

Learning to Decode Collaboratively with Multiple Language Models

ACL 2024

Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, David Sontag

Massachusetts Institute of Technology

Background

- **Combine the generations of multiple LLMs at decoding time**
 - Benefit
 - Faster decoding speed & More controllable generation
 - More coherent and less repetitive text generation
 - Previous Work: Speculative Decoding (Leviathan et al., 2023)
 - Enable a large model to be “tuned” by combining its generations with those of a smaller model from the same family

[START] japan : s benchmark bond n

[START] japan : s benchmark nikkei 22 75

[START] japan : s benchmark nikkei 225 index rose 22 76

[START] japan : s benchmark nikkei 225 index rose 226 . 69 7 points

[START] japan : s benchmark nikkei 225 index rose 226 . 69 points , or 0 1

[START] japan : s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 9859

[START] japan : s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 7 in

[START] japan : s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in tokyo late

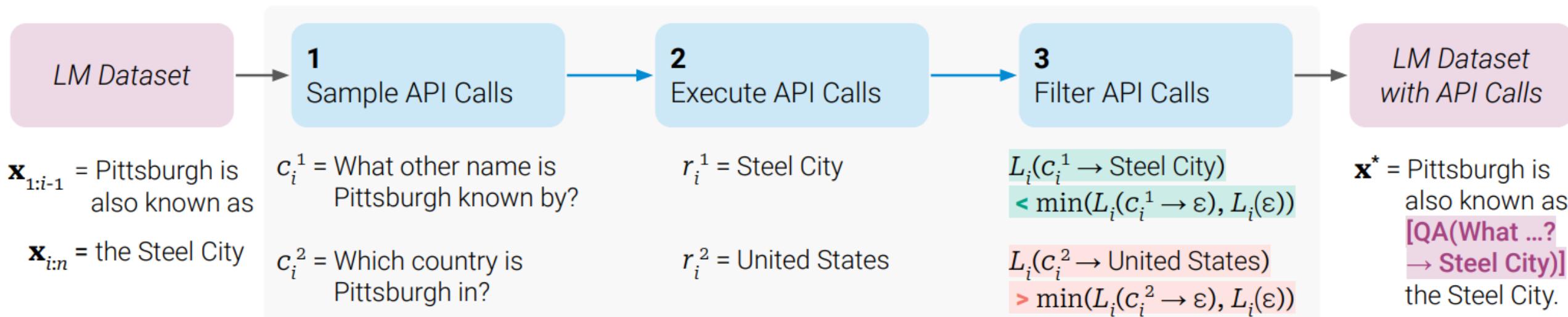
[START] japan : s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]

- **Green:** Suggestion made by approximation model (6M) that target model (97M)
- **Red:** Rejected suggestion
- **Blue:** Correction

Background

- **Equip LLM with ability to infuse external tools into generation**
 - Goal of incorporating outside knowledge and capabilities
 - External API
 - Search engines or retrievers
 - Code executors

Toolformer (Schick et al., 2023)



Introduction

- **Motive from related work**
 - With the goal of incorporating outside knowledge and capabilities
 - Prescription on how to combine the models and when to use the tools
 - Specific formulas for combining the logits of multiple models
 - (weak) supervision on where to insert tool/API calls in the training data

Introduction

- **Co-LLM: Learning to Decode Collaboratively with LMs**
 - Combine the generations of multiple large language models (LLMs) at decoding time
 - Explore a different type of model combination
 - Models learn to interleave their generations token-by-token
 - Each token is generated by one model, so the models collaborate to generate a token sequence together
 - Generates next token as latent variable, assuming no direct supervision at each decoding step
 - Effective collaboration pattern for a given task to be learned organically from data

Introduction

• Co-LLM: Learning to Decode Collaboratively with LMs



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 claws and large size.[...]

Method

- Use the larger model as an assistant allows the smaller model to make effective use of a larger knowledge base
- Focus its efforts on learning the correct “scaffolding” for instruction responses

Example generations of method

- Base Model generate the answer template
- Base Model (Smaller LLAMA) use Assistant (a larger LLAMA) to fill in factual knowledge

Introduction

• Co-LLM: Learning to Decode Collaboratively with LMs



User Prompt (Math Question)

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



Collaborative Response

MATH-finetuned Llama-7b + Lemma-34b

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

Method

- LLEMMA-34B as an API call to fill in parts of a LaTeX formula
- Model predicts when to call the assistant by itself
- Behavior it learns from training without direct supervision

Example generations of method

- Assistant is Math-specialized model as an "API"
- Assistant model generated the highlighted tokens
- Base model learned to defer generation at those locations

Latent-variable Framework for Collaborative Generation

- **Problem Definition**

- Given a set of LMs with different expertise or sizes,
propose a latent-variable framework that enables collaboration in a cost-efficient way
- Base model
 - Finetunable relatively small LM
- Assistant model
 - Typically larger and/or more specialized model
- Defers generation for the next token
 - Base model calls on an assistant to generate the next token

Latent-variable Framework for Collaborative Generation

- **Problem Definition**

- Defers generation for the next token
 - Base model calls on an assistant to generate the next token

- Sequence of tokens

$$(X_1, \dots, X_T)$$

- Discrete latent variable

$$Z_t \in \{0, 1, \dots, M\}$$

- Base model

$$i = 0$$

- Assistant model

$$i \in \{1, \dots, M\}$$

- Conditional distribution

$$P_i(X_t | X_{<t}), i \in \{1, \dots, M\}$$

- Joint sequence-level likelihood

$$P(X, Z) = \prod_{t=1}^T \left(P_\theta(Z_t | X_{<t}) P_{z_t}(X_t | X_{<t}) \right)$$

Latent-variable Framework for Collaborative Generation

- **Joint sequence-level likelihood**

$$P(X, Z) = \prod_{t=1}^T \left(P_\theta(Z_t | X_{<t}) P_{z_t}(X_t | X_{<t}) \right)$$

- Model generates token $\textcolor{brown}{X}_t$
- The token-level discrete decision Z_t
- Discrete latent variable $Z_t \in \{0, 1, \dots, M\}$
- Motive of Latent variable Z_t
 - Same spirit as the defer variable in Mozannar and Sontag (2020)
Classical Mixture-of-Experts models (Jordan and Jacobs, 1994)
Ensemble models (Saunders et al., 2019)

Latent-variable Framework for Collaborative Generation

- **Training Stage: Unsupervised Learning**

Joint sequence-level likelihood $P(X, Z) = \prod_{t=1}^T \left(P_\theta(Z_t | X_{<t}) P_{z_t}(X_t | X_{<t}) \right)$

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

- Token level decisions Z_t are unknown
 - Collecting human annotation is not scalable
 - A natural way of handling the issue with unsupervised learning
- Optimize the following marginal likelihood
 - Can be computed efficiently during training due to the conditional independence structure

Latent-variable Framework for Collaborative Generation

- Inference Stage: Collaborative Decoding**

- Find the best sequence X along with the best decision Z on which assistant LM to use

- Greedy decoding

Determine which model to decode from

Decode greedily from that model

$$\hat{X}, \hat{Z} = \arg \max_{X, Z} P(X, Z)$$

$$\hat{Z}_t = \arg \max_{Z_t} P_\theta(Z_t | X_{<t})$$

$$\hat{X}_t = \arg \max_{X_t} P_{\hat{Z}_t}(X_t | X_{<t})$$

Marginal likelihood training objective

Alternative strategy for inference

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

- Require calling all $M + 1$ models every token
- Slow down decoding

$$\hat{X}_t = \arg \max_{X_t} \sum_{Z_t} P_{Z_t}(X_t | X_{<t}) P_\theta(Z_t | X_{<t})$$

- Greedily choose Z_t performs well
- Interpretable collaboration, since each token is generated by a single model

Latent-variable Framework for Collaborative Generation

- **Question: Probabilistic modeling**

Marginal likelihood training objective

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

Alternative strategy for inference

$$\hat{X}_t = \arg \max_{X_t} \sum_{Z_t} P_{z_t}(X_t|X_{<t}) P_\theta(Z_t|X_{<t})$$

- How should we parameterize P_θ so that it can be learned in a data-efficient way?
- Can the base model learn how to cooperate with (larger or domain-specific) assistant models with only access to conditional probabilities (no internal hidden states or weights)?
- What kind of collaboration pattern can be induced if the collaboration is learned without supervision?

- **Answer: Latent-variable Framework**

- Answer these questions via the exposed interpretable variable Z_t

Co-LLM: Learning to Decode Collaboratively with LMs

- **Basic Case**

- Only have one base model and one assistant model
- Base model P_{base}
- Assistant model P_{asst}
- Make binary decision $Z_t \in \{0, 1\}$ of whether to generate from itself or defer to the assistant model

- **Explain Our Design**

- Modeling: Design for the parameterization of the model selector
- Training and inference procedure of the joint model

Co-LLM: Learning to Decode Collaboratively with LMs

- **Modeling: Model Selector** : $P_\theta(Z_t|X_{<t})$

- Represent θ as a linear binary classification head in the last layer

$$h_t(X_{<t}) \in \mathbb{R}^d$$

- Base model's last hidden state at time step t for inputs $X_{<t}$
- Introduce only d new parameters to the base model where d is the base model's hidden dimension size

$$P_\theta(Z_t|X_{<t}) = \sigma(\langle \theta, h_t(X_{<t}) \rangle)$$

- Sigmoid function $\sigma(\cdot)$
- The weight vector $\theta \in \mathbb{R}^d$

Co-LLM: Learning to Decode Collaboratively with LMs

- **Training Stage**

Standard negative log-marginal likelihood

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

- Align well with the typical pretrained objective of maximizing the next token probs

Likelihood of the next token after marginalizing out the latent variable Z_t

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

- Updates both θ and the base model
 - Forward Pass: P_{base} P_{asst} / Backward Pass: P_{base}
- Offloading “Difficult tokens” can potentially alleviate hallucination issues of the base model
 - Lead to better generalization even without much help from the assistant model

Co-LLM: Learning to Decode Collaboratively with LMs

- **Training Stage: Initialization of θ**

Set the pseudo labels for Z_t to initialize the parameters

$$\hat{Z}_t := \mathbb{1}[X_t = \arg \max_{v \in \mathcal{V}} P_{\text{asst}}(v | X_{<t}) \wedge X_t \neq \arg \max_{v \in \mathcal{V}} P_{\text{base}}(v | X_{<t})]$$

- The prime location for $Z_t = 1$ is when assistant model correctly predicts target token X_t but the base model does not

Initialize the d parameters θ by maximizing the likelihood of $P_\theta(\hat{Z}_t | X_{<t})$

- Hold the rest of the base model fixed

Experimental Result

- Lead to better performance by enabling base model to learn better Z_t values from data

Co-LLM: Learning to Decode Collaboratively with LMs

- **Inference Stage**

- Performance of the joint model is sensitive to the choice of Z_t in latent-variable
- Impose extra priors over Z_t for better performance
- Greedy decoding strategy : Set a threshold η for decoding Z_t
 - When $P_\theta(Z_t = 1) > \eta$, execute the assistant model to predict the next token

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

- η is picked via grid search on a small validation set per dataset
- The choice of threshold for the decoding probability
 - Allow for inference-time control over the amount of collaboration
 - Our performance degrades as threshold increases, in contrast with DExperts (Liu et al., 2021)

Experiment

- Model answers with different rates of deferral to the question
 - Optimal deferral threshold η

Algorithm 1: Find Optimal Deferral Th. η

Input : Base Model, Asst. Model, Model Selector ϕ , Validation Dataset \mathcal{D}

Let $\mathcal{P} = \{\}$

for $i = 0$ **to** $|\mathcal{D}|$ **do**

 Given the input prompt in \mathcal{D}_i , generate response $X^{(i)}$ using the base model

 Predict per-token deferral probability

$P_\phi(Z_t^{(i)} = 1 \mid X_{<t}^{(i)})$, and append it to \mathcal{P}

end

Sort \mathcal{P} in ascending order

Set current best threshold $\eta = 0$ and evaluation score $s = 0$

for $j = 0$ **to** 100 **by** 10 **do**

 Get the j -th quantile p_j in \mathcal{P} and use it as the deferral threshold η_j

 Generate responses $X^{(i)}$ for i in from 0 to $|\mathcal{D}|$ using the base and asst. model controlled by $P_\phi(Z_t^{(i)} = 1 \mid X_{<t}^{(i)}) > \eta_j$

 Evaluate the responses, and if the evaluation score is better than s , set $\eta = \eta_j$ and s to the new score

end

Return η

Experiment

- **Model Used**

- Llemma family (Azerbayev et al., 2023, finetuned for math and reasoning)
- The MEDITRON family (Chen et al., 2023, finetuned for biomedicine)
- The 7B and the 70B model from the same LLAMA2 family (Touvron et al., 2023) as the base and assistant model

- **Datasets**

- Instruction following
 - Full Tülu v2 mix data
- Reasoning and math problem solving
 - GSM8k, MATH
- Medical question answering
 - The BioASQ

Dataset Size

Dataset	# of Samples		
	Train	Dev	Test
Tülu-v2-mixture	326,115		
MT-Bench		24	
AlpacaEval			805
GSM8k	7,473	50	200
MATH	7,498	60	750
BioASQ	4,719	20	310
	(Factiod)		93
	(List)		61
	(Summary)		75
	(Yes/No)		81

Experiment

- **Baseline**
 - Single models
 - **0-shot performance** (For 70B models, we fine-tune using QLoRA)
 - Other collaborative models
 - **Contrastive Decoding (Li et al., 2022)**
 - **Proxy Tuning (Liu et al., 2024)**
 - Ablated Co-LLM
 - **Two simple heuristics as model selectors**
 - **Random:** Randomly chooses one of models P_θ produce a token with probability $p = 0.5$
 - **Greedy:** Selects the token with the higher probability for either model
 - Weakly-supervised Co-LLM
 - **It's inspired by process used to derive tool-use labels in Toolformer (Schick et al., 2023)**
 - $P_\theta(Z_t|X_{ trains the base model to defer to the assistant when pseudo-labels $\hat{Z}_t = 1$$

Experiment

- **Collaboration across domains**
 - Collaboration improves performance compared to the individual models themselves
 - Co-LLM with LLAMA-7B & 34B can achieve similar performance as fine-tuned LLEMMA-7B on GSM8k & MATH
- **Comparison against Proxy Tuning**
 - More effective at enabling collaboration between models from different domains
 - Requires more calls to the larger model, thus resulting in slower inference

<i>Math and reasoning tasks</i>	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

<i>BioASQ tasks</i>	Factoid	List	Yes/No	Summ.	Avg.
MEDITRON-7B	0.00	2.7	70.4	18.6	22.9
MEDITRON-70B	17.2	16.1	80.2	21.1	33.7
Finetuned LLAMA-7B	23.7	13.8	76.5	18.1	33.0
Finetuned LLAMA-70B (QLoRA)	24.7	20.7	75.3	21.1	35.5
PT (MEDITRON-70B + LLAMA-7B)	26.9	10.7	80.2	7.3	31.3
PT (MEDITRON-70B + MEDITRON-7B)	26.9	23.5	82.7	11.0	35.6
Co-LLM-7B + MEDITRON-7B	17.2	16.0	72.8	19.8	31.4
Co-LLM-7B + MEDITRON-70B	21.5	18.6	81.5	20.6	35.6

Experiment

- **Collaboration across scales**

- Co-LLM is better than the QLoRA finetuned assistant model (on MATH and BioASQ)
- Randomly switching between models lead to worse performances than Co-Random
- Co-Greedy does not consistently yield better performance

		AlpacaEval	GSM	MATH	BioASQ^a			
		(% Win)	(Acc.)	(EM)	Factoid (SAcc.)	List (F1)	Yes/No (Acc.)	Summ. (R2)
<i>Untuned</i>	LLAMA-7B	-	7.0	0.3	4.3	4.9	71.6	17.2
	LLAMA-70B	11.6	13.5	2.1	11.8	14.9	77.8	18.6
	LLAMA-70B+7B (CD)	-	11.5	1.3	11.8	9.0	71.6	17.5
<i>Finetuned</i>	LLAMA-7B (Finetuned)	69.3	34.5	7.6	23.7	13.8	76.5	18.1
	LLAMA-70B (QLoRA)	78.6 ^b	52.5	11.7	24.7	20.7	75.3	21.1
	LLAMA-70B+7B (PT)	72.3	52.5	17.3	29.0	16.8	85.2	21.3
<i>Collaboration</i>	Co-Random	46.3	17.0	6.1	6.5	1.9	30.9	17.5
	Co-Greedy	64.1	38.0	8.1	29.0	16.6	76.5	20.2
	Weak Supervision	56.7	40.0	12.3	22.6	14.6	80.2	17.5
	Co-LLM-7B (Base Only)	70.6	33.0	6.4	20.4	11.2	79.0	18.1
	Co-LLM-7B + LLAMA-70B	71.9	45.0	15.1	24.7	18.0	82.7	20.4

Experiment

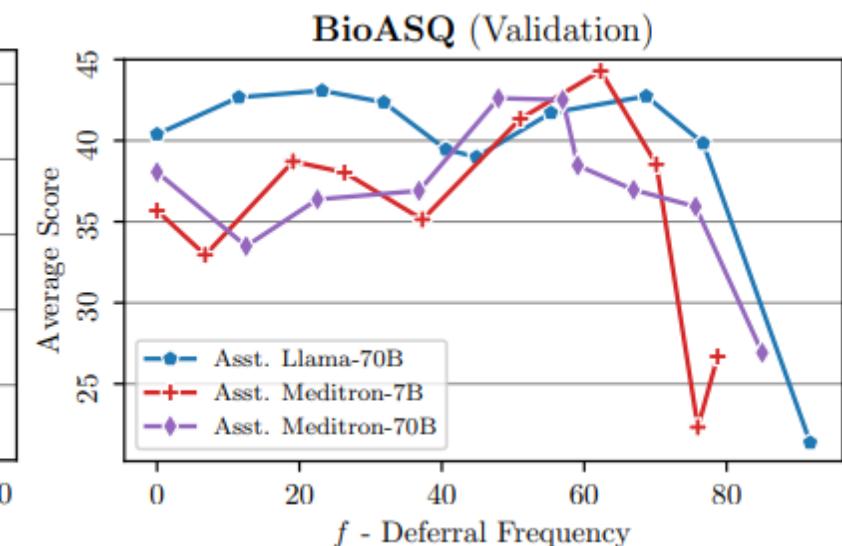
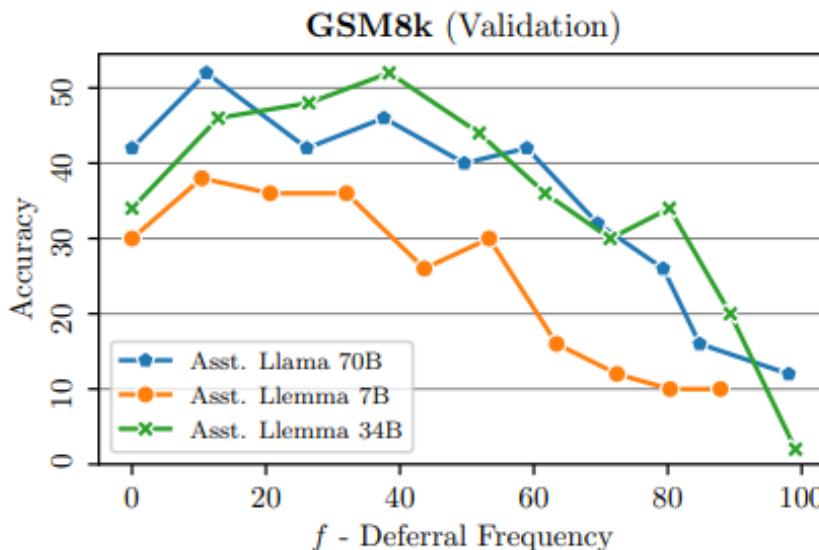
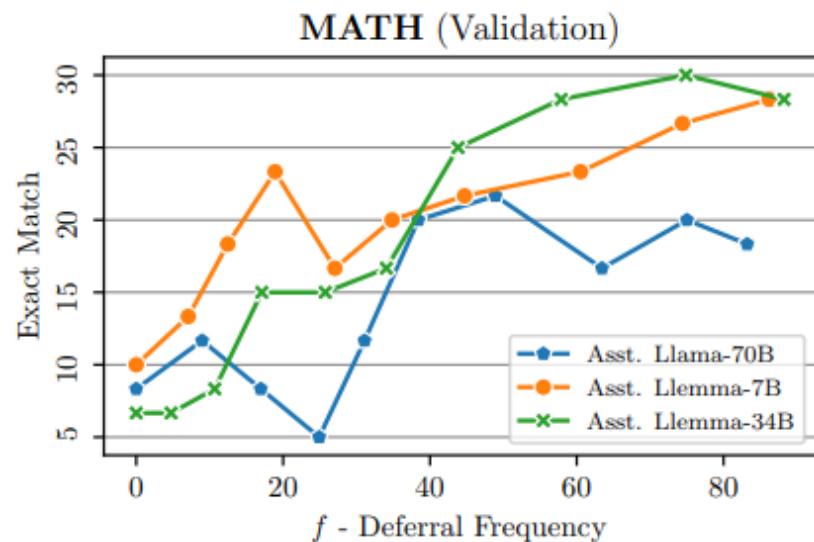
- **Collaboration across architectures**
 - Collaborate between dense model (MISTRAL-7B) and sparse model (MIXTRAL-8 \times 7B)
 - Joint model achieves strong accuracy gains compared to either the finetuned MISTRAL
- **Comparison with Mixture of Experts**

	Mixture of Experts (MoE)	Co-LLM
Inference Time	Train a model that can be partially executed	Combine the generations of multiple models
Size & Architecture	Subnetworks: The same size & architecture	Flexible collaboration between models of different sizes or even architectures
Gradient Access	Required	Not Required: Only fine-tune the base model

	<i>Math and reasoning tasks</i>	
	GSM	MATH
MISTRAL-7B	21.5	7.2
MIXTRAL-8 \times 7B (MoE)	38.5	16.2
Finetuned MISTRAL-7B	51.0	13.9
Co-LLM MISTRAL-7B + MIXTRAL-8 \times 7B	57.0	20.0

Experiment

- Model answers with different rates of deferral to the question
 - Optimal deferral threshold η / deferral frequency f



Experiment

- Collaborative Decoding



- Deferral control
 - Probability of switching to the assistant model to decode the next token given the context
 - When the probability is above some threshold η (indicated by A)
 - Use the decoded token as the context (highlighted with orange border)

Experiment

- Model answers with different rates of deferral to the question**

- Increase the frequency of deferral by lowering the threshold η

$f = 0$ It fails to produce a valid answer 125

$f = 0.4$ It is not tuned or aligned, it produces no helpful solutions

$f = 1$ Starts to invoke assistant model to generate latex code & compute results

Question:
Evaluate the expression $a^3 \cdot a^2$ if $a = 5$

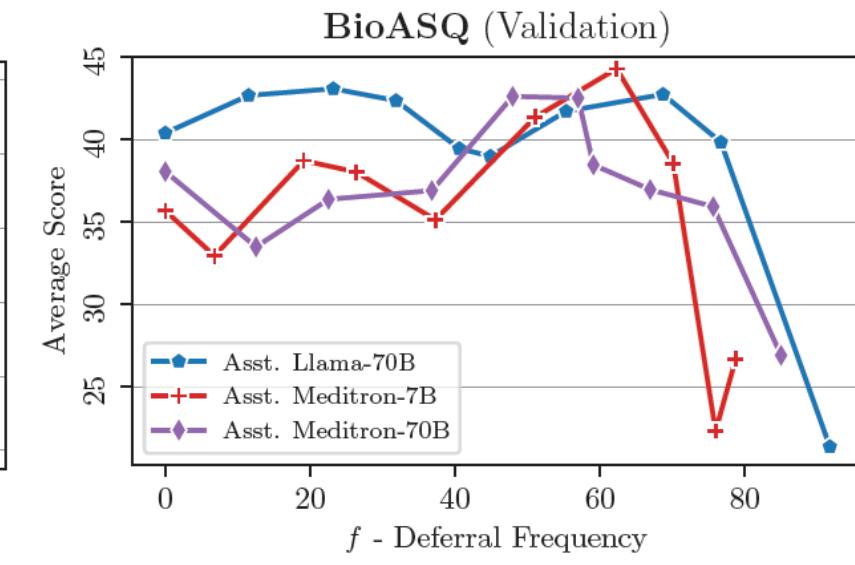
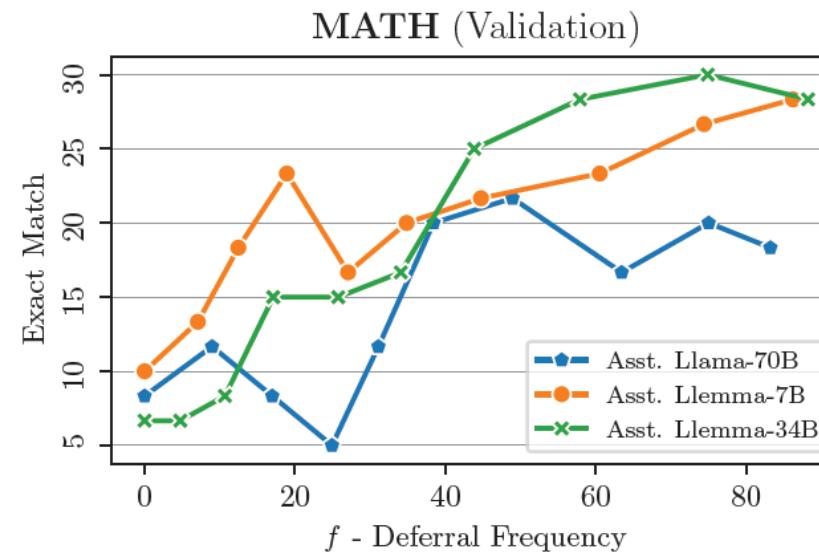
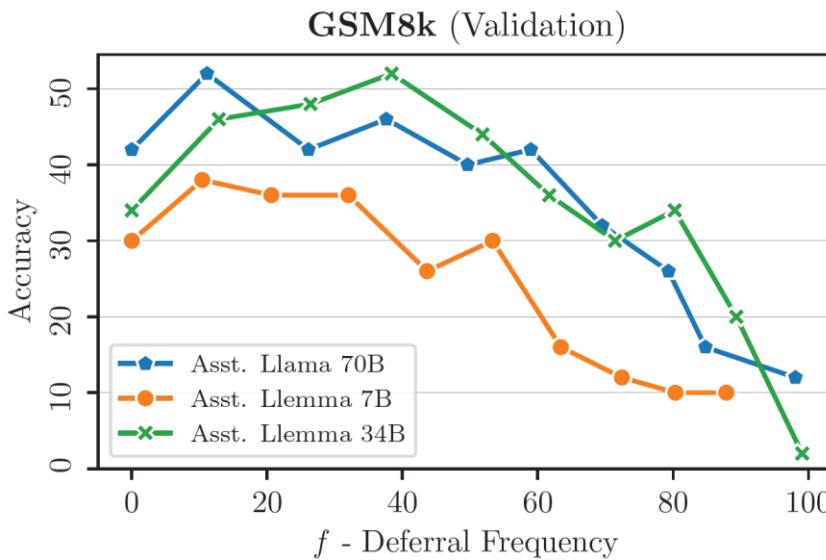
Answer:
3125

Deferral frequency $f = 0$	$f = 0.3$	$f = 0.4$	$f = 0.9$	$f = 1.0$
We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = 125$. The final answer is 125.	We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = 3000$. The final answer is 3000.	We have $a^3 = 5^3 = 125$, and $a^2 = 5^2 = 25$, so $a^3 \cdot a^2 = 3125$. The final answer is 3125.	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 3125. The final answer is 3125.	### 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:

Experiment

- **Model answers with different rates of deferral to the question**
 - Similar concave performance curve across different domains and scales
 - Optima vary across datasets, corresponding to different patterns of assistant models (e.g., API-call or “leading” style)
 - Pick the proper to balance the accuracy and the efficiency/cost

Performance of Co-LLM at different frequencies of deferral



Experiment

- Model answers with different rates of deferral to the question
 - Optimal deferral threshold η_* / deferral frequency f
 - Max number of generated token N

Co-LLM

	Base	Asst.	η_*	f
AlpacaEval $N = 2048$	LLAMA-7B	LLAMA-70B	0.80	0.1
GSM8k $N = 512$	LLAMA-7B	LLAMA-70B	0.17	0.1
	LLAMA-7B	LLEMMA-7B	0.08	0.1
	LLAMA-7B	LLEMMA-34B	0.05	0.3
MATH $N = 512$	MISTRAL-7B	MIXTRAL-8×7B	0.12	0.2
	LLAMA-7B	LLAMA-70B	0.57	0.6
	LLAMA-7B	LLEMMA-7B	0.05	0.9
	LLAMA-7B	LLEMMA-34B	0.30	0.8
BioASQ $N = 512$	MISTRAL-7B	MIXTRAL-8×7B	0.69	0.9
	LLAMA-7B	LLAMA-70B	0.38	0.2
	LLAMA-7B	MEDITRON-7B	0.07	0.5
	LLAMA-7B	MEDITRON-70B	0.20	0.5

Weakly-supervised Co-LLM

	Base	Asst.	η_*	f
AlpacaEval	LLAMA-7B	LLAMA-70B	0.11	0.1
GSM8k	LLAMA-7B	LLAMA-70B	1.00	0.0
MATH	LLAMA-7B	LLAMA-70B	0.44	0.1
BioASQ	LLAMA-7B	LLAMA-70B	1.00	0.0

Cocclusion

- **Co-LLM**
 - Teach multiple LLM to collaborate by interleaving their generations at the token level
 - Produce generations of better quality compared to using either of the models alone
 - Task: Math reasoning, Medical question answering, Instruction following
 - Latent ‘defer’ variable
 - Flexible and interpretable way:
Adjust the frequency for invoking other LMs at inference time without re-training
 - Future work
 - Integrate more than two LMs
 - Investigate potentially more complex collaboration strategies

Cocclusion

- **Limitations**

- Co-LLM thus requires picking the deferral threshold per task, which can be inconvenient in practice
 - However, the threshold also enables inference-time control over the amount of collaboration
- Assistant model may generate an identical token as the base model does
 - Not every deferral matters: for some position
 - Suggest the development of more fine-grained control of deferral strategies, potentially via more sophisticated modeling of deferral model parameters θ

Cocclusion

- **Limitations**

- Fully relying on an assistant model at some point in decoding
 - If the assistant model is not well-tuned or aligned, it may unintentionally break the generation due to occasional mistakes
 - One erroneous token might lead to a cascade of errors, causing repetition patterns or generating irrelevant content
- Future work is to develop a more robust deferral strategy
 - Allow backtracking when the assistant model fails to generate a proper response

Here's a recipe for Kubdari, a traditional Georgian dish:

Ingredients:

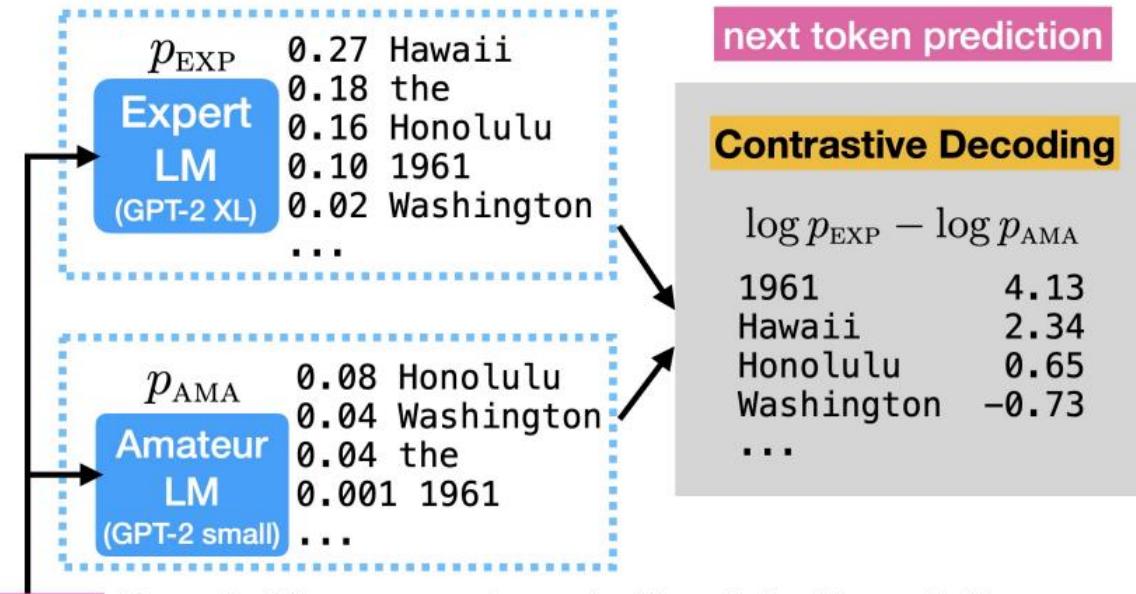
- * 1 lb ground beef
- * 1 onion, finely chopped
- * 2 cloves garlic, minced
- * 1 cup chopped parsley
- * 1 cup chopped cilantro
- * 1 cup chopped dill
- * 1 cup chopped ...

[...repeating the same pattern...]

Appendix

- **Contrastive Decoding (Li et al., 2022)**

- Generate fluent and lexically diverse text without compromising coherence
- Maximize difference between model log-probabilities
 - Plausibility constraints which restrict the search space to tokens with sufficiently high probability under the expert LM



Continuations:

Greedy: Hawaii. He was born in Hawaii. He was born in Hawaii...

Nucleus: Washington, D.C., to Barack Obama and Michelle Robinson...

CD: 1961 to a Kenyan father, Barack Hussein Obama and a mother of American descent, Stanley Ann Dunham...

Hawaii & Honolulu: Highly repetitive continuation

Washington: Incoherent continuation (0.02-> -0.73)

1961: Strongly preferred by CD (0.10 -> 4.13)

Appendix

- **Contrastive Decoding (Li et al., 2022)**

- Contrastive Decoding Objective
 - Factor out undesired behaviors highlighted by the smaller amateur LMs
 - Generate text from the remaining good behaviors of larger expert LMs
- Adaptive Plausibility Constraint
 - Hyperparameter in $[0, 1]$ that truncates the next token distribution of p_{EXP}
 - Larger α entails more aggressive truncation, keeping only high probability tokens

Contrastive decoding formulation

$$\max_{x_{\text{cont}}} \mathcal{L}_{\text{CD}}(x_{\text{cont}}, x_{\text{pre}})$$

subject to $x_i \in \mathcal{V}_{\text{head}}(x_{<i}), \forall x_i \in x_{\text{cont}}$

Objective to token level scores

$$\begin{aligned} & \text{CD-score}(x_i; x_{<i}) \\ &= \begin{cases} \log \frac{p_{\text{EXP}}(x_i | x_{<i})}{p_{\text{AMA}}(x_i | x_{<i})}, & \text{if } x_i \in \mathcal{V}_{\text{head}}(x_{<i}) \\ -\inf, & \text{otherwise.} \end{cases} \end{aligned}$$

Adaptive Plausibility Constraint $\mathcal{V}_{\text{head}}$

$$\mathcal{V}_{\text{head}}(x_{<i}) = \{x_i \in \mathcal{V} : p_{\text{EXP}}(x_i | x_{<i}) \geq \alpha \max_w p_{\text{EXP}}(w | x_{<i})\}$$

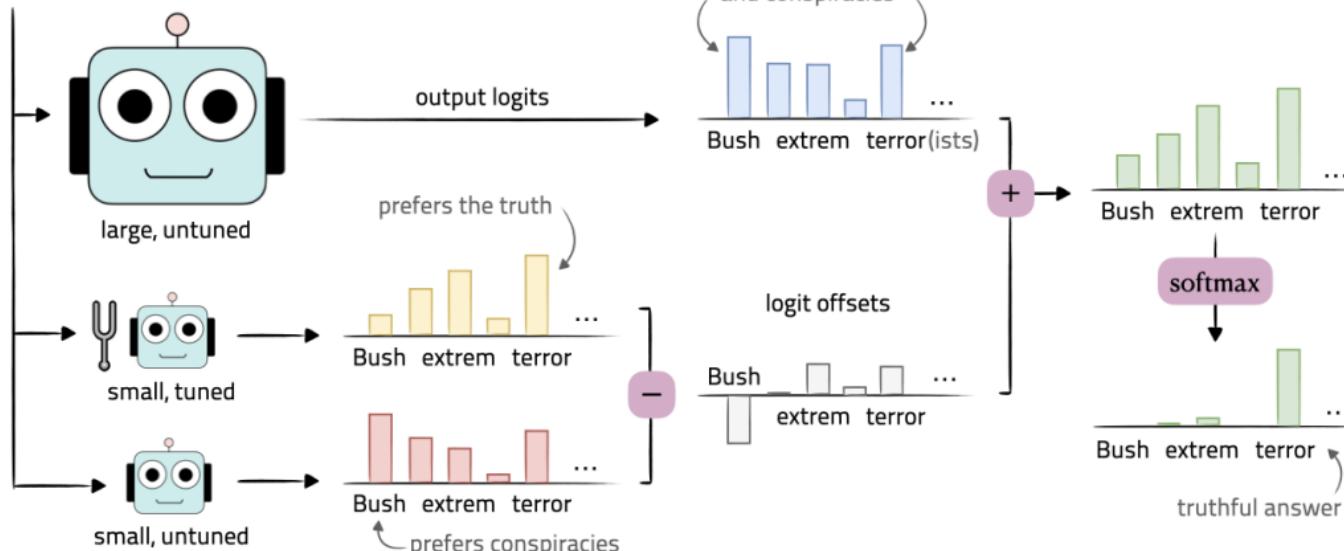
Appendix

- **Proxy Tuning (Liu et al., 2024)**

- Steer large pretrained (base) model (LLAMA2-13B or 70B)
 - Use small, cheaper-to-tune, (anti-)experts (based on LLAMA2-7B) for instruction-following, domain adaptation, and task finetuning
- Use the decoding-time experts equation (DExperts; Liu et al., 2021)
 - Shift original predictions of base model in the direction of the difference that tuning results

Who really caused 9/11?

Answer: 9/11 was really the doing of



Probability distribution from a proxy-tuned model $\tilde{\mathcal{M}}$

$$p_{\tilde{\mathcal{M}}}(X_t | x_{<t}) = \text{softmax} [s_{\mathcal{M}}(X_t | x_{<t}) + s_{\mathcal{M}^+}(X_t | x_{<t}) - s_{\mathcal{M}^-}(X_t | x_{<t})]$$

Appendix

- **Speculative decoding (Leviathan et al., 2023)**

- Reduce the inference cost for all inputs equally
- Sample generation from approximation model as speculative prefix for target model
 - **Green**: Suggestion made by approximation model (6M) M_q target model (97M) M_p
 - **Red**: Rejected suggestion / **Blue**: Correction
 - x_1 is rejected, when running M_p on $prefix + [x_1]$ (in case $q(x) > p(x)$)
 - Discard the computation of x_2
 - Re-sample x_1 from adjusted distribution $p'(x) = \text{norm}(\max(0, p(x) - q(x)))$

[START] japan ' s benchmark **bond** n

[START] japan ' s benchmark nikkei 22 75

[START] japan ' s benchmark nikkei 225 index rose 22 76

[START] japan ' s benchmark nikkei 225 index rose 226 : 69 7 points

[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points , or 0 1

[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points , or 1 : 5 percent , to 10 , 9859

[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points , or 1 : 5 percent , to 10 , 989 . 79 in

[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points , or 1 : 5 percent , to 10 , 989 . 79 in **tokyo** late

[START] japan ' s benchmark nikkei 225 index rose 226 : 69 points , or 1 : 5 percent , to 10 , 989 . 79 in late morning trading . [END]

Appendix

- **Toolformer (Schick et al., 2024)**

- Self-supervised way how to use different tools such as search engines, calculators, and translation systems via simple API calls
- Finetune the LM itself Using a self-supervised loss to determine which of these API calls

API Call

$$\mathbf{e}(c) = \langle \text{API} \rangle a_c(i_c) \langle / \text{API} \rangle$$

$$\mathbf{e}(c, r) = \langle \text{API} \rangle a_c(i_c) \rightarrow r \langle / \text{API} \rangle$$

Process

LM Dataset

1
Sample API Calls

2
Execute API Calls

3
Filter API Calls

LM Dataset
with API Calls

$\mathbf{x}_{1:i-1}$ = Pittsburgh is
also known as

c_i^1 = What other name is
Pittsburgh known by?

r_i^1 = Steel City

$L_i(c_i^1 \rightarrow \text{Steel City})$
 $\triangleleft \min(L_i(c_i^1 \rightarrow \varepsilon), L_i(\varepsilon))$

\mathbf{x}^* = Pittsburgh is
also known as

[QA(What ...?
→ Steel City)]
the Steel City.

$\mathbf{x}_{i:n}$ = the Steel City

c_i^2 = Which country is
Pittsburgh in?

r_i^2 = United States

$L_i(c_i^2 \rightarrow \text{United States})$
 $\triangleright \min(L_i(c_i^2 \rightarrow \varepsilon), L_i(\varepsilon))$

Appendix

- **Toolformer (Schick et al., 2024)**

Filtering API Calls

- Let $L_i(\mathbf{z})$ be the weighted cross entropy loss for M over the tokens x_i, \dots, x_n

$$L_i(\mathbf{z}) = - \sum_{j=i}^n w_{j-i} \cdot \log p_M(x_j \mid \mathbf{z}, x_{1:j-1}) \quad L_i^+ = L_i(\mathbf{e}(c_i, r_i)) \quad L_i^- = \min(L_i(\varepsilon), L_i(\mathbf{e}(c_i, \varepsilon)))$$

- Weighted loss if the API call and its result are given to M as a prefix L_i^+
- Minimum of the losses obtained from (i) doing no API call at all and (ii) doing an API call, but not providing the response L_i^-
- Given a filtering threshold τ_f , we thus only keep API calls for which $L_i^- - L_i^+ \geq \tau_f$
 - Add the API call and its result reduces the loss by at least τ_f
 - Compared to not doing any API call or obtaining no result from it

Appendix

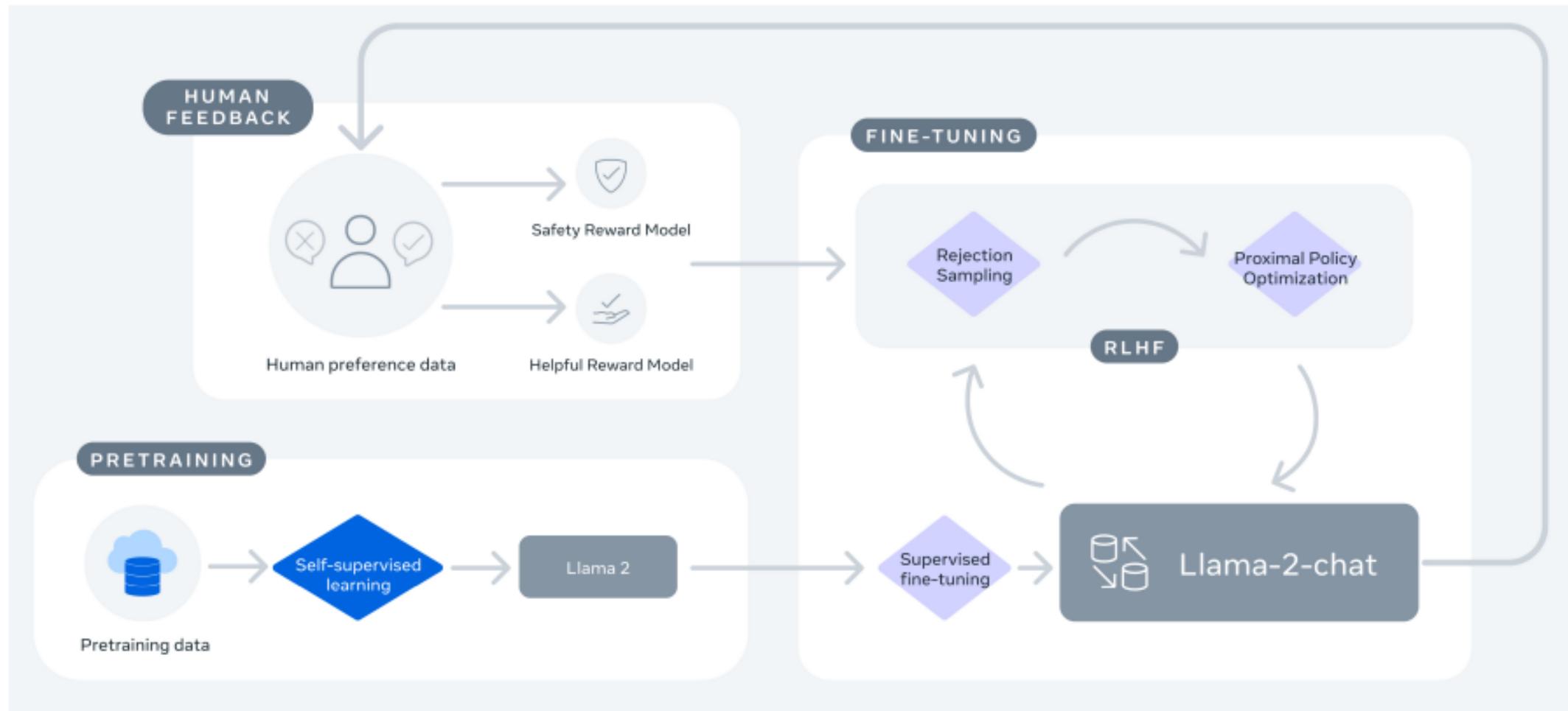
- **Toolformer (Schick et al., 2024)**

Filtering API Calls

Example	$L_i^- - L_i^+$	Useful
<p>The Flodden Window (a war memorial dedicated to The Middleton Archers), in the Grade I-listed Church of St Leonard in Middleton is said to be the oldest war memorial in the United Kingdom. <API> WikiSearch(War memorial Flodden) → Battle of Flodden > Commemoration > The stained-glass Flodden Window in Middleton Parish Church [...] was constructed by Sir Richard Assheton in memory of the Battle of Flodden and the archers from Middleton who fought in it. </API> Sir Richard Assheton of Middleton (who built St Leonard) was granted knighthood [...]</p>	5.49	✓
<p>Note: The WL will be open on Friday, <API> Calendar() → Today is Thursday, March 9, 2017. </API> March 10, and Sunday, March 19 for regular hours.</p>	2.11	✓
<p>The Nile has an approximate length of <API> QA(What is the approximate length of the Nile?) → 6,853 km </API> 6,853 kilometers, the White Nile being its main source.</p>	2.08	✓
<p>If Venus had an atmosphere similar to Earth's then you would expect Venus' mean temperature to be 499 K (1.74×287) rather than 735 K which is <API> Calculator($735 / 499$) → 1.47 </API> 1.47 ($735 / 499$) times hotter than it should be.</p>	1.59	✓

Appendix

- **Llama 2 (Touvron et al, 2023)**



Appendix

- **Llama 2 (Touvron et al, 2023)**

Overall performance on grouped academic benchmarks compared to open-source base models

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcon	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.3	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	65B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

Comparison to closed-source models

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	86.4	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	86.1	85.0
Natural Questions (1-shot)	–	–	29.3	37.5	33.0
GSM8K (8-shot)	57.1	92.0	56.5	80.7	56.8
HumanEval (0-shot)	48.1	67.0	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	65.7	51.2

Appendix

- **Llemma: An open language model for mathematics**

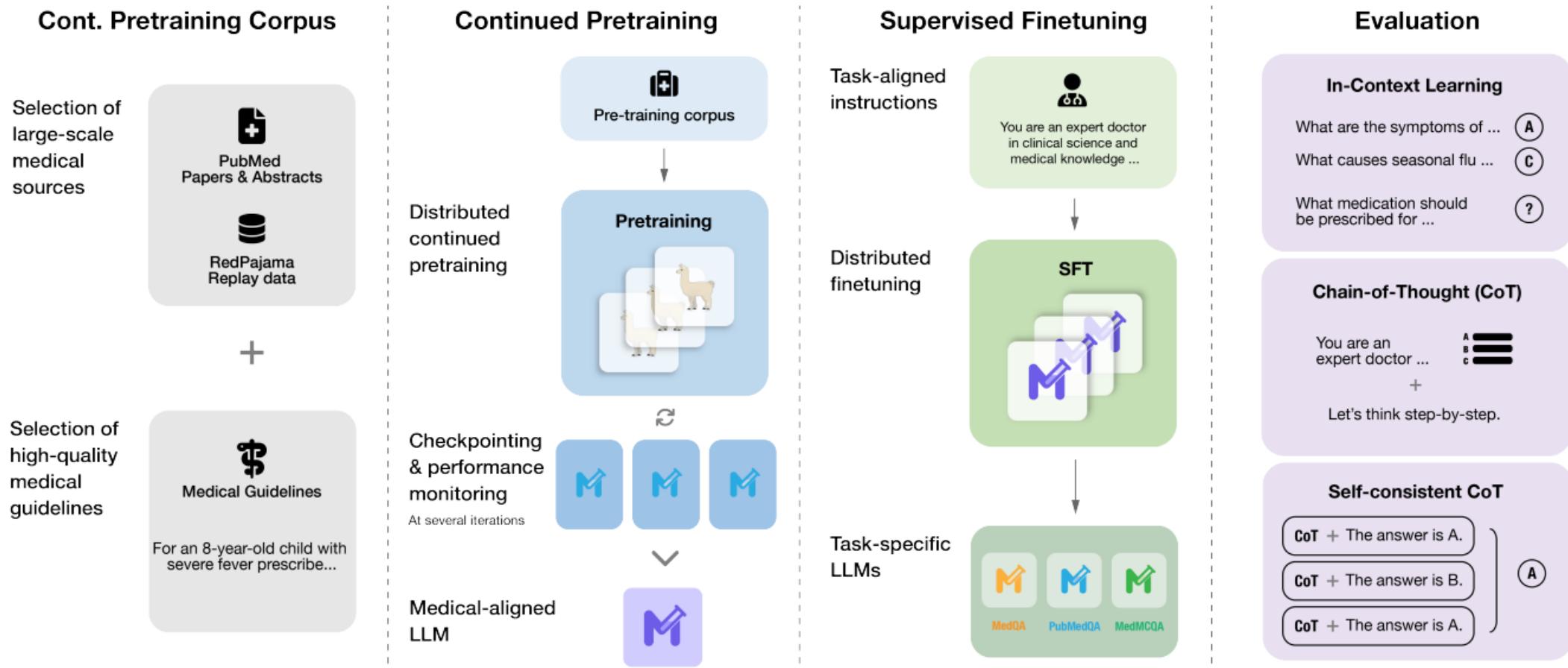
- 7B and 34B parameter language models specialized for mathematics
 - Use computational tool: Python interpreter & formal theorem prover
- 55b-token mixture of scientific papers, web data containing mathematics, and mathematical code
- Each model is initialized from Code Llama (Rozière et al., 2023)
- Reasoning: Chain of thought reasoning, Majority voting

Data source	Tokens	Weight
Proof-Pile-2	55B	–
Code (AlgebraicStack)	11B	1.00
Web (OpenWebMath)	15B	4.00
Papers (ArXiv)	29B	2.00
General code (RedPajama)	59B	0.22
General language (Pile)	300B	0.15

Mathematical problem solving with tool use			
	GSM8k+Python	MATH+Python	
	pass@1	pass@1	
Code Llama	7B	27.1%	17.2%
LLEMMA	7B	40.1%	21.5%
Code Llama	34B	52.7%	23.5%
LLEMMA	34B	62.6%	27.1%

Appendix

- **Meditron-7B & 70B (Chen et al, 2023)**
 - A pair of generative LLMs for medical reasoning, adapted from Llama-2



Appendix

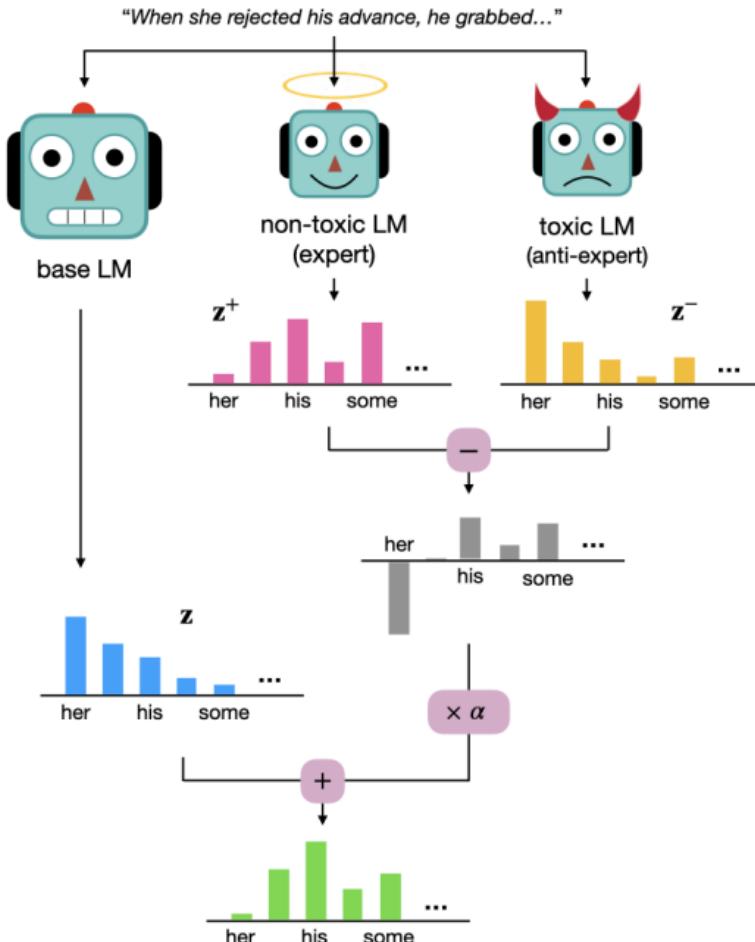
- **Meditron-7B & 70B (Zeming Chen et al, 2023)**
 - A pair of generative LLMs for medical reasoning, adapted from Llama-2

Main results of MEDITRON against open-source baselines

Model	Accuracy (↑)					
	MMLU-Medical	PubMedQA	MedMCQA	MedQA	MedQA-4-Option	Avg
Top Token Selection						
Mistral-7B*	55.8	17.8	40.2	32.4	41.1	37.5
Zephyr-7B- β^*	63.3	46.0	43.0	42.8	48.5	48.7
PMC-Llama-7B	59.7	59.2	57.6	42.4	49.2	53.6
Llama-2-7B	56.3	61.8	54.4	44.0	49.6	53.2
MEDITRON-7B	55.6	74.4	59.2	47.9	52.0	<u>57.5</u>
Chain-of-thought						
Llama-2-70B	76.7	79.8	62.1	60.8	63.9	68.7
MEDITRON-70B	74.9	81.0	63.2	61.5	67.8	<u>69.7</u>
Self-consistency Chain-of-thought						
Llama-2-70B	77.9	80.0	62.6	61.5	63.8	69.2
MEDITRON-70B	77.6	81.6	66.0	64.4	70.2	72.0

Appendix

• DExperts: Decoding-time controlled text generation with experts and anti-experts



- Product-of-experts ensemble $\tilde{P}(X_t | \mathbf{x}_{<t}) \propto P(X_t | \mathbf{x}_{<t}) \left(\frac{P^+(X_t | \mathbf{x}_{<t})}{P^-(X_t | \mathbf{x}_{<t})} \right)^\alpha$
 - Interpret the ratio as scaling coefficient for each token
- Sampling from DExperts $\mathbf{z}'[v] = \begin{cases} \mathbf{z}[v] & \text{if } v \in \mathcal{V}' \\ -\infty & \text{otherwise} \end{cases}$

$$\tilde{P}'(X_t | \mathbf{x}_{<t}) = \text{softmax}(\mathbf{z}'_t + \alpha (\mathbf{z}_t^+ - \mathbf{z}_t^-))$$

Threshold to \mathbf{Z}_t

Model	Toxicity (↓)		Fluency (↓)		Diversity (↑)		
	Avg. max. toxicity	Toxicity prob.	Output ppl.	Dist-1	Dist-2	Dist-3	
GPT-2	0.527	0.520	25.45	0.58	0.85	0.85	
PPLM (10%)	0.520	0.518	32.58	0.58	0.86	0.86	
Non-toxic expert	0.485	0.464	40.61	0.58	0.86	0.86	
DAPT	0.428	0.360	31.21	0.57	0.84	0.84	
GeDi	0.363	0.217	60.03	0.62	0.84	0.83	
DEXPERTS (anti-only)	0.352	0.191	52.02	0.58	0.80	0.73	
DEXPERTS (small)	0.302	0.118	38.20	0.56	0.82	0.83	
DEXPERTS (medium)	0.307	0.125	32.51	0.57	0.84	0.84	
DEXPERTS (large)	0.314	0.128	32.41	0.58	0.84	0.84	

Appendix

- Consistent estimators for learning to defer to an expert**

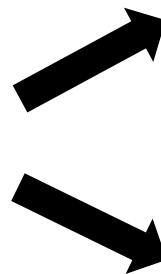
- Rejector decides whether the classifier or the expert should predict
- Expert may have access to side-information unavailable to the learner
- Example task: Chest X-ray diagnosis of pneumonia

Deferral Module (Rejector)

Q. How to combine model and expert to

- 1) Achieve better performance and
- 2) Ease the burden on the radiologist

A. Given only patient input, learn model to route decision to either expert or model in order to maximize system performance

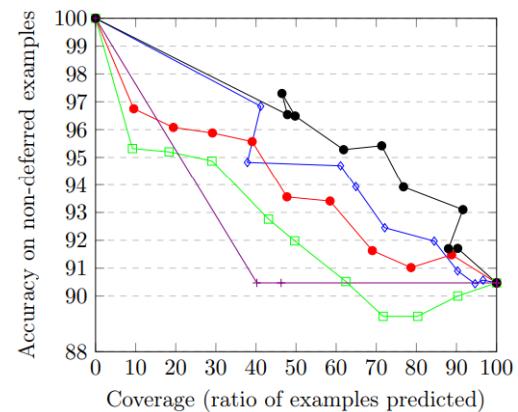


Defer to Expert

- Expert radiologist:
"Presence of pneumonia"

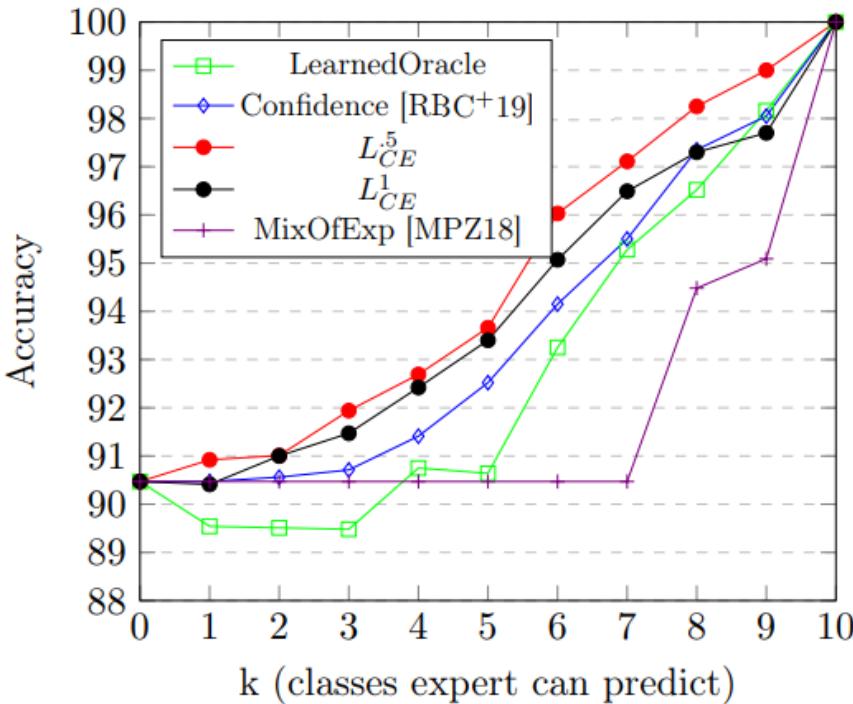
Classifier Prediction

- Classifier:
"Pneumonia: No"

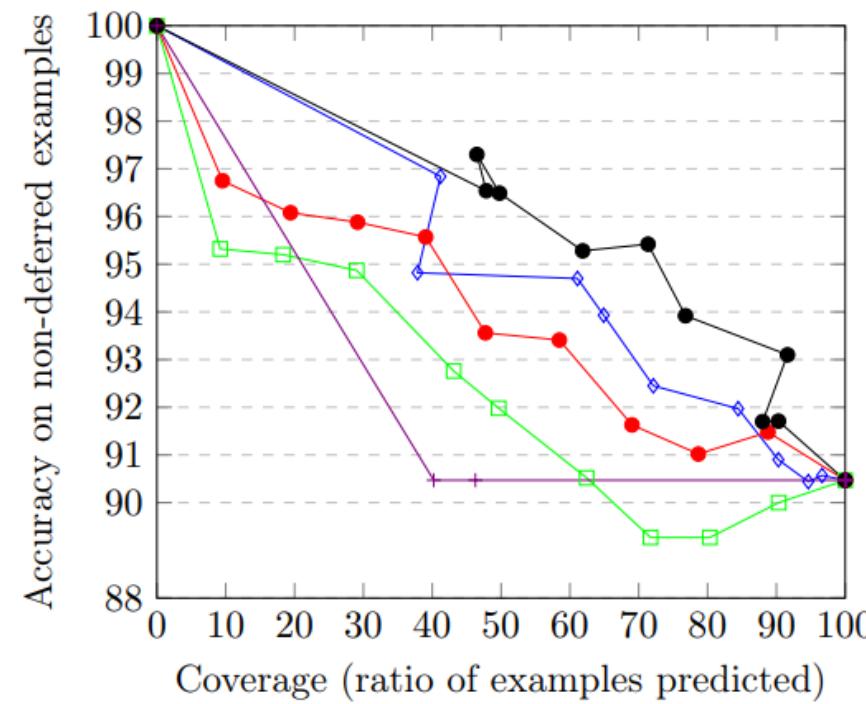


Appendix

- **Consistent estimators for learning to defer to an expert**
 - (a) overall system accuracy of our method and baselines (k is the number of classes the expert can predict)
 - (b) compares accuracy on the non-deferred examples versus the coverage for every k



(a)



(b)

Dataset: CIFAR-10

Appendix

- **Who Should Predict? Exact Algorithms For Learning to Defer to Humans**

- Jointly learn a classifier that can predict pneumonia and a rejector
 - Rejector decides on each data point whether the classifier or the human should predict

$$L_{\text{def}}(m, r) = \mathbb{E}_{X,Y,Z} [\ell_{\text{AI}}(X, Y, m(X)) \cdot \mathbb{I}_{r(X)=0} + \ell_{\text{HUM}}(X, Y, h(Z)) \cdot \mathbb{I}_{r(X)=1}]$$

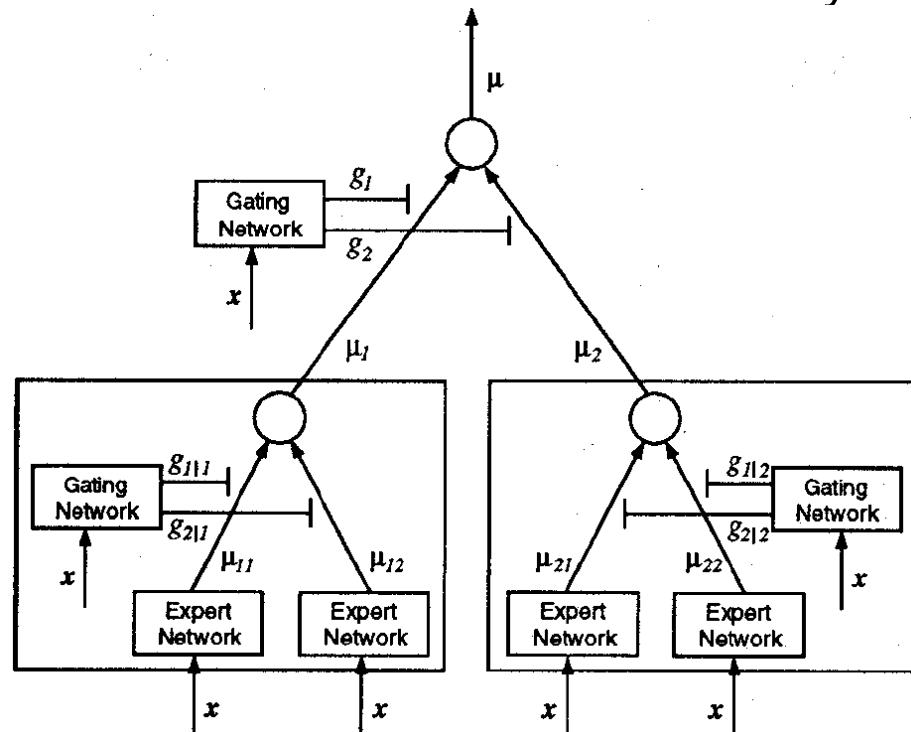
- Optimization
 - Focus mostly on the cost of misclassification with no additional penalty
 - Deferral loss becomes a misclassification loss $L_{\text{def}}^{0-1}(m, r)$ for the human-AI system

$$\underset{m,r}{\text{minimize}} \ L_{\text{def}}^{0-1}(m, r) := \mathbb{P} [((1 - r(X)) m(X) + r(X) h(Z)) \neq Y]$$

Appendix

- Hierarchical mixtures of experts and the EM algorithm**

- Tree-structured architecture for supervised learning
- Soft Boundaries: Boundaries between regions are parameterized surfaces
- Expectation-Maximization (EM) framework: Mixture model estimation and generalized linear model theory



Average Values of Relative Error

Architecture	Relative Error	# Epochs
linear	.31	1
backprop	.09	5,500
HME	.10	35
CART	.17	NA
CART (oblique)	.13	NA
MARS	.16	NA

Relative Error

Ratio between error and error obtained if the learner were to output the mean value of the outputs

Appendix

- **Domain Adaptive Inference for Neural Machine Translation**
 - Bayesian Interpolation (BI) (Allauzen and Riley, 2011) to sequence-to-sequence models

Adaptive training

$$L(\theta) = L_B(\theta) + \Lambda \sum_j F_j (\theta_j - \theta_{A,j}^*)^2$$

Test BLEU for en-de adaptive training

	Training scheme	News	TED	IT
1	News	37.8	25.3	35.3
2	TED	23.7	24.1	14.4
3	IT	1.6	1.8	39.6
4	News and TED	38.2	25.5	35.4
5	1 then TED, No-reg	30.6	27.0	22.1
6	1 then TED, L2	37.9	26.7	31.8
7	1 then TED, EWC	38.3	27.0	33.1
8	5 then IT, No-reg	8.0	6.9	56.3
9	6 then IT, L2	32.3	22.6	56.9
10	7 then IT, EWC	35.8	24.6	57.0

Adaptive decoding

- BI formalism:
Ensemble weights λ for each task
- Fixed weight ensemble:
Task t is not known

$$p(\mathbf{y}|\mathbf{x}, t) = \sum_{k=1}^K \lambda_{k,t} p_k(\mathbf{y}|\mathbf{x})$$

$$p(t|h_i, \mathbf{x}) = \frac{p(h_i|t, \mathbf{x})p(t|\mathbf{x})}{\sum_{t'=1}^T p(h_i|t', \mathbf{x})p(t'|\mathbf{x})}$$

Test BLEU for 2-model es-en and 3-model en-de model ensembling

Decoder configuration	es-en		en-de		
	Health	Bio	News	TED	IT
Oracle model	35.9	37.8	37.8	27.0	57.0
Uniform	36.0	36.4	38.9	26.0	43.5
BI + IS	36.2	38.0	38.7	26.1	56.4

Appendix

• EM Algorithm

- EM algorithm

- Finds the maximum likelihood solutions for models with latent variables

$$P(X|\theta) = \sum_Z P(X, Z|\theta) \rightarrow \ln P(X|\theta) = \ln\{\sum_Z P(X, Z|\theta)\}$$

- EM algorithm

- Initialize θ^0 to an arbitrary point
 - Loop until the likelihood converges
 - Expectation step

$$q^{t+1}(z) = \operatorname{argmax}_q Q(\theta^t, q) = \operatorname{argmax}_q L(\theta^t, q) = \operatorname{argmin}_q KL(q||P(Z|X, \theta^t))$$

$$q^t(z) = P(Z|X, \theta) \quad \text{Assign } Z \text{ by } P(Z|X, \theta)$$

- Maximization step

$$\theta^{t+1} = \operatorname{argmax}_{\theta} Q(\theta, q^{t+1}) = \operatorname{argmax}_{\theta} L(\theta, q^{t+1})$$

fixed Z means that there is no unobserved variables

Same optimization of ordinary MLE

Appendix

- **Tülu v2 mix data (Wang et al., 2023)**

- Collect a representative sample of different styles of datasets
- Number of conservation turns (\bar{N}_{rounds}), Length of prompts/completion (\bar{L}_{prompt} $\bar{L}_{\text{completion}}$)

Instruction datasets investigated in this work

Datasets	Sourced from	# Instances	\bar{N}_{rounds}	\bar{L}_{prompt}	$\bar{L}_{\text{completion}}$
SuperNI [48]	NLP datasets + Human-written Instructions	96,913	1.0	291.1	38.7
CoT [50]	NLP datasets + Human-written CoTs	100,000	1.0	266.0	53.2
Flan V2 [31]	NLP datasets + Human-written Instructions	100,000	1.0	355.7	31.2
Dolly [12]	Human-written from scratch	15,011	1.0	118.1	91.3
Open Assistant 1 [26]	Human-written from scratch	34,795	1.6	34.8	212.5
Self-instruct [47]	Generated w/ vanilla GPT3 LM	82,439	1.0	41.5	29.3
Unnatural Instructions [23]	Generated w/ Davinci-002	68,478	1.0	107.8	23.6
Alpaca [43]	Generated w/ Davinci-003	52,002	1.0	27.8	64.6
Code-Alpaca [6]	Generated w/ Davinci-003	20,022	1.0	35.6	67.8
GPT4-Alpaca [36]	Generated w/ Davinci-003 + GPT4	52,002	1.0	28.0	161.8
Baize [52]	Generated w/ ChatGPT	210,311	3.1	17.6	52.8
ShareGPT ³	User prompts + outputs from various models	168,864	3.2	71.0	357.8

Appendix

- **GSM8k (Cobbe et al., 2021)**

- 8.5k grade school math problems created by human problem writers
- Problems take between 2 and 8 steps to solve, and solutions for final answer
- Require a sequence of elementary calculations using basic arithmetic operations

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

Solution: Beth bakes 4 2 dozen batches of cookies for a total of $4 \times 2 = \textcolor{red}{<<4*2=8>>} 8$ dozen cookies

There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = \textcolor{red}{<<12*8=96>>} 96$ cookies

She splits the 96 cookies equally amongst 16 people so they each eat $96 / 16 = \textcolor{red}{<<96/16=6>>} 6$ cookies

Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?

Mrs. Lim got 68 gallons - 18 gallons = $\textcolor{red}{<<68-18=50>>} 50$ gallons this morning.

So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = $\textcolor{red}{<<68+82+50=200>>} 200$ gallons.

She was able to sell 200 gallons - 24 gallons = $\textcolor{red}{<<200-24=176>>} 176$ gallons.

Thus, her total revenue for the milk is $\$3.50/\text{gallon} \times 176 \text{ gallons} = \$\textcolor{red}{<<3.50*176=616>>} 616$.

Final Answer: 616

Appendix

- **MATH (Hendrycks et al., 2021)**

- High school math competitions problems with step-by-step solutions with **LATEX**
- Problems are tagged by difficulty from 1 to 5, span seven subjects including geometry
- Diagrams can be specified in text with the Asymptote language

Metamath Theorem Proving

$n \in \mathbb{N} \wedge \frac{n+1}{2} \in \mathbb{N} \implies \exists m \in \mathbb{N} : n = 2m + 1$.
GPT-f's generated proof:

```
| - ((N e. NN0 /\ ((N + 1) / 2) e. NN0) -> ((N - 1) / 2) e. NN0)
| - (N e. NN0 -> N e. CC)
| - 1 e. CC
| - ((N e. CC /\ 1 e. CC) -> (N - 1) e. CC )
:
:
```

DeepMind Mathematics Dataset

Divide 1136975704 by -142121963.
A: -8
Let k(u) = u**2+u-4. Find k(0).
A: -4
Sort 2, 4, 0, 6.
A: 0, 2, 4, 6
Solve 4 - 4 - 4 = 188*m for m.
A: -1/47

MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x + 1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt{2}$. The desired product is then $(-1 + \cos(\frac{\pi}{8})\sqrt{2})(-1 - \cos(\frac{\pi}{8})\sqrt{2}) = 1 - \cos^2(\frac{\pi}{8})\sqrt{2} = 1 - \frac{(1+\cos(\frac{\pi}{4}))}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

A subset of the topics

- Algebra
 - Conic sections, polynomial GCD, ...
- Calculus
 - Arclength, Jacobian, Laplacian, ...
- Statistics
 - Expectation, geometric mean, ...
- Geometry
 - Triangle area, triangle inradius ...
- Linear Algebra
 - Characteristic polynomials, ...
- Number Theory
 - Modular inverse, ...

Appendix

- **BioASQ (Tsatsaronis et al., 2015)**

- Task 1a: Large scale online biomedical semantic indexing
 - Training Data: Tune previously annotated PUBMED articles with MESH annotations
 - Large scale classification of biomedical documents into ontology concepts
- Task 1b: Biomedical SemanticQuestion Answering
 - Ideal answers for yes/no questions, factoid questions: Human would expect as an answer by a peer biomedical scientist

Format of training data of Task1b

```
{
  "abstractText": "From the above it is seen
    that the [...]
    scientific guidance of which lies wholly
    in the hands of scientists.",
  "journal": "Science (New York, N.Y.)",
  "meshMajor": ["Biomedical Research"],
  "pmid": "17772322",
  "title": "New Horizons in Medical
    Research.",
  "year": "1946"
}, {
```

Types of Q & A in Task 1b and respective examples

Question type	Required answer	Example question	Golden exact answer	Golden Ideal answer
Yes/No	Exact + Ideal	Is miR-21 related to carcinogenesis?	Yes	Yes. It has been demonstrated in several experimental studies that miR-21 has oncogenic potential, and is significantly deregulated in numerous types of cancer. Therefore, miR-21 is closely related to carcinogenesis.

Appendix

- Prompts used for formatting the datasets

Dataset	Prompt	Dataset	Prompt
AlpacaEval	<p>< user ></p> <p>What are some species of bears that are now extinct?</p> <p>< assistant ></p> <p>Please solve the following math problem with detailed steps.</p>	MATH	<p>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)\$.”</p> <p>Problem: Consider the function $g(x)=3x-4$. What is $g(0)$?</p> <p>Solution:</p>
GSM8k	<p>Question: Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?</p> <p>Answer:</p>	BioASQ	<p>You are a medical expert answering real-world biomedical questions. Based on your understanding of basic and clinical science, medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy, write a list of entity names, numbers, or short expressions that answer the following question.</p> <p>Question: Which genes are related to Emery-Dreifuss Muscular Dystrophy (EDMD)?</p> <p>Answer:</p>