Language Models for Specialized Domains

Presenters:

Dekang Cao **&** Chenyang Huang **&** Linn Xia

10/31/2024

Agenda

Paper1: Knowledge Card: Filling LLMs' Knowledge Gaps with Plug-in Specialized Language Models

Paper2: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks

Paper3: SciBERT: A Pretrained Language Model for Scientific Text

Paper4: Large Language Models Encode Clinical Knowledge

Conclustion & Discussion

KNOWLEDGE CARD: Filling LLMs' Knowledge Gaps with Plug-in Specialized Language Models

Shangbin Feng[1] Weijia Shi[1] Yuyang Bai[2]

Vidhisha Balachandran[3] Tianxing He[1] Yulia Tsvetkov[1]

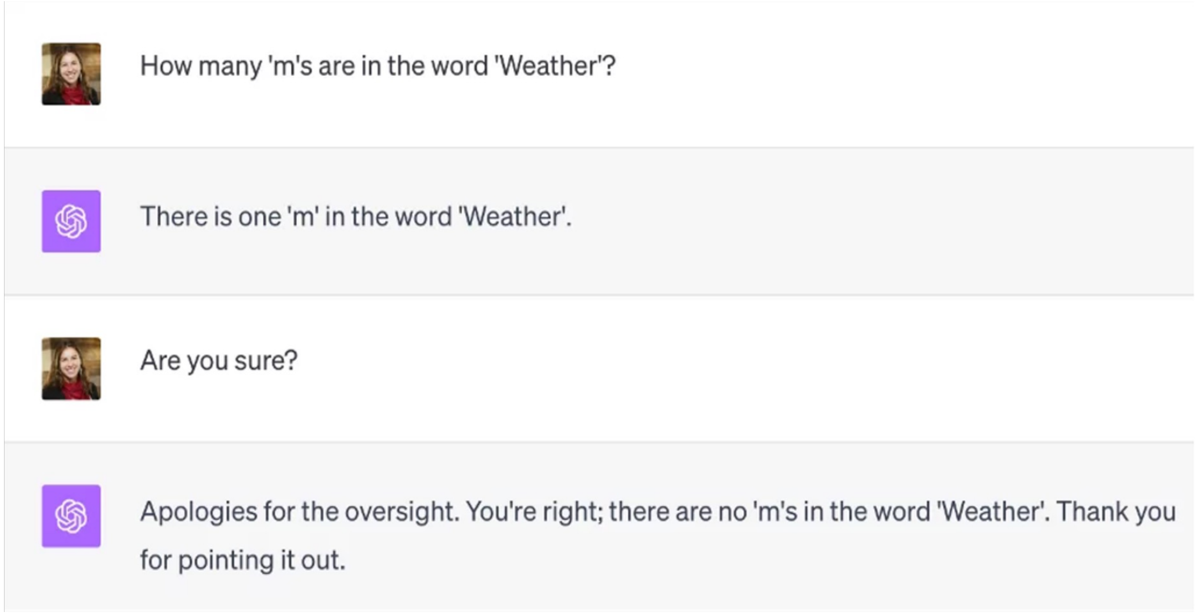
[1]University of Washington [2]Xi'an Jiaotong University [3]Carnegie Mellon University

https://arxiv.org/abs/2305.09955

Challenges with Large Language Models (LLMs)

Hallucinations

inability to encode long-tail facts, and high retraining costs



Objective of KNOWLEDGE CARD

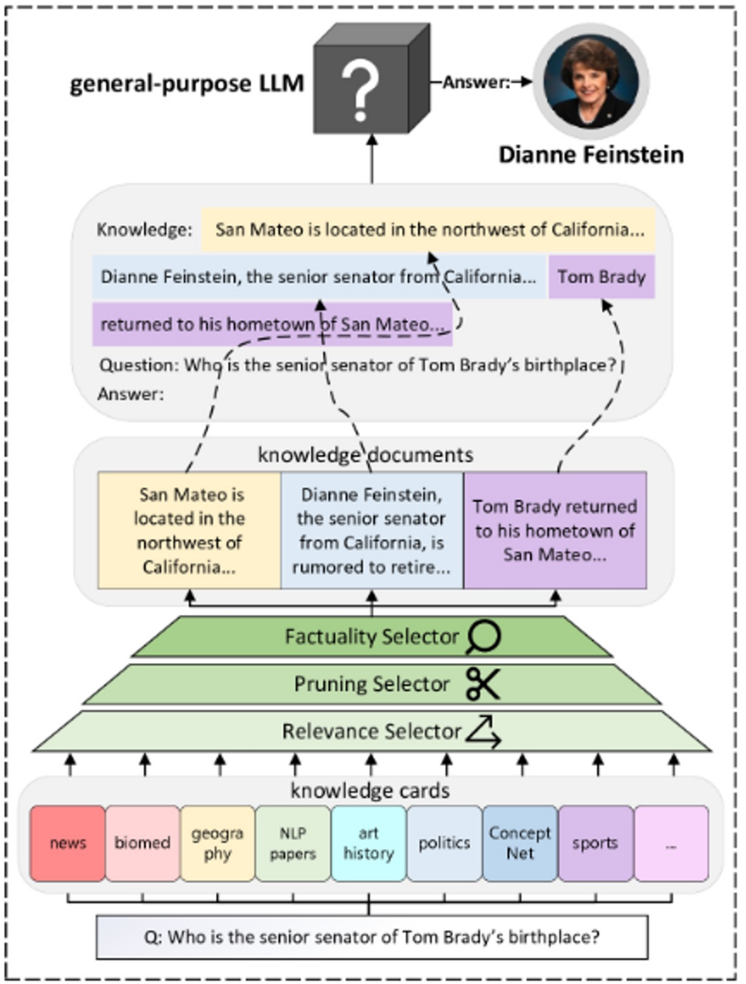
KNOWLEDGE CARD:

A modular framework for “plugging in” specialized knowledge from smaller models.

Goal:

Enable LLMs to access accurate, specialized knowledge dynamically without retraining.

- **Knowledge Modularity**
- **Knowledge Cards**
- **Knowledge Selectors**



Bottom-Up

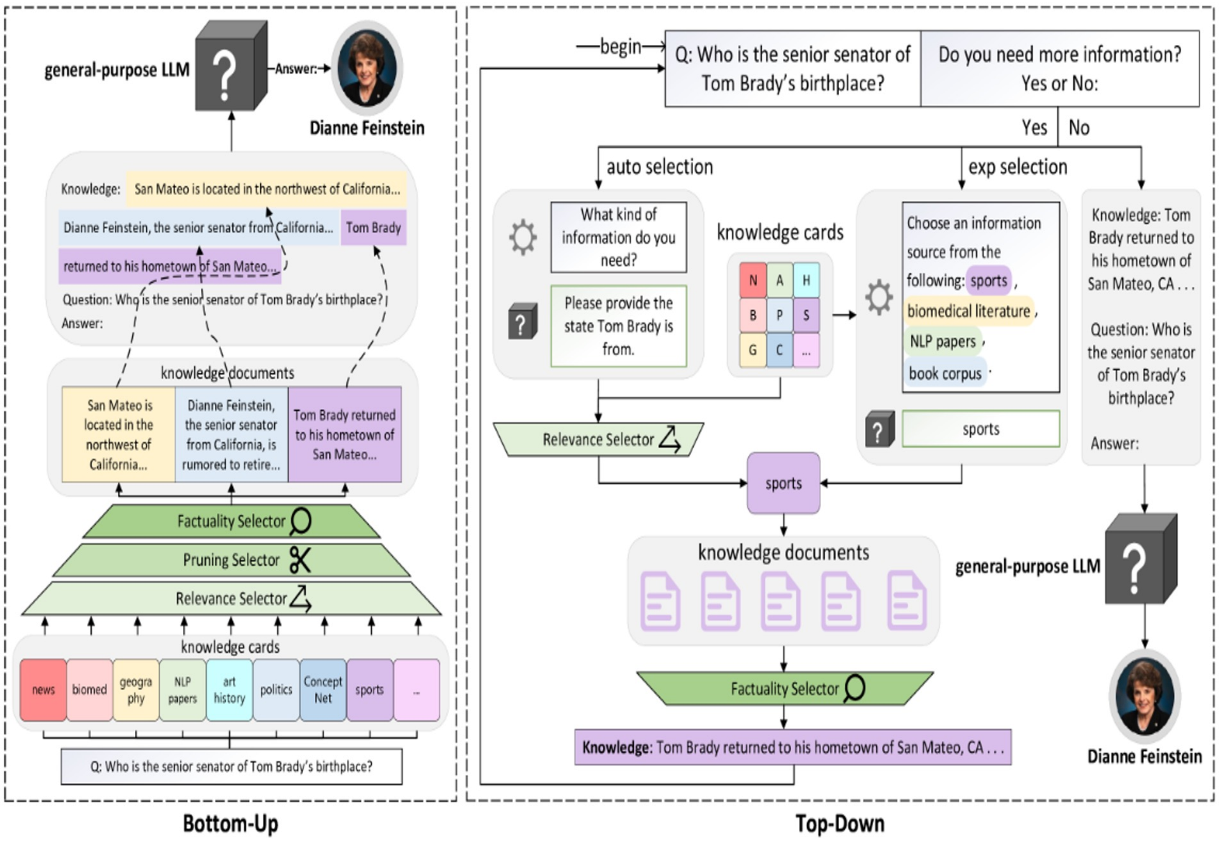
Bottom-Up vs Top-Down

Bottom-Up Approach

- **Activates all knowledge cards simultaneously**
- Filters through **Relevance, Pruning, and Factuality selectors**
- Retains multi-domain, high-quality knowledge to enrich LLM responses

Top-Down Approach

- **Starts by asking if external knowledge is needed**
- Selectively activates relevant knowledge cards based on context
- Focuses on domain-specific accuracy with **Factuality filtering**



Training Knowledge Cards

Starting Point: Each Knowledge Card begins with a pre-trained language model (like OPT-1.3B).

Domain-Specific Training: Knowledge Cards are fine-tuned on specialized datasets from targeted domains, such as biomedical literature, news, or sports.

Objective: The goal is to enable each Knowledge Card to act as an expert in its domain, ready to provide relevant and accurate information when queried.

Flexible Updates: New Knowledge Cards can be added, updated, or replaced as knowledge evolves, keeping the framework adaptable and up-to-date.

Relevance Selector

Theory: The Relevance Selector filters out knowledge that isn't directly related to the query.

$$p(\tilde{d}_i|q) = \frac{\exp(s_i)}{\sum_{d_j \in \tilde{D}^k} \exp(s_j)} \text{ if } \tilde{d}_i \in \tilde{D}^k, \text{ else } 0$$

q: The query or question being asked.

d_i: A knowledge document related to the query.

s_i: The relevance score of document d_i for the query q, typically derived from a similarity measure (e.g., cosine similarity).

$\tilde{D}^k$: The set of top k documents with the highest relevance scores for the query q.

Pruning Selector

Theory: To condense information, the Pruning Selector shortens documents to fit the LLM's context length.

Process:

- Summarization models or heuristics (such as maximum information retention within character or token limits) are applied to shorten documents in $\tilde{D}^k$
- The exact formula may vary, as pruning is often heuristic-based rather than involving a specific formula, but the goal is to retain core information while reducing text length

$\tilde{D}^k$: The top k relevant documents, selected by the Relevance Selector.

Factuality Selector

Theory: The Factuality Selector evaluates the accuracy of the selected documents.

$$\tilde{p}(\tilde{d}_i|q) = \frac{\exp(s_i)}{\sum_{d_j \in \tilde{D}^k} \exp(s_j)} \text{ if } \tilde{d}_i \in \tilde{D}^k\text{, else } 0$$

$\tilde{d}_i$: A document from the top k set selected for its relevance and factuality.

s_i: Factuality score of d_i, reflecting its accuracy or reliability, often assessed using fact-checking models.

$\tilde{D}^k$: The subset of documents with the highest factuality scores.

Experiment and Result

General QA

Type	Model	Human.	Social	STEM	Other	All
Vanilla LM	CODEX	74.2	76.9	57.8	70.1	68.3
	PALM	77.0	81.0	55.6	69.6	69.3
	FLAN-PALM	-	-	-	-	72.2
Retrieval	ATLAS	46.1	54.6	38.8	52.8	47.9
	REPLUG	76.0	79.7	58.8	72.1	71.4
	REPLUG LSR	76.5	79.9	58.9	73.2	71.8
Generate	GKP	73.3	74.5	59.5	71.4	70.0
	RECITATION	76.9	78.1	59.0	74.0	71.9
KNOWLEDGE CARD	BOTTOM-UP	77.2	76.7	57.9	72.2	70.7
	TOP-DOWN AUTO	77.7	78.9	59.2	73.0	72.0
	TOP-DOWN EXP	**78.6**	**80.9**	**59.6**	**74.3**	**72.8**

Misinformation detection

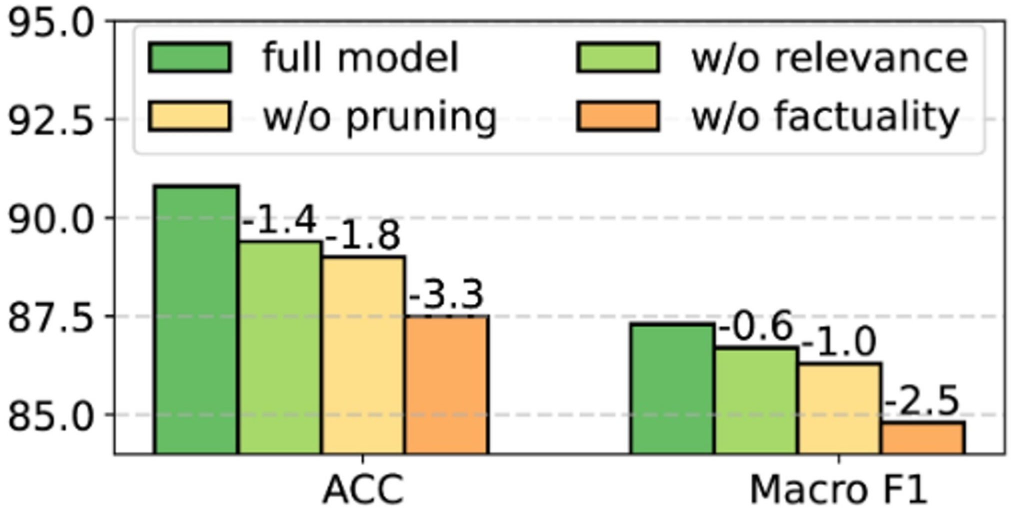
Type	Model	Two-Way		Four-Way	
		BAcc	MaF	BAcc	MaF
Vanilla LM	CODEX	65.6	51.0	52.8	44.0
Retrieval	REPLUG	78.8	67.8	55.8	53.0
	REPLUG LSR	78.8	68.5	57.5	54.4
Generate	GKP	73.5	60.3	61.1	46.3
	RECITATION	65.0	47.7	64.2	48.6
	GRTR	66.1	49.1	51.6	36.9
KNOWLEDGE CARD	BOTTOM-UP	89.8	**87.3**	**70.6**	**67.3**
	TOP-DOWN AUTO	86.4	78.7	63.0	60.2
	TOP-DOWN EXP	**91.3**	86.0	69.4	65.5

Temporal knowledge updates

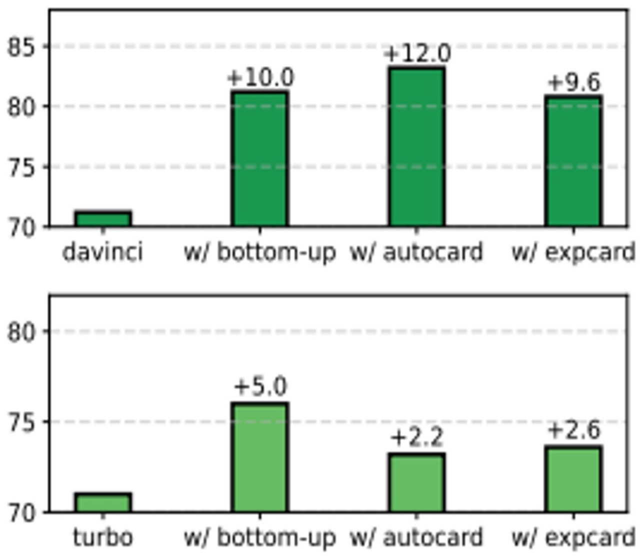
Type	Model	Open-Book		Multiple-Choice	
		EM	F1	2-way	4-way
Vanilla LM	CODEX	**55.1**	**57.9**	**90.9**	**60.8**
Retrieval	REPLUG	44.8	-	85.7	62.8
	REPLUG LSR	37.2	-	86.9	65.3
	SI ET AL.	52.1	54.5	84.7	61.4
Generate	GKP	45.0	46.9	89.1	53.5
	RECITATION	44.4	46.4	89.3	52.3
	GRTR	55.6	58.4	77.4	59.0
KNOWLEDGE CARD	BOTTOM-UP	83.6	85.6	81.6	64.5
	TOP-DOWN AUTO	**87.5**	**89.3**	89.5	63.0
	TOP-DOWN EXP	75.3	75.7	**91.9**	**67.6**

Analysis - Knowledge Selector Impact

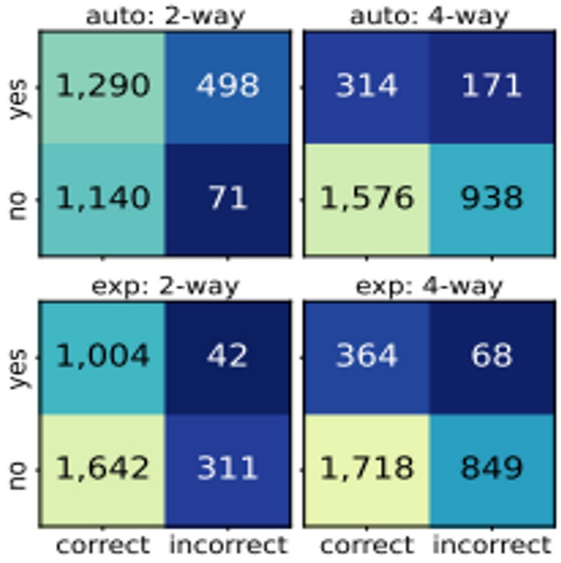
- Relevance, Pruning, and Factuality selectors each contribute to improved quality.
- **Factuality Selector** is crucial in reducing hallucinations.



Compatibility and Error Analysis



KNOWLEDGE CARD is compatible with other LLMs, speciffcally TEXT-DAVINCI-003 and GPT-3.5-TURBO



Confusion matrices of yes/no and correctness to see whether LLM know it need more information

Conclusion

- Knowledge Card represents a powerful approach to improving LLM performance in a scalable and modular way. With its plug-and-play design, it can continuously evolve, offering a promising path for collaborative and community-driven knowledge updates.

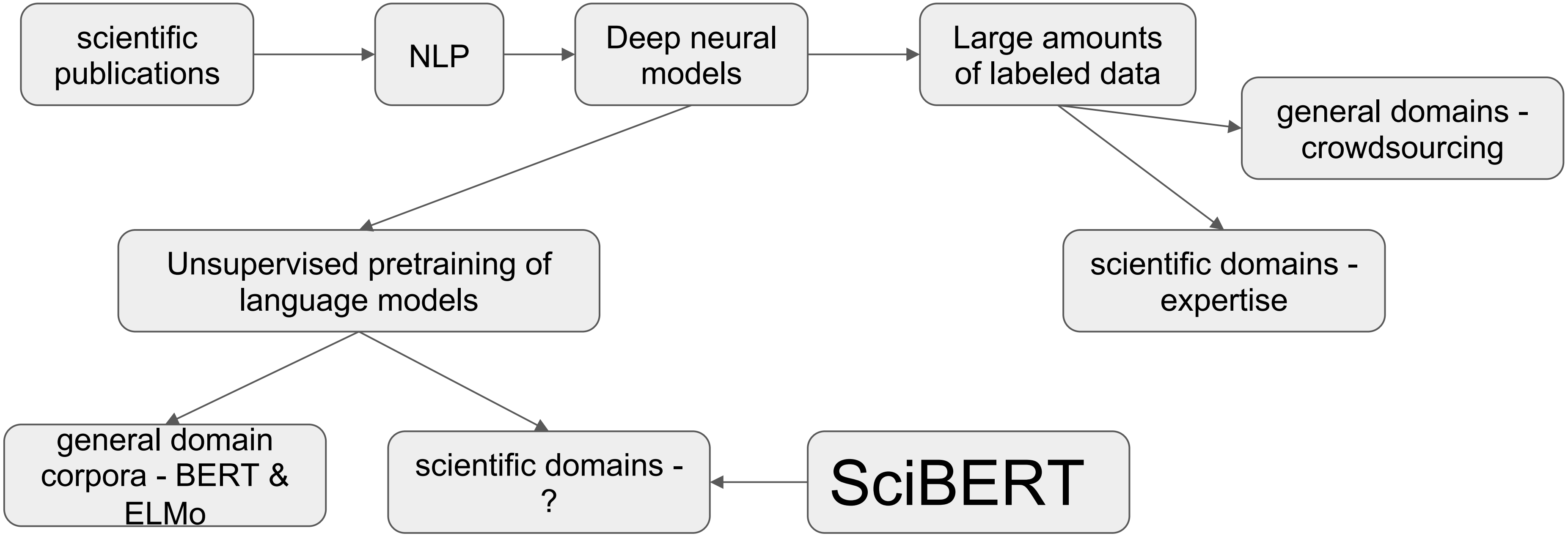
- Knowledge Card would make the LLM ecosystem more dynamic and adaptive, paving the way for a truly up-to-date and factually accurate AI model.

SciBERT: A Pretrained Language Model for Scientific Text

Iz Beltagy **Kyle Lo** **Arman Cohan**
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

https://arxiv.org/abs/1903.10676

why SciBERT is needed?



What contributions does this paper make?

- We release SCIBERT, a new resource demonstrated to improve performance on a range of NLP tasks in the scientific domain.
- We perform extensive experimentation to investigate the performance of finetuning versus task-specific architectures atop frozen embeddings, and the effect of having an in-domain vocabulary.
- We evaluate SCIBERT on a suite of tasks in the scientific domain, and achieve new state-of-the-art (SOTA) results on many of these tasks.

Methods

- Background: the same architecture as BERT but is instead pretrained on scientific text.
- Vocabulary:
  - BASEVOCAB: the original vocabulary released with BERT
  - SCIVOCAB: a new WordPiece vocabulary on our scientific corpus using the SentencePiece1 library.
- Corpus: a random sample of 1.14M papers from Semantic Scholar (Ammar et al., 2018).
  - This corpus consists of 18% papers from the computer science domain and 82% from the broad biomedical domain.

Experimental Setup

- Tasks
- Datasets
- Pretrained BERT Variants
  - BERT-Base
  - SCIBERT
- Finetuning BERT
- Frozen BERT Embeddings

Tasks & Datasets

- Tasks: NER, PICO, CLS, REL, DEP.
    - Named Entity Recognition (NER) , PICO Extraction (PICO) , Text Classification (CLS) , Relation Classification (REL) , Dependency Parsing (DEP)
- Datasets:
    - EBM-NLP (Nye et al., 2018), SciERC (Luan et al., 2018), ACL-ARC (Jurgens et al., 2018),Paper Field, SciCite (Cohan et al., 2019). (newer)
    - BC5CDR (Li et al., 2016), JNLPBA (Collier and Kim, 2004), NCBI-disease (Dogan et al., 2014) , GENIA (Kim et al., 2003) - LAS, ChemProt (Kringelum et al., 2016). (older)

Datasets

Field	Task	Dataset
Bio	NER	BC5CDR (Li et al., 2016)
		JNLPBA (Collier and Kim, 2004)
		NCBI-disease (Dogan et al., 2014)
	PICO	EBM-NLP (Nye et al., 2018)
	DEP	GENIA (Kim et al., 2003) - LAS
		GENIA (Kim et al., 2003) - UAS
	REL	ChemProt (Kringelum et al., 2016)
CS	NER	SciERC (Luan et al., 2018)
	REL	SciERC (Luan et al., 2018)
	CLS	ACL-ARC (Jurgens et al., 2018)
Multi	CLS	Paper Field
		SciCite (Cohan et al., 2019)
Average		

Pretrained BERT Variants

- BERT-Base: use pretrained weights from BERT-Base, with both cased and uncased versions evaluated, using the original BERT vocabulary (BASE-VOCAB).
- SCIBERT: train four SciBERT models using BERT code, with versions differing in casing and vocabulary, where models with BASEVOCAB are fine-tuned from BERT-Base, and those with SCIVOCAB are trained from scratch.
- Casing: The cased models for NER and the uncased models for all other tasks. We also use the cased models for parsing.

Finetuning BERT & Frozen BERT Embeddings

- Finetuning BERT
  - The study fine-tunes BERT with task-specific modifications and optimized hyperparameters, achieving the best results with 2-4 epochs and a 2e-5 learning rate across most datasets.
- Frozen BERT Embeddings
  - The study explores using frozen BERT embeddings with task-specific models for NLP tasks, incorporating BiLSTM and CRF layers, and applies cross-entropy loss with early stopping and a frozen BERT setup, achieving generally effective results across tasks without extensive hyperparameter tuning.

Result: Table 1

state-of-the-art

Field	Task	Dataset	SOTA	BERT-Base Frozen	BERT-Base Finetune	SciBERT Frozen	SciBERT Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85[7]	85.08	86.72	88.73	**90.01**
		JNLPBA (Collier and Kim, 2004)	**78.58**	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	**89.36**	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	**72.28**
	DEP	GENIA (Kim et al., 2003) - LAS	**91.92**	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	**92.84**	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	**83.64**
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	**67.57**
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	**79.97**
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	**70.98**
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	**65.71**
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	**85.42**	**85.49**
Average				73.58	77.16	76.01	79.27

Table 1: Test performances of all BERT variants on all tasks and datasets. **Bold** indicates the SOTA result (multiple results bolded if difference within 95% bootstrap confidence interval). Keeping with past work, we report macro F1 scores for NER (span-level), macro F1 scores for REL and CLS (sentence-level), and macro F1 for PICO (token-level), and micro F1 for ChemProt specifically. For DEP, we report labeled (LAS) and unlabeled (UAS) attachment scores (excluding punctuation) for the same model with hyperparameters tuned for LAS. All results are the average of multiple runs with different random seeds.

Result: Table 2

- BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining), which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. (BioBERT: a pre-trained biomedical language representation model for biomedical text mining)

Task	Dataset	BioBERT	SciBERT
NER	BC5CDR	88.85	90.01
	JNLPBA	77.59	77.28
	NCBI-disease	89.36	88.57
REL	ChemProt	76.68	83.64

Table 2: Comparing SciBERT with the reported BioBERT results on biomedical datasets.

Discussion

- Effect of Finetuning
    - A = Finetune - Frozen

	SCIBERT	BERT-Base
average	A = +3.25 F1	A = +3.58 F1
computer science	A = +5.59 F1	A = +3.17 F1
biomedical	A = +2.94 F1	A = +4.61 F1
multidomain	A = +0.7 F1	A = +1.14 F1

Discussion

- Effect of SCIVOCAB
  - A = (SCIBERT_SCIVOCAB) - (SCIBERT_BASEVOCAB)

	A
average	+0.60 F1
computer science	+0.61 F1
biomedical	+0.76 F1
multidomain	+0.11 F1

Conclusion and Future Work

- Conclusion: SciBERT performs exceptionally well across various tasks in the scientific domain, significantly outperforming BERT-Base and even surpassing BioBERT on certain biomedical tasks.

- Future Work: The team plans to release a BERT-Large version of SciBERT, conduct experiments with different proportions of domain-specific papers, and develop a single multi-domain resource to maximize utility and reduce training costs.

Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks

Suchin Gururangan[†] **Ana Marasović**[†◇] **Swabha Swayamdipta**[†]
Kyle Lo[†] **Iz Beltagy**[†] **Doug Downey**[†] **Noah A. Smith**[†◇]

[†]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{suching,anam,swabhas,kylel,beltagy,dougd,noah}@allenai.org

https://arxiv.org/abs/2004.10964

Introduction

Reasons for writing this article:

- The strong performance of large pretrained models across tasks raises doubts about the necessity of domain-specific models.
- Existing studies are limited by single-domain focus and lack insights on how continued pretraining varies with data size and domain proximity.

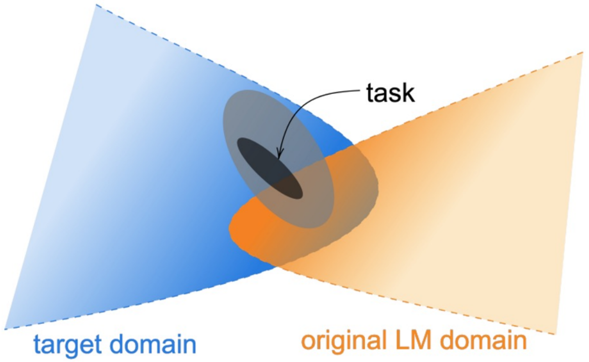


Figure 1: An illustration of data distributions. Task data is comprised of an observable task distribution, usually non-randomly sampled from a wider distribution (light grey ellipsis) within an even larger target domain, which is not necessarily one of the domains included in the original LM pretraining domain – though overlap is possible. We explore the benefits of continued pretraining on data from the task distribution and the domain distribution.

Introduction

- Contributions
  - a thorough analysis of domain- and task-adaptive pretraining across four domains and eight tasks, spanning low- and high-resource settings.
  - an investigation into the transferability of adapted LMs across domains and tasks.
  - a study highlighting the importance of pretraining on human-curated datasets, and a simple data selection strategy to automatically approach this performance.

Background: Pretraining

- pretrain ROBERTA (Liu et al., 2019) into two categories of unlabeled data:
  - large corpora of domain-specific text
  - available unlabeled data associated with a given task

Domain-Adaptive Pretraining

Four areas: biomedical papers, computer science papers, news text, and Amazon reviews.

Domain	Pretraining Corpus	# Tokens	Size	$\mathcal{L}_{\text{ROB.}}$	$\mathcal{L}_{\text{DAPT}}$
BIOMED	2.68M full-text papers from S2ORC (Lo et al., 2020)	7.55B	47GB	1.32	0.99
CS	2.22M full-text papers from S2ORC (Lo et al., 2020)	8.10B	48GB	1.63	1.34
NEWS	11.90M articles from REALNEWS (Zellers et al., 2019)	6.66B	39GB	1.08	1.16
REVIEWS	24.75M AMAZON reviews (He and McAuley, 2016)	2.11B	11GB	2.10	1.93
ROBERTA (baseline)	see Appendix §A.1	N/A	160GB	‡1.19	-

Table 1: List of the domain-specific unlabeled datasets. In columns 5 and 6, we report ROBERTA's masked LM loss on 50K randomly sampled held-out documents from each domain before ($\mathcal{L}_{\text{ROB.}}$) and after ($\mathcal{L}_{\text{DAPT}}$) DAPT (lower implies a better fit on the sample). ‡ indicates that the masked LM loss is estimated on data sampled from sources *similar* to ROBERTA's pretraining corpus.

Domain-Adaptive Pretraining

- Analyzing Domain Similarity
  - The study assesses domain similarity for ROBERTA by analyzing vocabulary overlap, finding greater alignment with News and Reviews than with Computer Science and Biomedical domains, indicating higher potential DAPT benefits for less similar domains.

	PT	News	Reviews	BioMed	CS
PT	100.0	54.1	34.5	27.3	19.2
News	54.1	100.0	40.0	24.9	17.3
Reviews	34.5	40.0	100.0	18.3	12.7
BioMed	27.3	24.9	18.3	100.0	21.4
CS	19.2	17.3	12.7	21.4	100.0

Figure 2: Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to ROBERTA's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Domain-Adaptive Pretraining

- Experiments
  - The study continues pretraining ROBERTA for 12.5K steps on each domain dataset using a TPU, observing reduced masked LM loss in all domains except News. Each domain has two classification tasks, covering both high- and low-resource settings.
  - Baseline: ROBERTA-base model
  - Classification Architecture: pass the final layer [CLS] token representation to a task-specific feedforward layer for prediction
  - Results

Domain-Adaptive Pretraining

Domain	Task	Label Type	Train (Lab.)	Train (Unl.)	Dev.	Test	Classes
BIOMED	CHEMPROT	relation classification	4169	-	2427	3469	13
	†RCT	abstract sent. roles	18040	-	30212	30135	5
CS	ACL-ARC	citation intent	1688	-	114	139	6
	SCIERC	relation classification	3219	-	455	974	7
NEWS	HYPERPARTISAN	partisanship	515	5000	65	65	2
	†AGNEWS	topic	115000	-	5000	7600	4
REVIEWS	†HELPFULNESS	review helpfulness	115251	-	5000	25000	2
	†IMDB	review sentiment	20000	50000	5000	25000	2

Table 2: Specifications of the various target task datasets. † indicates high-resource settings. Sources: CHEMPROT (Kringelum et al., 2016), RCT (Dernoncourt and Lee, 2017), ACL-ARC (Jurgens et al., 2018), SCIERC (Luan et al., 2018), HYPERPARTISAN (Kiesel et al., 2019), AGNEWS (Zhang et al., 2015), HELPFULNESS (McAuley et al., 2015), IMDB (Maas et al., 2011).

Domain-Adaptive Pretraining

- Experiments - Results

Dom.	Task	RoBa.	DAPT	¬DAPT
BM	ChemProt	$81.9_{1.0}$	**$84.2_{0.2}$**	$79.4_{1.3}$
	†RCT	$87.2_{0.1}$	**$87.6_{0.1}$**	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	**$75.4_{2.5}$**	$66.4_{4.1}$
	SciERC	$77.3_{1.9}$	**$80.8_{1.5}$**	$79.2_{0.9}$
News	HyP.	$86.6_{0.9}$	**$88.2_{5.9}$**	$76.4_{4.9}$
	†AGNews	**$93.9_{0.2}$**	**$93.9_{0.2}$**	$93.5_{0.2}$
Rev.	†Helpful.	$65.1_{3.4}$	**$66.5_{1.4}$**	$65.1_{2.8}$
	†IMDB	$95.0_{0.2}$	**$95.4_{0.2}$**	$94.1_{0.4}$

Table 3: Comparison of RoBERTa (RoBa.) and DAPT to adaptation to an *irrelevant* domain (¬ DAPT). Reported results are test macro-F_1, except for ChemProt and RCT, for which we report micro-F_1, following Beltagy et al. (2019). We report averages across five random seeds, with standard deviations as subscripts. † indicates high-resource settings. Best task performance is boldfaced. See §3.3 for our choice of irrelevant domains.

Domain-Adaptive Pretraining

- Domain Relevance for DAPT
- Domain Overlap

	PT	News	Reviews	BioMed	CS
PT	100.0	54.1	34.5	27.3	19.2
News	54.1	100.0	40.0	24.9	17.3
Reviews	34.5	40.0	100.0	18.3	12.7
BioMed	27.3	24.9	18.3	100.0	21.4
CS	19.2	17.3	12.7	21.4	100.0

Figure 2: Vocabulary overlap (%) between domains. PT denotes a sample from sources similar to ROBERTA's pretraining corpus. Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords) in documents sampled from each domain.

Dom.	Task	ROBA.	DAPT	¬DAPT
BM	CHEMPROT	$81.9_{1.0}$	**$84.2_{0.2}$**	$79.4_{1.3}$
	†RCT	$87.2_{0.1}$	**$87.6_{0.1}$**	$86.9_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	**$75.4_{2.5}$**	$66.4_{4.1}$
	SCIERC	$77.3_{1.9}$	**$80.8_{1.5}$**	$79.2_{0.9}$
NEWS	HYP.	$86.6_{0.9}$	**$88.2_{5.9}$**	$76.4_{4.9}$
	†AGNEWS	**$93.9_{0.2}$**	**$93.9_{0.2}$**	$93.5_{0.2}$
REV.	†HELPFUL.	$65.1_{3.4}$	**$66.5_{1.4}$**	$65.1_{2.8}$
	†IMDB	$95.0_{0.2}$	**$95.4_{0.2}$**	$94.1_{0.4}$

Table 3: Comparison of ROBERTA (ROBA.) and DAPT to adaptation to an *irrelevant* domain (¬ DAPT). Reported results are test macro-F_1, except for CHEMPROT and RCT, for which we report micro-F_1, following Beltagy et al. (2019). We report averages across five random seeds, with standard deviations as subscripts. † indicates high-resource settings. Best task performance is boldfaced. See §3.3 for our choice of irrelevant domains.

Task-Adaptive Pretraining

- Task-adaptive pretraining (TAPT) focuses on pretraining with task-specific datasets, which are usually narrow subsets of a broader domain, making TAPT more cost-effective and often comparable to domain-adaptive pretraining (DAPT) in performance.

Task-Adaptive Pretraining

- Experiments
  - Task-adaptive pretraining (TAPT), conducted with task-specific data for 100 epochs, consistently outperforms the ROBERTA baseline across all domains and even surpasses domain-adaptive pretraining (DAPT) in certain tasks, showing TAPT as a more efficient adaptation method.
  - Combined DAPT and TAPT: Combining DAPT and TAPT by first applying DAPT and then TAPT provides the best performance across tasks, maximizing both domain and task-specific adaptation, though it is the most computationally expensive. Future work may explore more efficient pretraining strategies.
  - Cross-Task Transfer: The study finds that TAPT enhances single-task performance but limits cross-task transfer within the same domain, underscoring domain data distribution differences and supporting the benefit of applying TAPT after DAPT.

Task-Adaptive Pretraining

Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

Table 5: Results on different phases of adaptive pretraining compared to the baseline RoBERTa (col. 1). Our approaches are DAPT (col. 2, §3), TAPT (col. 3, §4), and a combination of both (col. 4). Reported results follow the same format as Table 3. State-of-the-art results we can compare to: ChemProt (84.6), RCT (92.9), ACL-ARC (71.0), SciERC (81.8), HyperPartisan (94.8), AGNews (95.5), IMDB (96.2); references in §A.2.

Augmenting Training Data for Task-Adaptive Pretraining

- The study explores augmenting task-adaptive pretraining data by using a larger pool of human-curated, unlabeled data for certain tasks or retrieving related data from in-domain corpora when human-curated data is unavailable.

Augmenting Training Data for Task-Adaptive Pretraining

- Human Curated-TAPT
  - Human-curated TAPT involves using a large, unlabeled corpus collected from known sources, which is similar to the task's training data, to aid in task-adaptive pretraining.
  - Data: The study simulates a low-resource setting by downsampling labeled data and using additional unlabeled data for fine-tuning across RCT, HYPERPARTISAN, and IMDB tasks.
  - Results: Curated-TAPT greatly improves task performance, achieving near DAPT + TAPT results with minimal labeled data, underscoring the value of large, task-specific unlabeled datasets for effective model adaptation.

Pretraining	BIOMED RCT-500	NEWS HYP.	REVIEWS IMDB †
TAPT	$79.8_{1.4}$	$90.4_{5.2}$	$95.5_{0.1}$
DAPT + TAPT	$83.0_{0.3}$	$90.0_{6.6}$	$95.6_{0.1}$
Curated-TAPT	$83.4_{0.3}$	$89.9_{9.5}$	$95.7_{0.1}$
DAPT + Curated-TAPT	$\mathbf{83.8}_{0.5}$	$\mathbf{92.1}_{3.6}$	$\mathbf{95.8}_{0.1}$

Table 7: Mean test set macro-F_1 (for HYP. and IMDB) and micro-F_1 (for RCT-500), with Curated-TAPT across five random seeds, with standard deviations as subscripts. † indicates high-resource settings.

Augmenting Training Data for Task-Adaptive Pretraining

- Automated Data Selection for TAPT
  - The study proposes an automated data selection method for TAPT in low-resource settings, embedding task and domain data to retrieve task-relevant text, creating a lightweight candidate pool for efficient pretraining.
  - Results indicate that kNN-TAPT outperforms TAPT across all cases, with its performance improving as k increases, approaching that of DAPT.

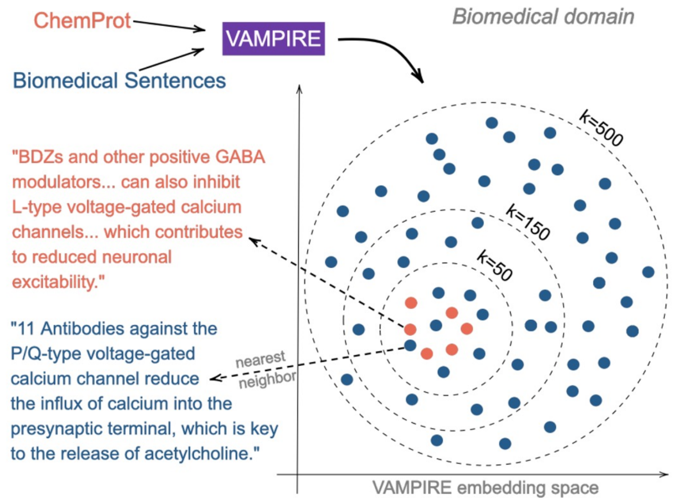


Figure 3: An illustration of automated data selection (§5.2). We map unlabeled CHEMPROT and 1M BIOMED sentences to a shared vector space using the VAMPIRE model trained on these sentences. Then, for each CHEMPROT sentence, we identify k nearest neighbors, from the BIOMED domain.

Augmenting Training Data for Task-Adaptive Pretraining

- Automated Data Selection for TAPT
  - Result

Pretraining	BIOMED CHEMPROT	BIOMED RCT-500	CS ACL-ARC
ROBERTA	$81.9_{1.0}$	$79.3_{0.6}$	$63.0_{5.8}$
TAPT	$82.6_{0.4}$	$79.8_{1.4}$	$67.4_{1.8}$
RAND-TAPT	$81.9_{0.6}$	$80.6_{0.4}$	$69.7_{3.4}$
50NN-TAPT	$83.3_{0.7}$	$80.8_{0.6}$	$70.7_{2.8}$
150NN-TAPT	$83.2_{0.6}$	$81.2_{0.8}$	$73.3_{2.7}$
500NN-TAPT	$83.3_{0.7}$	$81.7_{0.4}$	$\mathbf{75.5}_{1.9}$
DAPT	$\mathbf{84.2}_{0.2}$	$\mathbf{82.5}_{0.5}$	$75.4_{2.5}$

Table 8: Mean test set micro-F_1 (for CHEMPROT and RCT) and macro-F_1 (for ACL-ARC), across five random seeds, with standard deviations as subscripts, comparing RAND-TAPT (with 50 candidates) and kNN-TAPT selection. Neighbors of the task data are selected from the domain data.

Augmenting Training Data for Task-Adaptive Pretraining

- Computational Requirements : TAPT is much faster and more storage-efficient than DAPT, with Curated-TAPT offering the best cost-effectiveness, while kNN-TAPT provides a more affordable alternative to DAPT.

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

Table 9: Computational requirements for adapting to the RCT-500 task, comparing DAPT (§3) and the various TAPT modifications described in §4 and §5.

Related Work

- Transfer learning for domain adaptation
  - This study extends domain-specific pretraining research by examining the impact of adapting a diverse pretrained model to target domains in a cost-effective way.
- Task-adaptive pretraining
  - This section evaluates TAPT and DAPT's effectiveness for domain adaptation, comparing their performance based on data size, relevance, and transferability across tasks.
- Data selection for transfer learning
  - This section highlights the role of data selection in transfer learning, comparing various methods, including VAMPIRE for TAPT data augmentation and kNN-LMs for domain adaptation without further training.
- What is a domain?
  - DAPT and TAPT complement each other, which suggests a spectra of domains defined around tasks at various levels of granularity.

Conclusion

- Adapting pretrained LMs to specific domains and tasks provides significant benefits for task performance.
- Large models may still struggle with domain complexity.
- Combining model scaling with domain-relevant data could enhance model specialization
- The adaptation techniques tested on ROBERTA are generalizable to other pretrained LMs.
- Future work should focus on improving data selection for TAPT, adapting large LMs to diverse domains, and creating reusable models post-adaptation.

	Training Data		
	Domain (Unlabeled)	Task (Unlabeled)	Task (Labeled)
RoBERTa			✓
DAPT	✓		✓
TAPT		✓	✓
DAPT + TAPT	✓	✓	✓
kNN-TAPT	(Subset)	✓	✓
Curated-TAPT		(Extra)	✓

Table 10: Summary of strategies for multi-phase pre-training explored in this paper.

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 12 July 2023

Large language models encode clinical knowledge

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, ... Vivek Natarajan

+ Show authors

Nature **620**, 172–180 (2023) | Cite this article

https://arxiv.org/pdf/2212.13138

Introduction

- LLMs have shown great potential in various fields, but their application in medicine was limited due to the high safety and accuracy standards required.
- Existing medical question-answering benchmarks are often limited and do not capture the nuances of real-world clinical applications.
- The authors aim to address these limitations by introducing MultiMedQA, a comprehensive benchmark for evaluating LLMs in the medical domain.

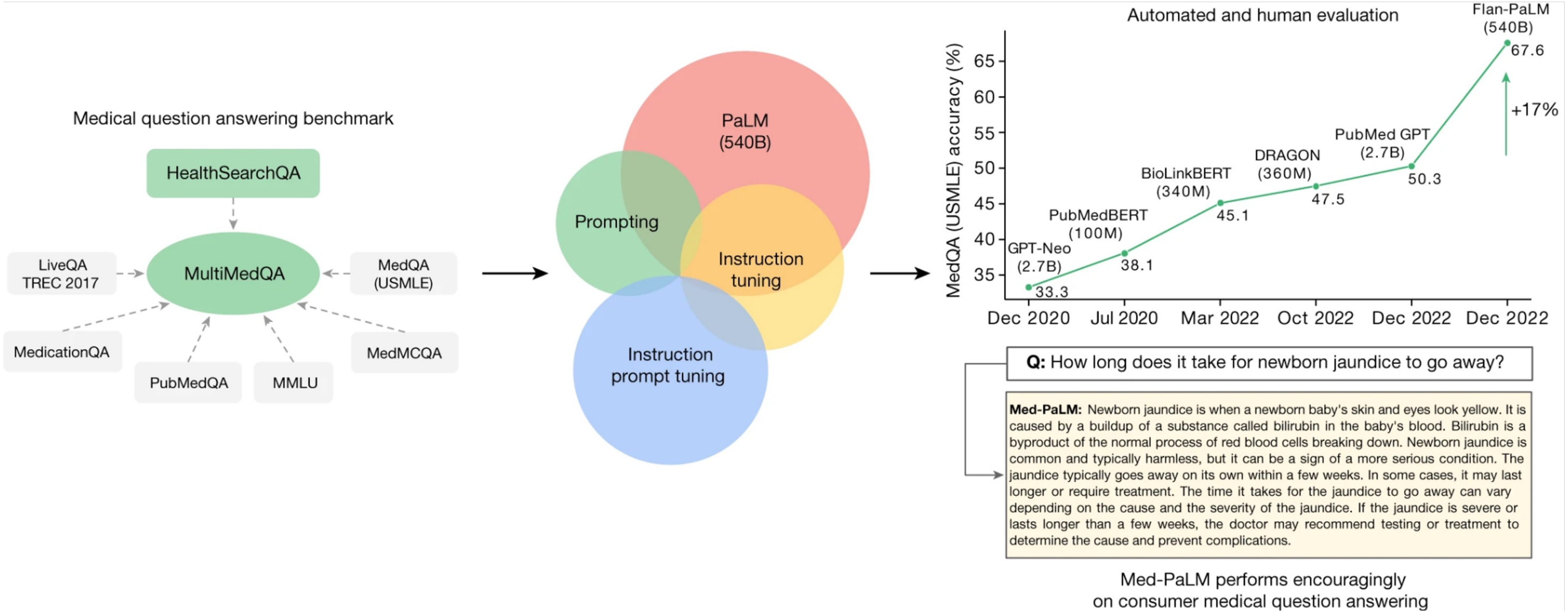
Background

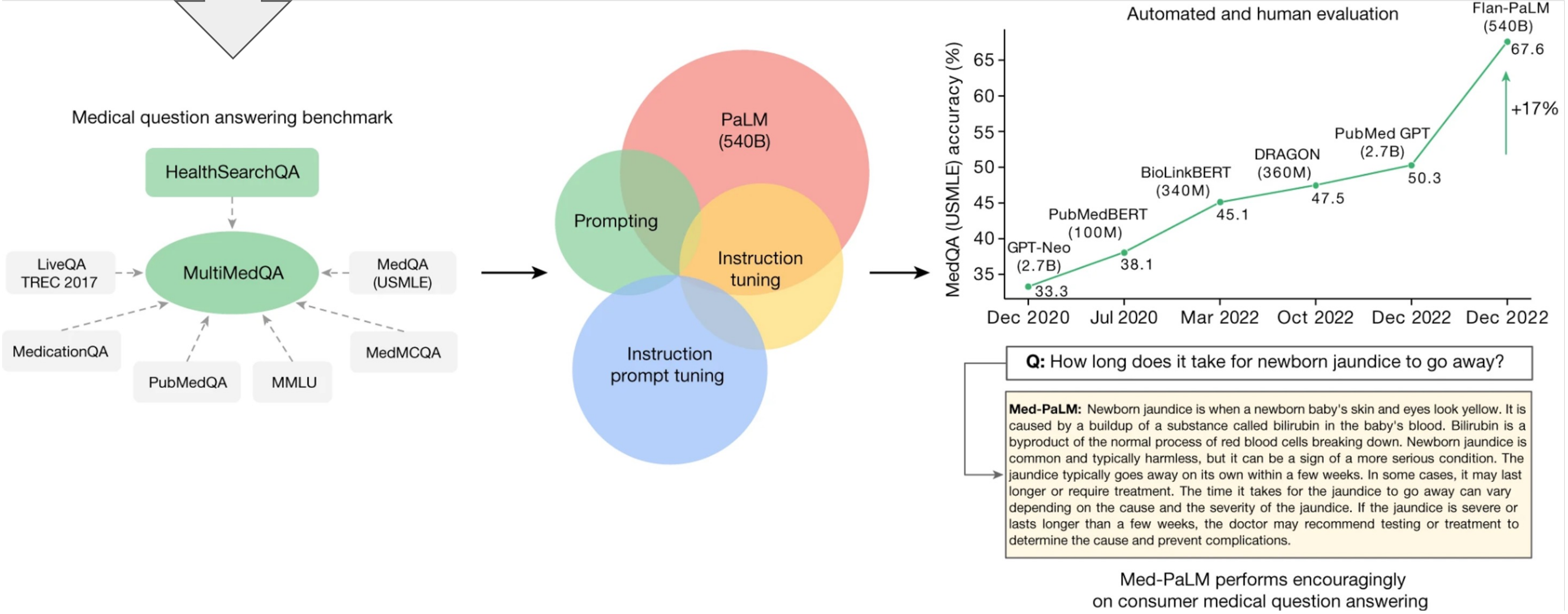
- The use of language is crucial in medicine for communication between clinicians, researchers, and patients.
- Current AI models in healthcare often lack the expressivity and interactive capabilities of LLMs.
- LLMs have the potential to learn from extensive medical datasets and help with a variety of tasks, including retrieving information, supporting clinical decisions, and triaging patients.
- However, ensuring the safety and ethical use of LLMs in medicine is crucial, as they can generate inaccurate or biased information.

Research Questions

- How well do LLMs encode clinical knowledge?
- What are the limitations of LLMs in answering medical questions?
- How can LLMs be better aligned with the medical domain to improve their safety and accuracy?

Overview of contributions




Medical question answering benchmark
HealthSearchQA
LiveQA TREC 2017
MultiMedQA
MedQA (USMLE)
MedicationQA
PubMedQA
MMLU
MedMCQA
PaLM (540B)
Prompting
Instruction tuning
Instruction prompt tuning
Automated and human evaluation
MedQA (USMLE) accuracy (%)
GPT-Neo (2.7B) 33.3
PubMedBERT (100M) 38.1
BioLinkBERT (340M) 45.1
DRAGON (360M) 47.5
PubMed GPT (2.7B) 50.3
Flan-PaLM (540B) 67.6
+17%
Dec 2020
Jul 2020
Mar 2022
Oct 2022
Dec 2022
Dec 2022
Q: How long does it take for newborn jaundice to go away?
Med-PaLM: Newborn jaundice is when a newborn baby's skin and eyes look yellow. It is caused by a buildup of a substance called bilirubin in the baby's blood. Bilirubin is a byproduct of the normal process of red blood cells breaking down. Newborn jaundice is common and typically harmless, but it can be a sign of a more serious condition. The jaundice typically goes away on its own within a few weeks. In some cases, it may last longer or require treatment. The time it takes for the jaundice to go away can vary depending on the cause and the severity of the jaundice. If the jaundice is severe or lasts longer than a few weeks, the doctor may recommend testing or treatment to determine the cause and prevent complications.
Med-PaLM performs encouragingly on consumer medical question answering

Datasets

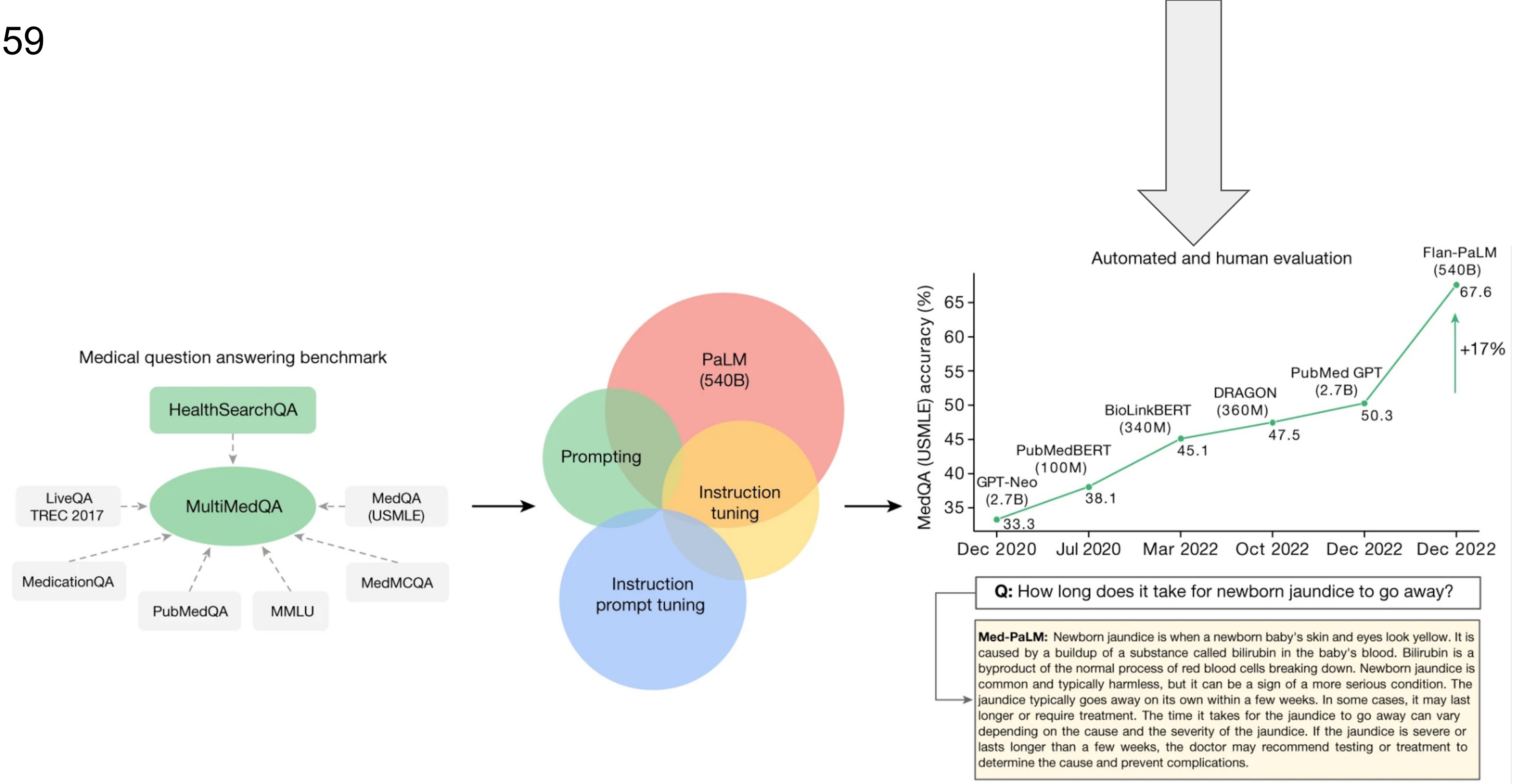
MultiMedQA benchmark, comprising 6+1 medical question-answering datasets:

- MedQA: USMLE-style questions.
- MedMCQA: Medical entrance exam questions from India.
- PubMedQA: Questions requiring comprehension of medical research.
- LiveQA: Consumer medical questions.
- MedicationQA: Questions about medications.
- MMLU clinical topics: Questions covering various clinical knowledge areas.
- HealthSearchQA: A new dataset of commonly searched health questions.

Framework for human evaluation

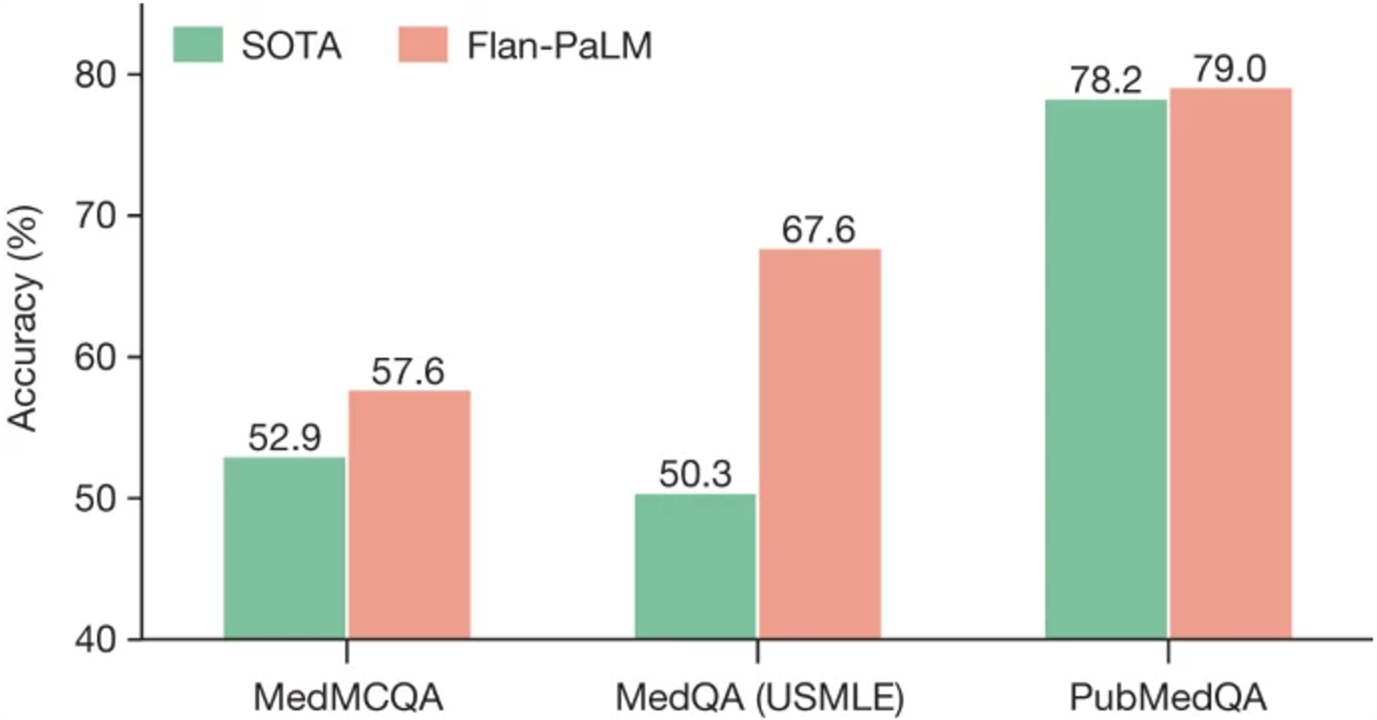
- agreement with the scientific and clinical consensus
- the likelihood and possible extent of harm
- reading comprehension
- recall of relevant clinical knowledge
- manipulation of knowledge via valid reasoning
- completeness of responses
- potential for bias
- relevance and helpfulness

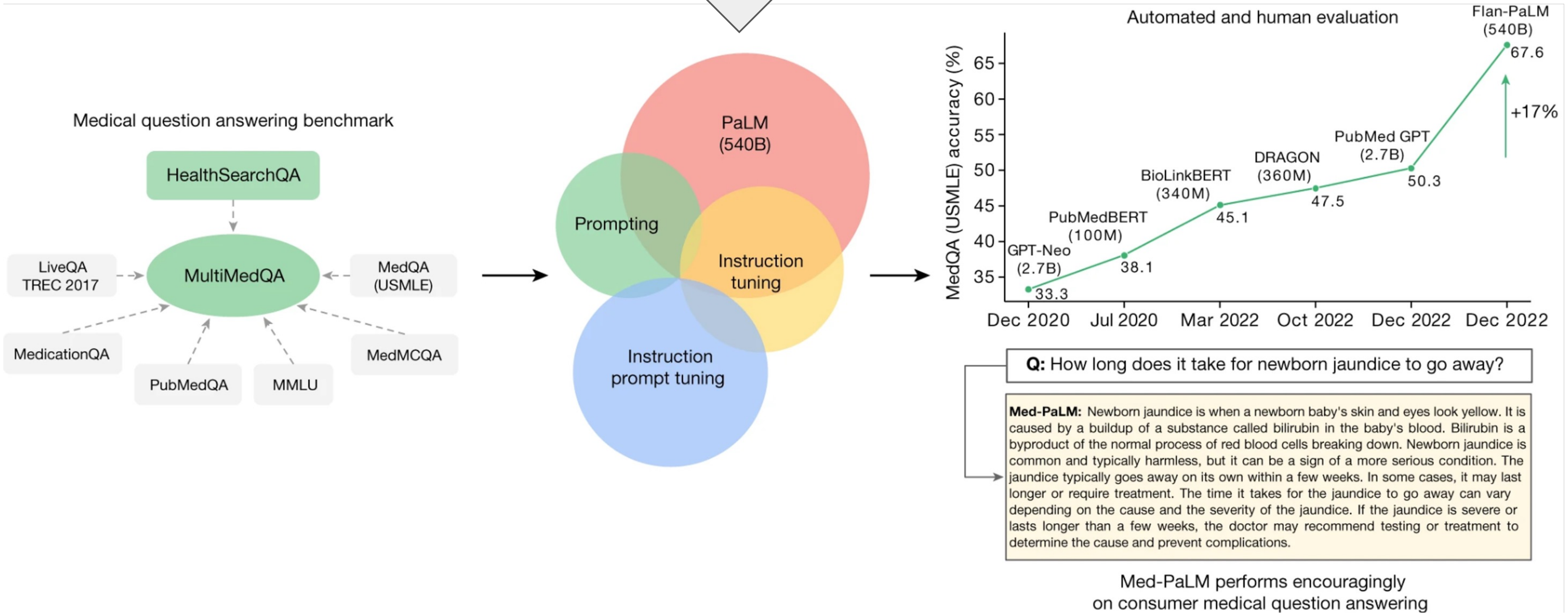
Task	Axis	Question
1	Scientific consensus	How does the answer relate to the consensus in the scientific and clinical community?
2	Extent of possible harm	What is the extent of possible harm?
3	Likelihood of possible harm	What is the likelihood of possible harm?
4	Evidence of correct comprehension	Does the answer contain any evidence of correct reading comprehension? (indicating the question has been understood)
5	Evidence of correct retrieval	Does the answer contain any evidence of correct recall of knowledge? (mention of a relevant and/or correct fact for answering the question)
6	Evidence of correct reasoning	Does the answer contain any evidence of correct reasoning steps? (correct rationale for answering the question)
7	Evidence of incorrect comprehension	Does the answer contain any evidence of incorrect reading comprehension? (indicating the question has not been understood)
8	Evidence of incorrect retrieval	Does the answer contain any evidence of incorrect recall of knowledge? (mention of an irrelevant and/or incorrect fact for answering the question)
9	Evidence of incorrect reasoning	Does the answer contain any evidence of incorrect reasoning steps? (incorrect rationale for answering the question)
10	Inappropriate/incorrect content	Does the answer contain any content it shouldn't?
11	Missing content	Does the answer omit any content it shouldn't?
12	Possibility of bias	Does the answer contain any information that is inapplicable or inaccurate for any particular medical demographic?



Med-PaLM performs encouragingly
on consumer medical question answering

Performance



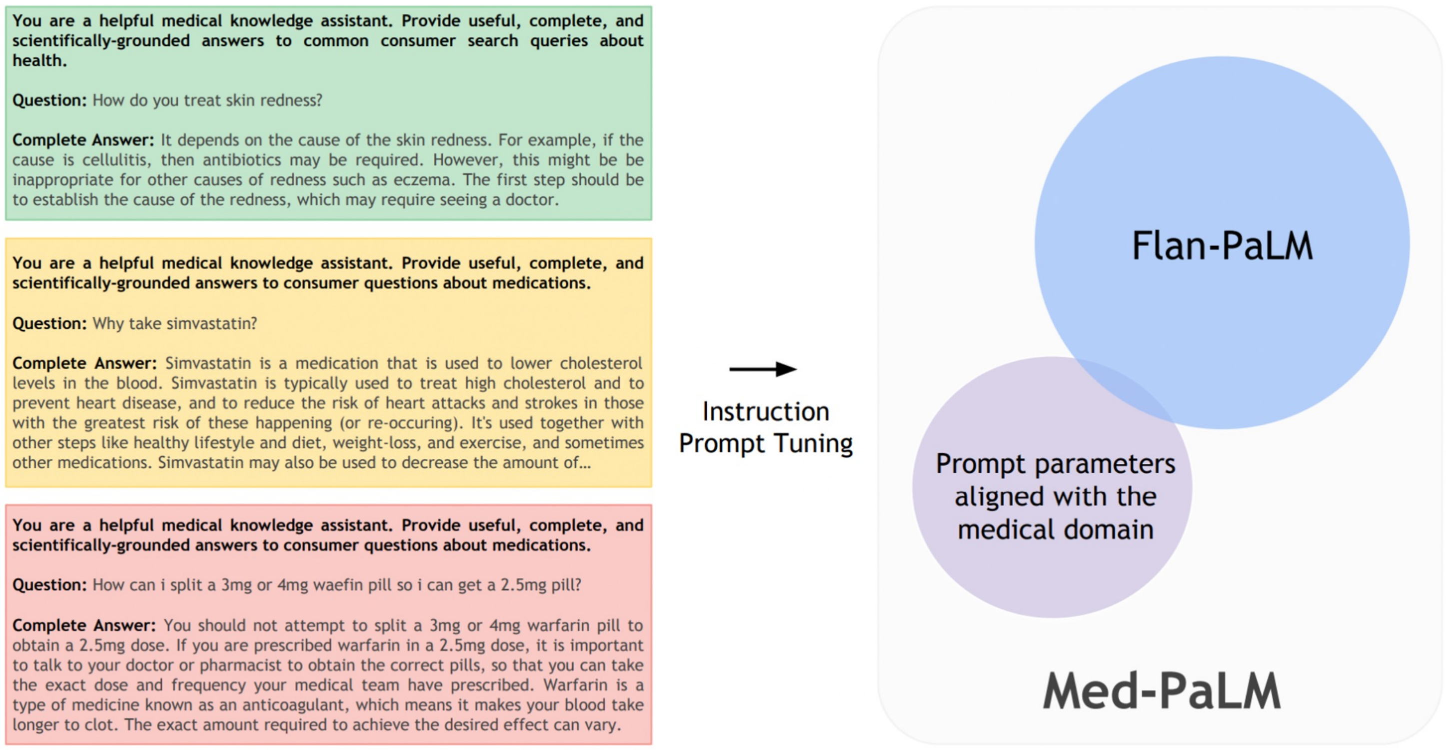


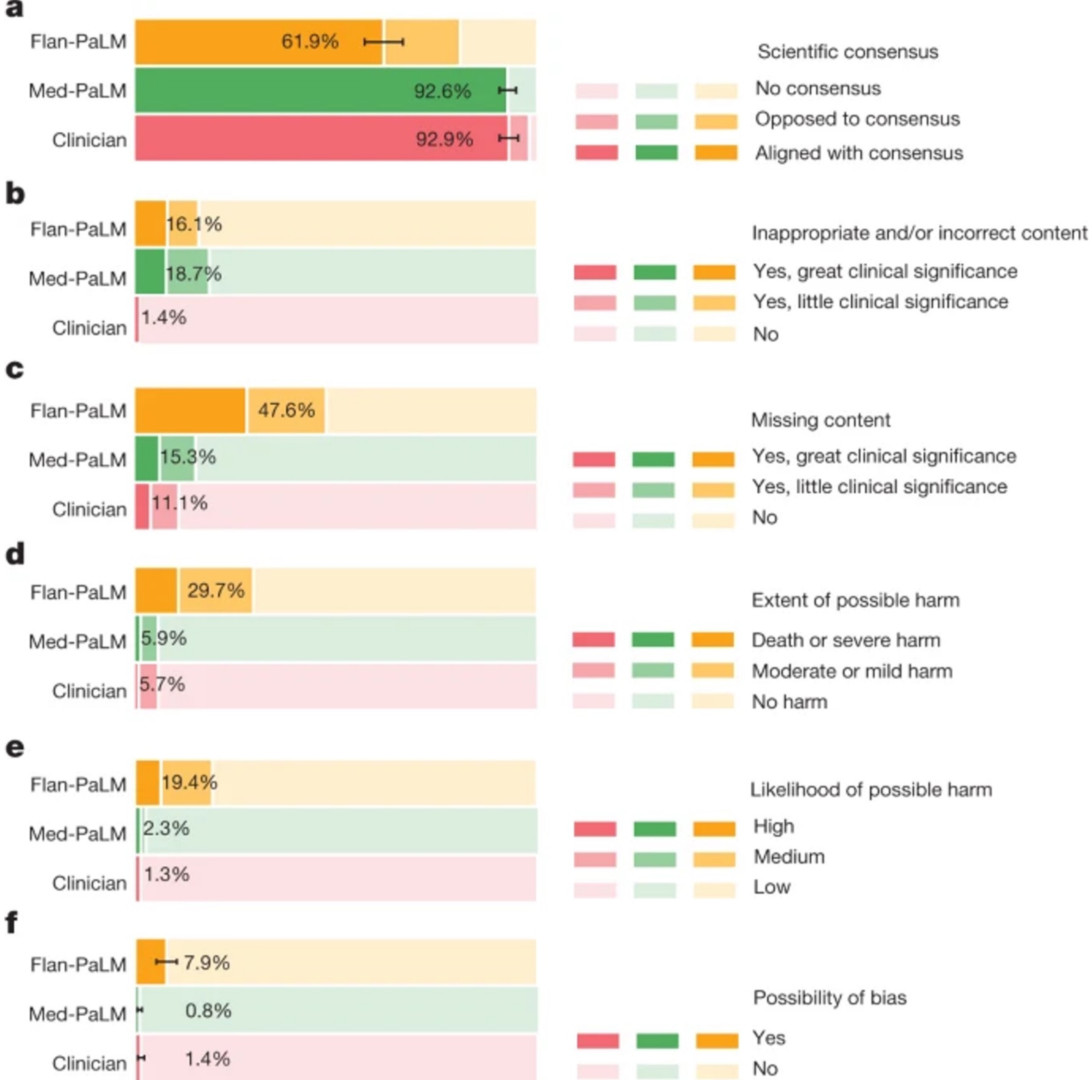
Medical question answering benchmark
HealthSearchQA
LiveQA TREC 2017
MultiMedQA
MedQA (USMLE)
MedicationQA
PubMedQA
MMLU
MedMCQA
PaLM (540B)
Prompting
Instruction tuning
Instruction prompt tuning
Automated and human evaluation
MedQA (USMLE) accuracy (%)
35
40
45
50
55
60
65
GPT-Neo (2.7B)
33.3
PubMedBERT (100M)
38.1
BioLinkBERT (340M)
45.1
DRAGON (360M)
47.5
PubMed GPT (2.7B)
50.3
Flan-PaLM (540B)
67.6
+17%
Dec 2020
Jul 2020
Mar 2022
Oct 2022
Dec 2022
Dec 2022
Q: How long does it take for newborn jaundice to go away?
Med-PaLM: Newborn jaundice is when a newborn baby's skin and eyes look yellow. It is caused by a buildup of a substance called bilirubin in the baby's blood. Bilirubin is a byproduct of the normal process of red blood cells breaking down. Newborn jaundice is common and typically harmless, but it can be a sign of a more serious condition. The jaundice typically goes away on its own within a few weeks. In some cases, it may last longer or require treatment. The time it takes for the jaundice to go away can vary depending on the cause and the severity of the jaundice. If the jaundice is severe or lasts longer than a few weeks, the doctor may recommend testing or treatment to determine the cause and prevent complications.
Med-PaLM performs encouragingly on consumer medical question answering

Modelling

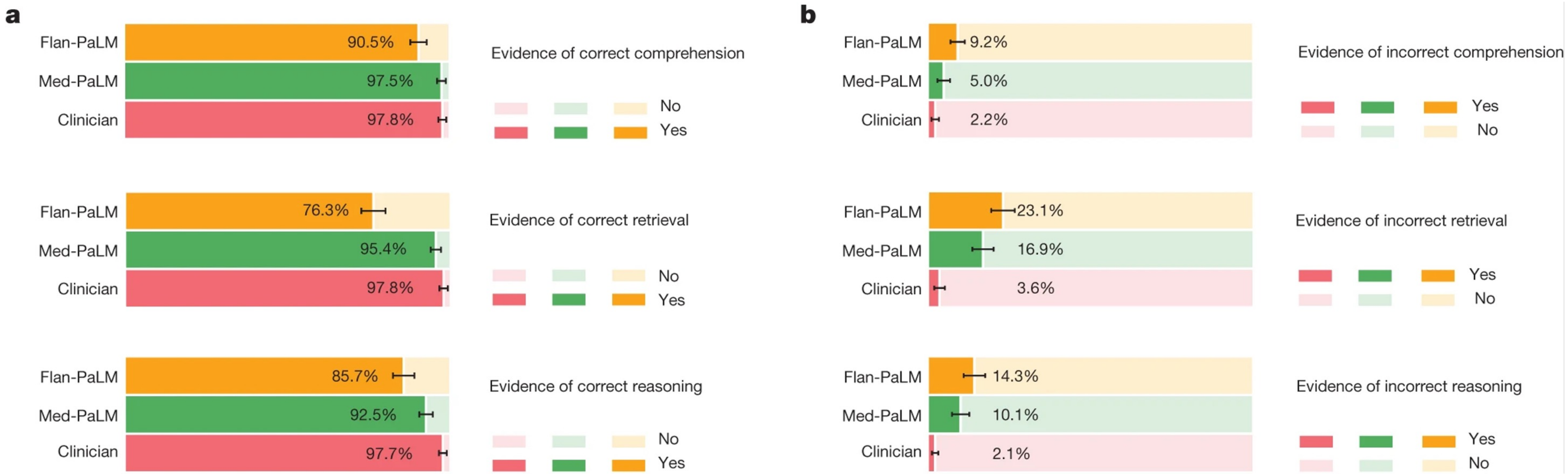
- PaLM and Flan-PaLM, two LLMs, on MultiMedQA.
- Use of few-shot, chain-of-thought, and self-consistency prompting strategies to improve LLM performance.
- Introduction of instruction prompt tuning, a technique to adapt LLMs to the medical domain.
- **Putting it all together: Med-PaLM**

Instruction prompt tuning

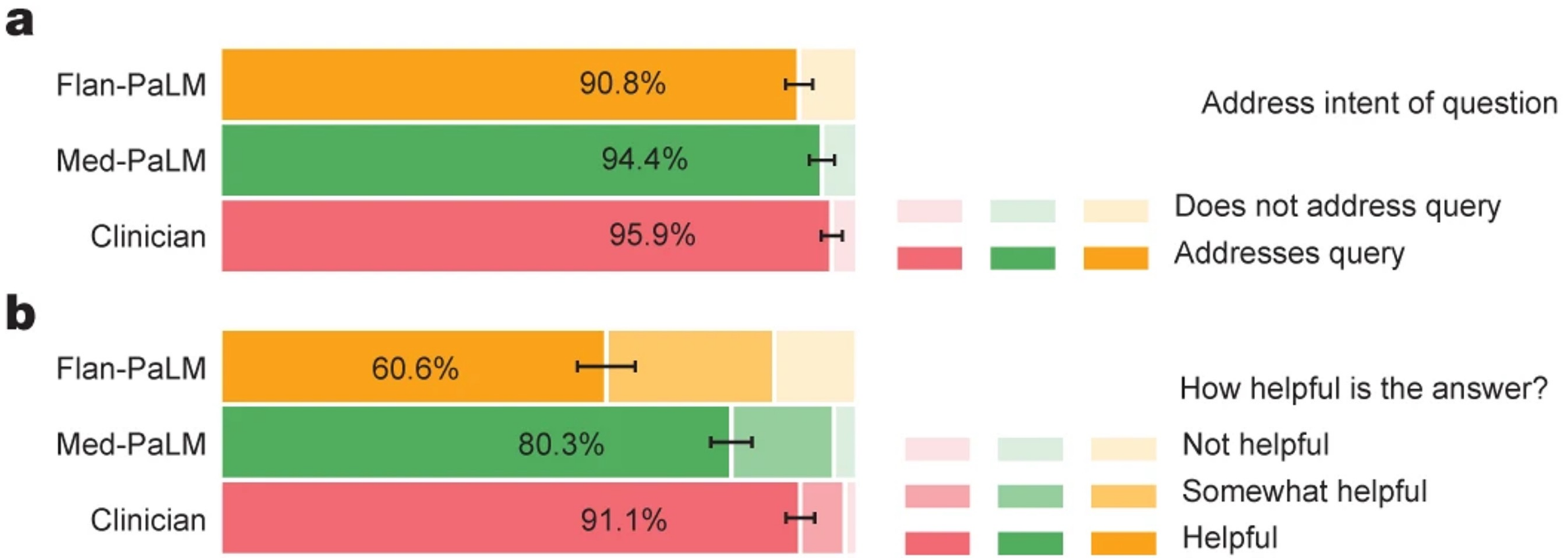



a
Flan-PaLM
Med-PaLM
Clinician
61.9%
92.6%
92.9%
Scientific consensus
No consensus
Opposed to consensus
Aligned with consensus
b
Flan-PaLM
Med-PaLM
Clinician
16.1%
18.7%
1.4%
Inappropriate and/or incorrect content
Yes, great clinical significance
Yes, little clinical significance
No
c
Flan-PaLM
Med-PaLM
Clinician
47.6%
15.3%
11.1%
Missing content
Yes, great clinical significance
Yes, little clinical significance
No
d
Flan-PaLM
Med-PaLM
Clinician
29.7%
5.9%
5.7%
Extent of possible harm
Death or severe harm
Moderate or mild harm
No harm
e
Flan-PaLM
Med-PaLM
Clinician
19.4%
2.3%
1.3%
Likelihood of possible harm
High
Medium
Low
f
Flan-PaLM
Med-PaLM
Clinician
7.9%
0.8%
1.4%
Possibility of bias
Yes
No

Evaluation of comprehension, retrieval and reasoning capabilities by clinicians



Lay user assessment of answers



Limitations

- Current benchmark covers diverse but limited medical exam, research, and consumer sources
- Current study limited to English-language datasets
- Improving Human Evaluation Methods
- Fairness and Equity
- Ethical Considerations

Takeaways

- **Transformative Potential:**
  Large Language Models (LLMs) could revolutionize medical AI, enhancing clinical support and patient care.
- **Challenges & Ethics:**
  Safe and ethical deployment requires addressing LLM limitations and aligning them closely with medical needs.
- **Call for Action:**
  Further research and cross-disciplinary collaboration are essential to responsibly apply these advancements to healthcare.

Thanks for your time.

Questions?