Adapting Large Language Models via Reading Comprehension

Daixuan Cheng, Shaohan Huang, Furu Wei

https://huggingface.co/AdaptLLM

Domain Adaptation of LLMs

- Training from Scratch[1]:

  Substantial computational and data costs[2]

- Instruction Fine-tuning[3]:

  Uncertainties about how well fine-tuned LLMs grasp generalizable domain knowledge[4]

- Retrieval-augmented Prompting[5]:

  An application of LLM rather than a direct enhancement to the LLM itself[6]

[1] Bloomberggpt: A large language model for finance
[2] FinGPT: Open-Source Financial Large Language Models
[3] Large Language Models Encode Clinical Knowledge
[4] The False Promise of Imitating Proprietary LLMs
[5] ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge
[6] Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey

Domain-Adaptive Pre-Training of LLMs

- Domain-Adaptive Pre-Training[1] (DAPT) of NLU models endows generalizable domain knowledge with reduced costs
- Does DAPT still hold for large-scale NLG models?

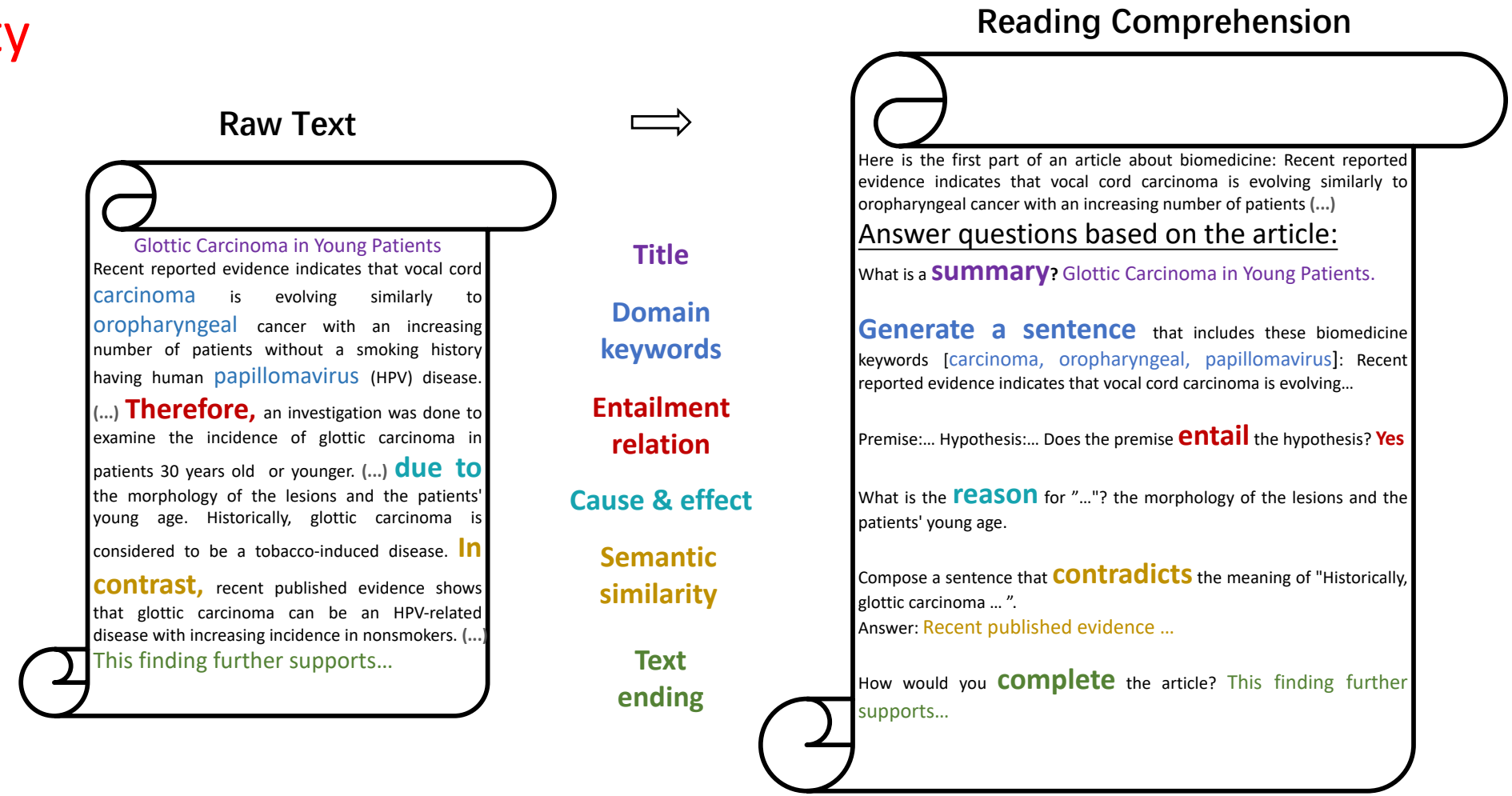
Method	Prompting			Fine-tuning			Knowledge Prob	
	BioMed.	Finance	Law	BioMed.	Finance	Law	BioMed.	Law
General LLM	**44.2**	**58.6**	34.2	64.2	79.9	42.0	36.5	45.0
DAPT	41.7	57.6	**35.0**	**66.5**	**80.9**	**45.4**	**36.9**	**45.6**

- Continued pre-training on the raw domain corpora can endow the LLM with new knowledge, but drastically hurts its prompting ability

[1] Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Adapting LLMs via Reading Comprehension

- Reading Comprehension: augment domain-specific pre-training corpora with relevant tasks, enabling the LLM to learn knowledge and improve prompting ability



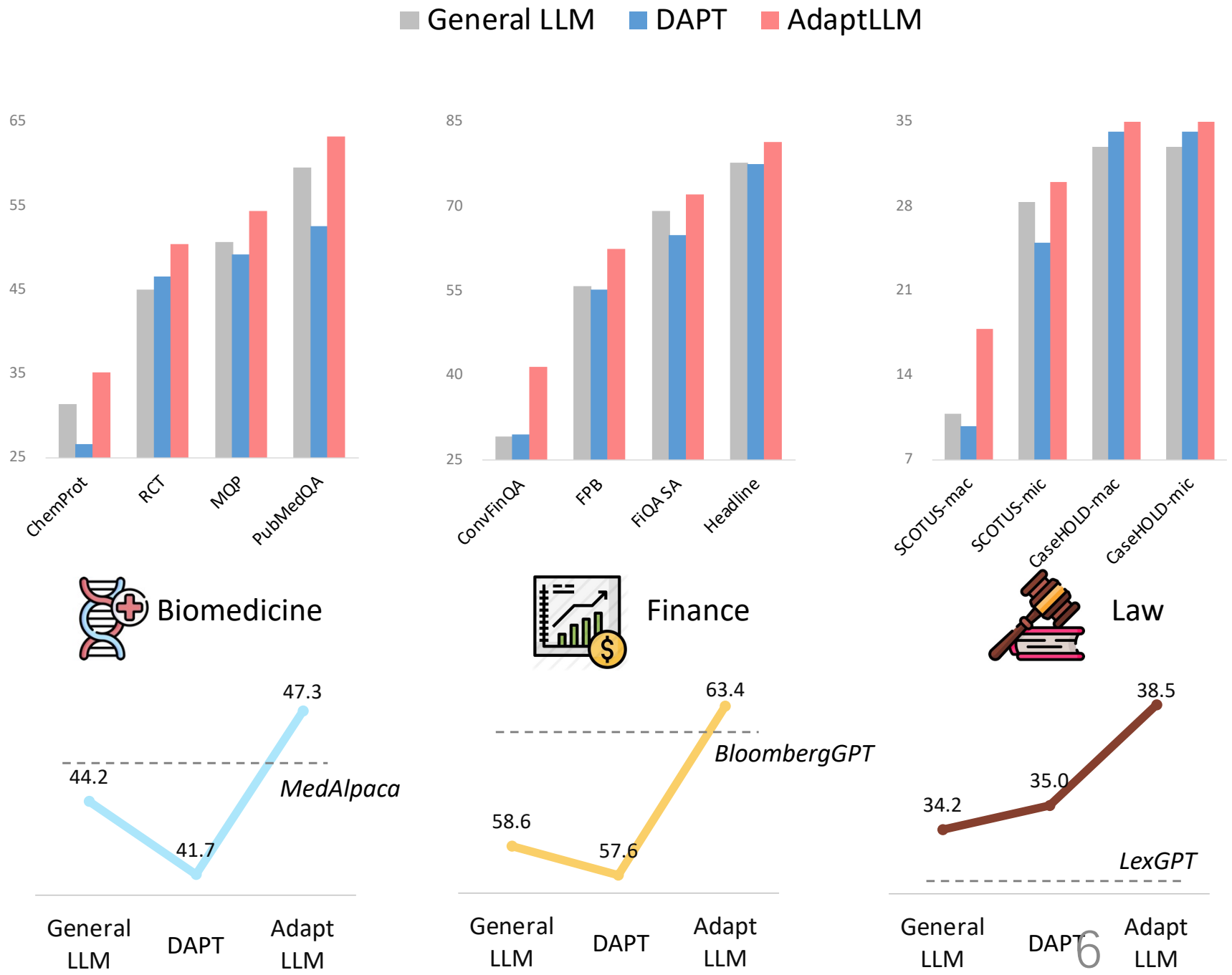
Method

- Mine tasks from raw texts
- Benefit from domain-specific knowledge embedded in raw texts and the enhanced prompting ability provided by the constructed tasks

Task Type	Mining Pattern	Input-output Template
Summarization		
Title Topic	Title as summary {SENT1} {VERBAL} {SENT2}	What is a summary? {TITLE} {SENT1} is about: {SENT2}
Word-to-Text		
Word-to-text	Domain keywords as input; sentence as output	Generate a sentence about these {DOMAIN} keywords [{WORD1}, {WORD2}, {WORD3}]: {SENT}
Definition	{WORD} {VERBAL} {SENT}	How to define {WORD}? {SENT}
Natural Language Inference		
Entail Neutral Contradict	{SENT1} {VERBAL}, {SENT2}	Does "{SENT1}" entail "{SENT2}"? {Yes/Maybe/No}
Commonsense Reasoning		
Cause-effect Effect-cause	{SENT1} {VERBAL}, {SENT2} {SENT1} {VERBAL} {SENT2}	What is the {effect/cause} of {SENT1}? {SENT2}
Paragraph Detection		
Similar Different	{SENT1} {VERBAL}, {SENT2}	Compose a sentence to {support/contradict} "{SENT1}". {SENT2}
Text Completion		
Text completion	Text ending as completion	How would you complete the article? {ENDING}

Main Results

- Consistent prompting gains cross three domains
- Our 7B model exhibits competitive performance with domain-specific models of much larger scales, such as BloombergGPT-50B



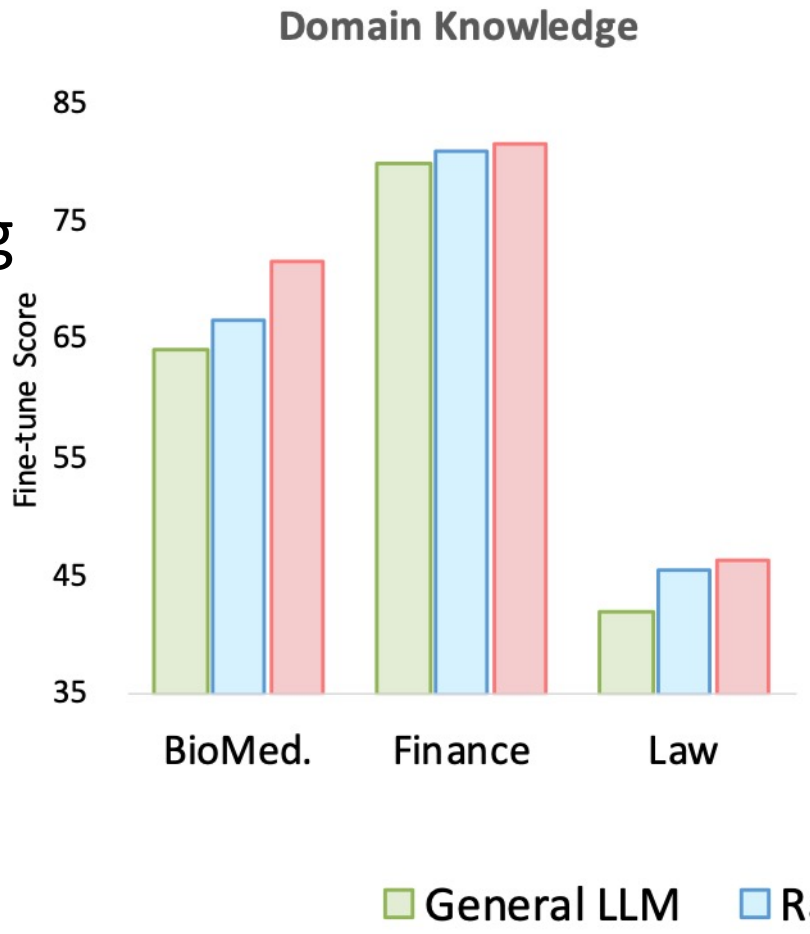
Analysis of Domain Knowledge and Prompting Ability

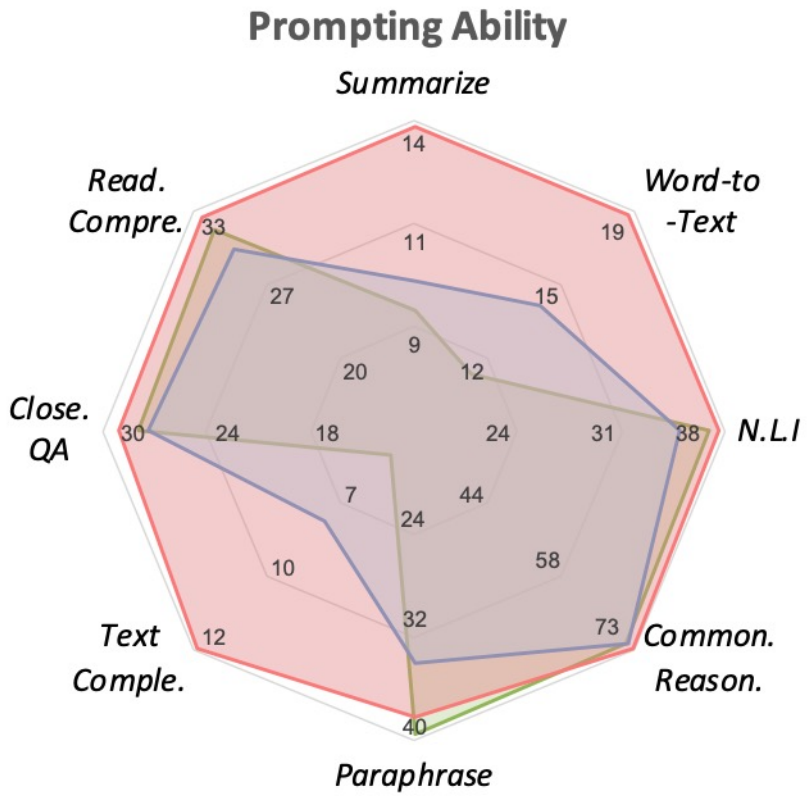
- Domain Knowledge

Consistent fine-tuning and knowledge probing improvements

- Prompting Ability

Training on domain-specific reading comprehension texts achieves better results than the general LLM





Thanks!

Daixuan Cheng, Shaohan Huang, Furu Wei

https://huggingface.co/AdaptLLM