Enhancing Free-Form Table Question Answering Models by Distilling Relevant-Cell-Based Rationales

Zhiyu Yang[1] **Shuo Wang**[2] **Yukun Yan**[2] **Pengyuan Liu**[1,3]* **Dong Yu**[1]
[1]School of Information Science, Beijing Language and Culture University
[2]Department of Computer Science and Technology, Tsinghua University
[3]National Language Resources Monitoring and Research Center for Print Media
202121198419@stu.blcu.edu.cn
liupengyuan@pku.edu.cn

Abstract

Free-form table question answering is a challenging task since tables contain structured contents compared to plain texts, which requires high-level reasoning abilities to effectively identify cells that are relevant to the question and produce a correct and faithful answer based on their relations. Large language models (LLMs) have exhibited remarkable reasoning capabilities in numerous NLP applications. However, in some specific tasks, specially-trained small models can still outperform LLMs. Furthermore, small models require extremely less computation costs compared to LLMs. To leverage the strengths of both types of models, we propose a **R**elevant-**C**ell-based **K**nowledge **D**istillation with inference-time **T**eacher **G**uidance **(RCKD-TG)** method. This approach aims to combine small free-form table question answering models' abilities to learn from human annotations and large language models' abilities to effectively reason from table contents, via applying Relevant-Cell-based rationales distilled from LLMs to small models' training and inference stages. Our experiments demonstrate the superiority of our method over vanilla small models in correctness, faithfulness, adequacy and fluency, also over general LLMs in adhering to the style of human annotations. We achieve state-of-the-art performance on FeTaQA, a representative free-form table question answering benchmark. Our result of a 41.3 BLEU score demonstrates the feasibility of effectively using small models' task-specific abilities and LLMs' reasoning capabilities at the same time. Additionally, our method exhibits high computation efficiency and data efficiency. Compared to strong baselines, we achieve better performance with significantly less training data.

1 Introduction

Tables are a prevalent form of structured data commonly found in databases and on the internet. Traditionally, tables were grounded using semantic parsing or by converting natural language queries into SQL language to obtain desired outputs. Representative benchmarks, like WTQ (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), WikiSQL (Zhong et al., 2017), Spider (Yu et al., 2018), TabFact (Chen et al., 2019), have been established for evaluating table grounding methods. Recent efforts have leveraged pre-training methods for table grounding tasks, with models like TAPAS (Herzig et al., 2020), TaBERT (Yin et al., 2020), TAPEX (Liu et al., 2021), and TableFormer (Yang et al., 2022) achieving promising results on existing benchmarks. However, these benchmarks primarily focus on providing short-form answers to queries, which are not challenging enough for models to perform complex reasoning.

With the emergence of large generative models, users increasingly prefer interacting with NLP systems through natural language queries and receiving responses in free-form answers (Chung et al., 2022; Ouyang et al., 2022; Thoppilan et al., 2022). As a result, traditional short-form table parsing benchmarks and methods fall short of meeting modern users' demands. To address this limitation, (Nan et al., 2021) introduced FeTaQA, a free-form table question answering benchmark, aiming to bridge the gap. FeTaQA challenges models to retrieve multiple relevant entities and reason over them to produce correct and

*Corresponding author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

faithful answers given more abstract and ambiguous queries. This benchmark is thus more challenging than previous short-form counterparts.

Researchers have continued to explore table pre-training methods to tackle free-form table question answering. For example, GenTaP (Shi et al., 2022) leverages large-scale synthesized training data with more than 500K training samples. However, such methods may not fully enhance a model's reasoning capabilities. Recently, Large Language Models (LLMs) have exhibited remarkable reasoning capabilities across various NLP tasks. (Brown et al., 2020; Wei et al., 2022) have demonstrated LLMs' ability to quickly generalize to unseen tasks and perform reasonably well given only a few prompts. However, developing and deploying LLMs for each specific scenario is computationally prohibitive. To address this challenge, the latest works (Li et al., 2023; Hsieh et al., 2023) have adopted Knowledge Distillation, using LLMs as Teacher models and small models as Student models. These methods train small models on LLMs' output to distill their knowledge and reasoning capabilities, and in the meantime, outperform their Teacher models. While successful in classification and multiple-choice QA tasks, their effectiveness in generation tasks and structured data remains rather unexplored.

It is harder to distill knowledge from larger teacher models in text generation settings because of inadequate supervision signal. Automatic evaluation metrics for language generation tasks are not robust enough to assess all important aspects pertinent to the quality of generated text. They succeed in evaluating the surface similarities between the generated text and the reference text, but fails in evaluating the accuracy, faithfulness and fluency of generated text. Therefore, training a small model using these metrics as supervision would constrain to the small model's output to be lexically and syntactically more similar to human annotations, but unable to guide the small model to generate semantically correct and grammatically fluent answers. Meanwhile, LLMs could provide mostly correct answers but fail to conform to human annotated preferences, and it is computationally prohibitive to fine-tune LLMs on human annotated data to mitigate this problem.

In light of this, we propose a simple method for free-form table question answering, **R**elevant-**C**ell-based **K**nowledge **D**istillation with inference-time **T**eacher **G**uidance **(RCKD-TG)**, to tackle the challenges of reasoning on table contents and generating free-from answers that are more preferred towards human annotations. We devise a specific training method, so that at inference time, LLM's abilities to reason and small model's abilities to produce answers more similar to human annotations can be combined effectively.

Firstly, on the training data, we prompt the LLM to generate information in cells that are relevant to answering the question. Then we use these rationales as references to teach the small model to produce crucial information autonomously. A multi-task learning framework is employed, training the small model on both the LLM's output and the QA data. Secondly, during inference, we augment the Student model's generation process with LLM's Relevant-Cell-based rationale, yielding results surpassing those using small models or LLMs alone. With less than 10K training data, our method achieves a state-of-the-art 41.3 BLEU score on FeTaQA, demonstrating the feasibility of combining LLMs' powerful reasoning and small models' task-specific capabilities while maintaining high data and computation efficiency compared to previous table pre-training methods.

In summary, the contributions of this paper are:

- Proposing Relevant-Cell-based Knowledge Distillation with inference-time Teacher Guidance to combine small models' capabilities of learning from human annotation and LLMs' capabilities of effectively reasoning and providing correct answers.
- Achieving state-of-the-art performance on FeTaQA through our proposed method.
- Demonstrating the feasibility of combining LLMs' reasoning strength and specially-trained small models' task-specific strength in a challenging downstream task.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

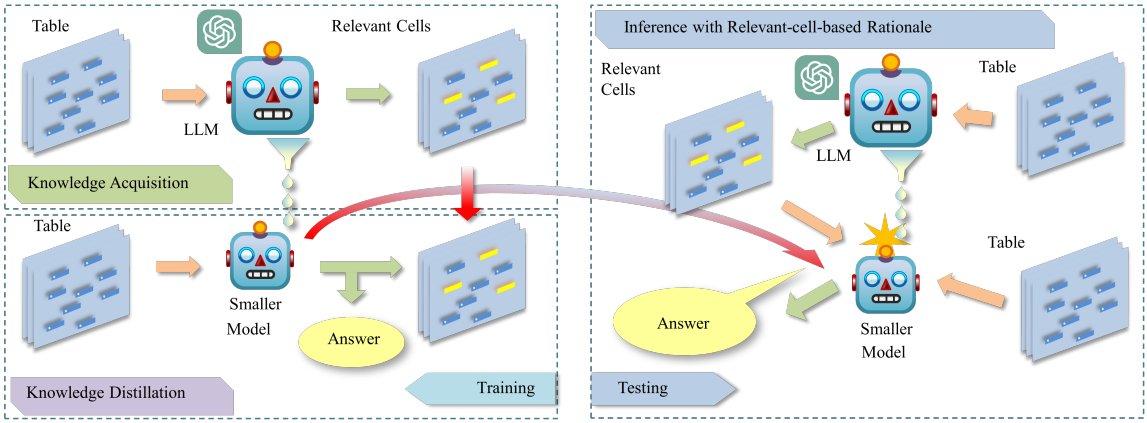

Figure 1: Overview of our proposed method. In Knowledge Acquisition phase, we prompt LLM to produce Relevant-Cell-based Rationales. In Knowledge Distillation phase, we distill LLM's knowledge to the Student model by using Relevant-Cell-based Rationales as target sequences for Student to generate. In the meantime, the Student model is trained on the table question answering data, forming a multi-task learning framework. At inference phase, we found that incorporating Teacher model's rationales as guidance for Student model can greatly improve the quality of Student model's answers.

2 Related Work

2.1 Table-to-Text Generation

To combine Natural Language Generation and table grounding, prior research has largely focused on the domain of table-to-text generation (Chen et al., 2020; Parikh et al., 2020; Cheng et al., 2021; Liu et al., 2022b; Nan et al., 2021; Zhao et al., 2023b). These works are more closely related to controllable text generation rather than question answering, as they often involve generating correct and faithful descriptions using highlighted table regions as controlling elements. For instance, datasets like ToTTo (Parikh et al., 2020) and HiTab (Cheng et al., 2021) require models to convert highlighted table segments into single-sentence summaries. LogicNLG (Chen et al., 2020) revolves around generating statements that can be logically derived from facts within provided table regions. RotoWire (Lebret et al., 2016) requires models to craft summaries based on basketball game tables, while SciGen (Moosavi et al., 2021) and NumericNLG (Suadaa et al., 2021) datasets demand arithmetic reasoning and description generation for tables from scientific papers. QTSumm (Zhao et al., 2023c) tasks models with generating comprehensive summaries encompassing key information from tables spanning diverse topics.

2.2 Table Question Answering

Distinct from table-to-text generation, Table QA involves responding to user queries based on information from a source table. Notable datasets such as WikiTableQuestions (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), and WikiSQL (Zhong et al., 2017) primarily emphasize short-form answers. To address these Table QA tasks, researchers have employed table-based pretraining techniques (Herzig et al., 2020; Eisenschlos et al., 2020; Shi et al., 2022; Yin et al., 2020; Liu et al., 2021). A more recent addition, FeTaQA (Nan et al., 2021), adopts ToTTo's (Parikh et al., 2020) statements, reformulating them into questions and utilizing the same statements as answers. In this work, we select FeTaQA as the benchmark to evaluate out methods.

2.3 Knowledge Distillation via LLMs

Ever since (Hinton et al., 2015) proposed the Knowledge Distillation method, many works have followed its white-box distillation paradigm. Notably, recent approaches have taken up symbolic knowledge distillation (West et al., 2021), also referred to as black-box knowledge distillation or sample matching.

In contrast to distilling soft representations such as logits from larger models, this approach employs a black-box model, often an LLM, to generate training samples for Student models (Petroni et al., 2019; West et al., 2021; Liu et al., 2022a; Bhagavatula et al., 2022). Recent progress in LLM knowledge distillation includes the works of (LI et al., 2022; Huang et al., 2022; Magister et al., 2022), and (Ho et al., 2022), all demonstrates that small models can gain insight from the Chain-of-Thought (CoT) rationales of larger models. However, these works focus on reasoning tasks that yield definite answers, such as answer choices or numerical values. In contrast, our work focuses on free-form question answering, where responses are in the form of natural language text.

3 Methodology

3.1 Relevant-Cell-based Rationale Generation

Existing LLM Knowledge Distillation method does not perform ideally on TableQA tasks. Existing methods rely on CoT prompting to distill knowledge from LLMs. Such knowledge fails to produce detailed and crucial information. We proposed a novel and effective method to prompt LLM to generate useful knowledge for TableQA task (see Figure 1, upper left).

Prior works on LLM Knowledge Distillation predominantly focus on reasoning tasks where the input consists solely of questions (Li et al., 2023; Hsieh et al., 2023; Wang et al., 2022). These tasks require models to answer questions using knowledge incorporated into their parameters during pre-training, or acquired from external sources such as knowledge graphs or databases. However, for Table QA tasks, all necessary information to answer questions is presented within the input tables. Traditional CoT prompting method fails to adequately ground table content as well as its structured information, as the knowledge is too vague and abstract for small models to effectively learn from. To bridge this gap, we introduce a novel prompting technique designed specifically for extracting Relevant-Cell-based rationales.

Past prompt templates typically follow (Wei et al., 2022)'s format, with which LLMs are prompted to generate step-by-step reasoning sequences given a question and answer choice. These templates consist of question x_i, answer choice y_i, and human-written rationales z_i. However, our task of Free-Form Table Question Answering involves questions and tables as inputs, free-form sentence answers as output. Consequently, we extend the template to

$$\mathcal{P} = \{x_p, t_p, y_p, r_p\} \tag{1}$$

where x_p represents questions, t_p represents tables, y_p denotes answer references, and r_p represents our hand-crafted exemplars for extracting Relevant-Cell-based rationales.

To construct our prompt template, we randomly select ten examples (x_p, t_p, y_p) from the training set

$$\mathcal{D}_{train} = \left\{\left(x_i^{train}, t_i^{train}, y_i^{train}\right)\right\}_{i=1}^{|D_{train}|} \tag{2}$$

and manually craft Relevant-Cell-based rationale r_p exemplars for the selected examples. These exemplars contain reasoning paths from question x_i to relevant table cells, coupled with explicit mention of the cells' corresponding column header names.

We argue that including these explicit column headers aids the small model in bridging the semantic gap between the question and cell content, thereby enhancing its ability to locate relevant information in the tables. Furthermore, these column headers denote the spatial positioning of table cells, acting as implicit guidance for the small model to comprehend table structural information. This collection of ten exemplars forms the seed set $\mathcal{P}$, with one exemplar $\mathcal{P}_k = (x_k, t_k, y_k, r_k) \sim \mathcal{P}$ randomly selected each time the LLM is prompted. This selected exemplar $\mathcal{P}_k$ is prepended to the question x_i and table input t_i, then the LLM generate the Relevant-Cell-based rationale r_i, and this process is repeated across the entire dataset, thus equipping each example with its distinct Relevant-Cell-based rationale. The result of this process is a corpus

$$\mathcal{C}_{train} = \{(x_i, t_i, y_i, r_i)\}_{i=1}^{|D_{train}|} \tag{3}$$

which contains training data and its corresponding rationales.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

3.2 Multi-Task Learning Framework Knowledge Distillation

In order to effectively utilize LLM generated Relevant-Cell-based rationales in both the training stage and inference stage, the following training framework is proposed.

There are two ways of incorporating rationales into the training process. The first is to put it in the input, and teach the small model to generate an answer given the question, table and rationale. $f(x_i, t_i, r_i) \rightarrow y_i$, where f denotes the small model.

The second way is to use the rationale as the target sequence, and ask the small model to generate it when given the question and table: $f(x_i, t_i) \rightarrow r_i$.

We argue that, in order to improve the small model's capabilities to reason and extract crucial information from structured data, it is imperative to use the rationale as the target sequence. In this way, the small model can effectively learn how to generate Relevant-Cell-based rationales, so that it can obtain the ability to perform appropriate reasoning steps when given a table and the question.

So that in the training phase, there are two target sequences for the small model, one is the Relevant-Cell-based rationale, the other is the answer to the question. While using the answers as the target sequences, the Relevant-Cell-based rationales are appended to the tables and questions to form the complete input sequences. By doing this, the small model can be trained to adopt Teacher's guidance at test time, as it has been familiarized with this input-output mapping during training.

So, in one optimizing step, the model is required to output model generated rationale and answer. Two separate losses

$$\mathcal{L}_{rationale} = \frac{1}{|C_{train}|} \sum_{i=1}^{|C_{train}|} l(f(x_i, t_i), r_i) \tag{4}$$

$$\mathcal{L}_{answer} = \frac{1}{|C_{train}|} \sum_{i=1}^{|C_{train}|} l(f(x_i, t_i, r_i), y_i) \tag{5}$$

are calculated between model generation and their respective ground-truth references. The total loss of

$$\mathcal{L} = \mathcal{L}_{rationale} + \mathcal{L}_{answer} \tag{6}$$

is then calculated. Consequently, a multi-task learning framework is formed (see Figure 1, lower left).

Additionally, we followed Distilling step-by-step and prepended task prefixes to the input of two different tasks. We prepended `[Relevant cells]` to the input to indicate the small model to generate Relevant-Cell-based rationale, and prepended `[Answer]` to indicate the small model to generate an answer.

3.3 Student Model Inference with Teacher Guidance

Using the multi-task learning framework, the small model alone can now be finetuned to produce answers that are similar to the surface forms of human annotation, but its ability to produce factually accurate, faithful and adequate answers remains limited. We propose a simple test-time augmentation to address this challenge. First, the Teacher model is prompted on the test set to produce Relevant-Cell-based rationales $r_{teacher}$ for the small model, serving as Teacher Guidance. When using the small model to generate answers $\hat{a}$ on the test set, the Teacher Guidance is appended to the tables t_i and questions x_i as input for the small model. We call this method **Inference with Teacher Guidance** (see Figure 1, right). The formulation for it is: $f(x_i, t_i, r_{teacher}) \rightarrow \hat{a}$, where f denotes the finetuned small model.

We also tested small model's performance to produce answers on its own, without using Teacher's Relevant-Cell-based rationales at test time. We let the finetuned small model directly generate answer $\hat{a}$ given the question x_i and table t_i in a test example. We call this method **Direct Inference**. This setting helps us better understand the effectiveness of our proposed Relevant-Cell-based Rationales compared to other baselines that does not involve the Teacher model at inference time.

Model	sacreBLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
T5-large (Nan et al., 2021)	30.5	0.63	0.41	0.53	0.49
TAPEX (Liu et al., 2021)	30.2	0.62	0.40	0.51	-
ReasTAP (Zhao et al., 2022)	30.4	0.63	0.40	0.51	-
UnifiedSKG (Xie et al., 2022)	32.4	0.64	0.42	0.54	0.51
GenTaP (Shi et al., 2022)	36.7	0.69	0.48	0.59	0.55
GPT-3.5-turbo (1-shot)	21.9	0.60	0.37	0.49	0.55
Ours, trained w/ CoT, **Direct Inference**	34.1	0.65	0.44	0.56	0.52
Ours, trained w/ RC, **Direct Inference**	34.3	0.66	0.44	0.56	0.52
Ours, trained w/ CoT, **Inference with Teacher Guidance**	31.5	0.63	0.41	0.53	0.50
Ours, trained w/ RC, **Inference with Teacher Guidance**	**41.3**	**0.70**	**0.50**	**0.61**	**0.58**

Table 1: Main results of our method comparing to baselines on the FeTaQA benchmark. CoT stands for Chain-of-Thought rationales. RC stands for Relevant-Cell-based Rationales. Due to the limitation of context window, only one dataset example can be fitted into the prompt template for GPT-3.5-turbo.

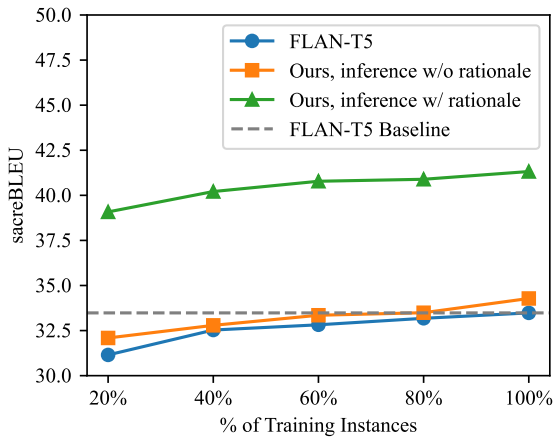

Figure 2: sacreBLEU score on FeTaQA benchmark with different amounts of training data.

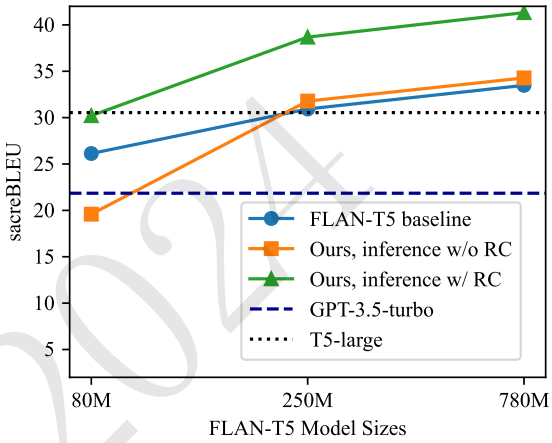

Figure 3: sacreBLEU score on FeTaQA benchmark with different sizes of FLAN-T5 model.

4 Experiments

In the upcoming experiments, we present a comprehensive analysis of our proposed **RCKD-TG** method. Firstly, we showcase the impressive performance of our method in comparison to previous works and the current state-of-the-art, highlighting its competitiveness across various evaluation metrics. Secondly, we explore the impact of dataset scale and model size, demonstrating our approach's remarkable data and computational efficiency. Thirdly, we conduct an in-depth evaluation of the answers generated by our method compared to those produced by LLMs. Also, we assess the quality of the Relevant-Cell-based rationales generated by both Teacher and Student model. Lastly, we delve into the capabilities of instruction-tuned models to effectively follow rationales and explore the impact of different types of instructions.

4.1 Datasets

We evaluate the performance of our method using the FeTaQA (Nan et al., 2021) benchmark, a comprehensive free-form table question answering benchmark comprising more than 10,000 (question, table, free-form answer, highlighted table cells) data samples. Unlike previous studies on Table Question Answering (Pasupat and Liang, 2015; Zhong et al., 2017; Yu et al., 2018) and LLM Knowledge Distillation (Li et al., 2023; Hsieh et al., 2023; Wang et al., 2022), FeTaQA presents unique challenges. The questions within this benchmark exhibit greater complexity, requiring intricate reasoning across multiple table cells. Additionally, the answers are expressed in a free-form manner, which demands the generation of

Method	sacreBLEU
Ours, **Direct inference**	34.3
Ours, **Inference with Teacher Guidance**	41.3
GPT-3.5-turbo, 0-shot	21.6
GPT-3.5-turbo, 1-shot	21.9
GPT-3.5-turbo, w/ surface form constraint	24.1

Table 2: Automatic evaluation results of GPT-3.5-turbo and our method's predicted answers.

in-depth elaborations and explanations as model predictions.

4.2 Implementation Details

In our implementation, the input is a pair of a question and its corresponding table. We followed UnifiedSKG to flatten the table as $\mathcal{T} = col : h, row_1 : r_1, ..., row_n : r_n$, where h is column header, r_i is the $i-th$ table row, "col" and "row" are natural texts indicating the region of table column headers and rows respectively. We also use a vertical bar "|" to distinguish between cells. The sequence begins with "`Table:` ", indicating the forthcoming linearized table information.

For generating Relevant-Cell-based rationales with LLM, we used GPT-3.5-turbo as our Teacher model. We sample from GPT-3.5-turbo with a one-shot prompt described in Methodology section and a temperature of 1.0. We sample one rationale from each data instance in FeTaQA training set and test set. For finetuning the small model on FeTaQA, we either use 8 NVIDIA RTX 3090 GPUs or 4 NVIDIA A40 GPUs. Initial learning rate is set to 5e-5 and a linear learning rate decay scheduler is employed. We use AdaFactor as the optimizer. We ran 40 epochs with the batch size of 64, and use early stopping with a patience of 5. The best checkpoint is selected according to the sacreBLEU score on the validation set.

4.3 Baselines

We conduct a comprehensive comparison of our method with several established baselines:

TAPEX (Liu et al., 2021): This method involves continuing pre-training of the BART model using synthetic SQL query execution data, enhancing model's table reasoning capabilities through SQL execution.

ReasTAP (Zhao et al., 2022): Based on the BART model, ReasTAP focuses on improving table understanding and reasoning abilities through pre-training on a synthetic corpus with questions that requires numerous types of reasoning skills to solve.

GPT-3.5-turbo: We perform one-shot testing on FeTaQA using our Teacher model alone and compare its performance with our proposed methods.

UnifiedSKG (Xie et al., 2022): This baseline unifies various structured knowledge grounding tasks with text-to-text language models, achieving remarkable performance across a range of table-to-text tasks.

It's noteworthy that ReasTAP and TAPEX did not originally evaluate their method on FeTaQA, we report their results from those published in this recent paper (Zhao et al., 2023a).

4.4 Models

Our standard Student model is FLAN-T5-large, which has 780M parameters. FLAN-T5 enhances T5 by scaling instruction fine-tuning and demonstrates better human-like reasoning abilities than the T5 model. We choose FLAN-T5 because it exhibited better rationale learning abilities compared with raw T5, as demonstrated in the following sections.

4.5 Main Results

Our primary results are obtained on the FeTaQA benchmark through training FLAN-T5-large Student model with Relevant-Cell-based rationales generated by GPT-3.5-turbo Teacher model. We also discuss the performance using regular CoT rationales during training, as shown in Table 1.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Ground Truth: Janice Hahn received 33.3% of the vote against Newsom's 55.5%.
Ours w/ CoT: Janice Hahn performed less favorably compared to Gavin Newsom in terms of both votes and percentage of votes. **Ours w/ RC:** Janice Hahn received 33.3% of the vote against Gavin Newsom's 55.5%.

Table 3: An example of the CoT-trained *Inference with Teacher Guidance* prediction that fails to provide a detailed answer.

Ground Truth: At the World Championships, Winnie Ng participated twice in the marathon finishing 23rd in 1993 and 30th in 1995.
T5-large: At the 1995 World Championships in Gothenburg, Sweden, Winnie Ng finished 30th with a time of 3:01:08. **Direct Inference:** Winnie Ng finished 23rd in the marathon at the 1993 World Championships and 30th at the 1995 World Championships. **Teacher Guidance:** Winnie Ng participated in the World Championships twice in the marathon event, finishing 23rd in 1993 and 30th in 1995.

Table 4: An example of our method's effectiveness on FeTaQA benchmark.

We report results from two testing modes: **Direct Inference** and **Inference with Teacher Guidance**. **Direct Inference** achieves a sacreBLEU score of 34.3, surpassing all methods except the existing state-of-the-art (SOTA), and outperforming the CoT rationale-trained model of 34.1 sacreBLEU score. **Inference with Teacher Guidance** achieves a sacreBLEU score of 41.32, establishing a new SOTA, while massively outperforming its CoT-trained counterpart which achieved 31.46. Using CoT-trained **Inference with Teacher Guidance** method could lead to small models generating vague and general answers, failing to point out specific details needed to thoroughly and correctly explain a query, as shown in Table 3.

In Table 2, we used 0-shot, 1-shot and 1-shot with surface form constraint settings to evaluate GPT-3.5-turbo on FeTaQA. For 1-shot with surface form constraint setting, we explicitly ask GPT-3.5-turbo to conform to the style of human annotated answer in the 1-shot example using detailed description. This method achieved a 24.1 sacreBLEU score compared to 0-shot setting's 21.6 and 1-shot setting's 21.9, showing marginal improvements as the efficacy of learning from in-context examples is limited.

Moreover, our method accomplishes this with just 7,326 training samples, along with an additional 7,326 rationales. Compared to previous SOTA's 500K synthetic pre-training corpus, our method demonstrates exceptional data efficiency. Furthermore, our approach outperforms the T5-3B model reported in UnifiedSKG, while utilizing a much smaller FLAN-T5 model with only 780M parameters, illustrating its computational efficiency.

4.6 Dataset Size Scaling and Model Size Scaling

We conducted experiments on datasets of varying proportions to investigate our method's performance in low-data regime. Training was performed on datasets comprising 20%, 40%, 60%, 80%, and 100% of the total data, as shown in Figure 2. Our Relevant-Cell-based rationale approach consistently outperforms the FLAN-T5 baseline. When using **Inference with Teacher Guidance**, our method trained on different

Source	Fluent	Correct	Adequate	Faithful
T5-large	85.5	60.3	71.8	69.5
Ours, trained w/ RC, **Direct inference**	88.4	67.3	76.9	79.7
Ours, trained w/ RC, **Inference with Teacher Guidance**	91.3	74.1	79.7	85.9
GPT-3.5-turbo	93.7	80.1	83.5	87.7

Table 5: Human evaluation results of T5-large baseline, our method and GPT-3.5-turbo's predicted answers on FeTaQA.

Source	Correct	Adequate	Faithful
LLM Generated Knowledge	84.3	81.3	97.9
FLAN-T5 Generated Knowledge	65.4	67.4	88.5

Table 6: Quality of Relevant-Cell-based rationales generated by either Teacher (LLM) or Student (FLAN-T5).

Method	sacreBLEU	
	result	Δ
T5	32.5	-
T5 w/ CoT	32.5	+0.0
T5 w/ Relevant Cells	33.4	+0.9
FLAN-T5	33.5	-
FLAN-T5 w/ CoT	34.1	+0.6
FLAN-T5 w/ Relevant Cells	34.3	+0.8

Table 7: Performance difference of T5 and FLAN-T5 trained on knowledge distilled from Teacher model and their respective baselines trained only on QA data.

dataset proportions all outperform the existing SOTA, which was achieved with 500K training samples, showcasing our method's data efficiency.

Similarly, we experimented with FLAN-T5 models with different sizes of full dataset to evaluate our method's effectiveness across model sizes, as shown in Figure 3. We found that our approach consistently surpasses the respective baseline of each model size, except for the 80M FLAN-T5-small model, where our method's performance drops. This phenomenon may be attributed to the small model's limited capacity to accommodate the complexity of two-task multi-task training. However, the performance significantly improves when using **Inference with Teacher Guidance**. This illustrates that even in situations where small models struggle to answer questions independently, introducing Teacher model knowledge can greatly enhance the small model's performance. This demonstrates the effectiveness of our combined small and large model approach.

4.7 Human Evaluation of Predictions and Rationales

We conducted comprehensive human and automatic evaluations of model predictions and Relevant-Cell-based rationales generated by the Student and Teacher models. We followed the human evaluation method in (Nan et al., 2021), and evaluated T5-large baseline, our proposed method and GPT-3.5-turbo's predictions in terms of Fluency, Correctness, Adequacy, and Faithfulness. Three NLP practitioners scored 50 model predictions on a scale of 0-5, which was then converted to a percentage scale, as shown in Table 5. Our methods consistently outperformed the T5-large baseline. **Inference with Teacher Guidance** method also exhibited notable gains in performance compared to **Direct Inference** method. Exemplified by Table 4, our methods exhibit significant improvements in answer completeness and correctness. Our best method achieved 97.3%, 92.5%, 95.4% and 97.9% of GPT-3.5-turbo's performance in terms of fluency, correctness, adequacy and faithfulness, while their automatic evaluation scores significantly surpassed those of GPT-3.5-turbo. This result demonstrates the feasibility of utilizing small model's ability of generating human annotated style text while retaining LLM's strong reasoning capabilities.

Similarly, we conducted human evaluations of the quality of Relevant-Cell-based rationales generated

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

Instruction Type	# of instructions					
	0	1	5	10	15	20
No Instruction	28.1	-	-	-	-	-
Task Prefix	-	**34.3**	-	-	-	-
NL Instructions	-	33.9	33.8	34.1	34.0	33.6

Table 8: Effect of different types and number of instructions on final performance on FeTaQA benchmark.

by FLAN-T5-large and GPT-3.5-turbo, as shown in Table 6. Three NLP practitioners evaluate 50 rationales based on three aspects: Correctness, Adequacy and Faithfulness. Our evaluations revealed that GPT-3.5-turbo's rationales significantly outperformed FLAN-T5-large's rationales in all aspects. Integrating FLAN-T5-generated rationales into the inference stage led to a decrease in performance, even lower than the baseline. This suggests that FLAN-T5's ability to self-generate rationales is limited, and directly using them to augment FLAN-T5's own inference stage hurts performance, demonstrating a disparity from previous works on classification and multiple choice tasks

4.8 Comparison of Instruction-Tuned FLAN-T5 and Raw T5's Rationale Following Abilities and the Effect of Different Types of Instructions

We separately trained raw T5 and instruction-tuned FLAN-T5 models using CoT and Relevant-Cell-based rationales, and compare the improvements over their respective baselines, as shown in Table 7. The results demonstrated that the results achieved by FLAN-T5 models through rationale-based training was comparable or better compared to raw T5 models, aligning with (Fu et al., 2023)'s observations. This suggests that, in future works that aim to enhance model reasoning and instruction-following abilities, utilizing instruction fine-tuned checkpoints of existing models could yield better performance.

We also examined the impact of different types of instructions, including "No Instructions", "Task Prefix instructions", and a different number of "Natural Language Instructions", as shown in Table 8. Results demonstrate that models trained without instructions achieved low performance of 28.1 sacreBLEU score, while those trained with 10 Natural Language instructions performed similarly to models trained with Task Prefix instructions, albeit slightly lower. Our Natural Language instructions were diverse and accurate descriptions of the two tasks in the multi-task learning framework, generated by the Teacher model and edited by the authors. The superior performance of the Task Prefix instruction might be attributed to T5's pre-training strategies, which employs similar task prefixes during pre-training on various tasks, such as `[summarize]`, `[snli sentence]`, `[cola sentence]`. This familiarity helps FLAN-T5 models better understand and distinguish the two tasks in our multi-task learning framework.

5 Conclusion

In conclusion, this paper introduces a novel Relevant-Cell-based Knowledge Distillation with inference-time Teacher Guidance method to combine small models' capabilities of learning from human annotation and LLMs' capabilities of reasoning effectively. Through a multi-task learning on instruction-tuned models and inferencing with Teacher model's guidance, we achieved a state-of-the-art BLEU score of 41.3 on the FeTaQA benchmark. This approach combines LLMs' robust reasoning strength with smaller models' task-specific capabilities, while maintaining high data and computational efficiency. Our work advances free-form table question answering and demonstrates a promising path of fusing general LLMs and task-specific models on downstream tasks. Future works could explore on how to more effectively combine and communicate between the Student model and Teacher model. Future works could also explore on extending our method to diverse downstream tasks, leveraging general LLMs as strong reasoners, and task-specific small models as response generators aligned with human annotated preferences.

References

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2022. I2d2: Inductive knowledge distillation with neurologic and self-imitation. In *Annual Meeting of the Association for Computational Linguistics*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, SHIYANG LI, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *ArXiv*, abs/1909.02164.

Wenhu Chen, Jianshu Chen, Yunde Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. *ArXiv*, abs/2004.10404.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation. *ArXiv*, abs/2108.06712.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Yao Fu, Hao-Chun Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *ArXiv*, abs/2301.12726.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Annual Meeting of the Association for Computational Linguistics*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. In *Annual Meeting of the Association for Computational Linguistics*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Annual Meeting of the Association for Computational Linguistics*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *ArXiv*, abs/2210.11610.

Mohit Iyyer, Wen tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Conference on Empirical Methods in Natural Language Processing*.

SHIYANG LI, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *ArXiv*, abs/2210.06726.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Annual Meeting of the Association for Computational Linguistics*.

Qian Liu, Bei Chen, Jiaqi Guo, Zeqi Lin, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *ArXiv*, abs/2107.07653.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022b. Plog: Table-to-logic pretraining for logical table-to-text generation. In *Conference on Empirical Methods in Natural Language Processing*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. In *Annual Meeting of the Association for Computational Linguistics*.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *ArXiv*, abs/2004.14373.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Annual Meeting of the Association for Computational Linguistics*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Peng Shi, Patrick Ng, Feng Nan, Henghui Zhu, J. Wang, Jiarong Jiang, Alexander Hanbo Li, Rishav Chakravarti, Donald Weidner, Bing Xiang, and Zhiguo Wang. 2022. Generation-focused table-based intermediate pre-training for free-form question answering. In *AAAI Conference on Artificial Intelligence*.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Pinto: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. In *North American Chapter of the Association for Computational Linguistics*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *ArXiv*, abs/2201.05966.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. In *Annual Meeting of the Association for Computational Linguistics*.

Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *ArXiv*, abs/2005.08314.

Tao Yu, Rui Zhang, Kai-Chou Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Z Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *ArXiv*, abs/1809.08887.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir R. Radev. 2022. Reastap: Injecting table reasoning skills during pre-training via synthetic reasoning examples. In *Conference on Empirical Methods in Natural Language Processing*.

Yilun Zhao, Boyu Mi, Zhenting Qi, Linyong Nan, Minghao Guo, Arman Cohan, and Dragomir R. Radev. 2023a. Openrt: An open-source framework for reasoning over tabular data. In *Annual Meeting of the Association for Computational Linguistics*.

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Yu Flores, and Dragomir R. Radev. 2023b. Loft: Enhancing faithfulness and diversity for table-to-text generation via logic form control. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Xiangru Tang, Yumo Xu, Arman Cohan, and Dragomir R. Radev. 2023c. Qtsumm: A new benchmark for query-focused table summarization. *ArXiv*, abs/2305.14303.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers