LREC-COLING 2024

MANCHESTER 1824
The University of Manchester

Effective Distillation of Table-based Reasoning Ability from LLMs

Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, Chenghua Lin

University of Manchester, United Kingdom
University of Surrey, United Kingdom
University of Pittsburgh, United State of America
University of Durham , United Kingdom

Outline

- Background and Motivation
- Methodology
- Experimental Results
- Conclusion

Background and Motivation

- Recent research reveals specific capabilities of LLMs can be transferred to smaller models through distillation.
- No prior work focuses on table reasoning skills in smaller models specifically tailored for scientific table-to-text generation tasks.
- We propose a novel table-based reasoning distillation approach, with the aim of distilling LLMs into smaller models.
- Experimental results show that a 220 million parameter model fine-tuned using distilled data surpasses specific LLMs on a scientific table-to-text generation dataset.

Novelty

- We explore the potential of using LLMs for the task of reasoning-aware scientific table-to-text generation.

- We propose a two-stage distillation framework containing data generation and finetuning stages.

Input Table:

Dataset	BERT (dev)	BERT (test)	BioBERT (test)	BioBERT (test)
MedNLI	79.56	77.49	82.15	79.04
MNLI (M)	83.52	-	81.23	-
SNLI (S)	90.39	-	89.10	-
M → MedNLI	80.14	**78.62**	82.72	80.80
S → MedNLI	80.28	78.19	83.29	81.29
M → S → MedNLI	80.43	78.12	83.29	80.30
S → M → MedNLI	**81.72**	77.98	**83.51**	**82.63**
MedNLI	79.13	77.07	**83.87**	79.95
S → M → MedNLI (expanded)	**82.15**	**79.95**	83.08	**81.85**

Distilled Reasoning:
Looking at the "**S → M → MedNLI**" row, we can see that the performance of **S → M → MedNLI** is higher on BioBERT compared to BERT.
Distilled Description:
"***BioBERT*** performs ***better than BERT*** on the ***S → M → MedNLI*** task."

Teach

T5-CoT:
BioBERT on S M MedNLI has a ***higher score*** than that of ***BERT***.

T5-traditional:
We can see that ***biobert outperforms*** bert by a large margin on all the datasets

Methodology

Our framework consists of synthesising data from LLMs and fine-tuning student models with the distilled data.

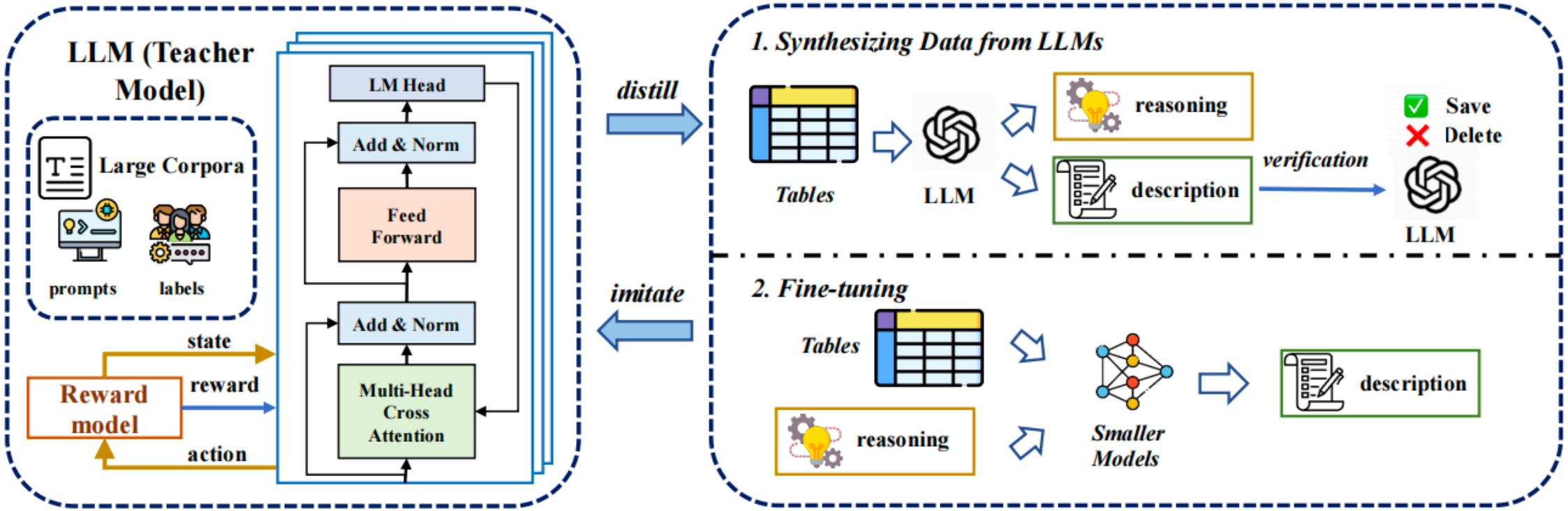
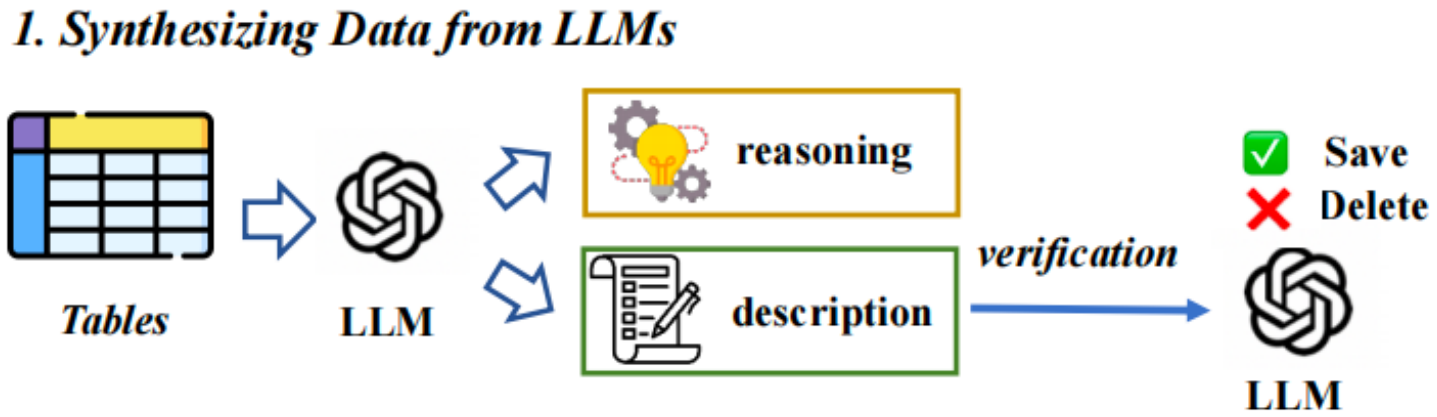


Table-based Reasoning Generation

- Data Synthesis

$$R_i, Y_i = \text{LLMs}(C, T_i)$$

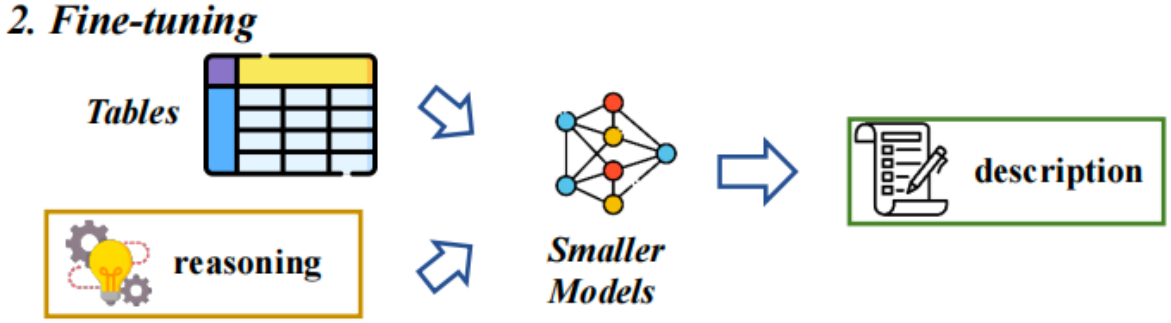


- Diverse reasoning

$$\{(R_1, Y_1), (R_2, Y_2)\} = \text{LLMs}(C, T_i)$$

- Data Filtering

Fine-tuning Small Models



Input Representation:

<R> <C> [BOLD] Dataset <C> [BOLD] BERT dev <C> [BOLD] BERT test <C> [BOLD] BioBERT dev <C> [BOLD] BioBERT test <R> <C> MedNLI <C> 79.56 <C> 77.49 <C> 82.15 <C> 79.04 <R> <C> MNLI (M) <C> 83.52 <C> - <C> 81.23 <C> - <R> <C> SNLI (S) <C> 90.39 <C> - <C> 89.10 <C> - <R> <C> M → MedNLI <C> 80.14 <C> [BOLD] 78.62 <C> 82.72 <C> 80.80 <R> <C> S → MedNLI <C> 80.28 <C> 78.19 <C> 83.29 <C> 81.29 <R> <C> M → S → MedNLI <C> 80.43 <C> 78.12 <C> 83.29 <C> 80.30 <R> <C> S → M → MedNLI <C> [BOLD] 81.72 <C> 77.98 <C> [BOLD] 83.51 <C> [BOLD] 82.63 <R> <C> MedNLI (expanded) <C> 79.13 <C> 77.07 <C> [BOLD] 83.87 <C> 79.95 <R> <C> S → M → MedNLI (expanded) <C> [BOLD] 82.15 <C> [BOLD] 79.95 <C> 83.08 <C> [BOLD] 81.85 <CAP> Table 4: All experiment results of transfer learning and abbreviation expansion (top-2 scores marked as bold). *<COT> Looking at the "S → M → MedNLI" row, we can see that the performance of S → M → MedNLI is higher on BioBERT compared to BERT.*

- Loss function

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N} \log P(Y \mid T, R)$$

Result

Dataset: SciGen

Metrics:
Surface-level:
- Meteor
- BERTScore
- BLUERT

Faithfulness-level:
TAPAS-Acc
TAPEX-Acc

Baselines:
- BART
- T5
- Flan-T5

Models	#Params	Faithfulness-level		Surface-level		
		TAPAS-Acc	TAPEX-Acc	Meteor	BERTScore	BLEURT
Teacher Model						
`text-davinci-002 (1-shot direct)`	175B	66.43	64.84	0.08	**0.82**	-0.97
`gpt-3.5-turbo (1-shot direct)`	175B	72.34	70.48	0.09	**0.85**	-0.91
`text-davinci-002 (1-shot CoT)`	175B	75.35	77.89	0.09	0.82	-0.94
`gpt-3.5-turbo (1-shot CoT)`	175B	82.53	84.99	0.09	0.83	-0.96
Medium Setting						
BART-large	0.40B	57.45	58.41	**0.23**	0.84	**-0.72**
T5-base	0.22B	53.27	52.45	0.15	0.82	-0.89
T5-large	0.77B	56.32	54.78	0.17	0.83	-0.77
Flan-T5-base	0.22B	54.78	56.25	0.16	0.84	-0.82
Flan-T5-large	0.77B	58.91	57.29	0.18	0.84	-0.80
Large Setting						
BART-large	0.40B	59.69	61.38	0.15	0.82	-0.89
T5-base	0.22B	55.32	53.76	0.15	0.82	-0.85
T5-large	0.77B	58.21	56.32	0.18	0.83	-0.79
Flan-T5-base	0.22B	56.41	55.37	0.16	0.82	-0.86
Flan-T5-large	0.77B	59.81	58.34	0.17	0.83	-0.83
CoT fine tuning						
T5-base-CoT	0.22B	78.16	82.30	0.08	0.83	-0.89
T5-large-CoT	0.77B	**80.62**	81.97	0.07	0.82	-0.89
Flan-T5-base-CoT	0.22B	78.72	**82.75**	0.08	0.82	-0.89
Flan-T5-large-CoT	0.77B	79.05	82.53	0.06	0.83	-0.89

Conclusion

In this paper, we propose

(1)a two-stage distillation framework that distils table-based CoT data from LLMs, which effectively transfer table reasoning abilities to smaller models in the scientific table-to text generation task.

(2) Our proposed method achieves comprehensive superiority in this specific task while requiring less data and smaller models.

Thanks!