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-Al (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 DAPT	44.2 41.7	58.6 57.6	34.2 35.0	64.2 66.5	79.9 80.9	42.0 45.4	36.5 36.9	45.0 45.6

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

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

Raw Text

Glottic Carcinoma in Young Patients
Recent reported evidence indicates that vocal cord
Carcinoma is evolving similarly to
Oropharyngeal cancer with an increasing
number of patients without a smoking history
having human papillomavirus (HPV) disease.

(...) **Therefore**, an investigation was done to examine the incidence of glottic carcinoma in patients 30 years old or younger. (...) **due to** the morphology of the lesions and the patients' young age. Historically, glottic carcinoma is considered to be a tobacco-induced disease. In

contrast, recent published evidence shows that glottic carcinoma can be an HPV-related disease with increasing incidence in nonsmokers. (...) This finding further supports...



Title

Domain keywords

Entailment relation

Cause & effect

Semantic similarity

Text ending



Here is the first part of an article about biomedicine: Recent reported evidence indicates that vocal cord carcinoma is evolving similarly to oropharyngeal cancer with an increasing number of patients (...)

Answer questions based on the article:

What is a **Summary?** Glottic Carcinoma in Young Patients.

Generate a sentence that includes these biomedicine keywords [carcinoma, oropharyngeal, papillomavirus]: Recent reported evidence indicates that vocal cord carcinoma is evolving...

Premise:... Hypothesis:... Does the premise **entail** the hypothesis? **Yes**

What is the **reason** for "..."? the morphology of the lesions and the patients' young age.

Compose a sentence that **Contradicts** the meaning of "Historically, glottic carcinoma ... ".

Answer: Recent published evidence ...

How would you **complete** the article? This finding further supports...

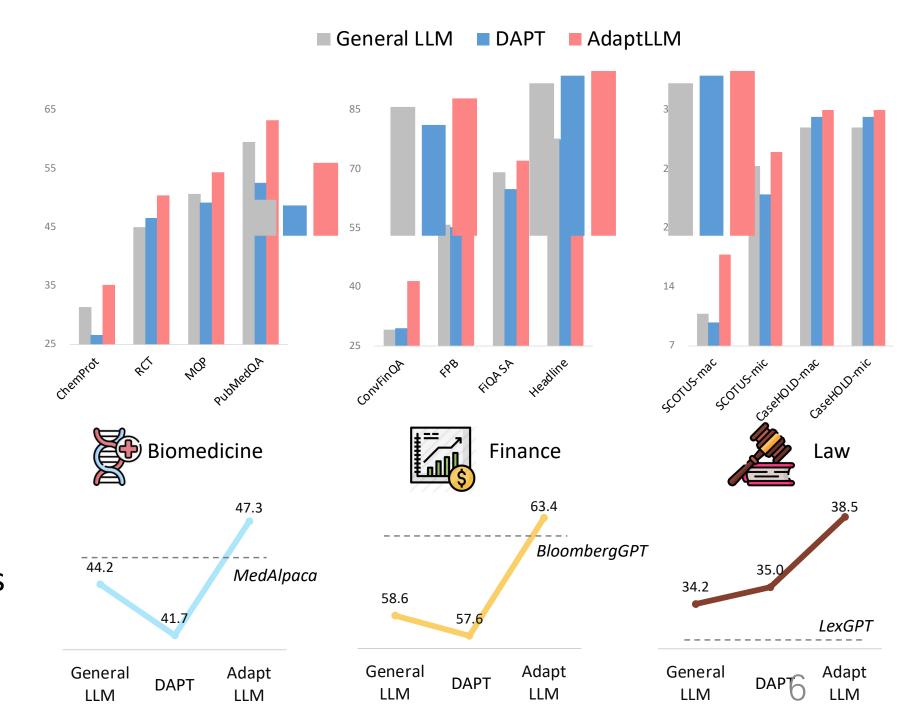
Method

- Mine tasks from raw texts
- Benefit from domainspecific 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 Rea	soning					
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}	<pre>Compose a sentence to {support/ contradict} "{SENT1}". {SENT2}</pre>				
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

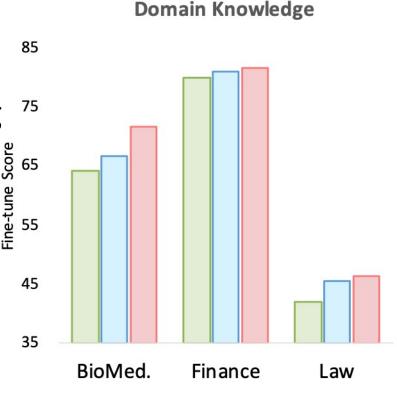


Analysis of Domain Knowledge and Prompting Ability

Domain Knowledge
 Consistent fine-tuning
 and knowledge probing
 improvements

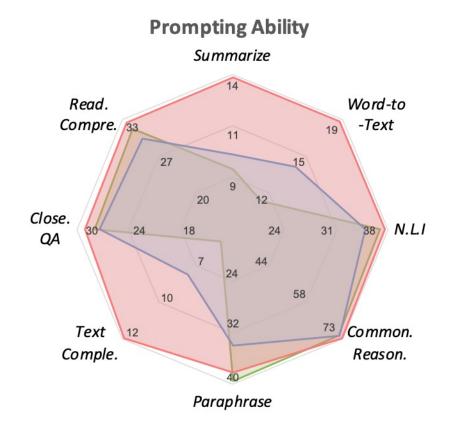
Prompting Ability

Training on domainspecific reading
comprehension texts
achieves better results
than the general LLM



General LLM

Raw Text



Read. Compre.

Thanks!

Daixuan Cheng, Shaohan Huang, Furu Wei

https://huggingface.co/AdaptLLM