From CoT to Collaborative Inference

Efficient Reasoning as Communication Compression

Ayush Sekhari

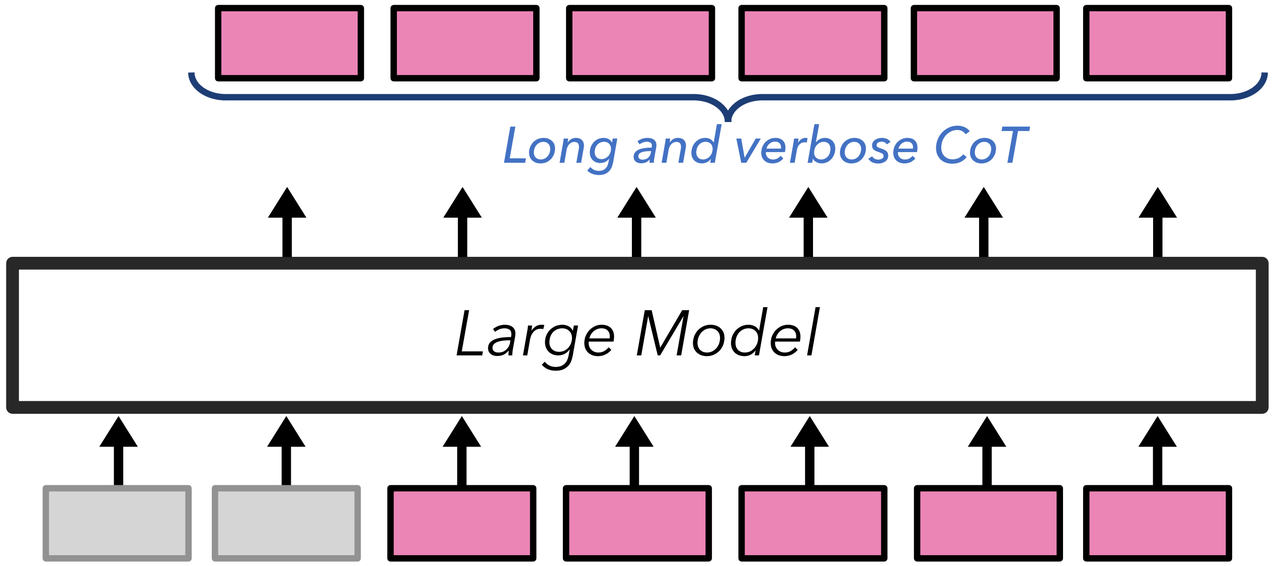
Joint work with Yilei Chen (BU), Sharut Gupta (MIT), Yannis Paschalidis (BU), and Aldo Pacchiano (BU)

The era of Automated Reasoning

- LLMs are everywhere and they are automating reasoning!
- Serving them is expensive; Reasoning has large overhead (COT):
  - Reasoning models produce responses 5-10x longer than equivalently-sized instruct models, even on questions that both can solve correctly [Fan et. Al. 2025].
  - DeepSeek-R1 generates 31x more intermediate tokens than Llama 3.1 8B while running 54x slower, requiring 16 H100 GPUs just to serve [Garvey, 2025].
- OpenAI spent **$8.67 billion** on inference through Q3 2025 for their reasoning models.

Can we keep the benefits of reasoning without paying the full cost of reasoning?

What today's reasoning pipeline looks like?

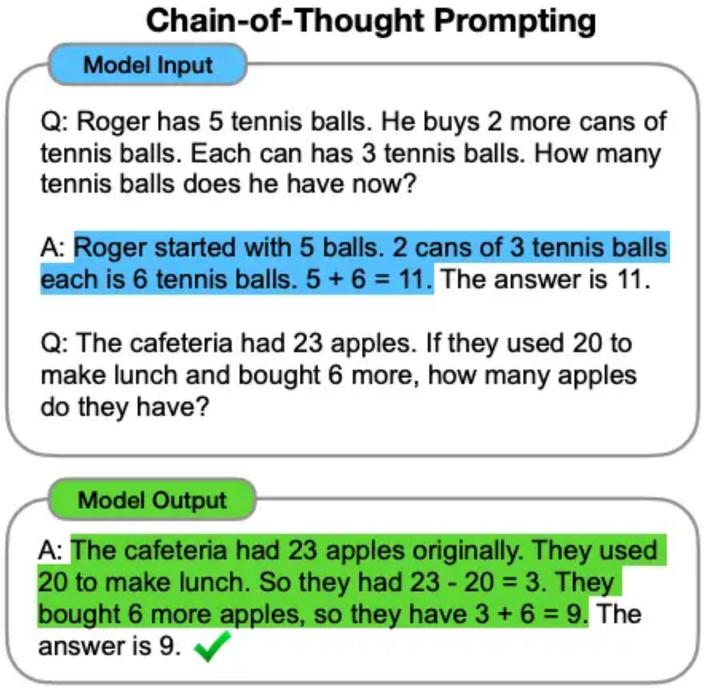


Chain-**o**f-**T**hought (CoT)

- All tokens are generated by the same expensive model
- Model handles: **planning, deduction, narration, formatting, ...**

Why Chain-of-Thought (CoT) Helps?

- Externalize intermediate computation
- Break hard reasoning into smaller steps
- Trade compute for reliability
- Improve performance on math / logic / science tasks



The Compute Cost of CoT

Reasoning quality has come with explosion in #Tokens

The Compute Cost of CoT

Reasoning quality has come with explosion in #Tokens

Costs

- Longer Outputs
- Slower Inference
- Cost for decoding $$$

Benefits

- Better accuracy on hard-tasks (math, coding, etc)
- More robust reasoning behavior

Not all tokens deserve the same model!

Key Problem: Monolithic Inference

The same large model is being used for both hard deduction and cheap response generation.

- Hard logical work and cheap rendering are entangled
- The architecture allocates equal compute across unequal subtasks

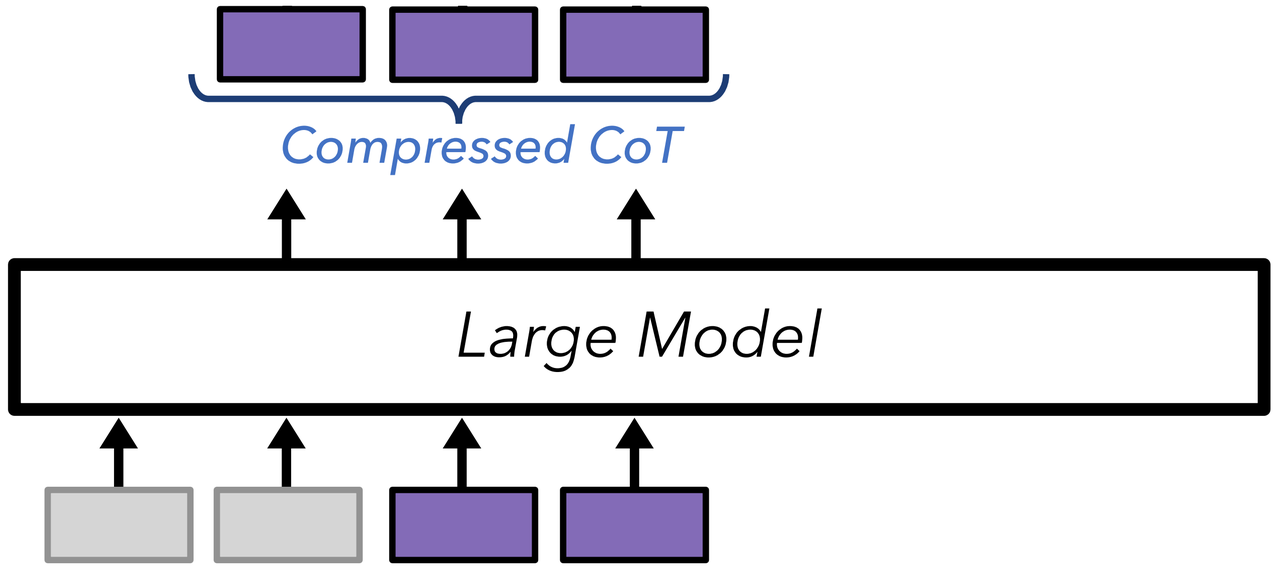
Reasoning as Communication

What if the large model only says the important part?

Compress the reasoning and hand off the rest to small model

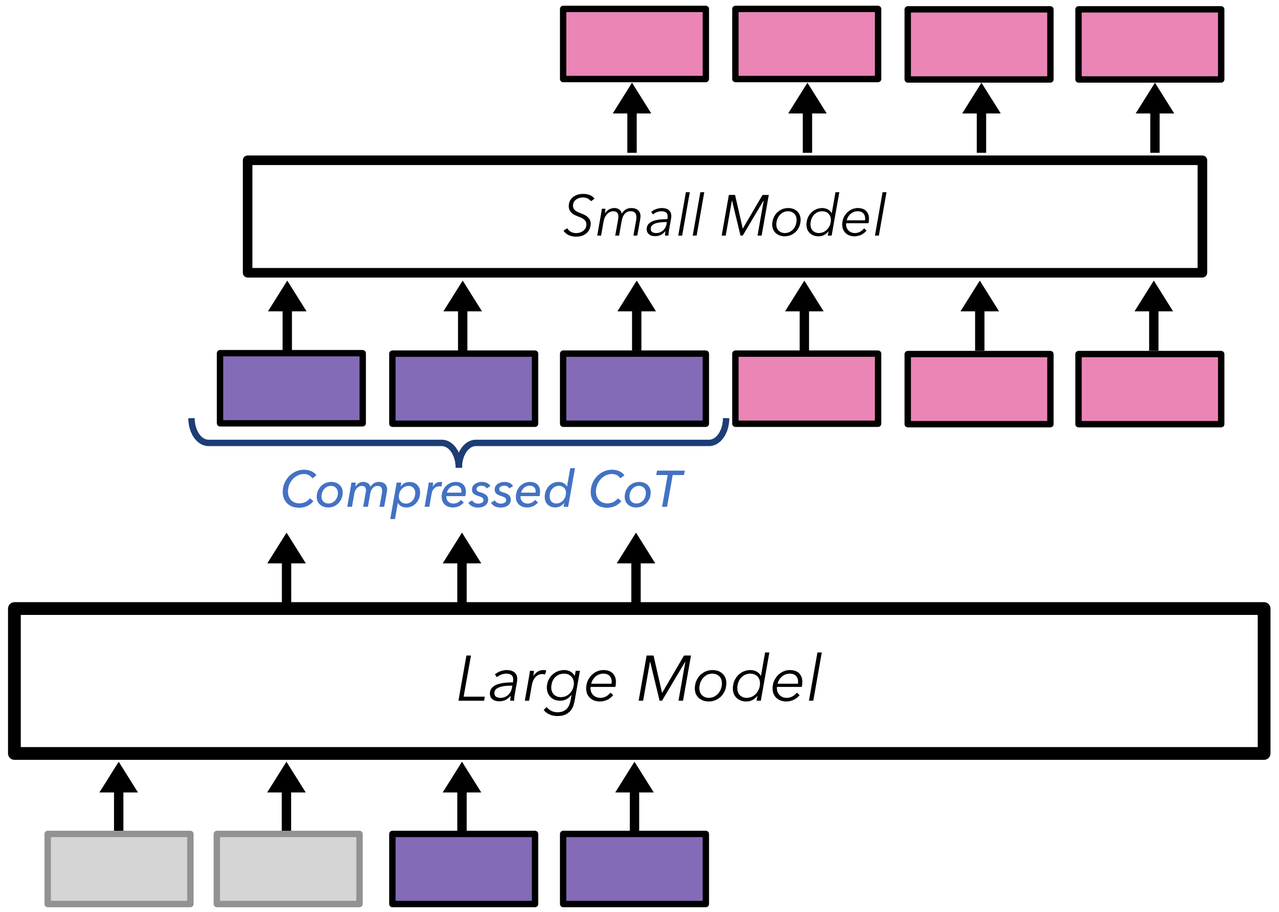
“Efficiency as a communication problem”

DUET: Dual-Model Two-stage Inference



Stage 1

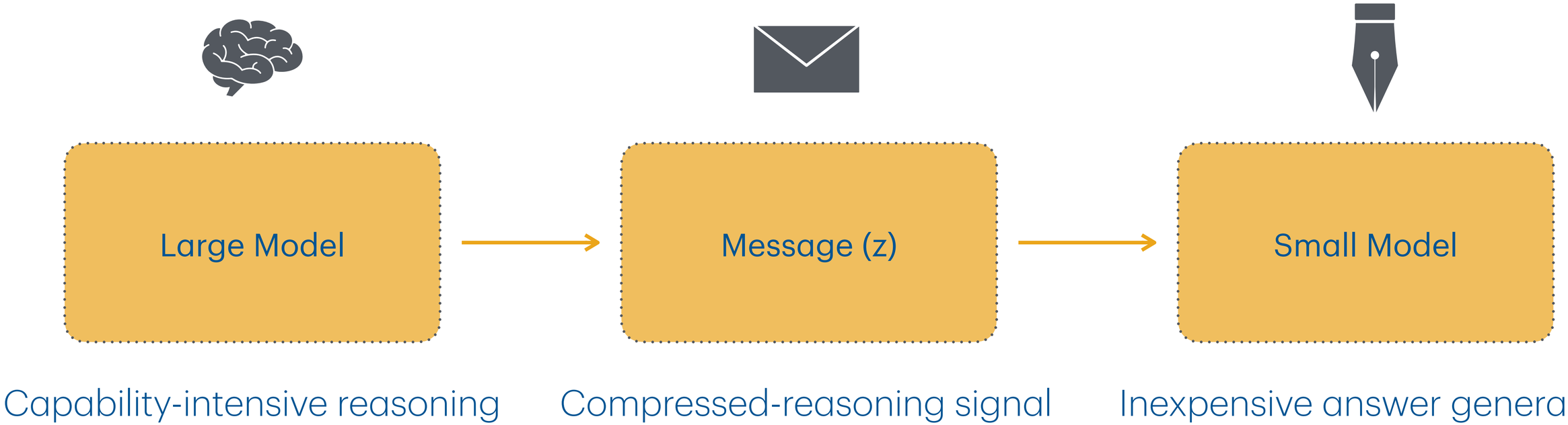
DUET: Dual-Model Two-stage Inference



Stage 2

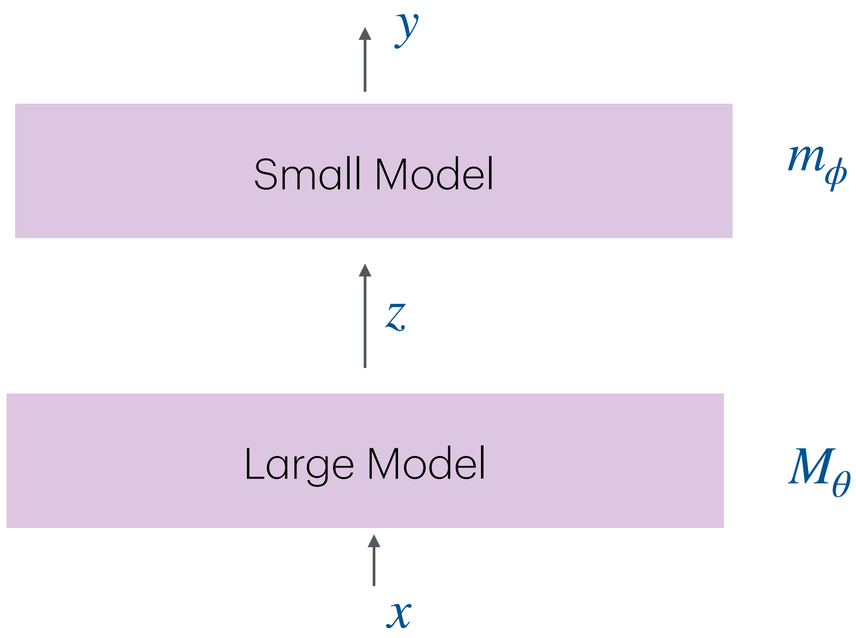
Stage 1

Who Does What?



Setup: The Optimization Problem

Reward Function $R(x, y)$



Learn two policies that solve the task while keeping the intermediate message (z) short

$$\max_{M_\theta, m_\phi} \mathbb{E}[R(x, y)] \quad \text{s.t.} \quad \mathbb{E}[\text{length}(z)] \leq B$$

↑ Performance

↓ Message Length under Budget

Practical Objective

Lagrangian Relaxation and Alternating Optimization

$$\max_{M_\theta} \max_{m_\phi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{E}_{z\sim M_\theta(\cdot|x)} [\mathbb{E}_{y\sim m_\phi(\cdot|z)} [R(x,y)] - \lambda \text{len}(z)] \right]$$

Collect Dataset $\{x, z^i, y^i\}_{i=1}^G$ using M and m

Big Model M

$$\max_{M_\theta} \mathbb{E}\left[R(x,y) - \lambda \text{len}(z)\right]$$

Alternate optimization using GRPO

Smaller Model m

$$\max_{m_\phi} \mathbb{E}\left[R(x,y)\right]$$

However, Naive Training Can Fail

Failure Mode: The Large model Learns to Say Almost Nothing

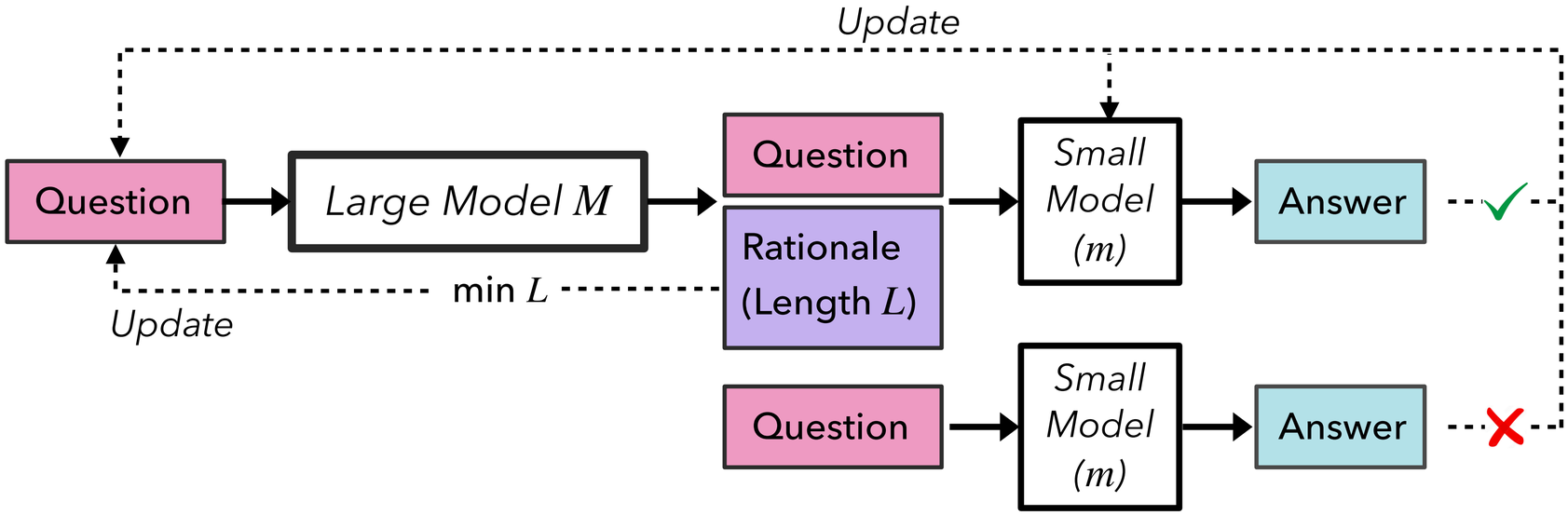
$$\max_{M_\theta} \max_{m_\phi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{E}_{z\sim M_\theta(\cdot|x)} [\mathbb{E}_{y\sim m_\phi(\cdot|z)} [R(x,y)] - \lambda \text{len}(z)] \right]$$

If the small model is decent on various problems

Then, the large model can game the objective by staying silent

Cheap communication is not the same as useful communication!

Key Idea 1: Marginal Utility Reward

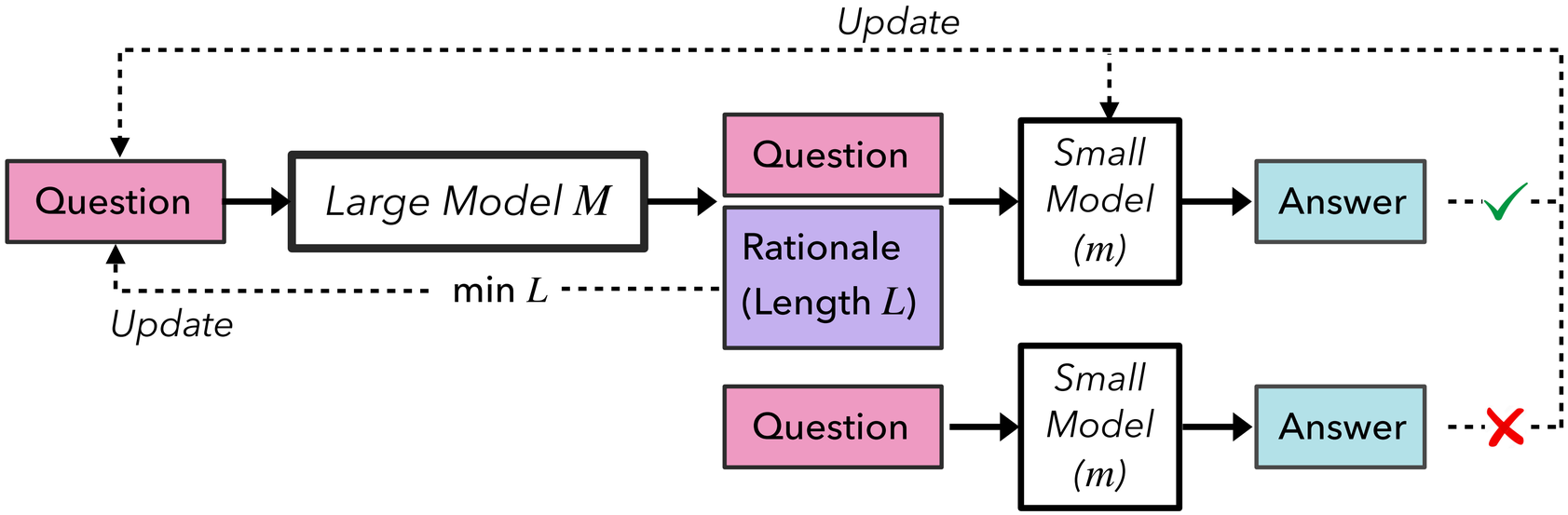


$y_{Mm} \sim m(\cdot \mid x, z)$

$y_m \sim m(\cdot \mid x)$

Marginal Reward: $R(x, y_{Mm}) - R(x, y_m)$.

Key Idea 1: Marginal Utility Reward



$$y_{Mm} \sim m(\cdot \mid x, z)$$

$$y_m \sim m(\cdot \mid x)$$

Optimization Objective: $\max_{M_\theta} \max_{m_\phi} \mathbb{E}_{x \sim \mathcal{D}} \left[\left(R(x, y_{Mm}) - R(x, y_m) \right) - \lambda \text{len}(z) \right]$

Key Idea 2: Adaptive Length Penalty

$$\lambda_t = \left[\lambda_{t-1} + \eta\left(\frac{\mathbb{E}_t[L(z)]}{B} - 1\right)\right]_+$$

- Strengthen the penalty when messages are too long.
- Relax the penalty when messages are under budget B.
- **Motivation**: Primal-dual update for inequality-constrained optimization.
- **Practical Benefits**: stabilize training and reduce manual tuning.

Experiments: Setup

Models

Large Model (M):
Qwen 3.4B (thinking)

Small Model (m):
Qwen 0.5B

Datasets

Training:
MATH-LightEval, DeepScaleR, and DAPO-Math-17k

Evaluation:
Math 500, AMC 23, AIME 2024, 2025, GPQA Diamond

Evaluations

Accuracy

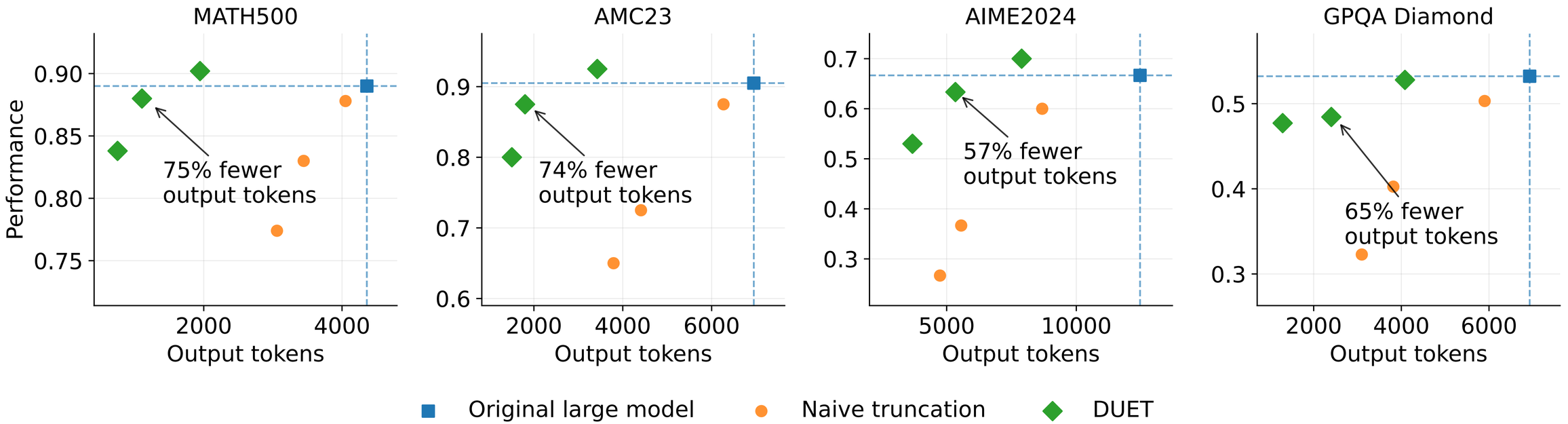
Tokens Produced

Intelligence Per Token:

$$\frac{\text{Accuracy}}{\text{\# Tokens Produced}}$$

Experiments: Main Result

Training Steps: (50, 100, 150)



Other baselines in the paper (Prompt-based, ThinkPrune, L-GRPO on M)

High Level Takeaway

Compression Beats Truncation

Duet

Learned Compression

Message Tailored to small model m

preserves task relevant signal

Truncation

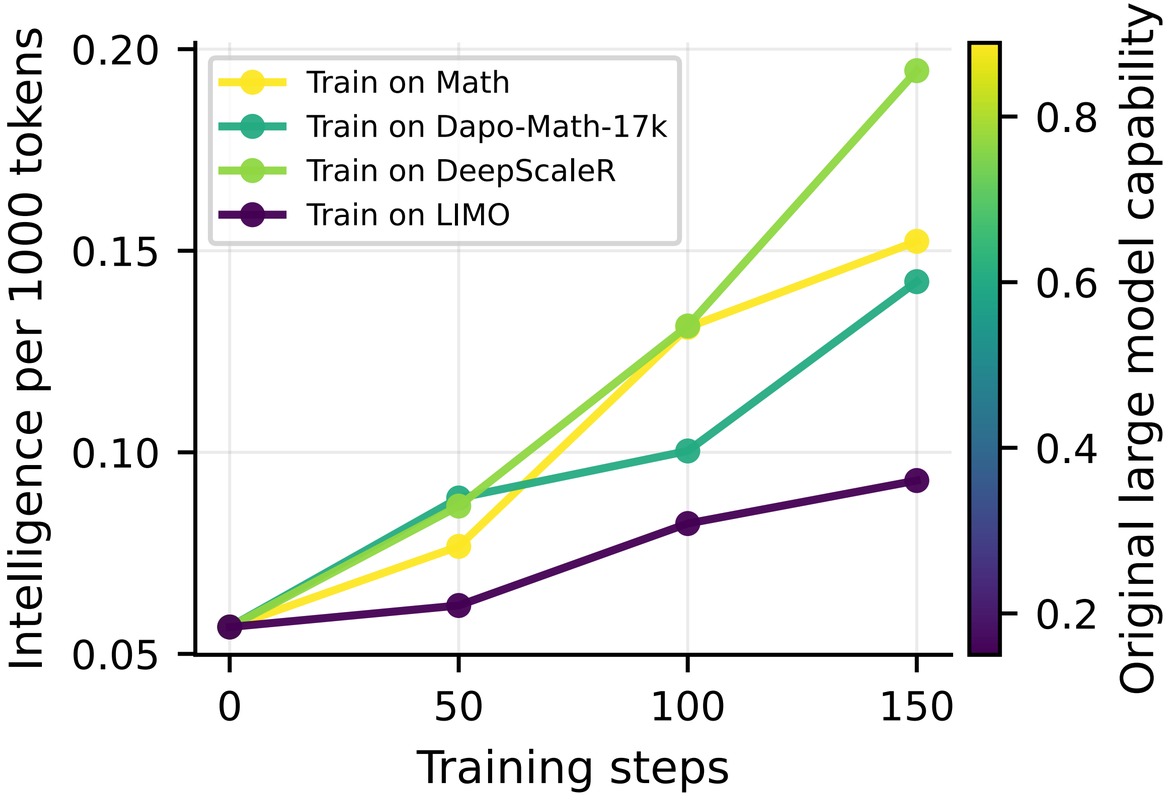
Fixed Budget

Can Loose Tokens Blindly

Hurts critical information

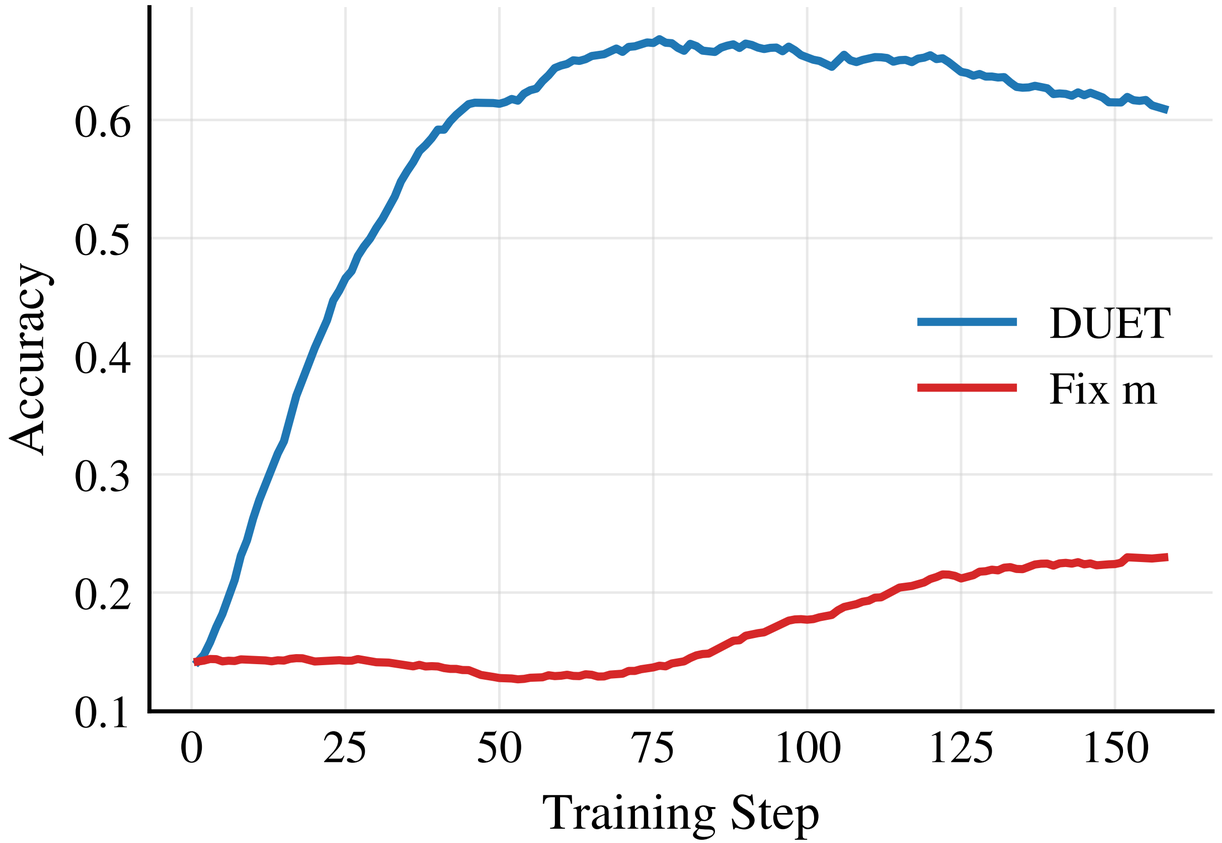
Experiments: Training Data Sensitivity

Compression requires something worth compressing



DUET works best when the large model is already competent on the training distribution.

Experiments: Is Joint Training Necessary?



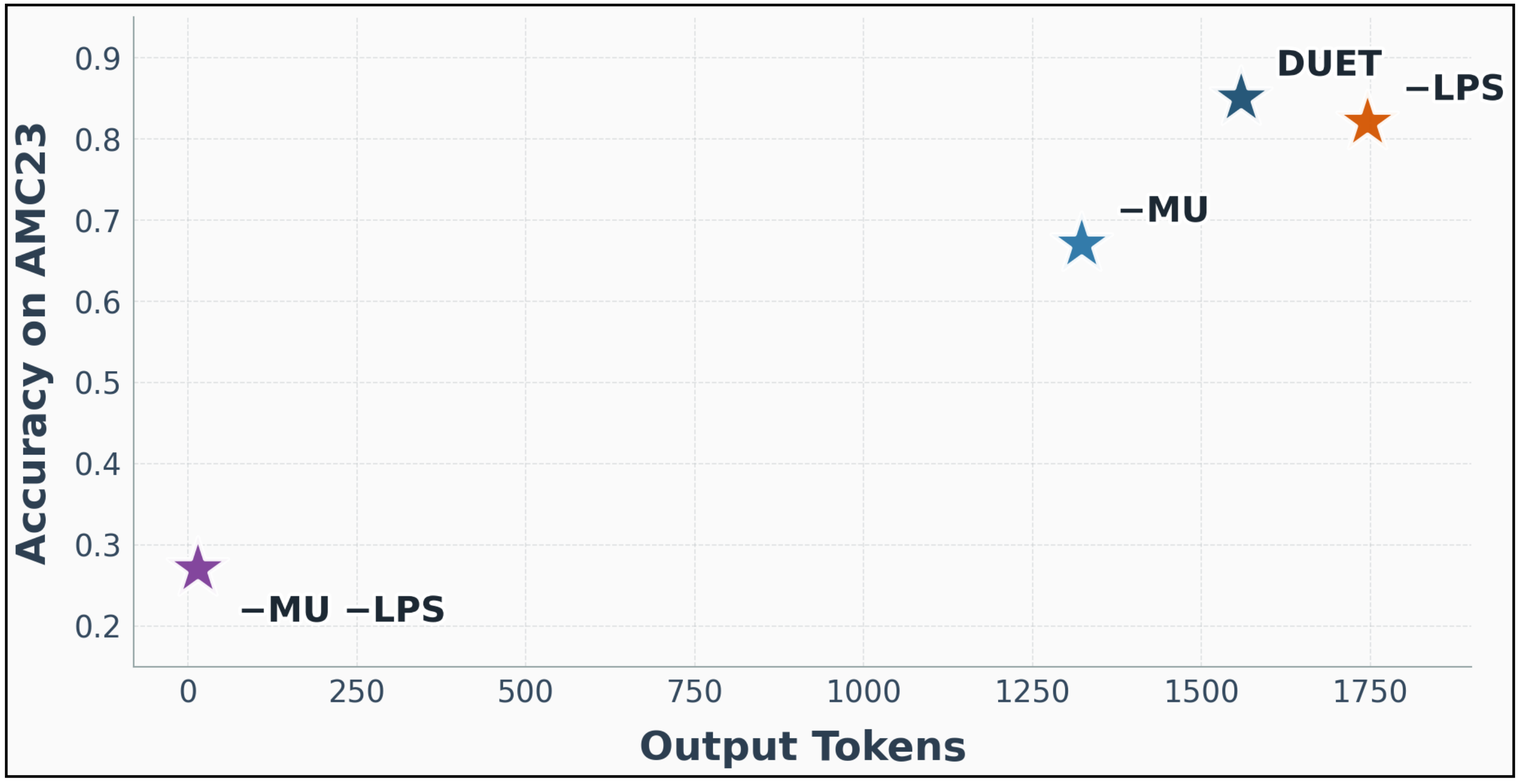
Better Tradeoff

Some compression

The communicated message gets more compact as both models adapt

Ablations on Marginal Utility and Length Penalty Schedule

Accuracy vs Generation length



Illustration

Prompt

Suzanne walks four miles every third day. What is the fewest number of miles she can walk in February? Let's think step by step and output the final answer within `\boxed{}`.

Large model reasoning

`<think>`

Okay, let's see. Suzanne walks four miles every third day. So, I need to figure out the fewest number of miles she can walk in February. Hmm. First, I should know how many days are in February. Well, depending on whether it's a leap year or not. But since the question doesn't specify, I think we should assume a common year, which has 28 days in February.

So, if she walks every third day, that means she walks on days 3, 6, 9, etc. So how many times does that happen in 28 days? Let's divide 28 by 3. 3 times 9 is 27, so that's 9 days. So she would walk 9 times, each time 4 miles. So 9 times 4 is 36 miles. So the fewest number of miles she can walk in February is 36.

`</think>`

Large Model M's Output

Small model structured solution

To find the fewest number of miles Suzanne can walk in February, follow these steps:

1. **Determine the number of days in February**: February has 28 days in a common year.

2. **Calculate the number of walking days**: Since she walks every third day, divide the number of days by 3.

$$\frac{28}{3} = 9$$

with a remainder of 1.

3. **Calculate the total miles**: Multiply the number of walking days by 4 (since each walk is 4 miles).

$$9 \times 4 = 36.$$

Final Answer:

$$\boxed{36}$$

Small Model m's output

Takeaways

- Efficient reasoning can be posed as a communication problem
- Inference can be reorganized as division of labor
- RL post-training can be used to minimize inference cost of trained models.

Next Steps....

- Text messages vs latent / continuous communication
- One-shot handoff vs Interactive Communication
- Confidence-aware querying of the large model
- Hierarchy of decoders; Theoretical foundations
- The usual interventions: better regularizers, length penalties, algorithms, etc

Thank You! Questions?