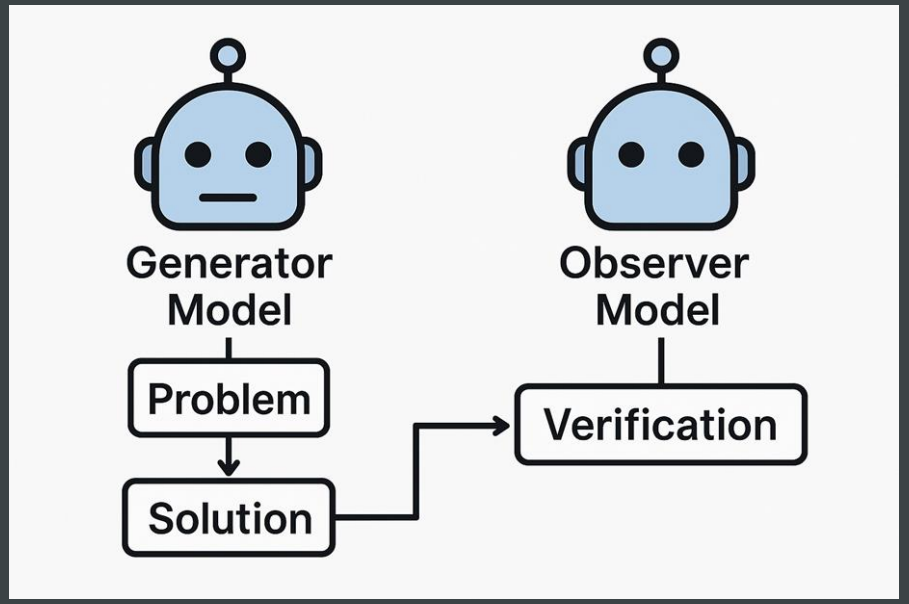
Solution / Verification Tradeoffs in Reasoning Models

CAISA NLP LAB – SUMMER SEMESTER 2025

FINAL PRESENTATION

UMUTCAN DOGAN 50324204



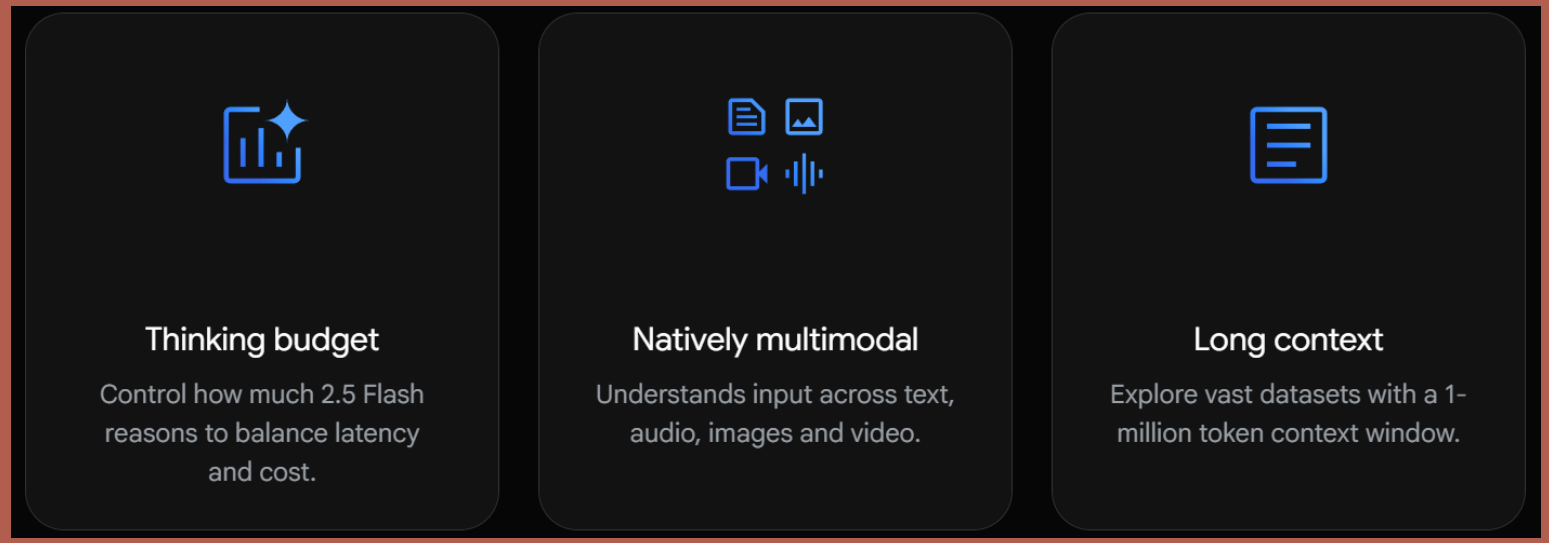
The task

Central Research Question: Can a machine learning model accurately predict the minimum reasoning token budget required for a verifier LLM to correctly assess solutions?

Why This Matters:

- Modern reasoning models use hidden "thinking" tokens that are computationally expensive
- Verification is crucial for AI safety and control
- Current models may waste computational resources on over-verification
- Efficient resource allocation could reduce inference costs significantly

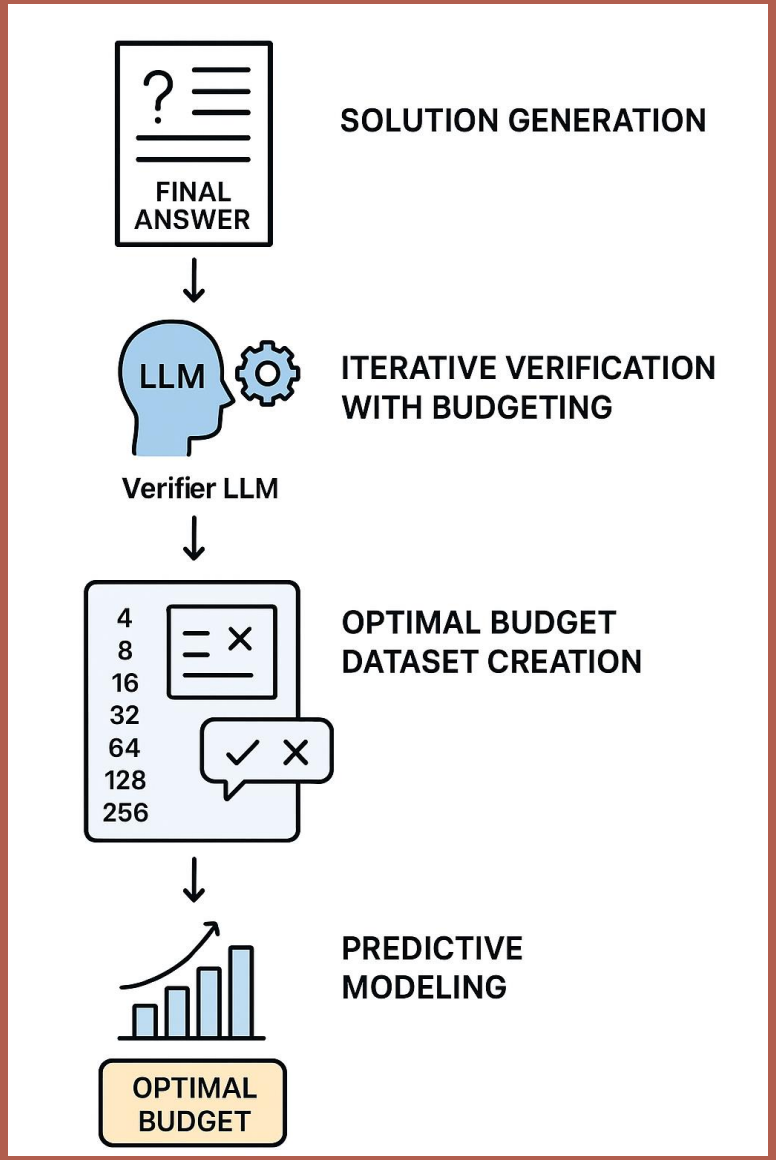
The model being used



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

Methodology

- **Solution Generation**
  - Used Gemini-2.5-flash on GSM8K dataset
  - Generated step-by-step solutions with final answers
- **Iterative Verification**
  - Tested 4 different reasoning budgets: [64, 128, 256, 512] tokens
  - Verifier outputs "true/false" judgments
- **Optimal Budget Dataset Creation**
  - Identified minimum budget for correct verification
  - Created novel "smallest_length_dataset"
- **Predictive Modeling**
  - Logistic Regression baseline on text embeddings
  - Fine-tuning an encoder only model



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. Also, provide a step-by-step explanation of your answer.
- **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?', '

'**step_by_step_explanation**': 'Natalia sold 48 clips in April. In May, she sold half as many, which is 48 / 2 = 24 clips. Altogether, she sold 48 + 24 = 72 clips.'

'ground_truth_answer': '72',

'**predicted_answer'**: '72',

'correct': True,

'reasoning_tokens': 127**,**

'verification_output': 'true',

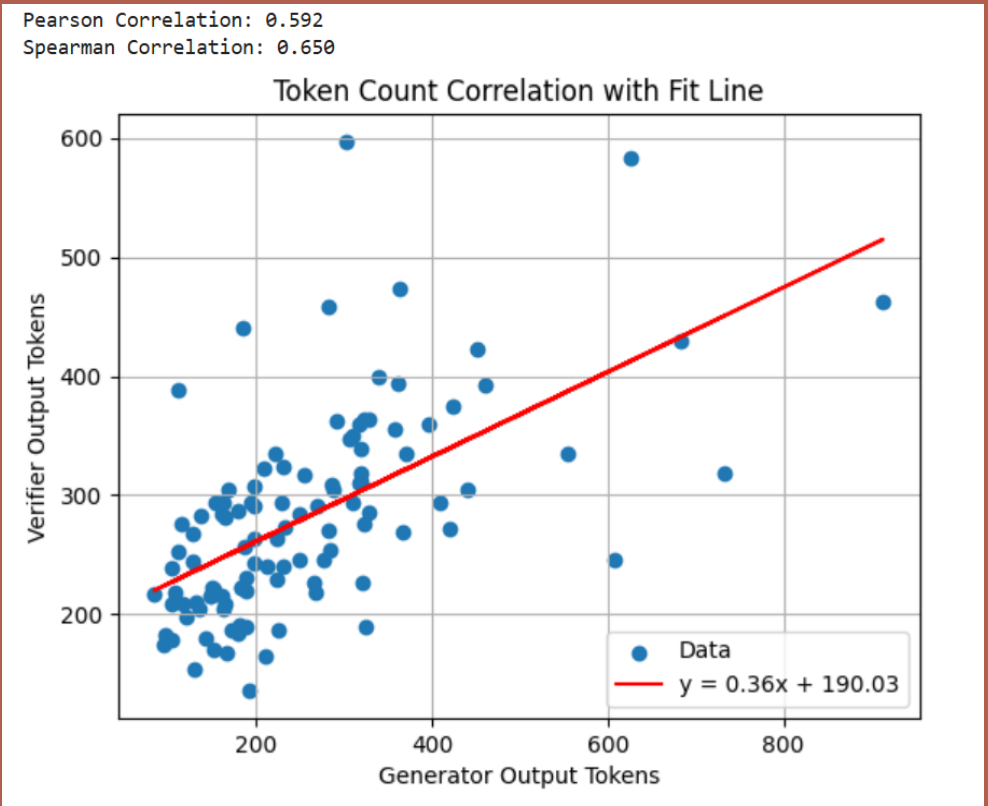
'**verification_reasoning_tokens'**: 244}

Creation of the new dataset

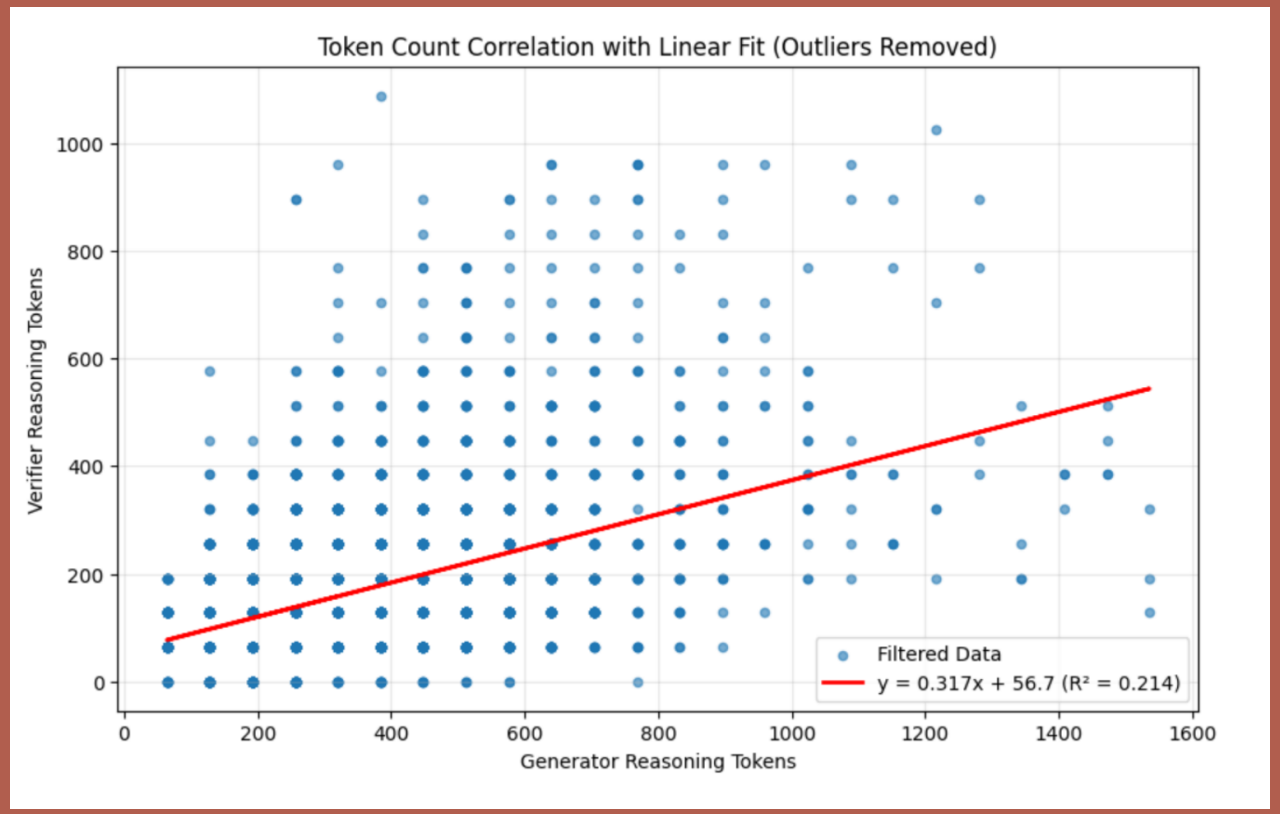
Table 1: Distribution of Optimal Reasoning Budgets

Reasoning Budget (Tokens)	**Count of Problems**
64	7111
128	39
256	25
512	28
Total	**7437**

Relation without the step-by-step explanation



Relation with step-by-step explanation



Generator vs. Verifier Analysis

Key Finding: Moderate Correlation

- Pearson correlation coefficient: 0.463
- When generators "think" longer, verifiers tend to "think" longer
- Relationship is not strictly linear - other factors influence verification effort

Compute Efficiency Analysis:

- **Optimal budgets** (our analysis): 95.5% need ≤64 tokens
- **Self-determined budgets (with dynamic thinking)**: Only 42.3% use ≤64 tokens
- **Implication:** Significant computational overuse in current systems

Results of the Baseline Model

Table 2: Per-Class Accuracy of the Baseline Logistic Regression Model

Reasoning Budget	**Correctly Predicted**	**Total in Test Set**	**Per-Class Accuracy**
64	545	546	99.8%
128	0	25	0.0%
256	0	5	0.0%
512	0	1	0.0%

Interpretation of the Results

- Model learned to be a majority-class classifier
- High overall accuracy misleading due to class imbalance
- Failed to distinguish complex cases requiring higher budgets

Why it happened?

We give the model step-by-step reasoning which reduces the need for reasoning substantially.

The dataset might be too easy for the model.

Key Insights & Implications

Theoretical Insights:

- Verification-generation asymmetry more pronounced than expected
- Current models systematically over-allocate verification resources

Practical Implications:

- Immediate efficiency gains possible through better budget allocation
- Resource optimization crucial for cost-effective deployment

Future Work

- **Complex Datasets:** Advanced mathematics, legal reasoning, logic puzzles
- **Non-linear Models:** Gradient boosting, neural networks, transformers
- **Regression Formulation:** Predict exact token counts vs. discrete classes

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.