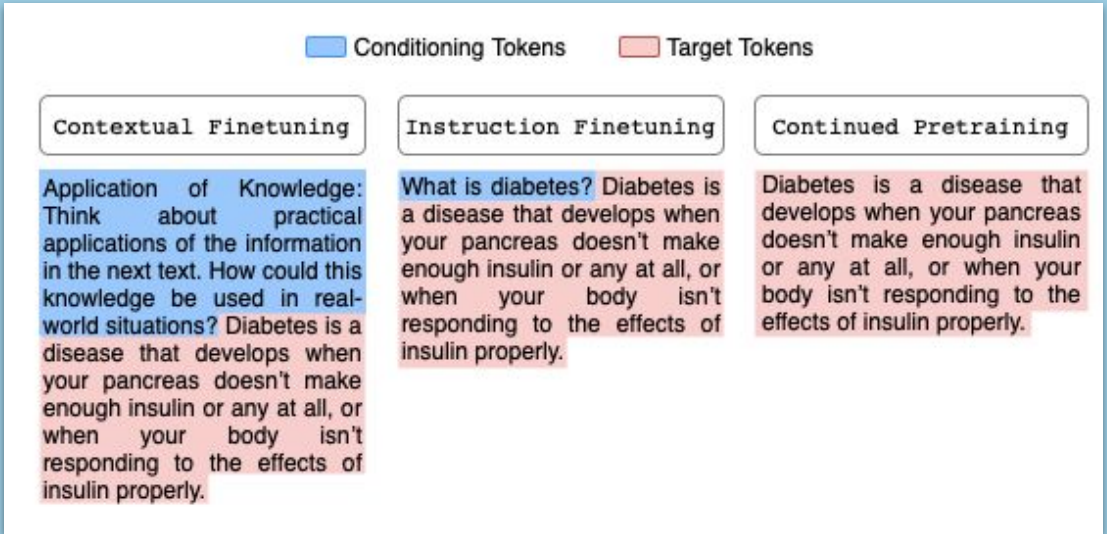
Teaching LLMs How to Learn with Contextual Fine-Tuning



Younwoo Choi*, Muhammad Adil Asif*, Ziwen Han, John Willes, Rahul G. Krishnan

University of Toronto & Vector Institute

*Equal contribution

VECTOR INSTITUTE

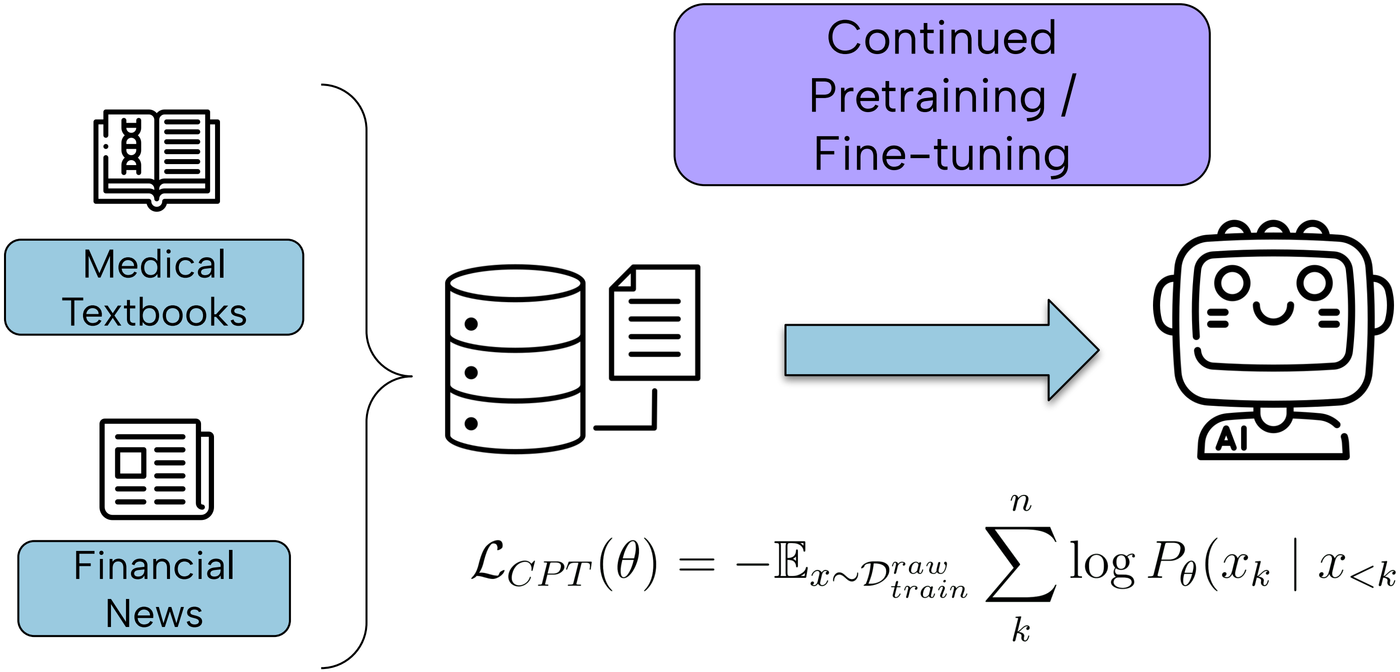
Motivation: Domain-Specific Fine-Tuning

- LLMs encode broad distributional knowledge across diverse domains.
  - Case 1: In fast-moving domains, models require periodic knowledge updates.
  - Case 2: Specialized domains benefit from targeted distribution alignment.

Potential solutions:

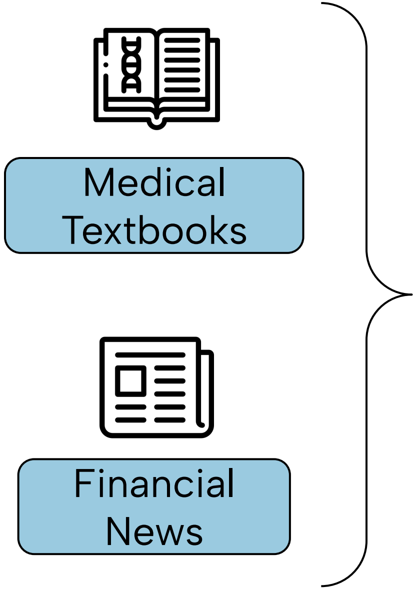
- RAG + in-context learning
  - Drawbacks: context length limitations constrain knowledge integration.

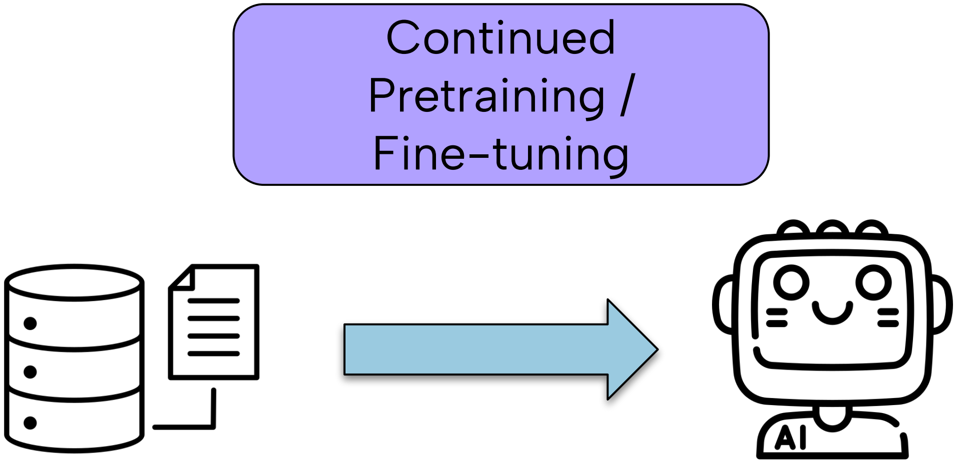
Motivation: Continued Pretraining



$$\mathcal{L}_{CPT}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}_{train}^{raw}} \sum_{k}^{n} \log P_\theta(x_k \mid x_{<k})$$

Motivation: Continued Pretraining

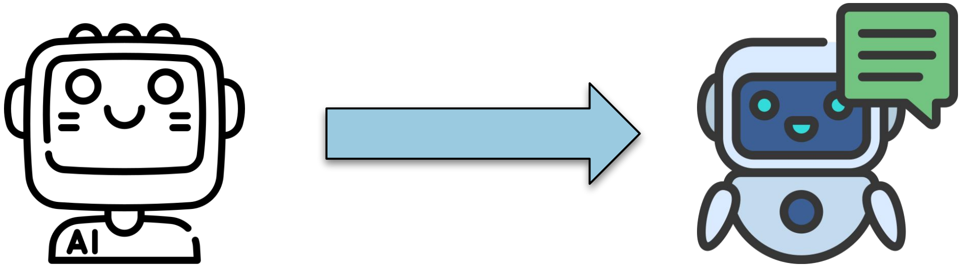




- We update all the parameters.
- Is there a more effective way to guide learning?

Motivation: Instruction Fine-Tuning & RLHF

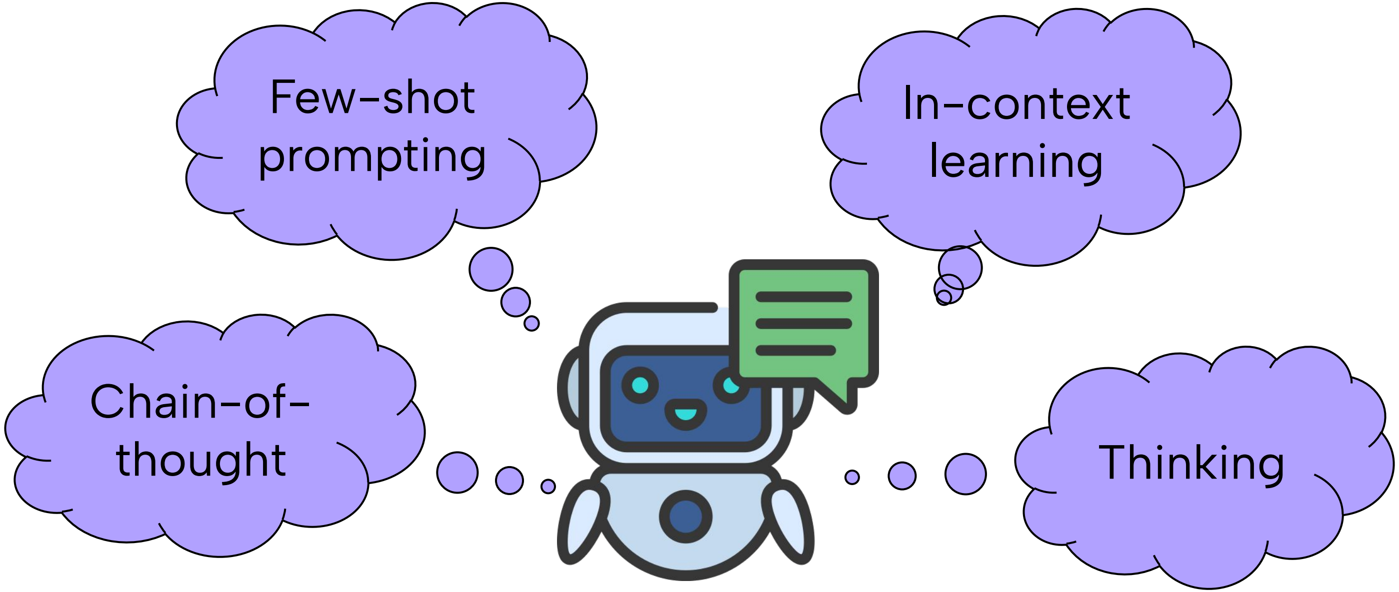
Instruction
fine-tuned & RLHF



$$\mathcal{L}_{IFT}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}^{IFT}_{train}} \sum_{k}^{m} \log P_\theta(y_k \mid x, y_{<k})$$

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

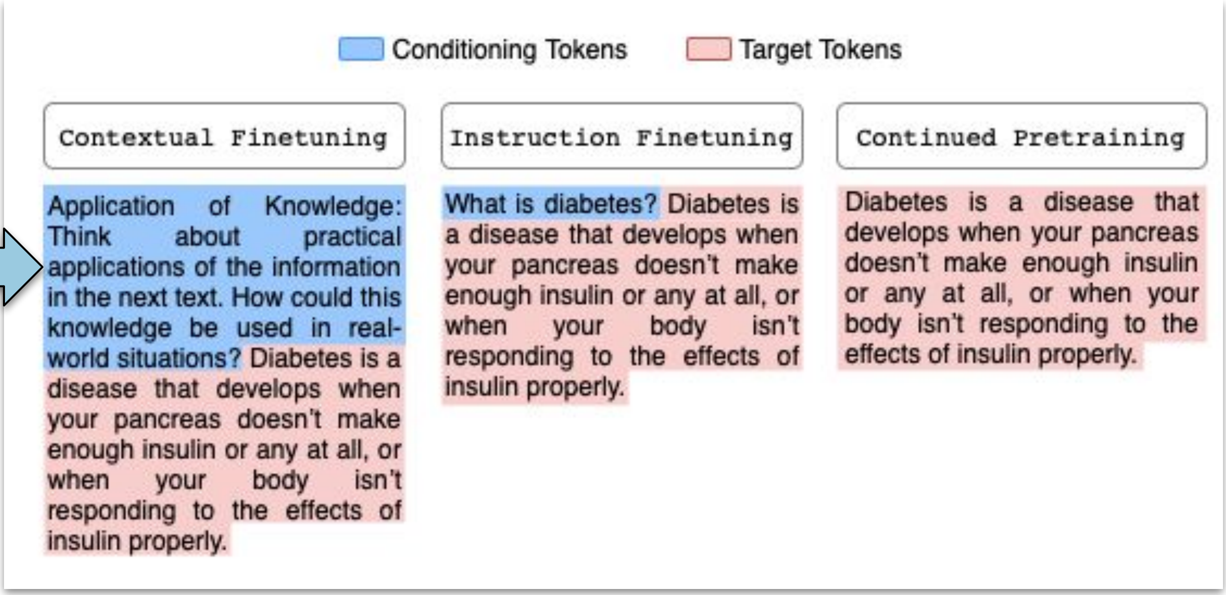
Motivation: Chat LLMs



Can prompting improve the efficacy of LLM fine-tuning?

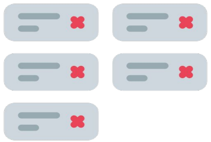
Designing Contextual Prompts

1. Manual Construction
2. Auto Construction



Designing Contextual Prompts: Manual

Educational Theories



Contextual Prompts

Example

The importance of reducing unnecessary cognitive load to facilitate learning. By focusing on essential information, learners can allocate their cognitive resources more effectively (Sweller, 2011)

"Concentrate on understanding the core principles and essential facts in the following text. Pay special attention to definitions, examples, and conclusions."

Designing Contextual Prompts: Manual

Contextual Prompts

$$\mathcal{C} = \{c^{(1)}, \ldots, c^{(10)}\}$$

- Application of Knowledge
- Reflective Thinking
- Creative Interpretation
- Summarization and Synthesis
- Focus on Key Concepts

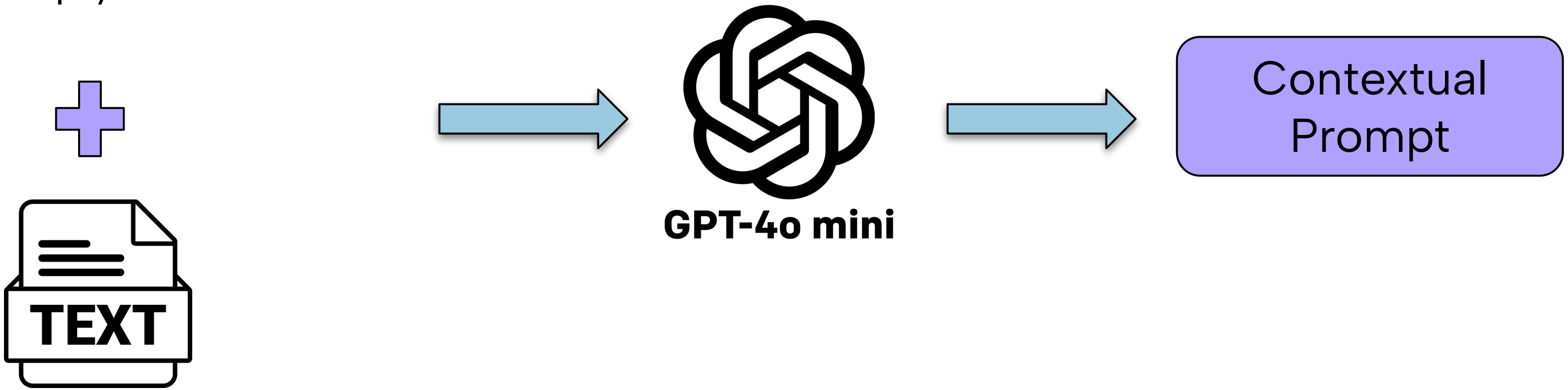
"Concentrate on understanding the core principles and essential facts in the following text. Pay special attention to definitions, examples, and conclusions."

- Contextual Understanding
- In-Depth Exploration
- Question-Based Learning
- Comparative Learning
- Critical Analysis

"Compare and contrast the upcoming information with what you have learned in similar topics. Look for differences, similarities, and connections."

Designing Contextual Prompts: Auto-Generated

"Given the following text, generate a contextual prompt that encourages a reader to focus on the main ideas and themes presented. The contextual prompt should be concise and help the reader engage deeply with the content."

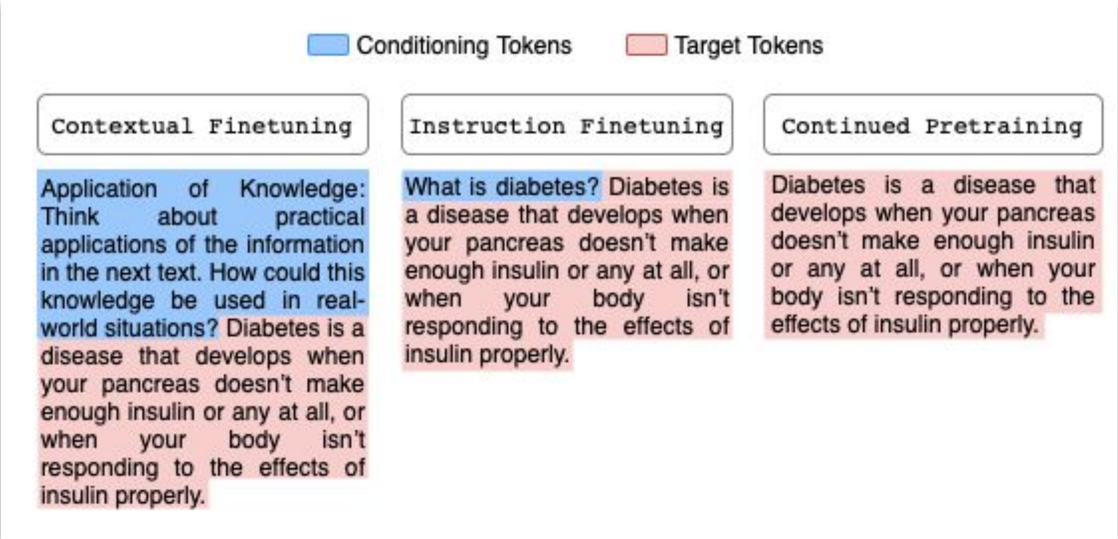


Learning with Contextual Prompts

We have

$$\mathcal{D}_{train}^{raw}$$

$$\mathcal{C} = \{c^{(1)}, c^{(2)}, \ldots, c^{(L)}\}.$$



Sample

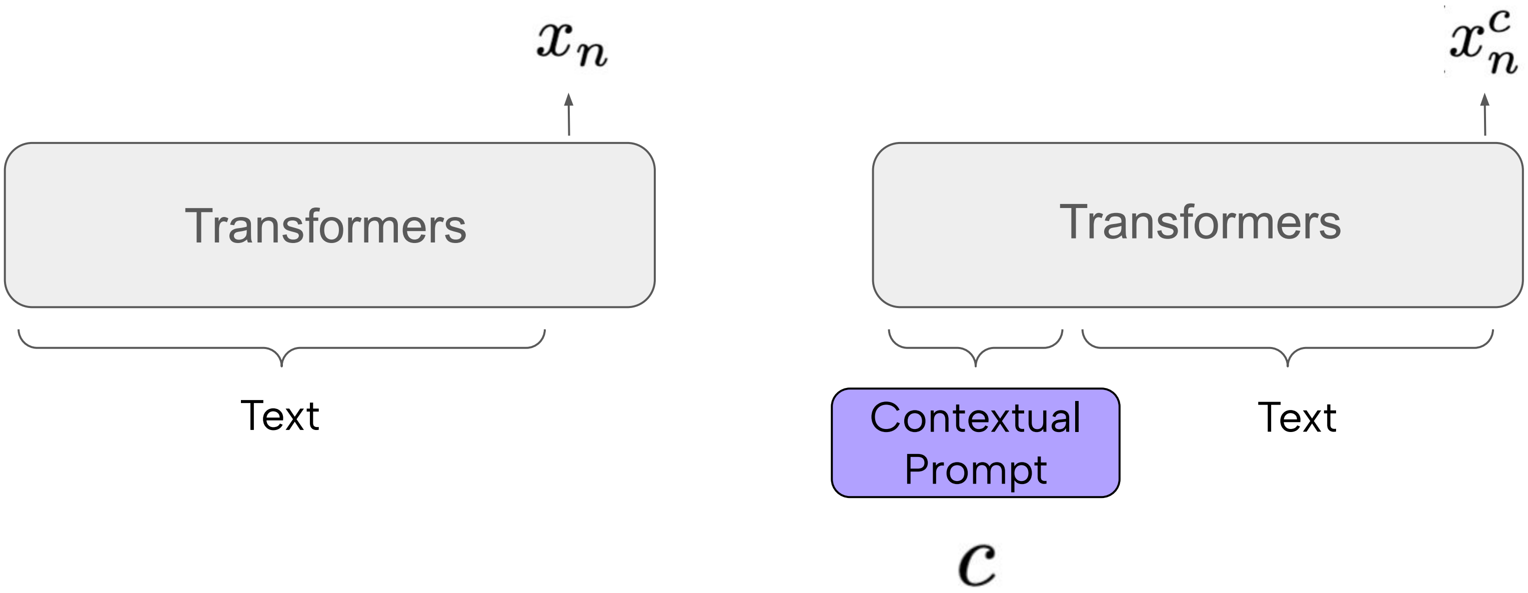
$$c = (c_1, c_2, \ldots, c_m)$$

$$x = (x_1, x_2, \ldots, x_n)$$

Loss

$$\mathcal{L}_{CFT}(\theta) = -\mathbb{E}_{\boxed{x}\sim\mathcal{D}_{train}^{raw}, \boxed{c}\sim\mathcal{C}} \sum_{k=1}^{n} \log P_\theta(x_k \mid c, x_{<k})$$

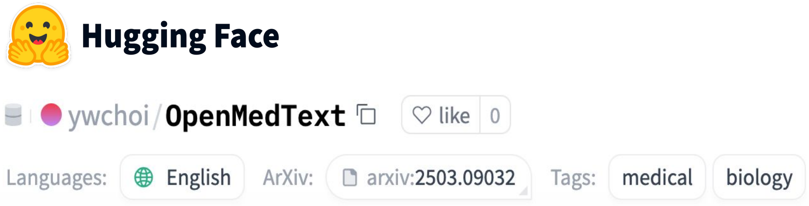
Hypothesis



$$\nabla_\theta f_\theta(x_n) \quad \text{VS} \quad \nabla_\theta f_\theta(x_n^c)$$

Experimental Setup

Models
- Llama-2 7B
  - Base
  - Chat
- Llama-2 13B
  - Chat



Fine-tuning datasets
- Biomedical domain
  - ***OpenMedText***
    - Our curated dataset of medical journals and textbooks.
    - 121,489 journals covering 37 topics and 29 medical textbooks.
- Financial domain
  - A collection of news articles.
  - Total 306,242 financial news articles.

Experimental Setup

- **Benchmarks**
  - Biomedical domain:
    - MMLU
      - Anatomy, Clinical Knowledge, College Biology, College Medicine, Medical Genetics, and Professional Medicine.
    - MedQA
  - Financial domain:
    - FiQA
      - Semantic analysis task.
      - “What is the sentiment of the following financial news?”
    - Causal20
      - Events classification.
      - “Classify each sentence into either ‘causal’ or ‘noise’”
    - Multifin
      - Headlines classification.
      - “Categorize each headline according to its primary topic”

Results

- **Contextual fine-tuning is effective across model scales.**

Biomedical Domain

	Accuracy (↑)							
Llama 2 7B	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Chat	44.07	46.79	48.61	39.02	49.00	**48.90**	38.96	45.05
Chat (CPT)	45.19	47.17	49.31	43.93	50.50	46.32	39.28	45.96
Chat (CFT)	**48.15**	**48.87**	**52.08**	**44.22**	**54.00**	46.69	**40.65**	**47.81**
Llama 2 13B	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Chat	51.85	56.60	54.17	46.82	**63.50**	56.99	**45.33**	53.61
Chat (CPT)	50.37	60.00	55.90	50.58	62.00	57.35	43.95	54.31
Chat (CFT)	**53.33**	**63.21**	**57.99**	**56.35**	62.50	**57.72**	44.85	**56.56**

Financial Domain

Llama 2 7B	FiQA F1	Causal 20 F1	Multifin F1	Average
Chat	56.40	**90.40**	38.74	61.48
Chat (CPT)	62.53	90.16	38.23	63.64
Chat (CFT)	**67.69**	90.17	**46.01**	**67.96**

Llama 2 13B	FiQA F1	Causal 20 F1	Multifin F1	Average
Chat	61.18	84.77	45.81	63.92
Chat (CPT)	66.96	**90.06**	45.33	67.45
Chat (CFT)	**70.55**	89.87	**50.94**	**70.45**

Results

- **Contextual fine-tuning is preferable to existing approaches for improving a model at a fixed scale.**

Biomedical Domain

Llama 2 7B	Accuracy (↑)							
	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Base	43.52	44.10	40.89	37.43	48.25	39.84	35.36	41.34
Base (CPT)	47.50	45.19	41.67	37.43	49.00	40.17	35.84	42.40
Base (CFT)	47.87	45.90	41.32	38.87	46.12	39.11	36.76	42.28
Base (CPT + IFT)	49.91	45.47	42.71	37.79	49.37	41.59	35.93	43.25
Base (CFT + IFT)	**51.11**	**46.37**	**42.80**	**40.10**	**50.00**	**42.74**	**36.99**	**44.29**
Chat	44.07	46.79	48.61	39.02	49.00	**48.90**	38.96	45.05
Chat (CPT)	45.19	47.17	49.31	43.93	50.50	46.32	39.28	45.96
Chat (CFT)	**48.15**	**48.87**	**52.08**	**44.22**	**54.00**	46.69	**40.65**	**47.81**

Against AdaptLLM

Llama 2 7B	Accuracy (↑)							
	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Chat	44.07	46.79	48.61	39.02	49.00	**48.90**	38.96	45.05
Chat (CPT)	45.19	47.17	49.31	43.93	50.50	46.32	39.28	45.96
Chat (CFT)	**48.15**	**48.87**	**52.08**	**44.22**	**54.00**	46.69	**40.65**	**47.81**
AdaptLLM	44.45	47.36	48.27	39.60	45.00	38.61	37.12	42.92

Results

- **The semantic content of the contextual prompts are important to improving performance.**

Biomedical Domain

	Accuracy (↑)							
Llama 2 7B	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Chat (CFT)	**48.15**	**48.87**	**52.08**	**44.22**	**54.00**	**46.69**	**40.65**	**47.81**
Chat (-CFT)	41.48	48.68	47.92	43.35	50.50	**46.69**	38.06	45.24
Llama 2 13B	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Chat (CFT)	**53.33**	**63.21**	57.99	**56.35**	**62.50**	**57.72**	**44.85**	**56.56**
Chat (-CFT)	50.00	59.62	**62.15**	52.89	61.50	57.17	43.09	55.20

Financial Domain

	FiQA	Causal 20	Multifin	Average
Llama 2 7B	F1	F1	F1	
Chat (CFT)	**67.69**	**90.17**	**46.01**	**67.96**
Chat (-CFT)	59.53	90.16	43.96	64.55

	FiQA	Causal 20	Multifin	Average
Llama 2 13B	F1	F1	F1	
Chat (CFT)	**70.55**	89.87	50.94	**70.45**
Chat (-CFT)	60.60	**90.13**	**53.45**	68.06

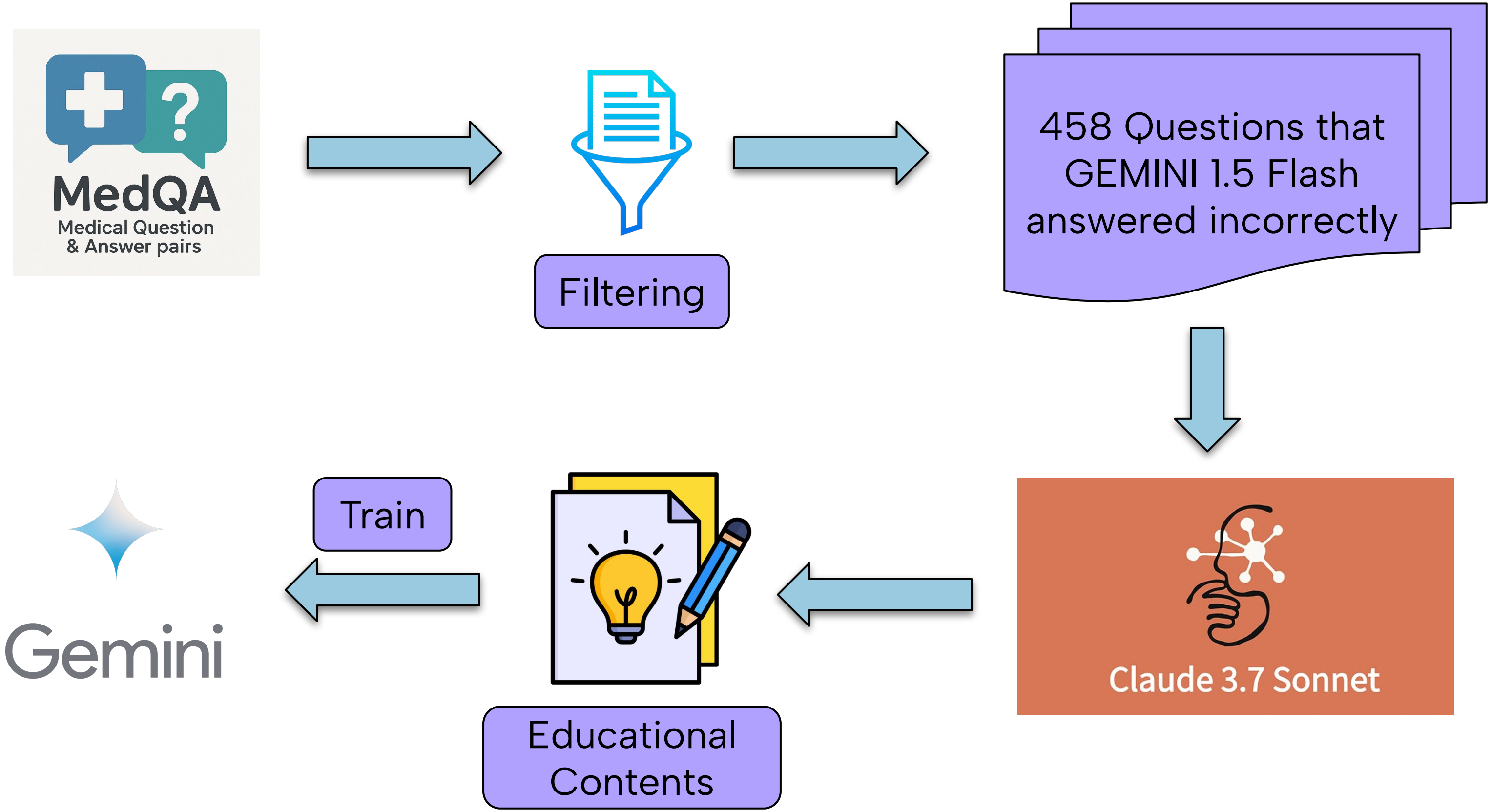
Results

- **The semantic content of the contextual prompts are important to improving performance.**

Auto-generated contextual prompts

Llama 2 7B	Accuracy (↑)							
	Anatomy	Clinical Knowledge	College Biology	College Medicine	Medical Genetics	Professional Medicine	MedQA	Average
Chat	44.07	46.79	48.61	39.02	49.00	**48.90**	38.96	45.05
Chat (CPT)	45.19	47.17	49.31	43.93	50.50	46.32	39.28	45.96
Chat (CFT)	**48.15**	**48.87**	**52.08**	44.22	**54.00**	46.69	**40.65**	**47.81**
Chat (TextAdaptCFT)	45.56	48.12	49.31	**44.80**	52.50	43.57	40.34	46.31

Improving GEMINI 1.5 Flash on MedQA



Improving GEMINI 1.5 Flash on MedQA

- **Contextual fine-tuning sees a 6% increase relative to continued pre-training.**

Model (Method)	**Accuracy (%)**
Gemini-1.5-Flash (CPT)	37.18
Gemini-1.5-Flash (CFT)	**43.89**

Understanding CFT with Synthetic Experiments

First, train a model that can learn a class of functions $\mathcal{F} = \{f \mid f(x) = w^\top x, w \in \mathbb{R}^d\}$.

Such that, for most functions, the model can approximate $f(x_{\text{query}})$

Then, consider learning a new class of functions $g \in \mathcal{G}$

That is a composition function $\mathcal{G} = \{g \mid g(x) = h(f(x)), h \in \mathcal{D}_{\mathcal{H}}\}$

Synthetic Experiments: Training

General prompt construction

$$P^i = (x_1, f(x_1), x_2, f(x_2), \ldots, x_i, f(x_i), x_{i+1})$$

Training Objective

$$\min_\theta, \mathbb{E}_P \left[\frac{1}{k+1} \sum_{i=0}^{k} \ell \left(M_\theta(P^i), f(x_{i+1}) \right) \right]$$

Synthetic Experiments: Fine-Tuning

Fine-tuning

1. Polynomial combination: $\mathcal{G} = \{g \mid g(x) = f(x) + f(x)^2\}$.
2. Multiple linear relationships: $\mathcal{G} = \{g \mid g(x) = f(x) + w_2^\top x, w_2 \in \mathbb{R}^d\}$

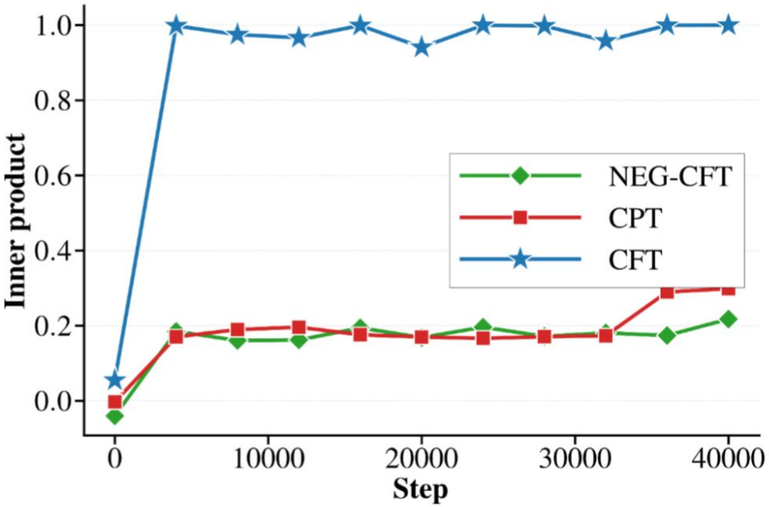
CPT $$P_{CPT} = (x_1, g(x_1), x_2, g(x_2), \ldots, x_k, g(x_k))$$

CFT $$P_{\text{CFT}} = (x_1, f(x_1), x_2, f(x_2), \ldots, x_k, f(x_k), x_1, g(x_1), x_2, g(x_2), \ldots, x_k, g(x_k))$$

NEG-CFT $$P_{\text{NEG-CFT}} = (x_1, r_1, x_2, r_2, \ldots, x_k, r_k, x_1, g(x_1), x_2, g(x_2), \ldots, x_k, g(x_k))$$

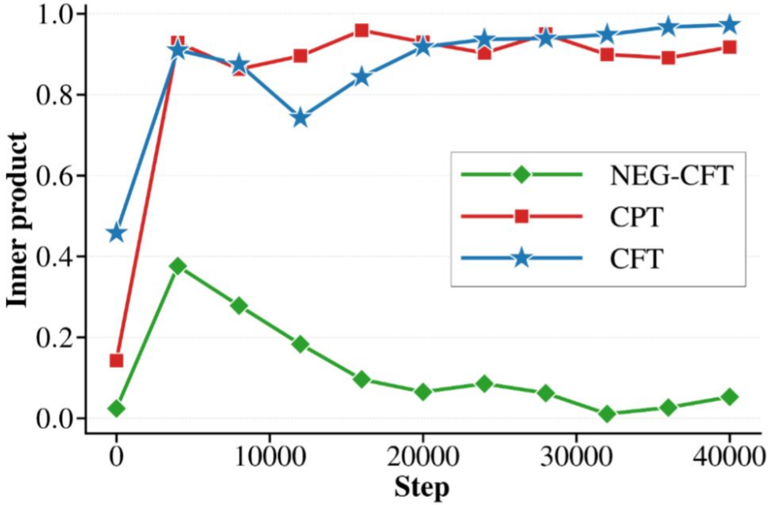
Synthetic Experiments: Results

- **Contextual prompts help the model capture the underlying functional relationships.**



(a) Inner product vs Step

$$\nabla_x g(x_{query}) = w_1 + 2(w_2^\top x_{query})w_2$$



(b) Inner product vs Step

$$\nabla_x g(x_{query}) = w_1 + w_2$$

Conclusion

- Introduced **contextual fine-tuning (CFT)**, a generalization of instruction fine-tuning.
- Leverages contextual gradients to guide learning through contextual prompts.
- Demonstrated improvement over traditional continued domain pre-training.
- Open-sourced a biomedical dataset curated from MDPI journals and open-source medical textbooks.

Future Work

- Hypothesis
  - Similar to Prystawski et al. (2023) findings on chain-of-thought prompting.
    - Local reasoning steps in pre-training corpora simulate step-by-step reasoning.
    - Contextual cues likely exist in pre-training data.
- Examine mechanisms by which prompts provide supervisory signals during learning.
- Test CFT on different data types:
  - Longer context lengths
  - Lower information density content (e.g., Reddit posts)
  - Currently validated only on high-density information (medical journals/textbooks)

Thank you!