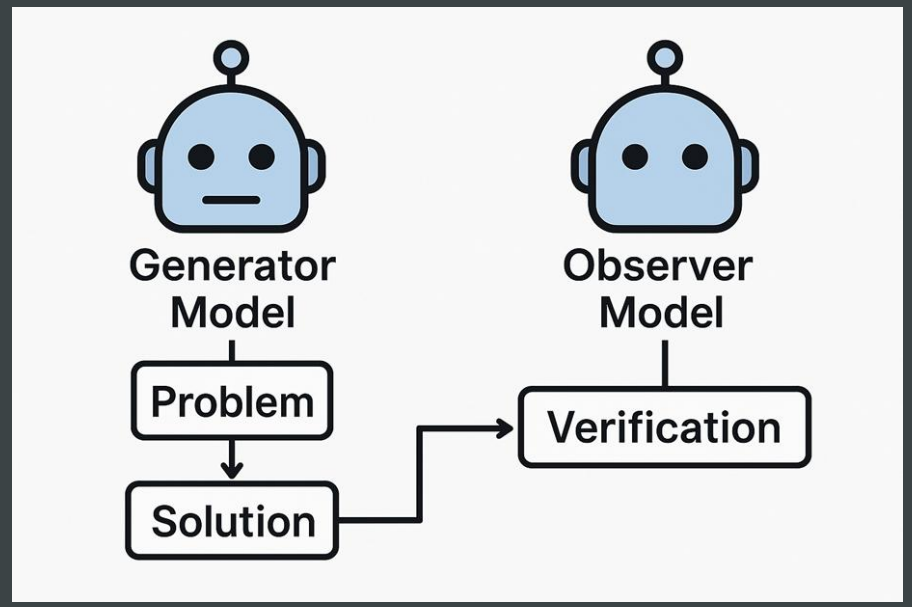
Solution / Verification Tradeoffs in Reasoning Models

CAISA NLP LAB – SUMMER SEMESTER 2025

MIDTERM PRESENTATION

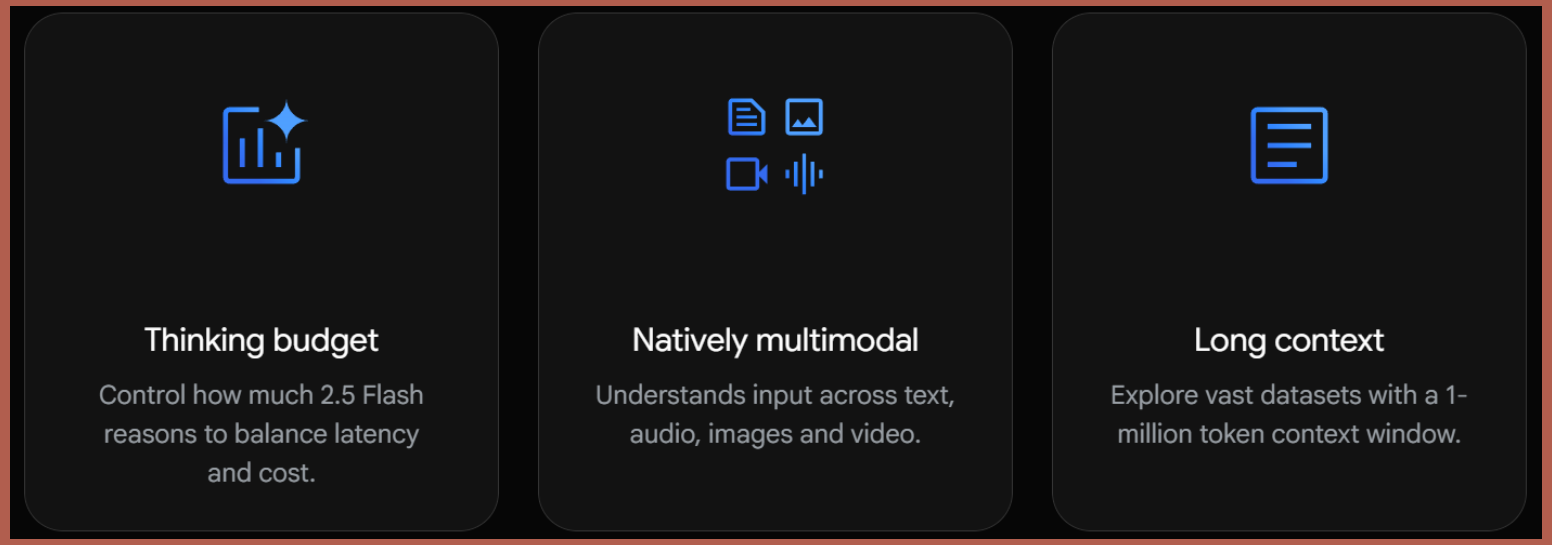
UMUTCAN DOGAN 50324204



The task

- Generator model
- Verification model
- Is there a relation between them?

The model being used



- Gemini 2.5 flash: thinkingBudget parameter (turned off)
- Dynamic thinking: Model decides when and how much to think

Prompts for the models

- **Generator model:** given the question below only output the final answer as a number only, nothing else, no symbols, no dots in the end, etc.\n\n\nquestion: {q}
- **Verification model:** You need to verify the following answer to the question. the question: {q} the answer: {ans}. You don't need to solve the question, just verify the answer as directly as possible. At the end of thinking, only output one of the words "true" or "false", indicating that the answer is correct or wrong. do not output anything else.

An example instance in the dataset

{**'question'**: 'Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?', **'**

'ground_truth_answer': '72',

'predicted_answer': '72',

'correct': True,

'total_output_tokens': 127,

'verification_output': 'true',

'verification_total_output_tokens': 244}

Verifying doesn't always mean that the answer is true

{ "question": "40% of the mosquitos in Jack's area are infected with malaria. 20% of the mosquitos are infected with Zika virus. Without a vaccine, the chances of getting infected with either virus after getting bitten by an infected mosquito are 50%. Jack is taking an experimental malaria vaccine that reduces the chances of getting infected after getting bitten by 50%. If Jack gets bitten by a random mosquito, what is the percentage chance he catches either Zika virus or malaria?",

"ground_truth_answer": "15",

"predicted_answer": "19",

"correct": **false**,

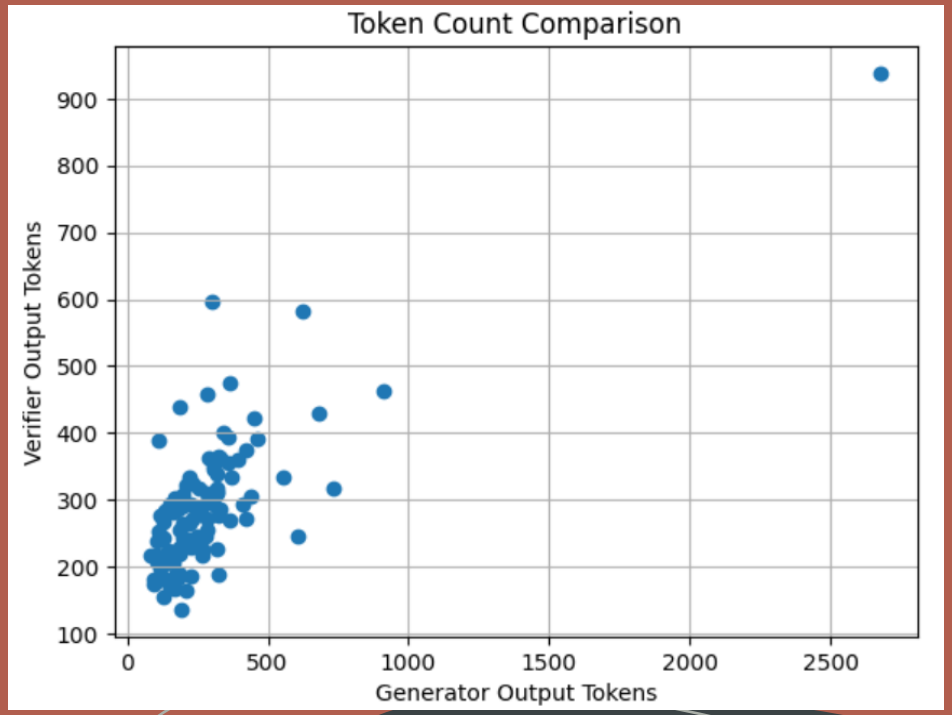
"total_output_tokens": 4831,

"verification_output": **"true"**,

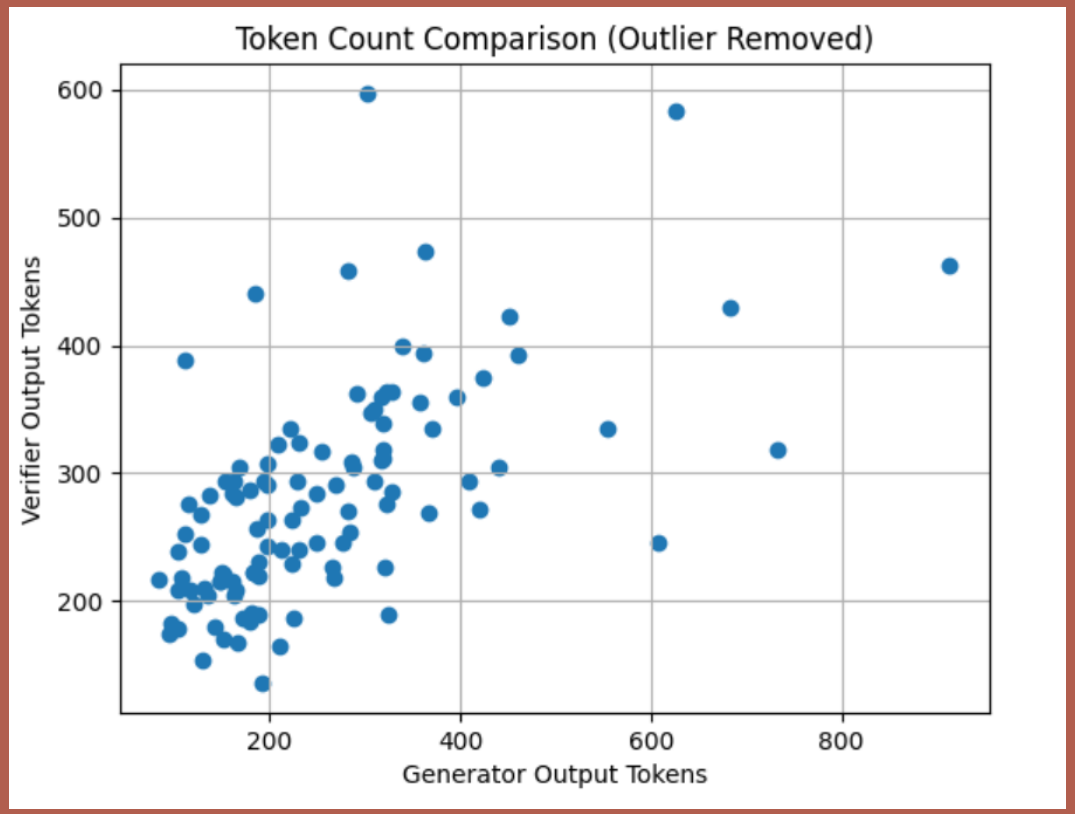
"verification_total_output_tokens": 4096 }

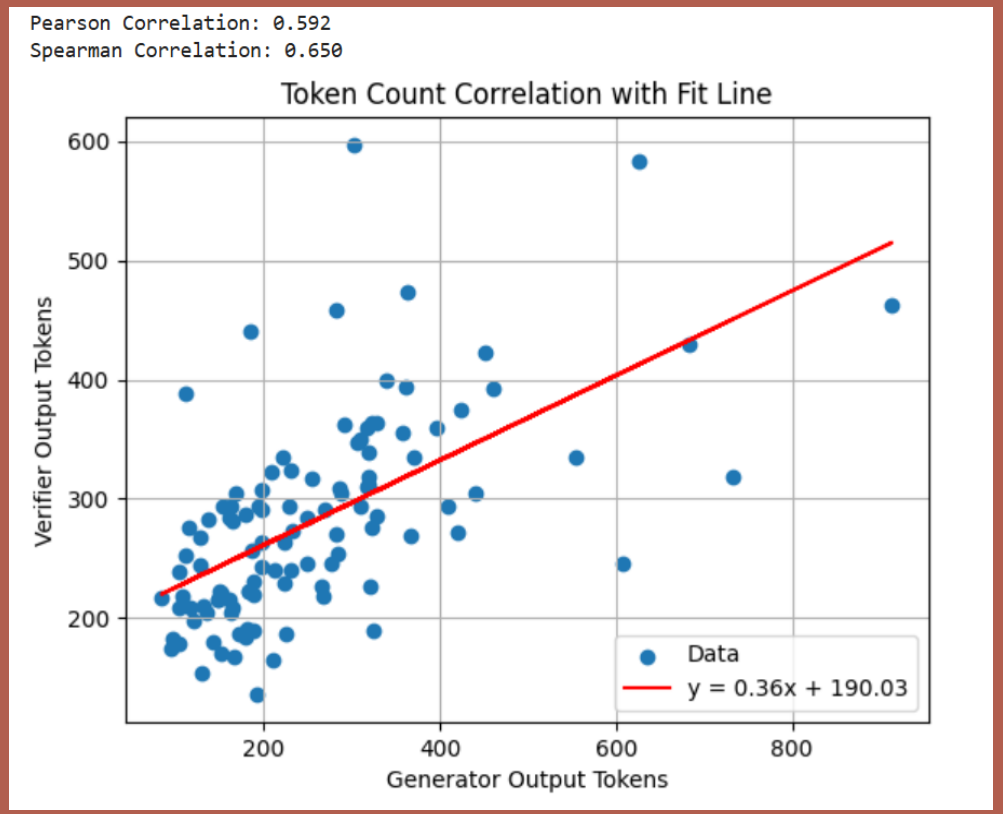
Early results


```
The number of data instances: 109
The instance with the max Generator Output Tokens:
{
  "question": "Tommy is making 12 loaves of bread. He needs 4 pounds of flour per loaf. A 10-pound bag of flour costs $10 and a 12-pound bag costs $13. When he is done making his bread, he has no use for flour and so he will throw away whatever is left. How much does he spend on flour if he buys the cheapest flour to get enough?",
  "ground_truth_answer": "50",
  "predicted_answer": "50",
  "correct": true,
  "compute_time": 14.372507333755493,
  "prompt_tokens": 119,
  "total_output_tokens": 2678,
  "verification_output": "true",
  "verification_compute_time": 4.166102647781372,
  "verification_prompt_tokens": 162,
  "verification_total_output_tokens": 938
}
```

With outlier removed





Future Work

One question that comes to mind: how does the accuracy of the verifier model come into play here?
In an ideal world, you'd want to answer a question like:
Given X tokens were spent by the generator to produce solution s, how many tokens Y do I need to spend on verification in order to be p% confident that the verifier is correct?
This question may require you to adjust the number of tokens that the verifier uses manually (i.e., option 2).

- Include more datasets
- Make the generator include the “proof of work” in the answer, and give it to verification model for further information.
- Experiment with more data
- Try to fit an ML model to the dataset

References

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.
- Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., et al. (2025). *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. arXiv.