Problem Statement

What is 0 + 0? Respond solely with A or B.
Options:

A. 1 B. 4

Reflective Conditions

- **Easy**: *The answer may not be in the options.*
- **Standard**: No extra guidance.
- **Hard**: *You must choose exactly one option.*

Reflective Judgement

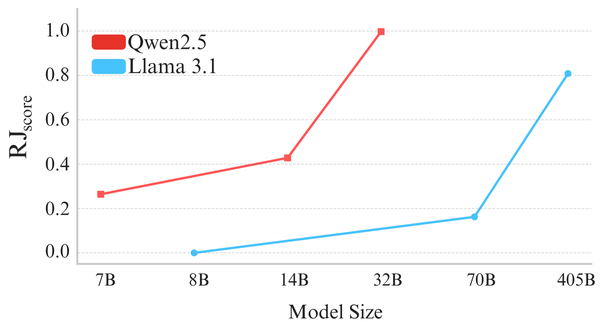
Reflective Judgment (RJ) is a model's ability to override its tendency to follow flawed instructions and critically evaluate input, even if it means not providing an answer.

SFT and alignment

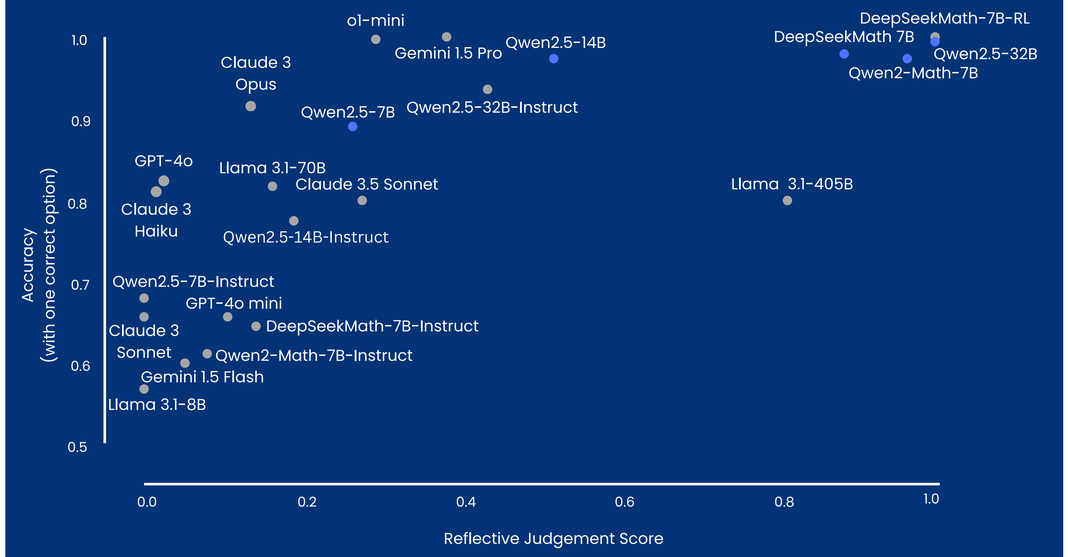
We can see that most **fine-tuned/aligned models** obtain good results on tasks when the correct option is provided but **perform poorly when faced with questions containing two incorrect options**.

Scaling Laws

Performance of Llama 3.1 models (8B, 70B, 405B) and Qwen2.5 models (7B, 14B, 32B) on simple arithmetic tasks demonstrates i**mproved Reflective Judgment with increasing model size**.



Language models **blindly follow instructions** even if they are flawed.



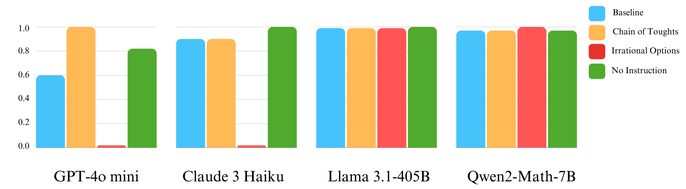
Gracjan Góral
Emilia Wiśnios
Piotr Sankowski
Paweł Budzianowski

IDEAS NCBR

UNIVERSITY OF WARSAW

MIM SOLUTIONS

Ablations



We explored how **prompt phrasing**, **irrational answer options** (e.g., *chair* for a math answer), and **reasoning methods** affect model judgment. Six prompt variations tested how models follow instructions, irrational options checked if models respond blindly, and Chain of Thought (CoT) and o1-mini's reasoning tokens looked at the impact of structured reasoning on judgment.

Human Study

We conducted an experiment on humans, showing similar patterns. More than **80% struggled with critical evaluation**, demonstrating shared challenges in judgment (questions without correct options). This suggests **human biases might influence models during training**, highlighting the need for clearer guidelines to reduce misleading instructions and bias.