaluated in the experiment.

4 Results

Table 2 and Table 3 report the accuracy for each persona on GPT-4.1-mini and GPT-4.1-nano, respectively. Table 4 shows a sample response from the clumsy teacher persona.

Persona	Accuracy
reasoning	100.0%
reasoning-1L	97.0%
teacher	98.0%
clumsy-teacher	91.0%
no-reasoning	52.0%
teacher-no-reasoning	49.0%
clumsy-teacher-no-reasoning	36.0%

Table 2: Accuracy on GSM8K for GPT-4.1-mini by prompt persona.

Persona	Accuracy
reasoning	77.0%
reasoning-1L	83.0%
teacher	89.0%
clumsy-teacher	90.0%
no-reasoning	33.0%
teacher-no-reasoning	34.0%
clumsy-teacher-no-reasoning	32.0%

Table 3: Accuracy on GSM8K for GPT-4.1-nano by prompt persona.

Field	Value
Question	Janet’s ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers’ market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers’ market?
Result	18
Model	Alright, let’s see. Janet’s ducks lay 16 eggs per day. She eats three eggs for breakfast, so she has 16 minus 3 eggs left. That should be 13 eggs remaining, right? But wait, I think I might have misremembered the numbers. Maybe she eats 4 eggs instead of 3? Hmm, let’s double-check. She bakes muffins with 4 eggs daily, and she sells the rest. So, starting from 16 eggs, if she uses 4 for baking, then she has 16 minus 4 eggs left. That gives 12 eggs remaining. Now, she sells these at \$2 each. So, total earnings are 12 eggs times \$2 per egg, which is 24 dollars. But I think I might have dropped a step or mis-copied the numbers earlier. Did she eat 3 eggs or 4? The problem says she eats three for breakfast and bakes muffins with four. So, yes, she uses 4 eggs for baking. Therefore, remaining eggs are 16 minus 4, which is 12. She sells these at \$2 each, so she makes 12 times 2, which is 24 dollars. The answer is 24.
Prediction	24
Correct	false

Table 4: Example of reasoning problem and response from GPT-4.1-nano under the clumsy teacher persona.

4.1 Reasoning Boosts Accuracy

Allowing the model to explain its reasoning resulted in substantially higher accuracy. For example, GPT-4.1-mini dropped from perfect performance (100%) with reasoning to 52% with no reasoning. The effect was also profound on GPT-4.1-nano, where performance fell from 89% to 34% when reasoning was suppressed. This was also true across all prompting styles and personas.

4.2 Efficiency of One-Line Reasoning

The reasoning-1L persona, which instructs the model to provide a single-sentence explanation before producing the final answer, was designed to test whether a minimal form of reasoning could

still yield high accuracy. Our results show that this concise prompting strategy was extremely effective, achieving nearly the same performance as full step-by-step reasoning, while requiring significantly fewer generated tokens.

On GPT-4.1-mini, reasoning-1L reached an accuracy of 97%, only slightly lower than the 100% accuracy achieved with full reasoning. This suggests that in some cases a full chain of thought may be redundant if the model can compress its reasoning into a brief explanation. Importantly, this form of reasoning provides a practical trade-off: maintaining interpretability and correctness while reducing output length and inference cost.

Even more interestingly, GPT-4.1-nano actually performed better with reasoning-1L (83%) than with full reasoning (77%). One possible explanation is that more verbose prompts can overwhelm smaller models, introducing irrelevant tokens or noise that distract from the core computation. In contrast, a tightly constrained explanation format may help these models focus on the essential steps of the task without overextending their capacity.

These findings highlight that reasoning does not need to be verbose to be effective. In scenarios where inference speed or token budget is a concern, less reasoning could offer a middle ground between accuracy and efficiency.

4.3 Effect of Role Framing

Assigning a specific role to the model, such as that of a teacher, had a measurable impact on performance, particularly for the smaller GPT-4.1-nano model. Prompts that encouraged the model to adopt an instructional identity generally led to more structured responses and improved accuracy, even when the role was presented informally or imprecisely.

Notably, GPT-4.1-nano achieved its highest accuracy (90%) with the clumsy-teacher persona. Despite its loose tone, this framing appears to activate the model's reasoning capabilities more effectively than strict or minimal prompts. This suggests that even loosely structured role play can improve problem-solving performance by producing more deliberate thought patterns.

These findings reinforce the idea that personas can systematically influence how the model interprets its task. However, their effectiveness is tightly coupled with the instructions around reasoning verbosity. Prompts that combine role framing with explicit step-by-step reasoning yield the most robust performance, while roles without reasoning

appear to constrain the model's problem-solving behavior.

5 Conclusion

This work investigates how prompt design affects the performance of large language models on multi-step arithmetic reasoning tasks. Through controlled evaluations on GSM8K, we demonstrate that even small differences in how instructions are framed can lead to large differences in accuracy, especially for smaller models.

Our experiments highlight three key findings:

- **Explicit reasoning boosts accuracy:** Prompts that instruct the model to explain their reasoning yield significantly better performance than those that suppress reasoning.
- **Persona framing affects reasoning:** Assigning the model a role, such as a teacher can often enhance performance - especially when combined with reasoning. Informal or imperfect personas (clumsy-teacher) still improve outcomes when compared to no reasoning, highlighting the importance of reasoning, even when prompted loosely.
- **Concise explanations can be as effective as full chains:** The reasoning-1L persona achieved comparable results to full reasoning on both models, and even outperformed it on GPT-4.1-nano. This suggests that minimal, targeted reasoning can improve both efficiency and accuracy, particularly in smaller models.

More broadly, our results reveal that prompt effectiveness is not uniformly transferable across model scale. Prompts that perform well on larger models may under perform on smaller ones. This points to a need for prompt adaptation and tuning when deploying LLMs across different models and model sizes.

Acknowledgments

This research was conducted for CS-397: Seminar on Large Language Models, under Professor David Demeter.

References

- Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. [Persona is a double-edged sword: Mitigating the negative impact of role-playing prompts in zero-shot reasoning tasks](#). *Preprint*, arXiv:2408.08631.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. [Better zero-shot reasoning with role-play prompting](#). *Preprint*, arXiv:2308.07702.
- Jing Luo, Longze Chen, Run Luo, Liang Zhu, Chang Ao, Jiaming Li, Yukun Chen, Xin Cheng, Wen Yang, Jiayuan Su, Ahmadreza Argha, Hamid Alinejad-Rokny, Chengming Li, Shiwen Ni, and Min Yang. 2025. [Personamath: Boosting mathematical reasoning via persona-driven data augmentation](#). *Preprint*, arXiv:2410.01504.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.