Co-LLM: Learning to Decode Collaboratively with Multiple Language Models

Shannon Shen shannons@csail.mit.edu

& Hunter Lang, Bailin Wang, Yoon Kim, David Sontag

ACL 2024

Motivation & Background

Why do we need multi-model collaboration at the token level?

Motivation & Background

Why do we need multi-model collaboration at the token level?

Inference Efficiency

Steering Generation

Combining Capabilities

Why do we need multi-model collaboration at the token level?

Inference Efficiency

Speculative Decoding

collaborates between a large and small model to speed up the decoding.

$$P_{\text{small}}(X_t \mid X_{<t}) \rightarrow P_{\text{large}}(X_t \mid X_{<t})$$

Leviathan, Yaniv, Matan Kalman, and Yossi Matias. "Fast inference from transformers via speculative decoding." International Conference on Machine Learning. PMLR, 2023.

Why do we need multi-model collaboration at the token level?

Steering Generation

Fudge

steers the generation to satisfy certain properties (a) using another model

$$P(X_t \mid X_{<t}, a) \propto P(a \mid X_{<t+1}) P(X_t \mid X_{<t})$$

Kevin Yang and Dan Klein. "FUDGE: Controlled Text Generation With Future Discriminators."
Conference of the North American Chapter of the Association for Computational Linguistics. NAACL, 2021.

Motivation & Background

Why do we need multi-model collaboration at the token level?

Combining Capabilities

MoE Methods, Branch Train Merge, Proxy Tuning, PHATGOOSE, ...

combine expertise from multiple (sub-)models for better generation.

$$P(X_t \mid X_{<t}) = \sum_{j=1}^{k} w_j(X_{<t}) \cdot P_i(X_t \mid X_{<t})$$

Jiang, Albert Q., et al. "Mixtral of experts." arXiv preprint arXiv:2401.04088 (2024).
Gururangan, Suchin, et al. "Scaling expert language models with unsupervised domain discovery." arXiv preprint arXiv:2303.14177 (2023).
Liu, Alisa, et al. "Tuning language models by proxy." arXiv preprint arXiv:2401.08565 (2024).
Muqeeth, Mohammed, et al. "Learning to route among specialized experts for zero-shot generalization." arXiv preprint arXiv:2402.05859 (2024).

A concrete example

What are some species of bears that are now extinct?

There have been many species of bears that have become extinct throughout history. Some examples include:

1. The giant short-faced bear (Arctodus simus) - This species went extinct around 10,000 years ago and was one of the largest bears to have ever lived. [...]

A concrete example

What are some species of bears that are now extinct?

There have been many species of bears that have become extinct throughout history.

Some examples include:

1. The giant short-faced bear (Arctodus simus) - **This species** went extinct around 10,000 years ago and was one of the largest bears to have ever lived. [...]

Scaffolding helpful & detailed responses

A concrete example

What are some species of bears that are now extinct?

There have been many species of bears that have become extinct throughout history. Some examples include:

1. **The giant short-faced bear (Arctodus simus)** - This species went extinct around **10,000 years ago** and was **one of the largest bears** to have ever lived. [...]

Factual information

A concrete example

What are some species of bears that are now extinct?

There have been many species of bears that have become extinct throughout history.

Some examples include:

1. **The giant short-faced bear (Arctodus simus)** - **This species** went extinct around **10,000 years ago** and was **one of the largest bears** to have ever lived. [...]

To compose a good response, different “expertise” is required, at different steps / stages of generation.

A concrete example

What are some species of bears that are now extinct?

There have been many species of bears that have become extinct throughout history.

Some examples include:

1. **The giant short-faced bear (Arctodus simus)** - **This species** went extinct around **10,000 years ago** and was **one of the largest bears** to have ever lived. [...]

To compose a good response, different “expertise” is required, at different steps / stages of generation.

It would be great if models, of different specialties, can collaborate during the course of generation.

Challenges

There have been many species of bears that have become extinct throughout history.

Some examples include:

1. **The giant short-faced bear (Arctodus simus)** - **This species** went extinct around **10,000 years ago** and was **one of the largest bears** to have ever lived. [...]

Challenges

The breakdown of needed expertise (red, orange, grey, ...?)

There have been many species of bears that have become extinct throughout history.

Some examples include:

1. **The giant short-faced bear (Arctodus simus)** - **This species** went extinct around **10,000 years ago** and was **one of the largest bears** to have ever lived. [...]

Challenges

The breakdown of needed expertise (red, orange, grey, ...?)

There have been many species of bears that have become extinct throughout history.

Some examples include:

1. **The giant short-faced bear (Arctodus simus)** - **This species** went extinct around **10,000 years ago** and was **one of the largest bears** to have ever lived. [...]

The patterns of switching (distribution of the colors)

Challenges

1. **How to learn the collaboration mode organically from the data without supervision?**

Challenges

1. **How to learn the collaboration mode organically from the data without supervision?**

   This is different from toolformer: it needs training data that prescribes tool use.

Challenges

1. **How to learn the collaboration mode organically from the data without supervision?**

   This is different from toolformer: it needs training data that prescribes tool use.

2. **Can we avoid training the models from scratch and leverage existing models?**

   As opposed to MoE like methods, where the experts need to be trained from scratch (and the experts are often not full-fledged models).

The latent variable model framework

The latent variable model framework

Opdualag contains two active components: 1) nivol

The latent variable model framework

Opdualag contains two active components: 1) nivol

$X_{<t}$ X_t

The latent variable model framework

Opdualag contains two active components: 1) nivol

$X_{<t}$ X_t

Language Model 1

$$P_1(X_t|X_{<t})$$

Language Model 2

$$P_2(X_t|X_{<t})$$

Language Model 3

$$P_3(X_t|X_{<t})$$

The latent variable model framework

Opdualag contains two active components: 1) nivol

$X_{<t}$ X_t

Language Model 1

$P_1(X_t|X_{<t})$

Language Model 2

$P_2(X_t|X_{<t})$

Language Model 3

$P_3(X_t|X_{<t})$

Collaboration Control

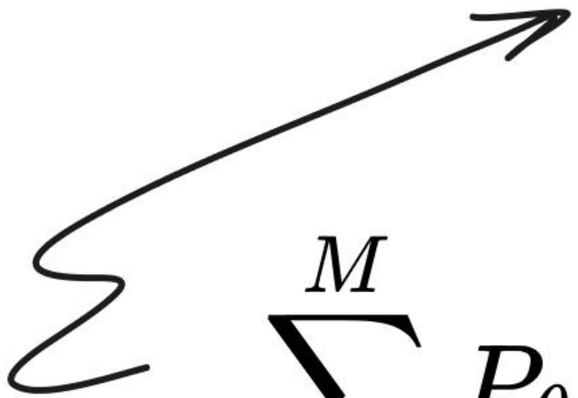
$P(Z_t|X_{<t})$

Using the Z_t-th model given the context

The latent variable model framework

Opdualag contains two active components: 1) nivol

$X_{<t}$ X_t



$$\sum_{Z_t=0}^{M} P_\theta(Z_t|X_{<t}) P_{Z_t}(X_t|X_{<t})$$

The latent variable model framework

Opdualag contains two active components: 1) nivol

$$P(X) = \prod_{t=1}^{T} \left(\sum_{Z_t=0}^{M} P_\theta(Z_t|X_{<t}) P_{Z_t}(X_t|X_{<t}) \right)$$

The latent variable model framework

$$P(X) = \prod_{t=1}^{T} \left(\sum_{Z_t=0}^{M} P_\theta(Z_t|X_{<t}) P_{Z_t}(X_t|X_{<t}) \right)$$

$$\text{NLL Loss} = -\log P(X) = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$

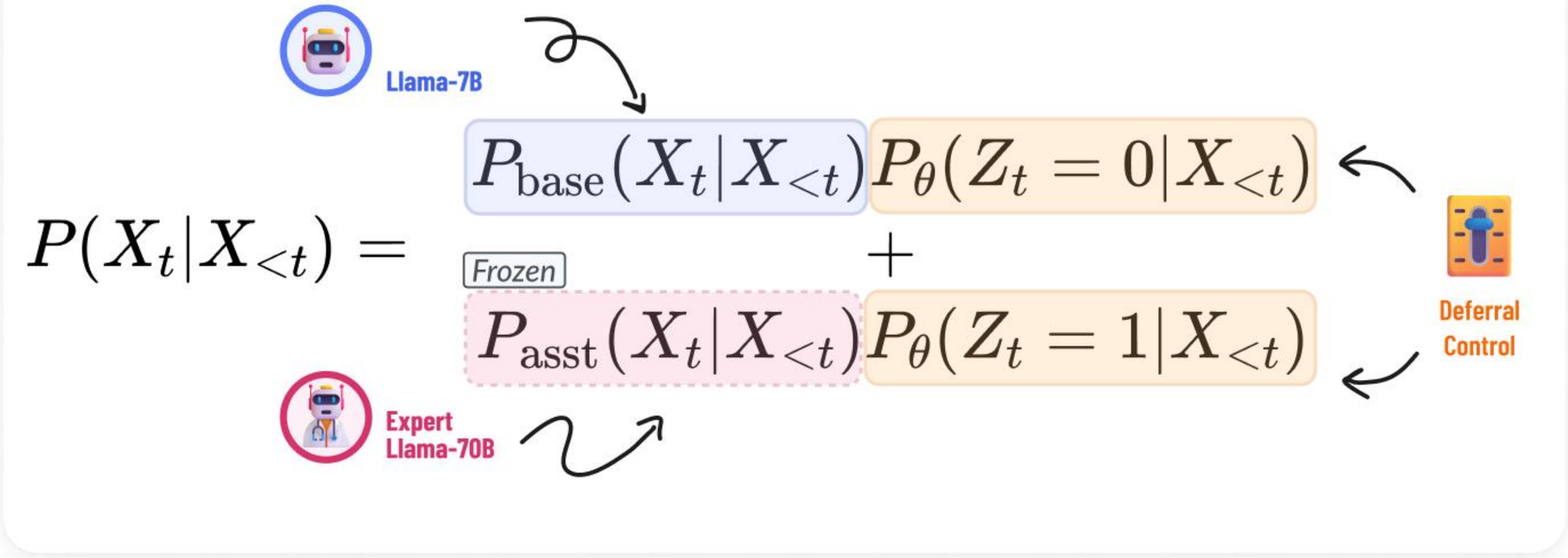
Two model case

$$\text{NLL Loss} = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$

$$P(X_t|X_{<t}) = \begin{array}{c} P_{\text{base}}(X_t|X_{<t})P_\theta(Z_t = 0|X_{<t}) \\ + \\ P_{\text{asst}}(X_t|X_{<t})P_\theta(Z_t = 1|X_{<t}) \end{array}$$

Two model case

$$\text{NLL Loss} = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$



Mozannar, Hussein, and David Sontag. "Consistent estimators for learning to defer to an expert." International Conference on Machine Learning. PMLR, 2020.

Two model case

$$\text{NLL Loss} = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$

$$P(X_t|X_{<t}) = P_{\text{base}}(X_t|X_{<t})P_\theta(Z_t = 0|X_{<t}) + P_{\text{asst}}(X_t|X_{<t})P_\theta(Z_t = 1|X_{<t})$$

Llama-7B

Frozen

Expert
Llama-70B

Collaboration
Control

Two model case

$$\text{NLL Loss} = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$

$$P(X_t|X_{<t}) = P_{\text{base}}(X_t|X_{<t})P_\theta(Z_t = 0|X_{<t}) + P_{\text{asst}}(X_t|X_{<t})P_\theta(Z_t = 1|X_{<t})$$

Llama-7B

Frozen

Expert Llama-70B

Collaboration Control

Run scoring on the training data and create a static dataset

Two model case

$$\text{NLL Loss} = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$

$$P(X_t|X_{<t}) = P_{\text{base}}(X_t|X_{<t})P_\theta(Z_t = 0|X_{<t}) + P_{\text{asst}}(X_t|X_{<t})P_\theta(Z_t = 1|X_{<t})$$

Frozen

If closer to 1, the loss becomes a typical NLL Loss for the base model.

Two model case

$$\text{NLL Loss} = -\sum_{t=1}^{T} \log P(X_t|X_{<t})$$

$$P(X_t|X_{<t}) = P_{\text{base}}(X_t|X_{<t})P_\theta(Z_t = 0|X_{<t}) + P_{\text{asst}}(X_t|X_{<t})P_\theta(Z_t = 1|X_{<t})$$

Llama-7B

Frozen

Expert Llama-70B

Classifying based on the hidden state from the last layer of base model.

Decoding

Greedy Decoding

$$\hat{Z}_t = \mathbf{1}[P_\theta(Z_t = 1 | X_{<t}) > \eta]$$

A fixed threshold deciding whether to switch the other model.

Experiments & Results

1. **Collaboration across domains**

2. **Collaboration across scales**

3. **Collaboration across architectures**

Training on Math and Reasoning Tasks

0-shot + CoT

	GSM	MATH
LLEMMA-7B	4.0	2.0
LLEMMA-34B	14.5	6.3
Finetuned LLAMA-7B	34.5	7.6
Finetuned LLAMA-70B (QLoRA)	52.5	11.7
PT (LLEMMA-34B + LLAMA-7B)	30.0	20.9
PT (LLEMMA-34B + LLEMMA-7B)	58.5	23.7
Co-LLM-7B + LLEMMA-7B	40.0	17.2
Co-LLM-7B + LLEMMA-34B	43.5	24.5

Collaboration across domains

Math and reasoning tasks	GSM	MATH
LLEMMA-7B	4.0	2.0
LLEMMA-34B	14.5	6.3
Finetuned LLAMA-7B	34.5	7.6
Finetuned LLAMA-70B (QLoRA)	52.5	11.7
PT (LLEMMA-34B + LLAMA-7B)	30.0	20.9
PT (LLEMMA-34B + LLEMMA-7B)	58.5	23.7
CO-LLM-7B + LLEMMA-7B	40.0	17.2
CO-LLM-7B + LLEMMA-34B	43.5	24.5

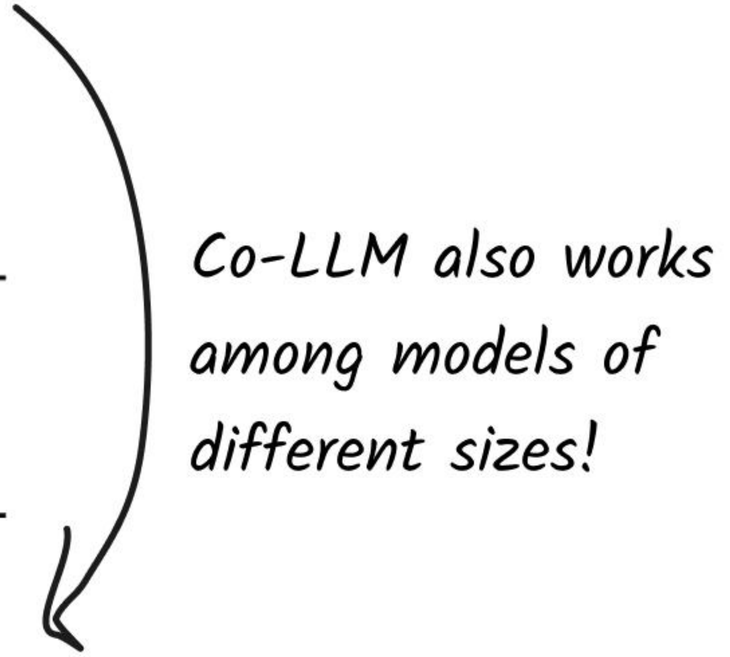
Collaboration across domains

Math and reasoning tasks	GSM	MATH
LLEMMA-7B	4.0	2.0
LLEMMA-34B	14.5	6.3
Finetuned LLAMA-7B	34.5	7.6
Finetuned LLAMA-70B (QLoRA)	52.5	11.7
PT (LLEMMA-34B + LLAMA-7B)	30.0	20.9
PT (LLEMMA-34B + LLEMMA-7B)	58.5	23.7
Co-LLM-7B + LLEMMA-7B	40.0	17.2
Co-LLM-7B + LLEMMA-34B	43.5	24.5

Co-LLM learns to leverage the math expert model and improves the performance compared to using either of them alone.

Collaboration across scale

Math and reasoning tasks	GSM	MATH
LLEMMA-7B	4.0	2.0
LLEMMA-34B	14.5	6.3
Finetuned LLAMA-7B	34.5	7.6
Finetuned LLAMA-70B (QLoRA)	52.5	11.7
PT (LLEMMA-34B + LLAMA-7B)	30.0	20.9
PT (LLEMMA-34B + LLEMMA-7B)	58.5	23.7
CO-LLM-7B + LLEMMA-7B	40.0	17.2
CO-LLM-7B + LLEMMA-34B	43.5	24.5



Comparing against proxy tuning

Math and reasoning tasks	GSM	MATH
LLEMMA-7B	4.0	2.0
LLEMMA-34B	14.5	6.3
Finetuned LLAMA-7B	34.5	7.6
Finetuned LLAMA-70B (QLoRA)	52.5	11.7
PT (LLEMMA-34B + LLAMA-7B)	30.0	20.9
PT (LLEMMA-34B + LLEMMA-7B)	58.5	23.7
Co-LLM-7B + LLEMMA-7B	40.0	17.2
Co-LLM-7B + LLEMMA-34B	43.5	24.5

Compare to Proxy Tuning (PT), Co-LLM can effectively collaborate with models trained differently.

Experiments & Results

1. **Collaboration across domains**
2. **Collaboration across scales**

We have more on the biomedical domain + collaboration between Llama-7B and 70B models in the paper.

Collaboration across architectures

Math and reasoning tasks	**GSM**	**MATH**
MISTRAL-7B	21.5	7.2
MIXTRAL-8×7B (MoE)	38.5	16.2
Finetuned MISTRAL-7B	51.0	13.9
Co-LLM MISTRAL-7B + MIXTRAL-8×7B	**57.0**	**20.0**

Experiments & Results

❶ Collaboration across domains

❷ Collaboration across scales

We have more on the biomedical domain + collaboration between Llama-7B and 70B models in the paper.

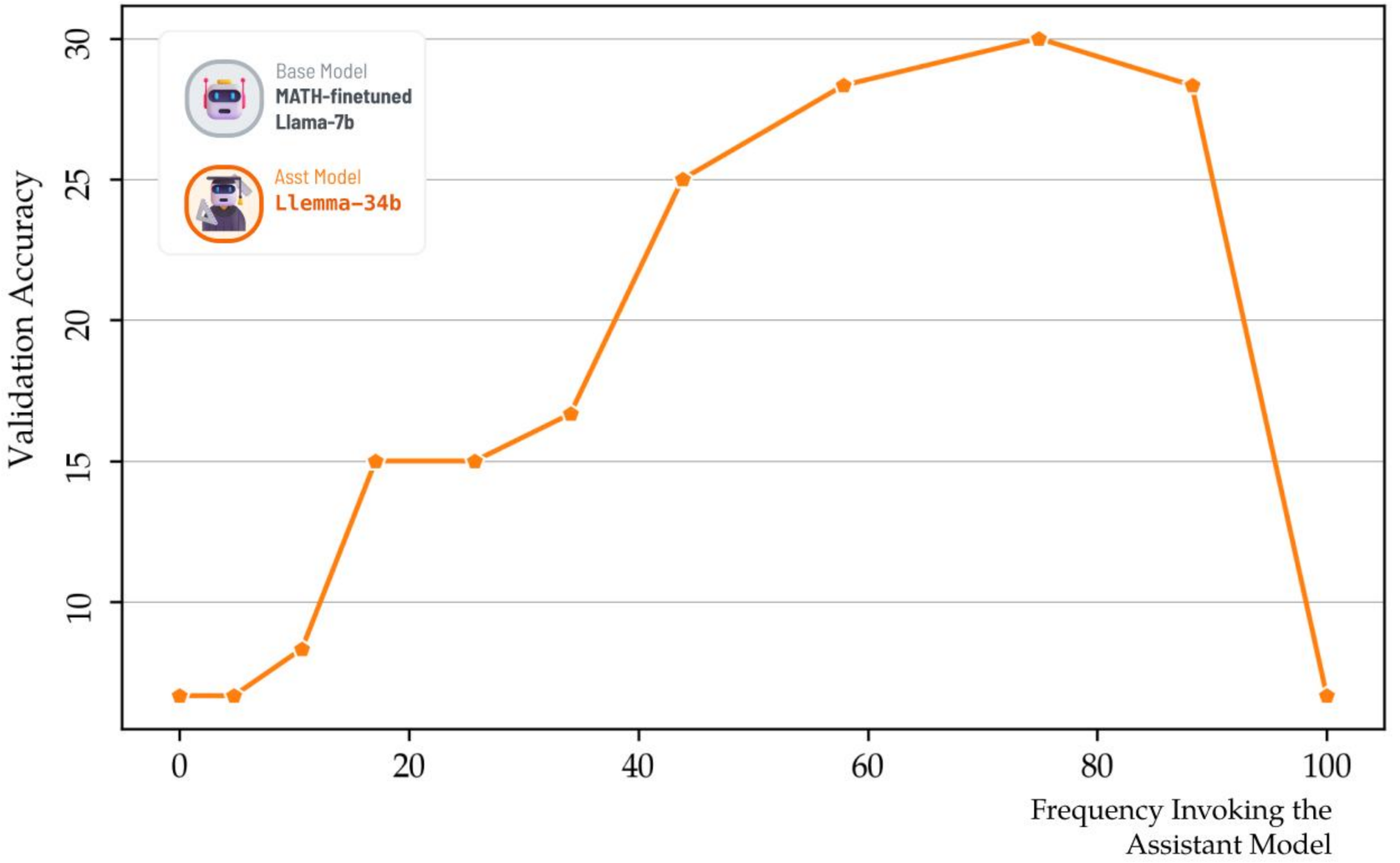
❸ Collaboration across architectures

We show that it works for collaboration between MoE and regular dense LLMs.

Ablation on deferral frequency

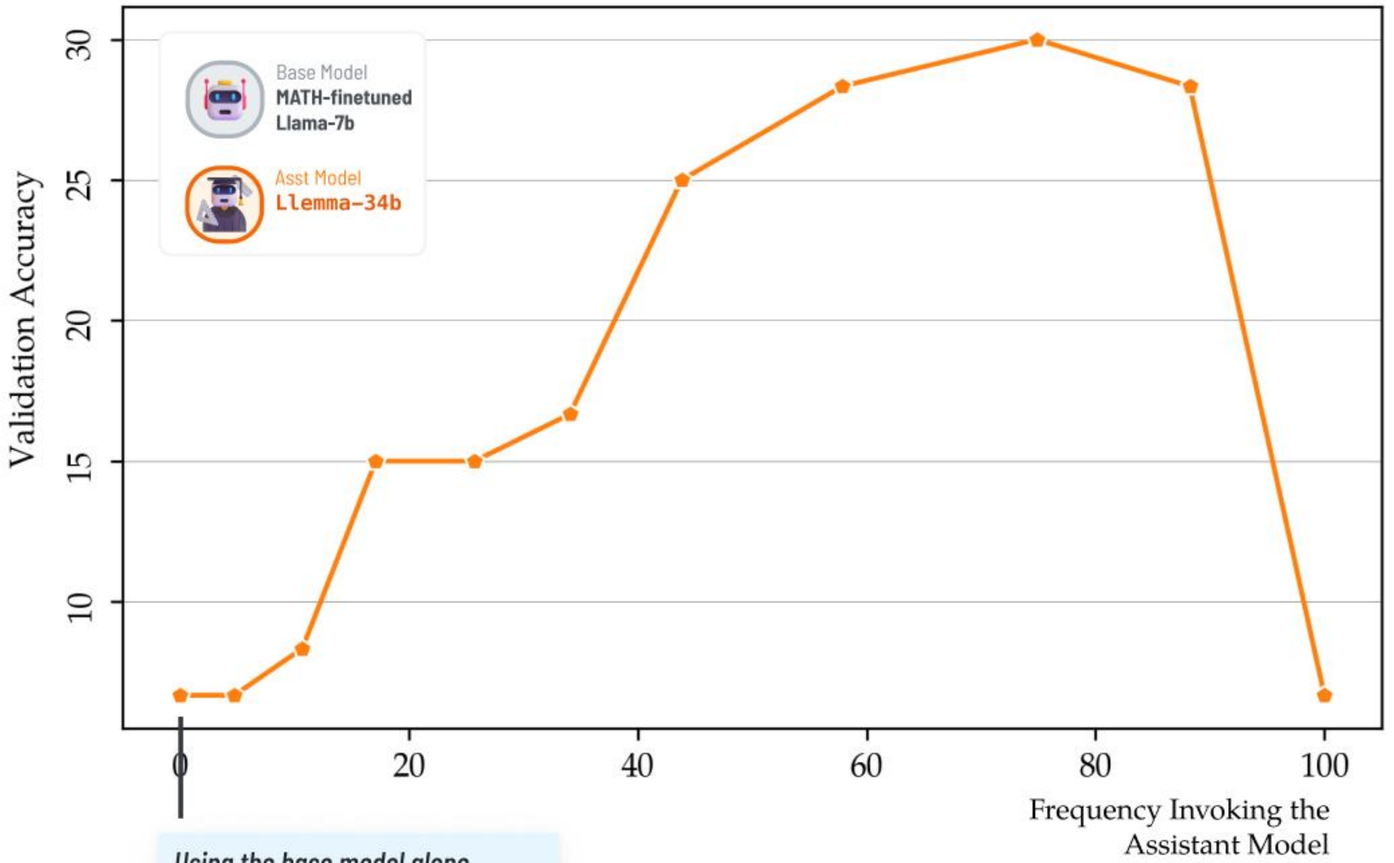
Accuracy of Co-LLM generations on the MATH Dataset

Example shown is the model generation responding to this question: Evaluate the expression $a^3 \cdot a^2$ if $a = 5$. The correct answer is 3125.



Accuracy of Co-LLM generations on the MATH Dataset

Example shown is the model generation responding to this question:
Evaluate the expression $a^3 \cdot a^2$ if $a = 5$. The correct answer is 3125.

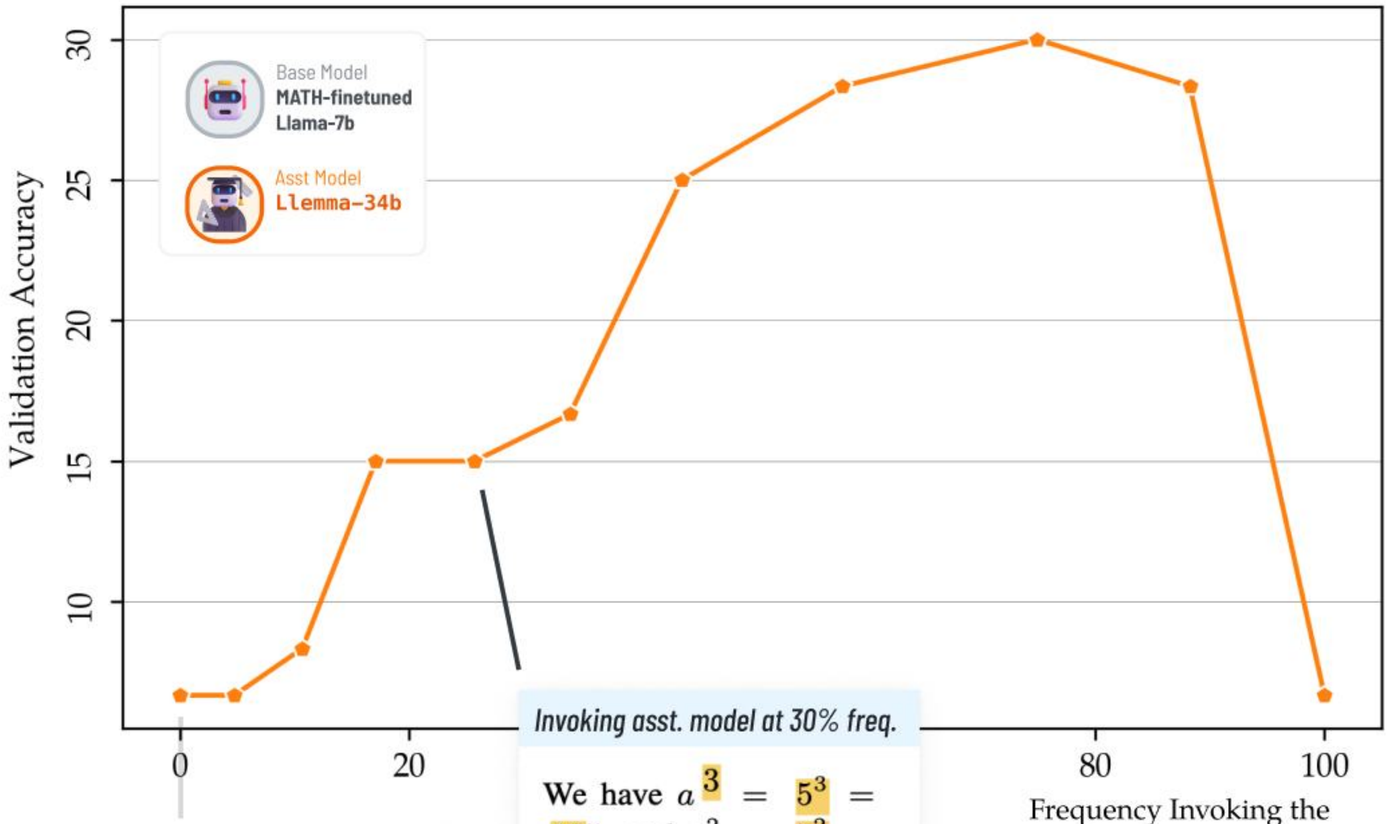


Using the base model alone

We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = \boxed{125}$. The final answer is 125.

Accuracy of Co-LLM generations on the MATH Dataset

Example shown is the model generation responding to this question: Evaluate the expression $a^3 \cdot a^2$ if $a = 5$. The correct answer is 3125.



Invoking asst. model at 30% freq.

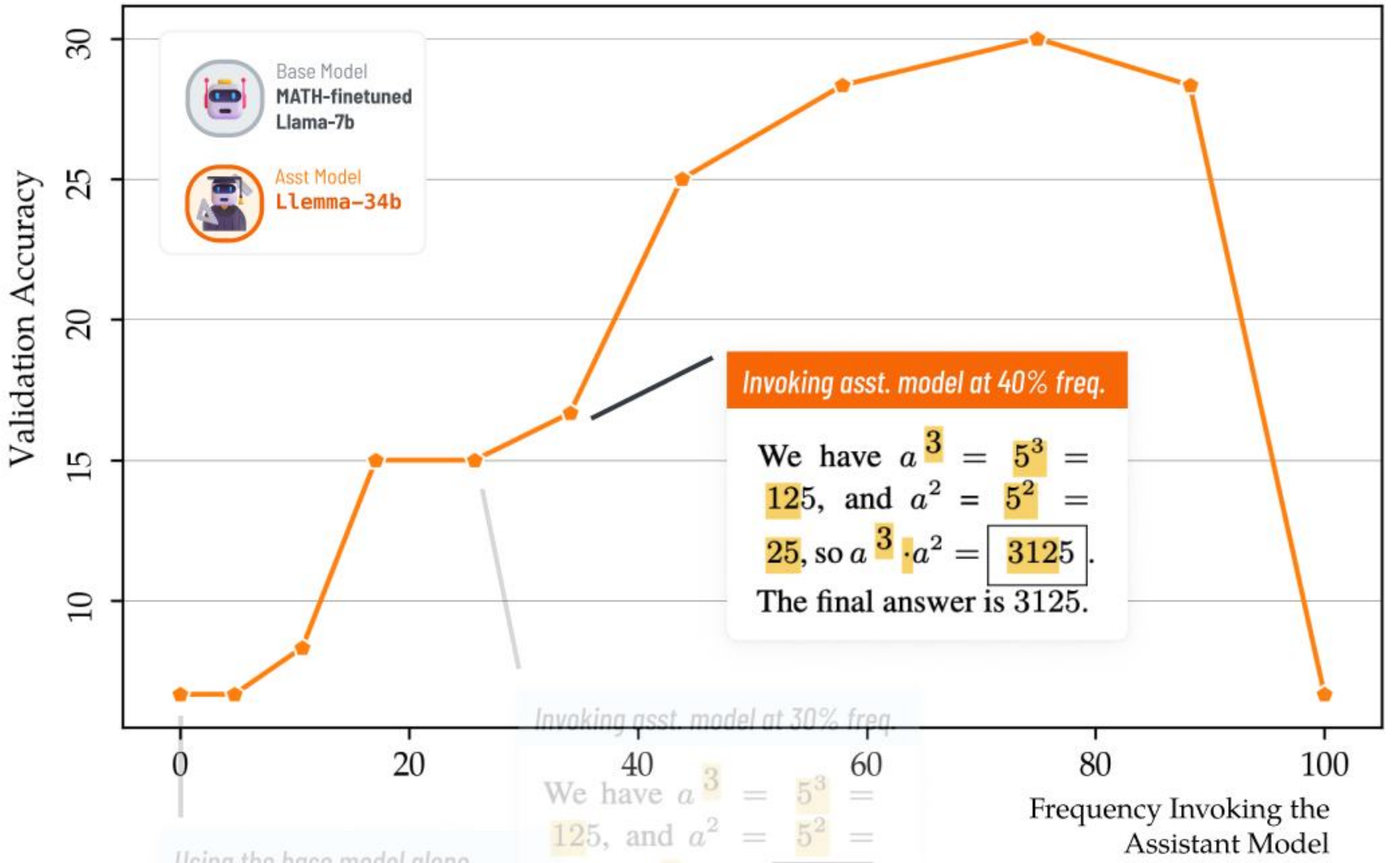
We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = \boxed{3000}$.
The final answer is 3000.

Using the base model alone

We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = \boxed{125}$. The final answer is 125.

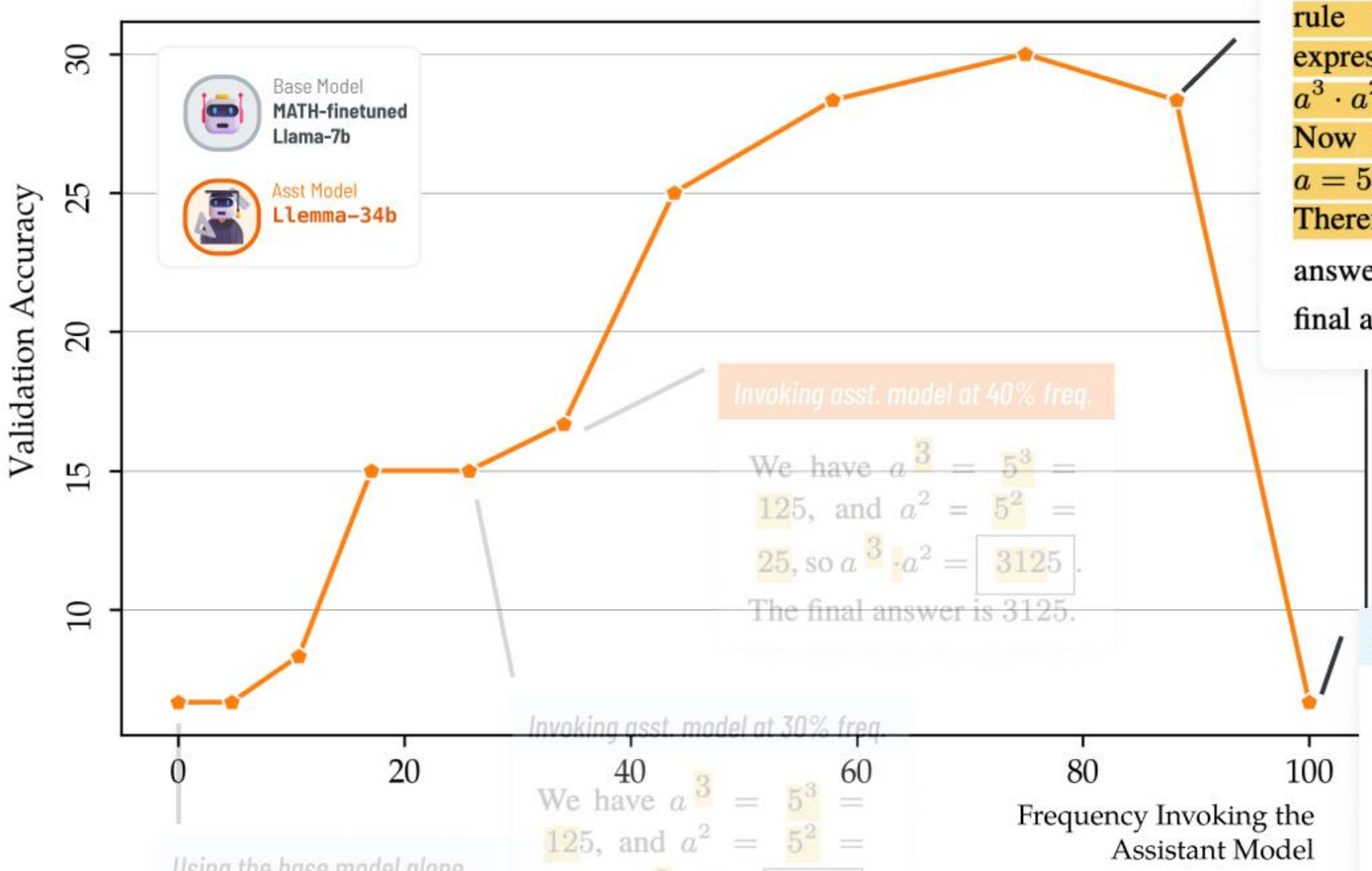
Accuracy of Co-LLM generations on the MATH Dataset

Example shown is the model generation responding to this question: Evaluate the expression $a^3 \cdot a^2$ if $a = 5$. The correct answer is 3125.



Accuracy of Co-LLM generations on the MATH Dataset

Example shown is the model generation responding to this question: Evaluate the expression $a^3 \cdot a^2$ if $a = 5$. The correct answer is 3125.



Invoking asst. model at 90% freq.

We can use the power rule to simplify this expression. We have that $a^3 \cdot a^2 = a^{3+2} = a^5$. Now we can substitute $a = 5$ to get $5^5 = 3125$. Therefore, the final answer is $\boxed{3125}$. The final answer is 3125.

Invoking asst. model at 40% freq.

We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = \boxed{3125}$. The final answer is 3125.

Invoking asst. model at 30% freq.

We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = \boxed{3000}$. The final answer is 3000.

Using the base model alone

We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = \boxed{125}$. The final answer is 125.

Using the asst model alone

1 Given a mathematics problem, determine the answer. Simplify your answer as much as possible. You can use latex to format your answer and you should state your final answer as "The final answer is $(final - answer)$." Problem:

Co-LLM teaches LLMs to **collaborate** by interleaving their generations **at the token level.**


Shannon Shen


Hunter Lang


Bailin Wang


Yoon Kim


David Sontag

Check our github repo

github.com/clinicalml/co-llm

User Prompt (Knowledge Seeking Question)

What are some species of bears that are now extinct?

Collaborative Response

`Instruct-finetuned Llama-7b + Llama-70b`

There have been many species of bears that have become extinct throughout history. Some examples include:

1. **The giant short-faced bear** (**Arctodus simus**) - This species went extinct around 10,000 years ago and was one of the largest bears to have ever lived.

2. The **cave bear** (**Ursus spelaeus**) - This species became extinct around **25**,000 years ago **and** was well-known for its long**, curved cla**ws and large size.[...]

User Prompt (Math Question)

Consider the function $g(x)=3x-4$. What is $g(0)$?

Collaborative Response

`MATH-finetuned Llama-7b + Llemma-34b`

We have $**g(0)=3(0)-4**=\boxed**{-4**}$. The final answer is -4.