TeacherLM：用于解析生成的通用推理语言模型

赵晨阳*
2020012363
计06
zhaochen20@mails.tsinghua.edu.cn

秦若愚
2019011115
计05
qry19@mails.tsinghua.edu.cn

程子睿
2019012490
计05
chengzr19@mails.tsinghua.edu.cn

1 简介

1.1 选题介绍

GPT-3 模型发布后，研究者已经证明了参数固定情况下的上下文学习（In-Context Learning）能够解锁大规模语言模型的推理能力，然而小规模语言模型难以在参数固定情况下利用少样本提示（Few-Shot Prompt）完成推理任务。放宽条件，在允许模型参数更新的情况下，带有推理的Chain-of-Thought Instruction Finetune（简记为CoT Finetune）也被证明能够提升大规模语言模型的推理能力，然而需要大量带有推理过程的数据才能提升小规模模型的推理效果。即便如此，CoT Finetune有一定可能提升小模型的推理能力，但其表现无论在准确率上还是在稳定性上仍旧不佳；甚至CoT Finetune的效果往往不如单纯使用Instruction Finetune而不加入推理过程。

因此，我们希望提出全新的专用于生成推理过程的语言模型TeacherLM，为现有推理任务提供模型生成的高质量的解析，取代人工标注的解析。最终，利用我们的解析进行CoT Finetune，以此提升小规模语言模型的推理能力。

1.2 工作亮点

我们提出并开源了第一组专用于解析生成的通用推理语言模型TeacherLM。利用TeacherLM能够为小规模数据集生成解析（Analysis）。利用这部分Analysis替换中人工标注的CoT，对小规模语言模型进行CoT Finetune能够达到媲美甚至超过人工标注的CoT的效果，而利用TeacherLM生成解析的成本远低于人工标注。

基于TeacherLM的CoT Finetune证明了小规模语言模型在少量带有模型生成解析的数据上利用Analysis进行Finetune也能解锁复杂的推理能力，而先前的工作认为只有大模型或者大规模数据才能解锁这一能力。

*组长

实际上，我们基于Teacher-Student Learning与知识蒸馏，提出利用TeacherLM生成解析并进行Anlysis Finetune这一训练友好且成本可控的解锁语言模型推理能力的新范式，并希望以此推动Reasoning这一领域的研究。

2 相关工作

2.1 Finetuned Language Models are Zero-Shot Learners

Jason Wei [1] 等人率先提出了Instruction Fintune方法（后文简称为FLAN），也即通过丰富的Instruction Template将大量的数据集串接为训练数据，对大规模语言模型进行Finetune。这一方法能有效提升模型在未知任务上的泛化能力。然而消融实验却发觉，FLAN方法只对大规模语言模型（130B及以上）搭配大规模语言数据集（工作中采用了62个大规模数据集，而且搭配了10个模板）有明显效果。而我们的工作主要致力于在小规模语言模型（10B以下）如何利用单一数据量不超过1万的小规模数据集提升推理泛化能力，因此FLAN的方法并不完全适用于我们的工作目标。

2.2 Scaling Instruction-Finetuned Language Models

基于FLAN方法，Hyung Won Chung [2] 等人提出了将FLAN与CoT Instruction Finetune融合的方法。在FLAN 62个数据集的基础上，研究者们进一步扩大数据集规模到达473个（这一部分数据集采用Instruction Finetune），同时加入了9个带有人工标注的CoT的数据集（将这部分数据集称为Reasoning，并对其采用了CoT Instruction Finetune）。经过这一规模数据量finetune后的语言模型在各类未知任务上都有了较好的提升；包括但不限于需要推理的Math Word Problem和不需要推理的Common Sense Answering。消融实验中发现，如果没有利用9个Reasoning数据集进行CoT Instruction Finetune，语言模型在推理任务上的表现会显著降低；而只有9个Reasoning数据集进行CoT Instruction Finetune，没有其他数据集进行Instruction Finetune，语言模型在非推理任务上的表现会大幅降低，甚至连推理任务本身都有可能出现CoT Hurt的情况。而我们的工作主要针对小规模语言模型在推理任务上的泛化能力，所有的下游小规模数据集均为推理任务，并且使用了CoT Instruction Finetune。由于没有Instruction Finetune来缓解潜在的CoT Hurt风险，实际上我们的训练任务难度更大，对于CoT的质量要求更高。值得强调的是，如图 1 所示，TeacherLM不仅能够输出Chain of Thought，而且能够输出Fundamental和Common Error，因此正如 3 一节所述，在我们的工作中，我们不仅采用了CoT Instruction Finetune，也采用了Fundamental Instruction Finetune和Erorr Instruction Finetune，并统称为Analysis Finetune。最后，我们将三者融合，进行了ACF实验。

2.3 Self-Consistency Improves Chain of Thought Reasoning in Language Models

为了缓解CoT Hurt的情况，Xuezhi Wang [3] 等人一方面提出将CoT Instruction Finetune混合Instruction Finetune，另一方面提出了Self-Consistence方法：在模型推理时进行投票，也能提升CoT Instruction Finetune的效果。具体而言，研究者们在CoT Instruction Finetune时，将训练数据组织为Question-Prompt-CoT-Answer Chain的形式输入模型；我们将这一形式称为Mid-Training，与之对应的Question-Answer-Prompt-CoT Chain我们称其为Tail-Training。测试时，模型将生成多条CoT推理路径，每条路径的结尾对应一个答案。最后，将出现次数最多的答案作为最后模型的推理结果，也即Multi-Vote。这一整

套包含Mid-Training和Muti-Vote的流程称为Self-Consistency，并且能够相较不采用Multi-Vote只使用Mid-Training有显著的提升。然而，生成多条CoT推理路径等效于一次推理过程扩展为了多次推理，这显著提升了模型推理过程中消耗的计算资源，生成最终答案的效率明显降低。我们的工作专注于在训练成本相对较低的情况下尽可能提升模型的推理效果，故而选择了Tail-Training的方法，并且没有采用计算成本高昂的Muti-Vote，而这无疑也对我们的Analysis质量提出了更高的要求。

3 方法

3.1 TeacherLM

在超大规模闭源中英文题库数据集上我们训练了TeacherLM推理语言模型组，含有7B1与176B两个大小。出于实验可复现性的考虑，未经特殊说明，本研究采用的模型都是7B1大小的TeacherLM。TeacherLM 可对输入输出对输出思维链解析（Chain of Thought, CoT）、易错点（Common Error，error）和相关知识（Fundamental），如图 1 所示。

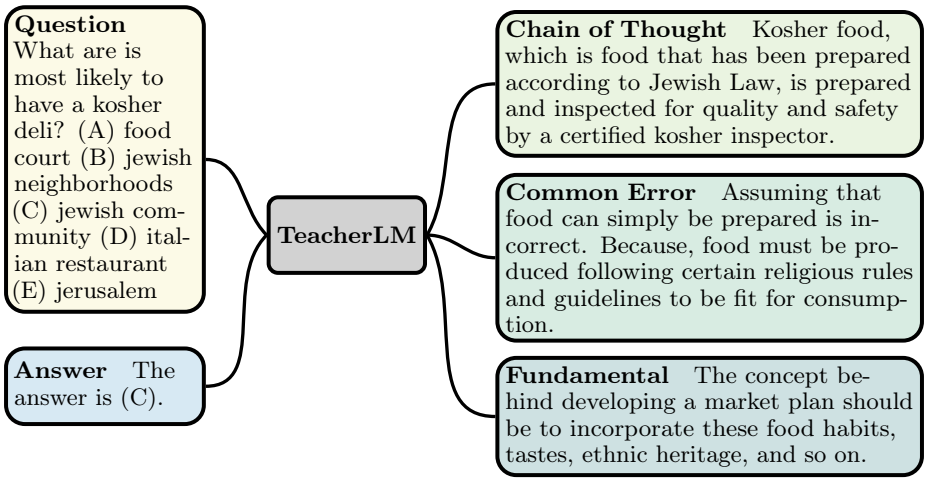

Figure 1: TeacherLM

3.2 Instruction Finetuning

如相关工作 2.1 中所述，Jason Wei [1] 等人提出了Instruction Finetune的方法，亦即通过固定的prompt模板将数据集的字段串接为完整且人类可理解的句子，输入到语言模型中进行finetune。在我们的工作中，我们对5个带有人工标注CoT的数据集均进行了Instruction Finetune，并以此作为baseline。

3.3 Analysis Instruction Finetune

如相关工作 2.2 中所述，Hyung Won Chung [2] 等人提出了Chain-of-Thought Instruction Finetune（简写为CoT Finetune）方法，亦即将传统Finetune时输入给模型的Input-Output Pair更换为Input-Output-Prompt-CoT Chain或者Input-Prompt-CoT-Output Chain。这一方法有一定概率提升预训练模型在数学问题、常识推理和符号处理等复杂推理任务上的表现。然而也有大量数据集采用CoT Instruction Finetune后效果反而不如未加入CoT的Instruction Finetune本身，研究者将这一现象称为CoT Hurt。

Plain Text	What are is most likely to have a kosher deli? (A) food court (B) jewish neighborhoods (C) jewish community (D) italian restaurant (E) jerusalem. The answer is (C).
Instruction with Manual Analysis	What are is most likely to have a kosher deli? (A) food court (B) jewish neighborhoods (C) jewish community (D) italian restaurant (E) jerusalem. The answer is (C). Let's think step by step. Jewish Community means a community of Jewish people originating from the Israelites and Hebrews of historical Israel and Judah.
Instruction with Chain-of-Thought Analysis	What are is most likely to have a kosher deli? (A) food court (B) jewish neighborhoods (C) jewish community (D) italian restaurant (E) jerusalem. The answer is (C). Let's think step by step. Kosher food, which is food that has been prepared according to Jewish Law, is prepared and inspected for quality and safety by a certified kosher inspector.
Instruction with Common Error	What are is most likely to have a kosher deli? (A) food court (B) jewish neighborhoods (C) jewish community (D) italian restaurant (E) jerusalem. The answer is (C). The common mistakes are: Assuming that food can simply be prepared is incorrect. Because, food must be produced following certain religious rules and guidelines to be fit for consumption.
Instruction with Fundamental	What are is most likely to have a kosher deli? (A) food court (B) jewish neighborhoods (C) jewish community (D) italian restaurant (E) jerusalem. The answer is (C). The fundamental of this question is: The concept behind developing a market plan should be to incorporate these food habits, tastes, ethnic heritage, and so on.

Figure 2: Analysis Instruction Finetune

考虑到TeacherLM输出的模型解析包括CoT、Error、Fundamental三种，因此我们分别将CoT Finetune时输入给模型的数据集自带的人工CoT更换为TeacherLM生成的CoT、Error和Fundamental，对应称为CoT Finetune（后文cot-tail）、Error Finetune（后文error-tail）和Fundamental Finetune（后文fundamental-tail），并且将每个数据集中三者中效果最好的一个称为Analysis Instruction Finetune（简称为Analysis Finetune），具体的prompt与加入的Analysis可以参考图 2 部分的示例。最后，实际上TeacherLM的训练过程输入的数据格式为Question-Answer-Prompt1-CoT-Prompt2-Error-Prompt3-Fundamental，因此我们对下游模型也采用了同样的输入方式进行了对比实验，并且将这一实验称为acf。

3.4 Evaluate Method

参考 4.1.1 中的数据集划分，我们利用训练集构造了训练数据，而余下的训练集构造了Evaluate方法。我们的测试方法基于lm-evaluation-harness仓库[2]。具体而言，我们为不同的数据集分别编写了新的Task，以Zero-Shot 单项选择题的方式进行测试。

测试能够得到每个checkpoint解决单项选择题的正确率（acc）与标准差（std），由于std的数值均分布在0.01附近，没有可比较性，因此我们在报告中仅以acc作为模型表现的指标。最后，小规模模型都是在对应finetune的数据集的测试集上完成测试，我们并未进行交叉测试。

[2]https://github.com/EleutherAI/lm-evaluation-harness

4 实验

4.1 实验设置

4.1.1 数据集设定

我们将在多种类型的基准数据集上探索TeacherLM的解析生成能力。(1) 基于真实场景理解的常识推理数据集**CREAK** [4]；(2) 带有正负属性标注的常识问答数据集**ECQA** [5]；(3) 多语句多步骤推理数据集**QASC** [6]；(4) 常识判断数据集**SenseMaking** [7]；(5) 隐式多步骤推理数据集**StrategyQA** 数据集 [8]。

我们将原始数据集中的数据进行了重新的组合，并且生成了适用于TeacherLM输入的问题、回答和分析三元组，数据集的样例如图 3 所示。考虑到原始数据集中训练集、测试集和验证集的划分方式、数据类型并不一致，我们统一抽取了原始数据集中带有完整的问题、回答和分析字段的数据划分了训练集（85%）和测试集（15%）。

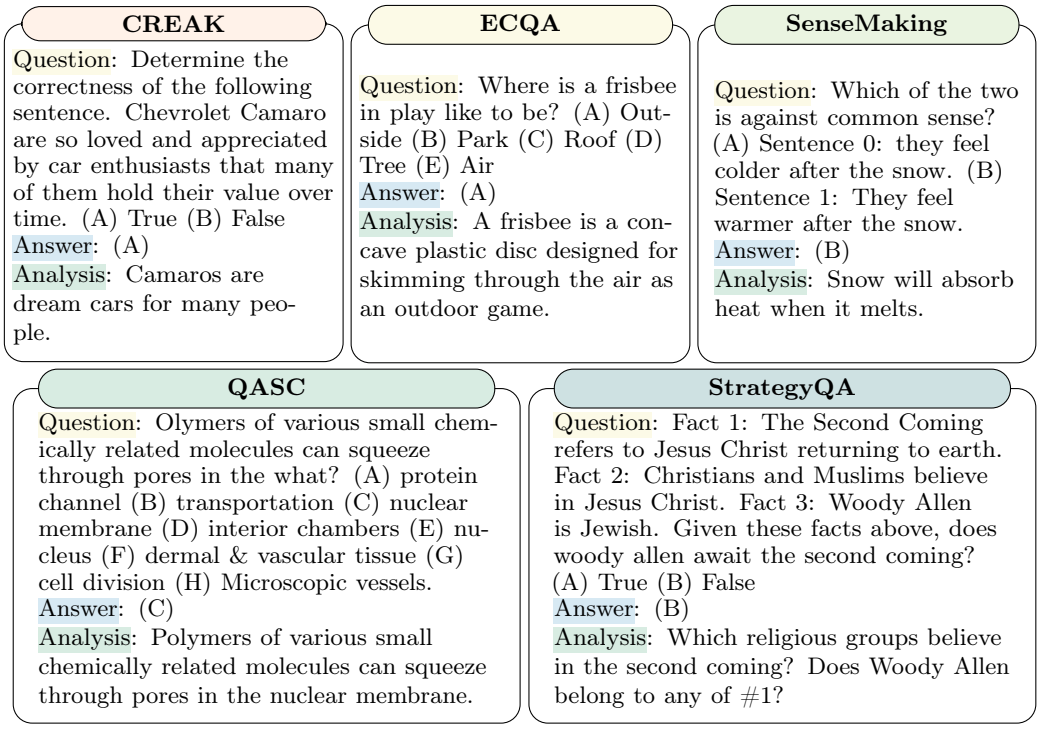

Figure 3: 数据集样例

4.1.2 模型设定

为了训练过程成本相对较低并且具有可扩展性，我们选择了更为训练友好的Decoder-Only模型作为被finetune的小规模模型。具体而言，我们在实验之初选择了facebook/opt-2.7b[3]、facebook/opt-6.7b[4]、bigscience/bloom-7b1[5]与bigscience/bloomz-7b1[6]四个模型。

[3]https://huggingface.co/facebook/opt-2.7b
[4]https://huggingface.co/facebook/opt-6.7b
[5]https://huggingface.co/bigscience/bloom-7b1
[6]https://huggingface.co/bigscience/bloomz-7b1

在对四个模型进行预实验后，我们发觉bigscience/bloomz-7b1的效果往往最佳，考虑到后续实验计算资源有限，最终在报告中的所有finetune小规模语言模型均选为bigscience/bloomz-7b1。

4.1.3 超参数设定

我们最为理想的结论是对于任意数据集与任意超参数，TeacherLM生成Analysis并进行Analysis Finetune的效果优于人工标注的解析（也即图标中的manual-tail）与单纯的Instruction Finetune（也即图表中的text）。然而，实验结论并没有如此理想。为此，我们进行了两组搜参实验。

思维链类型 我们首先利用五选项常识问答类数据集ECQA [5] 为例，对于batch size与learning rate两参数进行了搜索。考虑到原始的ECQA数据集具有如下3个字段：（1）正例解析：也即cot-pos，五个选项中唯一的正确答案的解析，也即正确答案为什么正确；（2）负例解析：也即cot-neg，五个选项中四个错误答案的解析，也即错误答案为什么错误；（3）全局解析：也即cot-ff，原数据集作者将所有选项的解析整合为一个语句，完整分析了整个问题各个选项的对错。

我们利用TeacherLM为ECQA生成了模型生成的CoT（标记为cot-our），而并未生成Error与Fundamental，如此以来得到4种解析：cot-pos、cot-neg、cot-ff与cot-our。接着，对于4种解析，我们设定了3组9个对比实验。第一组为text，也即Instruction Finetune而未计入任何解析；第二组为Mid，也即采用Question-Prompt-CoT-Answer的形式，将4种解析放置于输入数据的中部进行finetune；最后一组为Tail，也即采用Question-Answer-Prompt-CoT的形式，将4种解析放置于输入数据的尾部进行finetune。最后，在对于4种解析类型的搜索实验中，我们的batch size设定为{32, 64, 256}；而learning rate设定为{2e-5, 2e-6}，训练epoch为100，每个对比实验选择了测试效果（acc）最高的一个ckpt作为最终结果。最终得到的结果如表 1 所示，为了方便后文叙述，我们将每组实验的batch size, learing rate标记为<batch size, learing rate>的形式。

Batch Size		32		64		256	
Learning Rate		2×10^{-5}	2×10^{-6}	2×10^{-5}	2×10^{-6}	2×10^{-5}	2×10^{-6}
	text	61.76%	66.77%	56.02%	64.45%	43.53%	63.54%
Mid	cot-pos-mid	37.74%	44.85%	32.04%	46.86%	24.29%	37.78%
	cot-neg-mid	34.50%	38.51%	36.14%	41.48%	26.57%	39.33%
	cot-ff-mid	33.27%	42.07%	27.26%	46.67%	29.40%	41.93%
	cot-our-mid	56.56%	59.25%	44.71%	58.25%	44.62%	54.88%
Tail	cot-pos-tail	57.47%	52.37%	47.58%	52.60%	33.27%	54.92%
	cot-neg-tail	58.71%	63.31%	54.38%	58.98%	42.84%	60.71%
	cot-ff-tail	55.47%	55.42%	51.46%	50.18%	25.71%	54.10%
	cot-our-tail	63.99%	55.61%	54.56%	58.57%	35.14%	63.04%

Table 1: ECQA 9 种设定搜索参数

正如相关工作 2.2 一节中所述，此时出现了明显的CoT Hurt现象。人工标注的三种解析（cot-pos，cot-neg和cot-ff）的效果无论是mid还是tail，均是最差的。而不加入解析的text与TeacherLM的解析cot-our的表现优劣并不明显。譬如在<32, 2e-5>时，text的效果不如cot-our-tail；而<64, 2e-6>时，text的效果优于cot-our-tail。

因此，我们初步得出结论，ECQA的人工解析存在严重的CoT Hurt，而TeacherLM生成的cot-our能够缓解人工解析存在的严重的CoT Hurt；然而，在部分超参数下，cot-our能

够扭转CoT Hurt的情况，但是另一部分超参数则仍旧存在一定程度的CoT Hurt。综上，我们起初设定的最为理想的实验结论未能满足，并且考虑到text表现最佳的设定为<32, 2e-6>与<64, 2e-6>，我们进一步只使用text与cot-our-tail搜索batch size与learing rate。

Batch Size与Learing Rate 我们进一步搜索参数空间，设定batch size={32, 64, 256}，而learing rate={2e-5, 1e-5, 8e-6, 6e-6, 4e-6, 2e-6, 1e-6, 8e-7, 6e-7, 4e-7, 2e-7, 1e-7}，实验结果如图 4 所示。

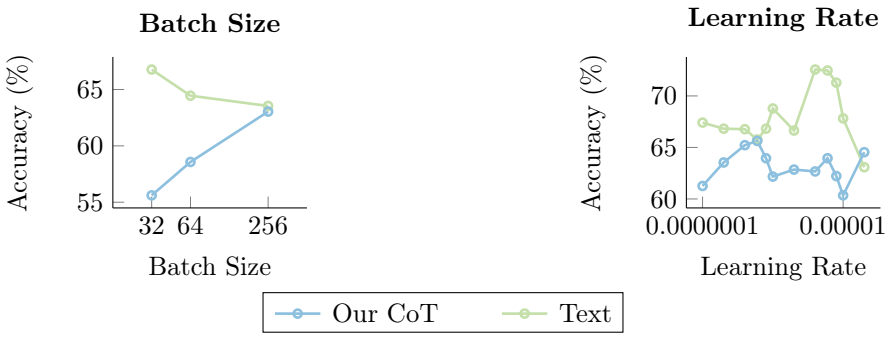

Figure 4: ECQA Parameter

图 4 左侧展示了固定学习率为6e-6时acc关于batch size的曲线，可以发现在这一学习率下，所有的batch size均存在CoT Hurt；而图右侧展示了固定batch size为64时，acc关于learning rate的曲线，同样在这一batch size下，大多learing rate均存在显著的CoT Hurt。最终，我们认为ECQA数据集CoT Hurt的问题不可避免，而TeacherLM CoT Hurt的情况相较人工解析CoT Hurt的情况更为轻微，但是仍旧无法逆转。

基于对于ECQA的搜参实验，我们最终选定后续训练的超参数为：batch size=64，learing rate=6e-6，training epoch=50。此外，我们没有使用dropout、few shot等其他训练方法，因此最终对于所有5个数据集均采用此超参数。

4.1.4 主实验方法与Baseline设定

如 3.3 节所述，TeacherLM 生成的解析分为CoT、Error 和Fundamental 三种，因此我们设置了CoT Instruction Finetune、Error Instruction Finetune 和Fundamental Instruction Finetune三种实验。此外，我们还将上述三种解析组合为acf作为第四种解析，采用与其它三种解析相同的方式做Instruction Finetune实验。我们将这四种解析（中效果最好者）称为analysis。

此外，我们设置了利用人工标注的CoT（以下称为Banual）的CoT Instruction Finetune实验，以及不加入CoT （以下称为text）仅进行Instruction Finetune的实验；这两部分实验作为baseline。

4.2 主实验结果

我们对上述5个数据集分别使用6种Finetune方式 （text，manual-tail，cot-tail，fundamental-tail，error-tail与acf）进行了实验，分别测试小模型在各个数据集下的准确率，结果如表 2 和图 5 所示，对于每一数据集的分析如下：

CREAK 在**CREAK**数据集中，analysis好于manual好于text，并且以TeacherLM生成的Common Error为解析的Error Instruction Finetune效果最好，考虑到CREAK为常识推理数据集，我们认为这是因为回答CREAK题目时需要更多题目之外的推理信息，而这部分信息恰好被analysis所补充。

ECQA 如前文所述，**ECQA**存在显著的CoT Hurt现象。具体而言，text好于analysis好于manual，且manual方法CoT Hurt极为严重，准确率极低，我们认为ECQA需要的常识知识过多，小模型仅仅使用相比预训练时极其少的数据量进行finetune，无法很好地掌握这些常识。而且，Analysis Finetune的方法可能与小模型预训练时的语言范式有很大不同，模型在如此有限的finetune数据下难以迁移到新的语言范式，反而连最基础的问答式语言范式都会受到影响。

QASC 在**QASC**数据集中，analysis 好于text，而manual 出现了CoT Hurt的现象，这是因为QASC的题目范围为科学常识，TeacherLM 由于训练数据质量更好、包含范围更广，可以更好地适应该数据集。

SenseMaking 在**SenseMaking**数据集中，manual好于analysis好于text，这表明推理解析信息对小模型解答题目是有帮助的，但TeacherLM生成的解析与人工解析相比不太适合该类型题目。这也体现了TeacherLM的解析效果能够在多数时候媲美甚至超过人工，然而部分数据集人工解析的质量非常之高，TeacherLM仍然不能完全取代人工解析。

StrategyQA 在**StrategyQA**数据集中，analysis 好于text好于manual，以TeacherLM生成的Fundamental为解析的Fundamental Instruction Finetune效果最好，并且manual出现了CoT Hurt的现象。实际上，StragedyQA数据集几乎是常识推理类数据集中难度最高且最负盛名的数据集，其中的问题要求更深的推理步骤和更多的推理信息，TeacherLM的推理能力足够惊艳，能够生成质量超过人工标注的解析，也能够逆转CoT Hurt的情况。

通过表 2 与图 5 的结果与上述分析，我们得出以下结论：

（1）在大部分情况下，Analysis Finetune的效果要优于Manual Cot Instruction Finetune和Text Finetune。我们认为这是由于训练TeacherLM的题目数据质量更高、数量更多、包含范围更广，因此TeacherLM生成的解析在通常意义下具有更高的专业性和可靠性，能够提供成本远低于人工标注的全新解析生成方式。此外，对于较为复杂的问题，人工标注解析往往需要对应领域的专家才能完成，而TeacherLM的泛化性远较强，相较单一人类专家可以适应更多类型的数据集。

（2）在SenseMaking数据集中出现了manual的效果好于analysis的现象。这自然体现了TeacherLM也存在一定的局限性。TeacherLM用于为各类问答数据集生成解析，在某些题目类型上并不能发挥优势；而人工标注的方式成本更高，在不考虑成本的前提下，更多专家的参与能够适应不同的题目类型，因此人工解析在SenseMaking数据集上表现更优。

（3）最后，我们还发现在一些数据集上出现了普遍的CoT Hurt的现象，即text的效果好于analysis与manual。考虑到相关工作 2.2 和 2.3 中的分析，CoT Hurt可能是由于我们的实验设置是对小规模语言模型在单一下游小规模数据集上进行Finetune，因此极有可能出现知识迁移不充分的情况。此外，我们的实验中并没有使用CoT Instruction Finetune和Instruction Finetune混合以缓解CoT Hurt，也没有使用Self-Consistence方法，实验难度较高，所以会出现部分数据集CoT Hurt的现象也符合情理。

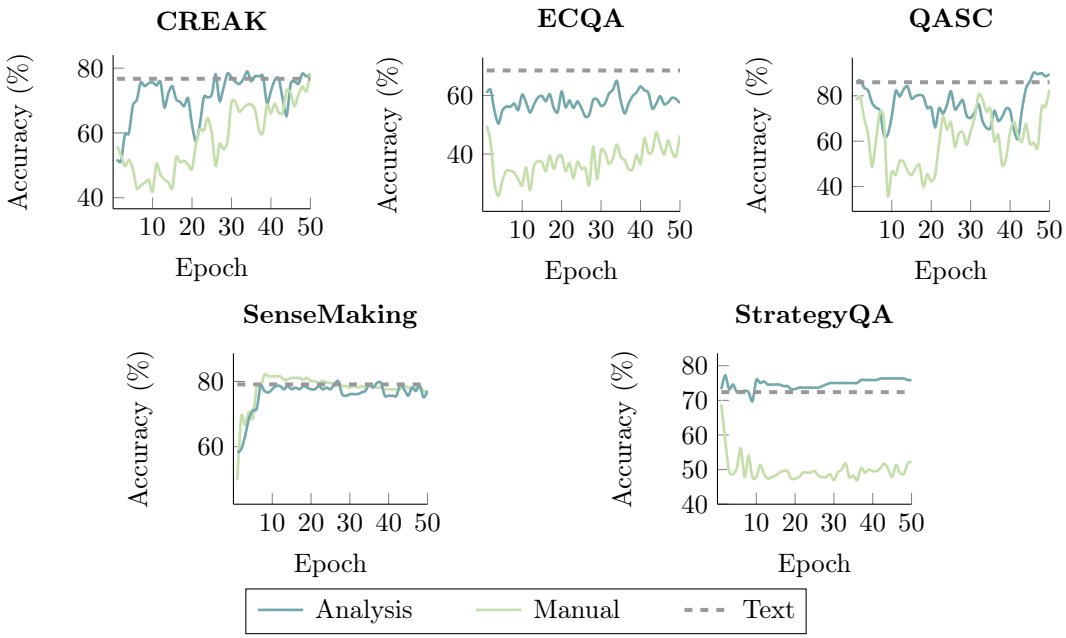

Figure 5: 全部数据集训练全过程波动示意图

	CREAK	ECQA	QASC	SenseMaking	StrategyQA
Text	76.73%	**68.55%**	85.96%	79.10%	72.37%
Manual	78.56%	49.64%	82.72%	**82.09%**	68.86%
Chain-of-Thought	75.86%	60.48%	89.96%	79.10%	74.12%
Common Error	**78.99%**	64.95%	**90.17%**	80.10%	74.12%
Fundamental	75.78%	58.30%	84.13%	71.14%	**77.19%**
All	65.28%	51.96%	83.05%	78.11%	68.86%

Table 2: 全部数据集最佳效果

4.3 对标ChatGPT与人工评测

4.3.1 对标ChatGPT

为了进一步证明TeacherLM强大的推理能力，我们将其与ChatGPT进行对标实验。具体而言，对于某一数据集下的同一问题与答案，我们采用相同的输入方式（Question-Answer-Prompt）分别输入给TeacherLM与ChatGPT生成解析，并将得到的解析分别标记为cot与chat，对应构造cot-tail与chat-tail两种训练方式。

我们以CREAK和ECQA为例，直接对比在前文所述超参数下的cot-tail与chat-tail的效果，其效果如图 6 所示：

分析上图，我们能够观察到，cot-tail的效果在大多时候优于chat-tail，并且二者的表现存在显著差异。此外，chat-tail的准确率波动远远大于cot-tail。因此，我们认为TeacherLM的推理能力优于ChatGPT。

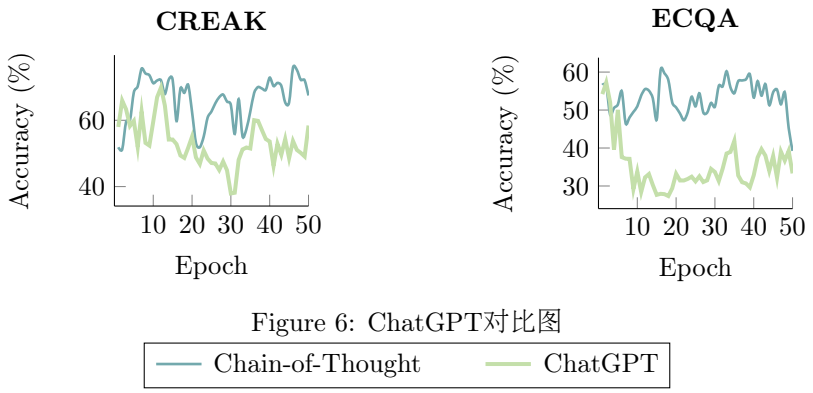

Figure 6: ChatGPT对比图

4.3.2 人工评测

为了进一步分析TeacherLM的解析质量，我们设置了如下6 个指标评测TeacherLM生成的解析和人工解析的质量：(1) 流畅性；(2)一致性；(3)正式性；(4)充分性；(5)非冗余性；(6)整体质量。每一部分总分为5分，并计算平均分。

最终在ECQA 数据集上，TeacherLM得分3.79，而人工解析得分4.01，ChatGPT得分3.96。这似乎与前文实验得到的准确率结果不符合，然而考虑到人工解析实际上较为通顺，而且更加符合人类的语言习惯，但是信息丰富度和有效性不足；ChatGPT的解析则极度顺畅，然而其中会出现一些错误的信息。TeacherLM的解析在流畅度和信息丰富度、准确度之间达到了某种权衡，在ChatGPT和人工解析之间平衡了解析质量，经过finetune后的效果最佳也可以理解。

5 总结

我们提出并开源了第一组专用于解析生成的通用推理语言模型TeacherLM，证明了小规模语言模型在少量带有模型生成解析的数据上进行Analysis Finetune也能解锁复杂的推理能力。而Analysis Finetune实际上是基于Teacher-Student Learning与知识蒸馏提出的训练友好的解锁语言模型推理能力的新范式。

总体而言，7B1大小的TeacherLM就可以生成绝大多时候媲美甚至超出人工的解析，即便存在CoT Hurt的情况，其解析质量已然足够惊艳。尽管TeacherLM的解析经过人工评测看来仍然不足以完全取代人工标注的解析，但是其较小的体积和生成成本仍然为问答类数据集的数据增强提供了成本可控且方便易用的新方法。

接下来，我们预计将TeacherLM的相关工作投稿至重要学术会议。此后将进一步探索TeacherLM对数学问题的推理能力、TeacherLM的解析可解释性以及伦理问题。

References

[1] Wei, J., M. Bosma, V. Y. Zhao, et al. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021.

[2] Chung, H. W., L. Hou, S. Longpre, et al. Scaling instruction-finetuned language models, 2022.

[3] Wang, X., J. Wei, D. Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models, 2022.

[4] Onoe, Y., M. J. Q. Zhang, E. Choi, et al. Creak: A dataset for commonsense reasoning over entity knowledge, 2021.

[5] Aggarwal, S., D. Mandowara, V. Agrawal, et al. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065. Association for Computational Linguistics, Online, 2021.

[6] Khot, T., P. Clark, M. Guerquin, et al. QASC: A dataset for question answering via sentence composition. *CoRR*, abs/1910.11473, 2019.

[7] Wang, C., S. Liang, Y. Zhang, et al. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026. Association for Computational Linguistics, Florence, Italy, 2019.

[8] Geva, M., D. Khashabi, E. Segal, et al. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*, 2021.