Large Language Models are few(1)-shot Table Reasoners

Wenhu Chen
University of Waterloo, Vector Institute
wenhuchen@uwaterloo.ca

Abstract

Recent literature has shown that large language models (LLMs) are generally excellent few-shot reasoners to solve text reasoning tasks. However, the capability of LLMs on table reasoning tasks is yet to be explored. In this paper, we aim at understanding how well LLMs can perform table-related tasks with few-shot in-context learning. Specifically, we evaluated LLMs on popular table QA and fact verification datasets like WikiTableQuestion, FetaQA, TabFact, and FEVEROUS and found that LLMs are competent at complex reasoning over table structures, though these models are not pre-trained on any table corpus. When combined with 'chain of thoughts' prompting, LLMs can achieve very strong performance with only a 1-shot demonstration, even on par with some SoTA models. We show that LLMs are even more competent at generating comprehensive long-form answers on FetaQA than tuned T5-large. We further manually studied the reasoning chains elicited from LLMs and found that these reasoning chains are highly consistent with the underlying semantic form. We believe that LLMs can serve as a simple yet generic baseline for future research. The code and data are released in https://github.com/wenhuchen/TableCoT.

1 Introduction

The problem of structured knowledge grounding has been extensively studied for many years. Tables, as one of the most popular (semi)-structured forms to store world knowledge receive significant attention from the natural language processing (NLP) community. Traditional approaches mostly rely on synthesizing executable languages like SQL or SPARQL to access the information inside the table. However, these symbolic languages normally make a rigid assumption about the table and cannot capture the semantics of text chunks inside the table. Such issues are even more pronounced with web tables due to their irregular forms. To fully understand web tables, both structured reasoning and textual reasoning are required. Such challenges have attracted many researchers to work in the field. Recently, a wide range of table-based tasks have been proposed like table question answering (Pasupat and Liang, 2015; Chen et al., 2020c; Zhu et al., 2021; Chen et al., 2021b; Talmor et al., 2020; Chen et al., 2020a; Nan et al., 2022), table fact verification (Chen et al., 2019; Aly et al., 2021), table-based generation (Chen et al., 2020b; Parikh et al., 2020; Nan et al., 2021), and table-grounded conversation (Budzianowski et al., 2018; Nakamura et al., 2022). This wide range of table-based tasks all come with different input-output formats and domains. Due to the heterogeneity of these tasks, models achieving the best results on these tasks normally need to be fully fine-tuned on the specific downstream dataset with 10K-100K examples to achieve reasonable performance.

Recently, there have been efforts like UnifiedSKG (Xie et al., 2022) aiming to unify these heterogeneous table-based tasks as a generic text-to-text format. UnifiedSKG has shown that using T5-3B (Raffel et al., 2020) with the text-to-text format can already achieve state-of-the-art performance on almost all the table-based tasks without task-specific designs. However, the proposed text-to-text models still need to be fully fine-tuned on the downstream tasks. UnifiedSKG also identified that T0-style (Sanh et al., 2022) cross-task transfer can only achieve almost random performance.

Wei et al. (2022); Wang et al. (2022); Zhou et al. (2022); Drozdov et al. (2022) have recently discovered that large language models (Brown et al., 2020; Chowdhery et al., 2022; Ouyang et al., 2022) can be used to solve complex mathematical and commonsense reasoning tasks with few-shot in-context learning. Inspired by this discovery, we aim at understanding whether these LLMs can also solve complex table-based reasoning tasks. Though the LLMs are not specifically designed to encode ta-

Findings of the Association for Computational Linguistics: EACL 2023, pages 1120–1130

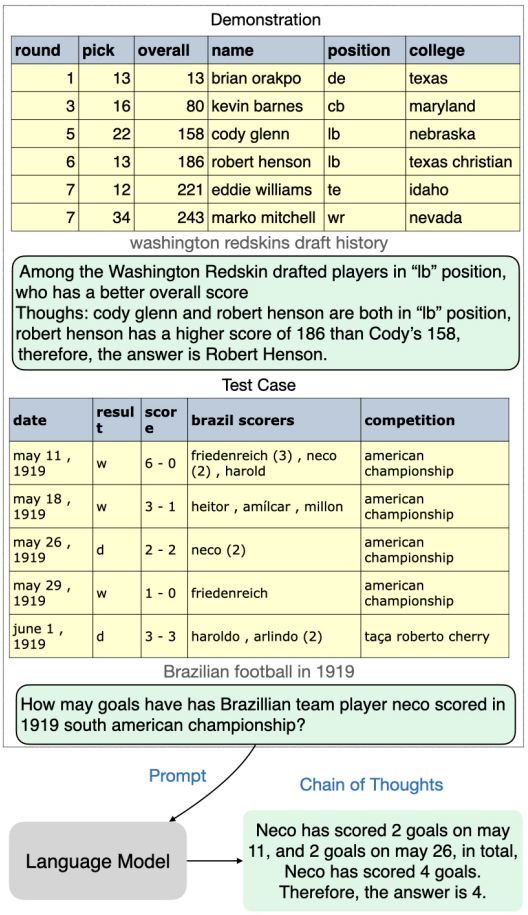

Figure 1: In-context learning for table-related tasks with chain-of-thoughts reasoning.

bles, given the enormous number of tables present in the pre-training corpus, we believe they are also competent at reasoning over table information.

In this paper, we experimented with few-shot in-context learning for LLMs as depicted in Figure 1. Instead of fine-tuning the model, we only provide a few examples to showcase the desired input-output format as the condition for the model to follow to solve unseen test examples. We experiment with several prompting variants including (1) direct prediction, (2) Chain of Thoughts (Wei et al., 2022) (CoT), (3) Chains of thoughts with self-consistency (Wang et al., 2022) (CoT+SC). We evaluate these methods on WikiTableQA (Pasupat and Liang, 2015), FetaQA (Nan et al., 2022), TabFact (Chen et al., 2019) and FEVEROUS (Aly et al., 2021). Our results reveal that LLMs (Ouyang et al., 2022; Chen et al., 2021a; Chowdhery et al., 2022) can achieve striking performance with only 1 or 2 demonstrations, e.g. 48.8% on WikiTableQuestions and 78.8% on TabFact, which are on par some near-SoTA models (Yu et al., 2021; Eisenschlos et al., 2020). On other datasets like FetaQA with long-form answers, our human evaluation reveals that GPT-3 can significantly outperform the fine-tuned T5-large by more than 30% in terms of correctness and adequacy.

Furthermore, we manually studied the chain of thoughts elicited from LLMs and found that the rationale is highly consistent with the 'ground truth' semantic forms when the model predictions are correct. We found that these models are surprisingly competent at performing symbolic operations over the table, like maximum, minimum, counting, comparison, addition, and difference. However, we also identify several issues of the LLMs on these table reasoning tasks: (1) due to the token limitation, the model is unable to generalize to 'huge' tables with 30+ rows, which is the major error source, (2) LLMs can sometimes make simple mistakes when performing symbolic operations.

Due to the simplicity and generality, we believe LLMs with CoT should be used as an important baseline for any future table-related research.

2 Related Work

2.1 Reasoning over Tables

Table-based reasoning is traditionally accomplished by semantic parsing to execute commands on tables like WikiTableQuestions (Pasupat and Liang, 2015), WikiSQL (Zhong et al., 2017), and Spider (Yu et al., 2018). These models aim to synthesize SQL/SPARQL to interact with tables. However, these machine languages have a rigorous requirement regarding the tables, e.g. the value in the same column should follow the same data type. Such rigorous assumptions are frequently violated by web tables containing unnormalized free-form text in cells. Therefore, language understanding inside the table is essential to achieve a better score. Recently, Yin et al. (2020); Herzig et al. (2020); Liu et al. (2021); Deng et al. (2022) have proposed to pre-train table and text to learn joint representation. These pre-trained models can use joint representation to perform reasoning implicitly without relying on symbolic execution. By pre-training the model on large-scale crawled or synthesized data, these models can normally achieve the best-known performance on table tasks. However, these models still require a significant amount of fine-tuning on the downstream datasets. Unlike these methods, we are interested in in-context learning, where the model can only learn with a

few examples (demonstration) without any fine-tuning. One contemporary work similar to ours is BINDER (Cheng et al., 2022), which utilizes Codex to synthesize SQL to execute logical forms against tables for question answering. One big difference is that BINDER (Cheng et al., 2022) involves logical form execution, if the execution fails, BINDER will fall back to using language models to answer the question, which is more similar to ours.

2.2 In-context Learning with LLMs

GPT-3 (Brown et al., 2020) and other large language models demonstrated strong abilities to perform few-shot predictions without fine-tuning, where the model is given a description of the task in natural language with few examples. Scaling model size, data, and computing are crucial to enable this learning ability. Recently, (Rae et al., 2021; Smith et al., 2022; Chowdhery et al., 2022; Du et al., 2022) have proposed to train different types of large language models with different training recipes. The LLMs have demonstrated a striking capability utilizing the few-shot prompts to accomplish unseen tasks without any fine-tuning, which is found to be an emergent capability not presented in smaller language models.

2.3 Chain of Thoughts Reasoning

Although LLMs (Brown et al., 2020; Chowdhery et al., 2022) have demonstrated remarkable success across a range of NLP tasks, their ability to demonstrate reasoning is often seen as a limitation. Such capability cannot be acquired simply by scaling up the model size. Recently, the 'chain of thoughts' prompting (Wei et al., 2022) has been discovered to empower LLMs to perform complex reasoning over text. By providing the model with several exemplars of reasoning chains, LLMs can learn to follow the template to solve difficult unseen tasks. Later, Wang et al. (2022) propose to use self-consistency with CoT to further improve performance. Later on, Kojima et al. (2022) discovered that LLMs can even perform reasoning without any demonstration by using appropriate prompts. These recent findings reveal the strong capability of LLMs to perform complex reasoning. However, the current studies are still heavily focused on text-based tasks like question answering, common sense reasoning, etc. The models' capability to reason over tables is yet unknown. In this paper, we are specifically interested in understanding LLMs' capability to reason over web tables with CoT prompting.

Question Answering (WikiTableQA, FeTaQA)

Read the table below regarding "2008 Clásica de San Sebastián" to answer the following questions.

Rank | Cyclist | Team | Time | UCI ProTour Points
1 | Alejandro Valverde (ESP) | Caisse d'Epargne | 5h 29' 10 | 40
2 | Alexandr Kolobnev (RUS) | Team CSC Saxo Bank | s.t. | 30
3 | Davide Rebellin (ITA) | Gerolsteiner | s.t. | 25
.....
7 | Samuel Sánchez (ESP) | Euskaltel-Euskadi | s.t. | 7
8 | Stéphane Goubert (FRA) | Ag2r-La Mondiale | + 2 | 5
9 | Haimar Zubeldia (ESP) | Euskaltel-Euskadi | + 2 | 3
10 | David Moncoutié (FRA) | Cofidis | + 2 | 1

Question: Which country had the most cyclists finish within the top 10?
Explanation: ITA occurs three times in the table, more than any others. Therefore, the answer is Italy

Question: How many cyclist has achieved more or equal than 30 ProTour Points?
Explanation: Alejandro Valverde and Alexandr Kolobnev obtain at least 30 points, therefore, the answer is 2

Fact Verfication (TabFact, FEVEROUS)

Read the table below regarding "2008 Clásica de San Sebastián" to verify whether the provided claims are true or false.

Rank | Cyclist | Team | Time | UCI ProTour Points
1 | Alejandro Valverde (ESP) | Caisse d'Epargne | 5h 29' 10 | 40
2 | Alexandr Kolobnev (RUS) | Team CSC Saxo Bank | s.t. | 30
3 | Davide Rebellin (ITA) | Gerolsteiner | s.t. | 25
.....
7 | Samuel Sánchez (ESP) | Euskaltel-Euskadi | s.t. | 7
8 | Stéphane Goubert (FRA) | Ag2r-La Mondiale | + 2 | 5
9 | Haimar Zubeldia (ESP) | Euskaltel-Euskadi | + 2 | 3
10 | David Moncoutié (FRA) | Cofidis | + 2 | 1

Claim: There are three palyers in 2008 Clásica de San Sebastián obtaining less than 10 points.
Explanation: Samuel Sánchez (ESP), Stéphane Goubert (FRA), Haimar Zubeldia (ESP) and David Moncoutié (FRA) are all getting less than 10 UCI ProTour points in the game, therefore, there are 4 players instead of 3 players, hence the claim is false

Claim: Only 2 cyclists have achieved at least 30 UCI ProTour Points.
Explanation: Alejandro Valverde and Alexandr Kolobnev are the two cyclists with at least 30 points, therefore the claim is true

Figure 2: Prompts used for question answering and fact verification tasks.

3 Method

We experiment with different in-context learning methods to solve the table-based reasoning tasks. To formulate the prompt, we linearize the table and concatenate it with a few examples as demonstrations of the language model to predict the output from an unseen test example. The format is described in Figure 2. We mainly investigate three different variants for language model prompting, including (1) Direct Prediction, (2) Chain of Thoughts (CoT), and (3) Chain of Thoughts + Celf-Consistentcy decoding (CoT+SC). For self-consistency methods, we use LLMs to generate five diverse reasoning paths and then use majority voting to select the most voted answer.

To limit the budget and constrain the input token length, we truncate the input tables to contain only

the first 22 rows and the first 8 columns. For each cell, we truncate the word length to contain only the first 10 words. Through such truncation, we can restrict the input token length to within 2000 tokens. We will talk about the impact of input token length on the final performance.

4 Experimental Results

For the GPT-3 experiments, we used the four provided models, Ada, Babbage, Curie, and Davinci with 350M, 1.3B, 6.7B, and 175B parameters respectively. We mainly use Davinci-text-002 (Ouyang et al., 2022) in our experiments. We also report results for Codex (Chen et al., 2021a) (Davinci-code-002) on some datasets. We use a temperature of 0.7 without any frequency penalty and without top-k truncation. We found that the model performance is robust to the sampling strategies and the hyper-parameters. These models are mainly trained on web-crawled data and code data, without any specialized training on table corpus.

4.1 Datasets

Here we list all of our datasets as follows:

WikiTableQuestions Pasupat and Liang (2015) consists of complex questions annotated based on Wikipedia tables. Crowd Workers are asked to compose a series of complex questions that include comparisons, superlatives, aggregation, or arithmetic operations. The annotated dataset is cross-validated by other crowd workers. In our experiments, we use the unseen test set for evaluation. We evaluate the standard test set with roughly 4000 questions. In this dataset, we adopt the answer exact match as our evaluation metric.

FetaQA Nan et al. (2022) consists of free-form table questions. These questions are mostly complex questions that require integrating information from discontinuous chunks in the table. Instead of having short answers, the dataset annotates long free-form answers. Unlike other datasets using copies of short text spans from the source, the questions in FetaQA require a high-level understanding. We adopt sacre-BLEU and human evaluation as our evaluation metrics. The evaluation set contains a total of 2003 examples.

TabFact Chen et al. (2019) consists of both simple and complex claims annotated by crowd workers based on Wikipedia tables. In the simple subset, the claims normally do not involve higher-order operations like max/min/count, etc. While the complex subset mainly contains claims involving higher-order operations. We evaluate the original test set containing 12,779 examples. We report binary classification accuracy on the set.

FEVEROUS Aly et al. (2021) consists of compositional claims annotated by crowd workers regarding Wikipedia tables. Since the dataset contains both table-supported and text-supported claims. We filter out text-supported claims and only keep the 2,295 table-supported claims as our test set. Different from TabFact, FEVEROUS consists of more complex tables with irregular structures like multi-row, multi-column, multi-table, etc. We report dev-set accuracy.

4.2 Baselines

In these experiments, we mainly consider the following baseline models.

Pre-trained Encoder-Decoder Model Pre-trained encoder-decoder model is one of our competitors, which aims to encode the table as a plain sequence into the encoder, and then apply the decoder to generate either an answer or a verdict. In this paper, we mainly compare against T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) as our baselines.

Pre-trained Table Understanding Model This family of models is specifically pre-trained on the table-related corpus, which utilizes specific architecture to encode table structure and handle symbolic computation. In this paper, we mainly consider TAPAS (Herzig et al., 2020), TABERT (Yin et al., 2020), and TAPEX (Liu et al., 2021).

Neural Symbolic Model This family of models includes a non-pre-trained neural symbolic model, which can synthesize machine language to interact with the table. This line of work includes Logic-FactChecker (Zhong et al., 2020), Neural-Symbolic Machine (Liang et al., 2018), etc.

4.3 Main Results

Here we show our main results for different datasets as follows.

WikiTableQuestions As can be seen from Table 1, directly asking GPT-3 to generate answers can only lead to 26% EM score. However, if we prompt the model with the CoT demonstrations,

Type	Model	Test EM
Train	Pasupat and Liang (2015)	37.1
Train	Zhang et al. (2017)	43.7
Train	Liang et al. (2018)	43.7
Train	Agarwal et al. (2019)	44.1
Train	Wang et al. (2019)	44.5
PT + FT	Herzig et al. (2020)	48.8
PT + FT	Yu et al. (2021)	**52.7**
1-shot	GPT-3 Direct	24.0
2-shot	GPT-3 Direct	27.3
1-shot	GPT-3 CoT	44.2
2-shot	GPT-3 CoT	45.7
2-shot	Codex CoT	48.8

Table 1: Experimental Results on WikiTableQuestions. PT means pre-training and FT means fine-tuning.

GPT-3 is more likely to follow the logical operation to derive the answers. With two demonstrations, GPT-3 can achieve roughly 46% EM score. By switching from GPT-3 to Codex, we are able to further improve the EM score to over 48.8%. These results are particularly surprising given that TAPAS has a built-in module to complete symbolic operations, while GPT-3 was not trained on any table-specific dataset. These results demonstrate GPT-3's built-in capabilities to perform diverse types of reasoning over tables.

FetaQA As demonstrated in Table 2, we compare GPT-3 with different fine-tuned models from Nan et al. (2022). Unlike the other datasets with short phrase answers, the goal of this dataset is to generate a complete long-form answer. Unlike WikiTableQuestion, the questions normally do not involve complex operations like max, min, compare, average, etc. The long-form answer is similar to the role of CoT. Therefore, we only applied 'direct generation' in this experiment. In terms of BLEU score (Papineni et al., 2002), GPT-3 is still a bit behind the fine-tuned T5-large. However, the BLEU score cannot reflect the faithfulness and correctness of the model generation. Thus, we follow Nan et al. (2021) to do human evaluation over the four aspects: (1) fluency (whether the generated sentence contains the linguistic error), (2) correctness (whether the generated sentence answers the question correctly), (3) faithfulness (whether the generated sentence is grounded on the input table), and (4) adequacy (whether the generated sentence is comprehensive enough to cover all the answers). We list our results in Table 3. Similarly, we also sample 100 model predictions and manually evaluate their quality and adopt binary scores for each

Type	Model	sacreBLEU
zero-shot	Pipeline (Nan et al., 2022)	9.16
FT	Pipeline (Nan et al., 2022)	11.00
FT	T5-small (Nan et al., 2022)	21.60
FT	T5-base (Nan et al., 2022)	28.14
FT	T5-large (Nan et al., 2022)	**30.54**
1-shot	GPT-3 Direct	26.88
2-shot	GPT-3 Direct	**27.02**

Table 2: Experimental Results on FetaQA. PT means pre-training and FT means fine-tuning.

Source	Fluency	Correct	Adequate	Faithful
Pipeline	85.2	25.4	23.6	23.6
T5-large	94.6	54.8	50.4	50.4
Human	95.0	**92.4**	**95.6**	**95.6**
GPT-3	**98.0**	84.0	78.0	90.0

Table 3: Human Evaluation Results on FetaQA.

example. As can be seen, GPT-3 can significantly outperform T5-large over all the aspects, i.e. more than 30% improvement over correctness, adequacy, and faithfulness. The evaluation indicates that the model output is almost on par with the average human performance on this dataset.

TabFact As demonstrated in Table 4, we compare GPT-3 against the other pre-trained and fine-tuned models including TAPAS (Eisenschlos et al., 2020), TAPEX (Liu et al., 2021), etc. We show that GPT-3 direct prediction is already getting a decent accuracy of 72%, which is slightly higher than Logic FactChecker (Zhong et al., 2020). When combined with CoT reasoning, the model accuracy increases to over 77%. Similar to before, we found that Codex can generate more accurate reasoning chains, thus achieving better accuracy of 78.8%, which is only 2% lower than pre-trained table understanding model TAPAS (Eisenschlos et al., 2020). The more intriguing property about LLM + CoT is that the intermediate rationale can be produced without any training. All the existing trained models do not have the capability to produce the intermediate reasoning steps due to the lack of annotation in the dataset.

FEVEROUS We demonstrate our results on FEVEROUS dev-set in Table 5 and compare different-sized UnifiedSKG models (built with T5). We found that GPT-3's performance with direct prediction is similar to UnifiedSKG-base. Similar to TabFact, we found that the model performance can be boosted with 'chain of thoughts' prompt-

Type	Model	Overall
FT	Chen et al. (2019)	65.1
FT	Zhong et al. (2020)	71.1
FT	Zhang et al. (2020)	73.2
FT	Yang et al. (2020)	74.4
FT	Lewis et al. (2020)	82.5
PT + FT	Eisenschlos et al. (2020)	81.0
PT + FT	Liu et al. (2021)	**84.2**
1-shot	GPT-3 Direct	72.0
2-shot	GPT-3 Direct	73.9
1-shot	GPT-3 CoT	75.5
2-shot	GPT-3 CoT	76.0
1-shot	GPT-3 CoT+SC	77.3
2-shot	Codex CoT	78.8

Table 4: Experimental Results on TabFact. PT means pre-training and FT means fine-tuning.

Type	Model	Dev Set
FT	Aly et al. (2021)	82.23
FT	UnifiedSKG-base (Xie et al., 2022)	75.05
FT	UnifiedSKG-large (Xie et al., 2022)	79.81
FT	UnifiedSKG-3B (Xie et al., 2022)	**82.40**
1-shot	GPT-3 Direct	74.20
2-shot	GPT-3 Direct	75.22
1-shot	GPT-3 CoT	75.70
2-shot	GPT-3 CoT	76.44
1-shot	GPT-3 CoT+SC	**77.22**

Table 5: Experimental Results on FEVEROUS. PT means pre-training and FT means fine-tuning.

ing. The best-performing model is roughly between UnifiedSKG-base and UnifiedSKG-large. Compared to TabFact, the model's overall performance is weaker mainly because the table structure in FEVEROUS is more irregular, containing lots of segments and subtables. Such structural difficulties pose great challenges to GPT-3.

Model Scaling We investigate the model scaling's impact on the final performance and plot our findings in Figure 3. On the WebTableQuestions dataset, we found that model size is essential for achieving the best performance. As can be seen, the 6.7B GPT-3 model is only achieving half of the performance of the 175B GPT-3 model. Similarly, on TabFact, we found that the smaller models with 6.7B or fewer parameters are almost getting random accuracy, which is even worse than QA tasks. This again suggests that LLMs' reasoning ability over web tables is emergent as the model scales up.

4.4 Case Study

We demonstrate a few examples in Figure 4 where GPT-3 makes correct predictions. In the first example, GPT-3 is able to first identify all the Belgian

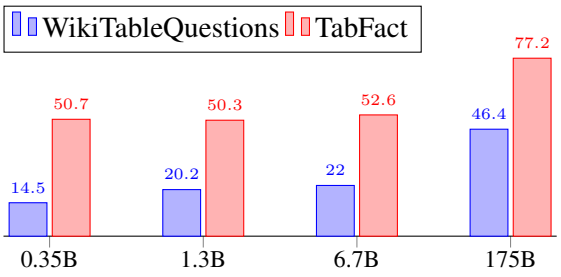

Figure 3: The model performance with respect to model size on WikiTableQuestions and TabFact.

riders from the table and then perform the addition of 3+3+1=7 precisely. In the second example, GPT-3 can identify the players with the position of 'd' and count the number correctly to refute a false claim. In the third example, we can see that GPT-3 is able to associate multiple blocks of information to generate a comprehensive long-form answer. The elicited 'chain of thoughts' in these examples are highly aligned with the underlying semantic forms. These findings suggest that LLMs like GPT-3 can provide high-quality explanations to justify their decision-making.

We also provide a few mistakes made by GPT-3 in Figure 5. In the first example, GPT-3 miscounts the 'number of countries above 1 billion box office' because it misidentifies 'world' also as a country. In the second example, GPT-3 misunderstood '2nd highest' as 'highest', which leads to prediction error. In the last example, GPT-3 misunderstands the semantics of the question and answers 'left office time' instead of 'took office time'. These examples show the typical errors of grounding the inputs to the wrong rows or columns of the table.

4.5 Analysis

Impact of Number of Shots First of all, we conduct an ablation study to understand the impact of a number of shots in the final performance. In order to control the budget, we only sample 200 samples from WikiTableQuestions, TabFact and FEVEROUS for this ablation study. As can be seen from Figure 7, GPT-3 is not quite sensitive to the number of provided demonstrations. Increasing from 1-shot to 2-shot can often benefit the model, however, increasing the shot number further does not yield more performance gain. We conjecture that instruct fine-tuning used in GPT-3 (Ouyang et al., 2022) can easily extrapolate the task meaning, thus, having a single demonstration is already enough for the model to understand the task.

Place	Rider	Country	Team	Points	Wins
1	Sylvain Geboers	Belgium	Suzuki	3066	3
2	Adolf Weil	Germany	Maico	2331	2
3	Torlief Hansen	Sweden	Husqvarna	2052	0
4	Roger De Coster	Belgium	Suzuki	1865	3
5	Joel Robert	Belgium	Suzuki	1730	1
6	Heikki Mikkola	Finland	Husqvarna	1680	2
7	Willy Bauer	Germany	Maico	1276	0
8	Gaston Rahier	Belgium	ČZ	1112	0
9	Pierre Karsmakers	Netherlands	Husqvarna	1110	0
10	Dave Bickers	United Kingdom	ČZ	1076	0

Question: total wins by belgian riders?

According to the table, Sylvain Geboers, Roger De and Joel Robert are Belgian riders and they have totally 3+3+1=7 wins. Therefore, the answer is 7.

name	pos	height	2012 club
Merrill Moses	gk	m	New York athletic club
Peter Varellas	d	m	the Olympic club
Peter Hudnut	cb	m	Los Angeles wp club
Jeff Powers	cf	m	Newport wp foundation
Adam Wright	d	m	New York athletic club
Shea Buckner	d	m	New York athletic club
Layne Beaubien	d	m	New York athletic club
Tony Azevedo	d	m	New York athletic club
Ryan Bailey	cf	m	Newport wp foundation
Chay Lapin	gk	m	long beach shore aquatics

Claim: more players have the position of gk than any other position.

There are 2 players with the position of gk and 5 players with the position of d. 2 is less than 5, therefore, the claim is false.

Year	Title	Role	Notes
1970	House of Dark Shadows.	Willie Loomis	-
1971	Daughters of Darkness	Stefan	-
1971	Night of Dark Shadows	Alex Jenkins	-
1976	A Small Town in Texas	Deputy Lenny Lutz	-
1978	Killer's Delight	Danny	-
1981	Pennies from Heaven	The Detective	-
1984	Racing with the Moon	Mr. Nash	-
1984	Impulse	Bob Russell	-
1984	Gimme an 'F'	Bucky Berkshire	-
1986	Native Son	Max	-
1991	The Dark Wind	Jake West	-
1993	Surf Ninjas	Mac	-
2016	The Job Interview	The Applicant	Short film

Question: What were John Karlen's roles in 1970 and 1978?

John Karlen's roles in 1970 were Willie Loomis in House of Dark Shadows and Stefan in Daughters of Darkness. His role in 1978 was Danny in Killer's Delight.

Figure 4: 'Correct' predictions from WikiTableQuestions, TabFact, and FetaQA datasets, where the 'blue' text are the outputs from the GPT-3, 'red' means the correct rows to reference.

Rank	Country	Box Office	Year	office from national
1	Canada/United States	$10.8 billion	2012	–
2	China	$3.6 billion	2013	59% (2013)
3	Japan	$1.88 billion	2013	61% (2013)
4	United Kingdom	$1.7 billion	2012	36.1% (2011)
5	France	$1.7 billion	2012	33.3% (2013)
6	South Korea	$1.47 billion	2013	59.7% (2013)
7	India	$1.4 billion	2012	–
8	Germany	$1.3 billion	2012	–
9	Russia	$1.2 billion	2012	–
10	Australia	$1.2 billion	2012	4.1% (2011)
11	Italy	$0.84 billion	2013	30% (2013)
12	Brazil	$0.72 billion	2013	17% (2013)
-	World	$34.7 billion	2012	–

Question: how many countries had at least $1 billion in box office?

Canada/United States, China, Japan, United Kingdom, France, South Korea, India, Germany and Russia, Australia and world all had at least $1 billion in box office. Therefore, the answer is 11.

no in series	title	directed by	original air date
47	knight of the drones (part 1)	sidney hayers	september 30 , 1984
48	knight of the drones (part 2)	sidney hayers	september 30 , 1984
49	the ice bandits	georg fenady	october 7 , 1984
...	...	...	...
61	buy out	jeffrey hayden	february 10 , 1985
62	knightlines	charles watson sanford	march 3 , 1985
63	the nineteenth hole	georg fenady	march 10 , 1985
64	knight & knerd	georg fenady	march 17 , 1985
65	ten wheel trouble	bob bralver	march 24 , 1985
66	knight in retreat	roy campanella ii	march 29 , 1985
67	knight strike	georg fenady	april 5 , 1985

Claim: knight in retreat is the 2nd highest no in series.

Kight in strike has no of series of 67, which is the highest no in series, therefore, the claim is false.

Took office	Left office	President	Party
7 Oct 1985	2 May 1988	Dick Ukeiwé	The Rally–UMP
2 May 1988	11 June 1989	Albert Etuve	The Rally–UMP
28 June 1989	7 July 1995	Simon Loueckhote	The Rally–UMP
31 July 1995	16 July 1997	Pierre Frogier	The Rally–UMP
16 July 1997	3 June 1998	Harold Martin	The Rally–UMP
3 June 1998	9 May 2004	Simon Loueckhote	The Rally–UMP
21 May 2004	3 July 2007	Harold Martin	L'Avenir Ensemble
31 July 2007	10 May 2009	Pierre Frogier	The Rally–UMP
22 May 2009	3 March 2011	Harold Martin	L'Avenir Ensemble
1 April 2011	29 August 2012	Roch Wamytan	Kanak and Socialist
29 Aug 2012	7 August 2013	Gérard Poadja	Caledonia Together
23 May 2014	17 July 2015	Gaël Yanno	Popular Caledonian
17 July 2015	30 July 2018	Thierry Santa	The Rally–UMP
30 July 2018	Incumbent	Gaël Yanno	Popular Caledonian

Question: On which dates did Harold Martin take office as President of the Congress of New Caledonia?

Harold Martin took office as President of the Congress of New Caledonia on 3 June 1998, 3 July 2007 and 3 March 2011.

Figure 5: 'Wrong' predictions from WikiTableQuestions, TabFact, and FetaQA datasets, where 'blue' text are the outputs from the GPT-3, 'red' means the region of the correct cell to reference, and 'green' means the reference trusted by GPT-3.

Quality Evaluation of Reasoning Chains We conduct a human evaluation to assess whether GPT-3 is making the correct prediction with the correct reasons. Specifically, we sample 100 reasoning paths from the correctly predicted examples and manually study whether these reasoning chains are grounded on the table or simply 'hallucination'. As can be seen from Figure 7, we found that around 90% of reasoning chains are faithful to the information in the table, and only less than 10% of the reasoning chains are hallucinated. Based on this

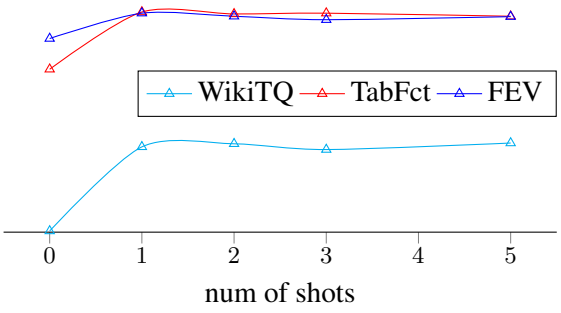

Figure 6: k-shot ablation study over WikiTableQuestions and TabFact and FEVEROUS.

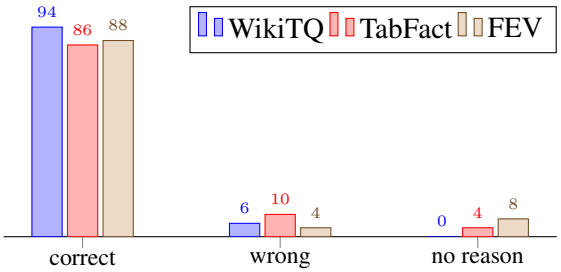

Figure 7: human evaluation of 'reasoning chains' in WikiTableQuestions, TabFact, and FEVEROUS.

evaluation, we believe that LLMs are not guessing the answers correctly by chance.

We believe these 'reasoning chains' are useful in many aspects: (1) the chains can provide a rationale to humans to justify the decision-making process. (2) one of the notorious annotation tasks is to annotate the 'underlying' semantic form for many NLP tasks, which require expertise for human annotators, on the other hand, the annotation cost is huge. Using GPT-3 to demonstrate useful natural language 'semantic forms' could potentially greatly lower the annotation burden of these tasks.

Impact of Table Size An important factor for model performance is the size of the table. Here we want to understand how relevant the model performance is w.r.t the input table length. We group the table token length into different groups like '0-100', '100-200', etc, and plot the group-wise accuracy for WikiTables and TabFact in Figure 8. As can be seen from the table, we found that GPT-3's performance is highly sensitive to the table size. As the table size grows, the accuracy almost decreases monotonically. After the table size exceeds 1000 tokens (e.g. 1500 word pieces), GPT-3's performance almost degrades to random guesses. This ablation study reveals one of the drawbacks of using LLMs for table reasoning. To further enhance LLMs' performance, we need to develop better methods to maintain more consistent performance across different-sized tables.

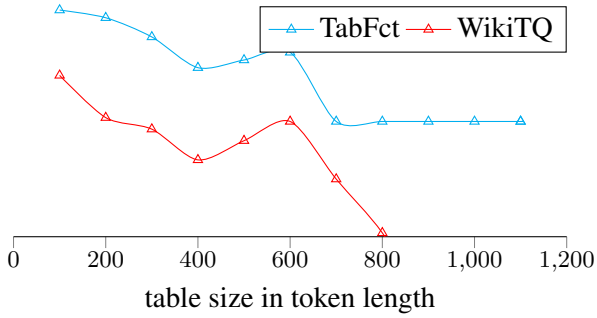

Figure 8: Model performance on WikiTableQuestions and TabFact w.r.t the input table size.

Discussions In this study, we investigate the possibilities of prompting LLMs to perform complex reasoning tasks over tables. However, we do not believe LLM prompting can replace the existing symbolic methods. LLMs have several favorable properties: (1) no annotation is needed, and (2) the functional coverage is broader than symbolic methods. However, LLM prompting exhibits unpredictable randomness and cannot generalize to large tables. In contrast, symbolic models are (1) agnostic to the table size, and (2) can reliably perform designed functions without much randomness. But they in general require a significant amount of annotated data to learn.

In conclusion, these two types of models are complementary to each other. To push the limit forward, we need to investigate how to combine the merits of these two types of methods. For example, the symbolic methods can perform certain operations to narrow down to a targeted region in the table, and then LLMs can be used to reason over the limited information.

5 Conclusion

In this paper, we investigate whether the current LLMs (GPT-3) can be directly utilized to perform table reasoning tasks. Surprisingly, though LLMs are not optimized for table-based tasks, we found these models highly competent in performing complex table reasoning tasks, especially when combined with 'chain of thoughts' prompting. We believe this study can open new possibilities for LLM application in table-related tasks to either directly predict the output or to serve as an auxiliary tool for annotating complex intermediate forms.

Limitations

Our approach has several limitations: (1) the proposed approach is still far from state-of-the-art performance, and there is still room for improve before it can be used as an alternative. (2) the method is still costly, we show that the model can only achieve superior performance when scaling up. Smaller-sized models are still weak at table reasoning. Therefore, we need to consider how to empower smaller models with such reasoning capabilities.

References

Rishabh Agarwal, Chen Liang, Dale Schuurmans, and Mohammad Norouzi. 2019. Learning to generalize from sparse and underspecified rewards. In *International conference on machine learning*, pages 130–140. PMLR.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and verification over unstructured and structured information (feverous) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W Cohen. 2020a. Open question answering over tables and text. In *International Conference on Learning Representations*.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020b. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942.

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. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021b. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Julian Eisenschlos, Syrine Krichene, and Thomas Mueller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Mueller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chen Liang, Mohammad Norouzi, Jonathan Berant, Quoc V Le, and Ni Lao. 2018. Memory augmented policy optimization for program synthesis and semantic parsing. *Advances in Neural Information Processing Systems*, 31.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. In *International Conference on Learning Representations*.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, et al. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2021. Dart: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2020. Multimodalqa: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Bailin Wang, Ivan Titov, and Mirella Lapata. 2019. Learning semantic parsers from denotations with latent structured alignments and abstract programs. 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 3774–3785.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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 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 preprint arXiv:2201.05966*.

Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Program enhanced fact verification with verbalization and graph attention network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7810–7825.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir R Radev, Richard Socher, and Caiming Xiong. 2021. Grappa: Grammar-augmented pre-training for table semantic parsing. In *ICLR*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.

Hongzhi Zhang, Yingyao Wang, Sirui Wang, Xuezhi Cao, Fuzheng Zhang, and Zhongyuan Wang. 2020. Table fact verification with structure-aware transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1629.

Yuchen Zhang, Panupong Pasupat, and Percy Liang. 2017. Macro grammars and holistic triggering for efficient semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

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

Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin. 2020. Logicalfactchecker: Leveraging logical operations for fact checking with graph module network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6065.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. 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 3277–3287.