Enhancing Tabular Reasoning with Pattern Exploiting Training

Abhilash Shankarampeta[1*], **Vivek Gupta**[2*†]
[1]IIT Guwahati; [2]University of Utah
sareddy53@gmail.com; vgupta@cs.utah.edu;

Abstract

Recent methods based on pre-trained language models have exhibited superior performance over tabular tasks (e.g. tabular NLI), despite showing inherent problems with reasoning over the tabular data (Gupta et al., 2021). In this work, we utilize Pattern-Exploiting Training (PET) to strengthen pre-existing knowledge and reasoning abilities of these tabular reasoning models'. Compared to current baselines, our upgraded model exhibits a superior understanding of knowledge facts and tabular reasoning. Additionally, we demonstrate that such models are more effective for underlying downstream tasks of tabular inference on InfoTabS. Furthermore, we also show our model's robustness against adversarial sets generated through various character and word level perturbations on InfoTabS.

Breakfast in America	
Released	29 March 1979
Recorded	May–December 1978
Studio	The Village Recorder in LA
Genre	Pop, art rock, soft rock
Length	46:06
Label	A&M
Producer	Peter Henderson, Supertramp

H1: Breakfast in America is a pop album with a duration less than 50 minutes.
H2: Peter Henderson produces only rock albums.
H3: Breakfast in America was released towards the end of 1979.
H4: Breakfast in America is recorded in California.
H5: Supertramp is an English band.
H6: The album was released on 29 March 1978.

Table 1: An example of tabular premise with from InfoTabS (Gupta et al., 2020).The hypotheses **H1, H4** is entailed by it, **H2, H5** is a neutral and **H3, H6** is a contradiction.

1 Introduction

Natural Language Inference is the problem of categorizing a hypothesis into an entailment or contradiction, or neutral based on the given premise (Dagan et al., 2013). Large language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b) have been applied to large datasets like SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), where they have shown performance equivalent to that of humans. In this work, we focus on the task of tabular reasoning through table inference on InfoTabS (Gupta et al., 2020).

The existing models are shown to be ineffective for reasoning over semi-structured data (Gupta et al., 2021). These models often ignore relevant rows and use spurious correlations in hypothesis or pre-training information for making inference (Neeraja et al., 2021; Poliak et al., 2018; Gururangan et al., 2018; Jain et al., 2021; Gupta et al., 2021). Due to existing biases in training data (Rajpurkar et al., 2018; Zhou and Bansal, 2020) with hypothesis having annotation artifacts (Gururangan et al., 2018), often models trained on such data lack generalizability and robustness (Glockner et al., 2018). Furthermore, the absence of comprehensiveness test sets hinders robust model evaluation. Thus, evaluating models based only on accuracy does not reflect their reliability and robustness (Ribeiro et al., 2020; Moradi and Samwald, 2021).

In this paper, we investigate if the current models' reason, i.e., can extract the right knowledge and correctly infer from that extracted knowledge. For example, to classify H1 (from table 1), a model needs to filter out the relevant rows, i.e., extract knowledge and perform appropriate reasoning using it. Some of the reasoning involved include numerical reasoning like count, sort, compare, arithmetic (H1: 46 < 50), commonsense knowledge (H3: December occurs at the end of the year), and factual knowledge (H4: LA is short for Los Angeles). Several studies have shown that while learning linguistic knowledge, these models also store factual and relational knowledge present in the training data (Petroni et al., 2019).

[*] Equal Contribution [†] Corresponding Author

To extract the information from LMs, many strategies based on probing classifiers (Hewitt and Liang, 2019; Voita and Titov, 2020, and others), attention (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), and prompting (Petroni et al., 2019; Shin et al., 2020, and others) have been proposed. We use prompting to extract knowledge since it does not adds additional parameters. To incorporate knowledge implicitly through the unstructured or semi-structured text (Xiong et al., 2020; Roberts et al., 2020; Eisenschlos et al., 2020) and to train the model on reasoning task, we reformulate tabular NLI task as cloze-style questions and perform gradient-based finetuning using a more supervised Pattern-Exploiting Training (Schick and Schütze, 2021a,b; Tam et al., 2021). Our PET approach also outperforms (+2.96 on α_1, +3.64 on α_2, +2.24 on α_3, see § 4) the existing methods on the downstream task of tabular reasoning on InfoTabS (Gupta et al., 2020). We also created challenging adversarial datasets to investigate the model's reliance on word overlap, understanding names, numbers, locations, and counterfactual statements. Our experiments show that our approach outperforms existing methods on adversarial test sets. Our contributions are following:

1. We propose a method for generating prompts for determining if current models can infer from knowledge.

2. We enhanced the model's reasoning by reformulating the tabular NLI tasks with Pattern-Exploiting Training (i.e., cloze style questions answering task).

3. Our experiments show that our proposed approach preserves knowledge and improves performance on downstream NLI tasks.

4. We also created adversarial datasets to assess the model's robustness. Our findings indicate that our approach enhances the model's robustness to the character and word-level perturbations.

The dataset and associated scripts, is available at `https://infoadapet.github.io/`.

2 Motivation

Tabular Reasoning: A wide range of reasoning skills such as arithmetic and commonsense are required to comprehend semi-structured data. Reasoning with this kind of data requires understanding the types of text in the cells, and information may be aggregated over numerous rows if necessary. To judge the H1, (see table 1) model needs to understand *duration* and *length* are the same in the context of the table, which is about a music album. Not only that, but numerical reasoning is also required to compare 46:06 mins is less than 50 mins. At the same time, the model should understand that the premise (table) is about a music album, so to classify the H1 model needs to understand the information present in 2 rows (*Genre, Length*) and perform numerical reasoning on top of that. For a hypothesis like *The album was produced by one person* which is a contradiction, the model needs to count from the row *Producer* and understand that the count is greater than one. Hence the hypothesis is a contradiction.

Knowledge and Reasoning: To reason about the H3 (table 1) model need to first extract the relevant row i.e. "*Released*" row from the table. Then it needs to compare the phrase "end of 1979" with the "*Released*" row value "29 March 1979". The model needs to perform temporal reasoning to know that year 1979 is correct. However, the month "March" is not the "end of the year," i.e., November or December as known from the commonsense knowledge. Similarly, classifying the H4 (table 1) model requires knowing that LA is short for Los Angeles, located in California. So to classify H3 and H4 apart from reasoning skills, there is a requirement for knowledge. While previous works tried to incorporate knowledge via pre-training (Eisenschlos et al., 2020; Neeraja et al., 2021), we integrate knowledge and reasoning ability simultaneously using Pattern Exploiting Training (Tam et al., 2021) in this work. This approach improves the existing knowledge and enhances reasoning compared to existing methods.

Robustness: In this work, we propose adversarial test sets to evaluate various aspects of understanding and reasoning. Using character and word level perturbations that might change the sense of the hypothesis, we evaluate various characteristics like Word Overlapping Bias, understanding of names, numbers, location, and counter-facts. For example, if H1 (table 1) is changed to *Breakfast in Wales is a pop album with a duration of less than 50 minutes.* now the label of H1 is changes from **entailment** to **neutral** since we do not know any information of *Breakfast in Wales* from table 1. Similarly, if we make the hypothesis H6 an entailment using a negation, i.e., to *The album was not released on 29 March 1978.* these small changes changed the entire meaning of the hypothesis. Ideally, a robust model with better reasoning ability should perform well on these adversarial sets, which is also observed with our method.

3 Our Approach

In this section we describe our approach for **a)** evaluating the knowledge used for reasoning **b)** enhancing reasoning using PET **c)** assessing the robustness of tabular reasoning task. We address the difficulties raised in section 1 and 2 and provide ways to alleviate them.

3.1 Evaluation of Knowledge

To assess the effect of pre-training on tabular reasoning and determine if the models can infer from knowledge. We evaluate factual and relational knowledge in the language model before and after training for the downstream task like reasoning. We evaluate knowledge by querying the model using "fill-in-the-blank" cloze statements (prompts). As gauging knowledge using prompts is limited by how the prompts are constructed. We used some simple techniques based on parts of speech to design these prompts. These prompts are generated using hypotheses from the α_1, and dev sets as these sets have similar distribution as the training data (Gupta et al., 2020). To study the effect of the premise, we also query the model in the presence of the premise. To do this we modify the input as *premise + prompt*.

Prompts for Factual Knowledge Evaluation As most factual knowledge is present in the proper nouns and numbers in a sentence, we randomly mask proper nouns or numbers in the hypothesis to generate a prompt and query the LM to fill the masked tokens in the prompt. For example *Duration of Breakfast in America is 46 minutes* (table 1), *Breakfast in America*, *46* are the factual information present in the sentence and they connected by *duration*. We randomly mask either *Breakfast in America* or *46* to make the prompt *Duration of Breakfast in America is <mask> minutes*.

There are also cases when a masked word is a number in the numeric form (e.g., 2), but the model predicted as "two" we solved this issue by converting the predicted word into its numeric form or vice versa if possible. *Breakfast in America is produced by <mask> producers. (<mask> = two)*

Prompts for Relational Factual Knowledge Evaluation. In addition to evaluating the model's factual knowledge, it is also crucial to evaluate its relational knowledge. For example, *Breakfast in America was <mask> towards the end of 1979. (<mask> = released)*. The model needs to understand that *Breakfast in America* is a music album to predict *released* instead of *eaten* which is highly probable. We also use WordNet (Miller, 1995) to find the synonyms of the masked word to check if the predicted word is in them.

3.2 Enhancing Reasoning with incorporating Knowledge

The issue of deducing inferences from tabular premises is similar to the typical NLI problem, except that the premises are tables rather than sentences in this case. When evaluating the reasoning skills, we use a variety of representations of the tabular premise (see A.1). We also study the effect of pretraining on an NLI task before training on InfoTabS.

Pattern-Exploiting Training. Using Pattern-Exploiting Training (PET) (Schick and Schütze, 2021a), NLU tasks are reformulated as cloze-style questions, and fine-tuning is performed using gradient-based methods. We use ADAPET (A Densely-supervised Approach to Pattern-Exploiting Training) (Tam et al., 2021), which increases supervision by separating the label token losses and applying a label-conditioned masked language modeling (MLM) objective to the entire original input.

The input to the language model is converted into a cloze-style form with the pattern *<premise> ? <mask>, <hypothesis>*. The model is tasked to predict the masked word from the vocabulary. The model computes each token's probability as a softmax normalized overall tokens, allowing the logits of all vocabulary tokens to impact each likelihood, similar to the regular MLM objective. While in PET, the masked word is forced to predict from the output space *{Yes, Maybe, No}* which mapped to labels *{Entailment, Neutral, Contradiction}*. As a result, there will never be a gradient signal for non-label tokens.

Inverting the query to the model to *"In light of the answer, what is the appropriate context?"* from *"What is the appropriate label based on the input?"*

Perturbation	Original text	Perturbed text
Character	Peter Henderson produces only rock albums	Peter Henbgderson produces only rock albsums Peter Hendersno produces only rokc albums Pter Henderson produces onl rock abus Petqr Henkerson prgduces only rock alocms
Location	Breakfast in America is recorded in California Breakfast in America is recorded in USA Breakfast in America is by an English rock band.	Breakfast in America is recorded in Florida. Breakfast in America is recorded in Syria. Breakfast in America is by an Mexican rock band.
Name	Peter Henderson produces only rock albums	John Doe produces only rock albums
Numbers	The album was released on 29 March 1978.	The album was released on 29 March 346. The album was released on 1 March 1978.
Negation	The genres of the album are pop and rock.	The genres of the album are not pop and rock.
Paraphrase	The album was recorded in the last half of 1979.	In the second part of 1979, the album was recorded.

Table 2: Examples of various perturbations used to generate the adversarial test sets based on table 1

label conditioned mask language modeling is introduced by randomly masking out tokens from the context. If the label is "true," during training, the model is obligated to predict the original token; however, if the label is incorrect, the model is forced to ignore the original token.

3.3 Robustness using Adversarial Data

We use a variety of character-level and word-level perturbations on hypotheses to replicate circumstances in which the input is somewhat noisy or diverges from the training data distribution. These adversarial sets will test the model's reliance on word overlap, understanding numbers, and counterfactual statements. We use TextAttack (Morris et al., 2020), NLP Checklist (Ribeiro et al., 2020), QuillBot [1] for generating the adversarial data. (Refer tables 2 and 10 for examples). The perturbations include:

Character-level perturbation: A random word is selected, then perturbations are applied to its characters. We used perturbations like inserting random characters, swapping characters in the selected word, deleting a randomly selected character from the word, and replacing a randomly selected character. This perturbation does not affect the label of the hypothesis as it does not change the meaning of the sentence.

Location perturbation: This word-level perturbation changes the recognized locations (countries, cities, and nationalities) of a sentence to another location given in the location map. The NER model detects the location for a given sentence, swapping it to another location sampled from a dictionary. Here cities are swapped with other cities, not with countries same goes for nationalities and countries. This perturbation changes the entailed sentences into contradictions but does not affect the labels of neutral and contradictions.

Name perturbation: The NER model detects the names of persons present in the sentence. A person's name is replaced from a list of names randomly. This perturbation alters the label of every hypothesis into a neutral because the perturbed hypothesis and premise mention different persons.

Perturbing Numbers: This transformation recognizes numbers (numeric and alphabetic form) in the sentence and returns sentences with altered numbers. This perturbation changes the entailed sentences into contradictions but does not affect the labels of neutral and contradictions. Contradictory statements remain contradictory because it is implausible that a randomly sampled number will be the actual number in the premise, making the hypothesis entailed.

Negation: This perturbation generates counterfactual sentences by negating the given sentence. This perturbation transforms entailment into a contradiction and vice versa, but neutrals remain the same.

Paraphrasing: This transformation paraphrases the given sentences without the loss of meaning using QuillBot.This perturbation does not affect the label of the hypothesis as it does not change the sense of the hypothesis.

Composition of Perturbations: We also perturbed sentences by applying various distinct perturbations sequentially. For example, in **num+para+name** we perturbed a sentence *Supertramp, produced an album that was less than 13 minutes long*, with premise table 1 to *Supertramp, produced an album that was less than 13 minutes*

[1] https://quillbot.com

long (number) then *Supertramp released an album with a running time of less than 13 minutes.* (paraphrase) then *James released an album with a running time of less than 13 minutes* (name). For examples refer to table 10.

4 Experiments and Analysis

Our experiments answers the following questions:

1. **RQ1:** Can the large language model reason with existing knowledge? Does our adaptive training approach enhance model reasoning ability?

2. **RQ2:** Does fine-tuning on downstream tasks benefit model reasoning? Can our approach of adaptive training benefit the model by incorporating knowledge for better reasoning on downstream tabular NLI task on InfoTabS?

3. **RQ3:** Lastly, can models trained with our adaptive approach be more robust to spurious correlations? Does our approach enhance the model's ability to effectively comprehend semantic and syntactic alternations?

Dataset: Our experiments use *InfoTabS*, a tabular inference dataset introduced by (Gupta et al., 2020). The dataset is diverse in terms of the tables and keys it contains, dependent on prior knowledge and common sense. Given the premise (table), every hypothesis in the dataset is labeled as either an Entailment (E), Contradiction (C), or Neutral (N) statement. In addition to the conventional development set and test set (referred to as α_1), an adversarial test set (α_2) lexically equivalent to α_1 but with minor changes in the hypotheses to flip the entail-contradict label and a zero-shot cross-domain test set (α_3) containing large tables from other domains that are not in the training set are used for evaluation. For all of our experiments, we use the accuracy of classifying the labels as our primary metric for evaluation. The domain of tables in both training set and α_1,α_2 are similar. However, the training and fine-tuning tables are exclusive. The test sets $\alpha_1, \alpha_2, \alpha_3$ has 1800 table-hypothesis pairs each with 200 unique tables. Table 3 depict the sample size of adversarial sets.

Models: We use the pre-trained RoBERTa-Large (RoBERTa$_L$) (Liu et al., 2019b) language model from Hugging Face (Wolf et al., 2020) for all of our investigations. We employ various configurations of language models to assess knowledge in two different cases. These configurations include RoBERTa$_L$, RoBERTa$_L$ finetuned on InfoTabS (RoBERTa$_L$+CLS) i.e, Knowledge InfoTabS (KI) (Neeraja et al., 2021), RoBERTa$_L$ trained for tabular inference using PET (ADAPET), and finetuning InfoTabS on ADAPET (ADAPET+CLS). Here we define finetuning as training a classifier head (CLS). We also investigate the effect of NLI pre-training using RoBERTa$_L$ pretrained on MNLI (Williams et al., 2018), and mixed dataset (mixNLI) containing ANLI+MNLI+SNLI+FeverNLI [2] (Nie et al., 2020; Bowman et al., 2015; Nie et al., 2019). For a fair comparison with the Knowledge InfoTabS (KI) baseline, we only consider the BPR and DRR baseline without any implicit and explicit knowledge addition. All models are trained on 16538 table-hypothesis pairs (1740 unique tables) for ten epochs with a learning rate of 1e-5.

Peturb Type	Size	Peturb Type	Size
character	1800	negation+char	1726
location	1229	negation+name	1677
name	1646	number+char	837
negation	1726	number+name	776
number	837	number+negation	817
paraphrase	1800	num+paraphrase	837
num+para+name	776	paraphrase+name	1721

Table 3: Number of examples for each perturbation type in the adversarial set.

Table Representation: We explored two ways to represent table (a.) *Table as paragraph* which uses Better Paragraph Representation for table representation, (b.) and *Distracting Row Removal* which prune table based on the similarity between hypothesis and tables rows. We explore pruning of top 4 (DRR@4) and top 8(DRR@4) rows for our experiments. Both representation methods are adapted from Neeraja et al. (2021). For more details on table representation, refer to Appendix A.1.

4.1 Results and Analysis

Below we describe the results of a) extracting and evaluating knowledge, b) Tabular NLI on InfoTabS, c) robustness to adversarial data

4.1.1 Models Knowledge Evaluation

To answer RQ1, we evaluate the knowledge in the presence and absence of premise with Entail and Contradictory hypotheses as they are judged based on the information tables. In contrast, for Neutral,

[2] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

Type	Input	RoBERTa$_L$		ADAPET	
		w/o	+CLS	w/o	+CLS
Factual	only E/C	39.6	27.9	38.9	31.3
	prem + E/C	60.3	29.4	57.6	43.8
	only E	39.1	27.8	38.8	32.8
	prem + E	61.7	29.1	61.9	46.1
Relational	only E/C	43.2	24.2	46.1	31.8
	prem + E/C	50.4	21.6	51.4	34.2
	only E	45.8	25.4	50.3	33.2
	prem + E	54.2	20.7	56.8	41.4

Table 4: Top 1 Accuracy of Factual & Relational Knowledge Evaluation on DRR@4.(w/o - no CLS, RoBERTa$_L$+CLS - Knowledge InfoTabS

it is not always the case.

Type	Input	RoBERTa$_L$		ADAPET	
		w/o	+CLS	w/o	+CLS
Factual	only E/C	56.2	41.1	55.8	49.9
	prem + E/C	74.3	43.3	73.4	60.4
	only E	56.1	42.1	55.6	50.4
	prem + E	74.2	46.1	76.2	63.9
Relational	only E/C	60.9	47.7	62.9	55.1
	prem + E/C	67.5	47.8	68.2	60.3
	only E	60.5	49.2	65.4	56
	prem + E	67.8	47.1	69.1	64.7

Table 5: Top 5 Accuracy of Factual & Relational Knowledge Evaluation on DRR@4. (w/o - no CLS, RoBERTa$_L$+CLS - Knowledge InfoTabS

In all the settings (tables 4 and 5) with and without premise, our model outperformed Knowledge InfoTabS (Neeraja et al., 2021), at the same time our model can infer from the knowledge which can be seen from the improved performance on Entailed data in the presence of premise. From table 4, it is clear that our approach performance on relational knowledge evaluation is more than double of Knowledge InfoTabS in every setting. Our approach also outperforms Knowledge InfoTabS by a significant margin on factual knowledge evaluation. Even training a classifier on top of ADAPET outperforms Knowledge InfoTabS. We evaluated on contradiction hypothesis to assess if the model can rightly identify false claims despite having correct entity types.

In almost all the settings, our approach performs almost comparable to RoBERTa$_L$, and it even outperforms RoBERTa$_L$ in only Entail, and Premise+ Entail settings. Training a classifier on top of RoBERTa$_L$ decreases the performance knowledge evaluation but training a classifier head on top of ADAPET still tops RoBERTa$_L$+CLS (KI).

4.1.2 Knowledge Incorporation and Reasoning on InfoTabS

To answer RQ2, we experimented with various premise representations of tables as paragraphs (BPR, DRR@4, DRR@8) (see table 6). We observed that Roberta-Large, with ADAPET, improves performance in all premise representations except for α_3 with BPR compared to Knowledge InfoTabS due to an increased number of keys in the tables (13.1 per table in α_3 when compared to 8.8 per table in α_1 and α_2).

With ADAPET we are also able to improve performance using linearized table (see table 8) compared to Gupta et al. (2020) (+1.04 in α_1, +0.58 in α_2, +0.69 in α_3). Our ADAPET (no pre-training) tops Knowledge InfoTabS (Neeraja et al., 2021) in every premise representation and test split. +0.72 in α_1, +0.68 in α_2, +1.52 in α_3 with DRR@4.

We noticed that the DRR@8 representation of the table outperforms every other representation, especially in α_3 due to the removal of the irrelevant rows (+2.46 over BPR, +1.46 over DRR@4). The zero-shot test set α_3 which has a significant proportion of unseen keys (different domain tables) when compared to other test sets (number of unique keys intersection with train is 312, 273, 94 for α_1, α_2 and α_3 respectively) has seen a substantial improvement with the use of NLI pre-trained model. When compared to ADAPET (no pretraining), there has been an improvement of +4 units (no CLS) and +2.61 units (with CLS).

We finetune our model with MNLI and mixNLI to improve performance due to the model's exposure to diverse data (Andreas, 2020; Pruksachatkun et al., 2020). We notice that utilizing an NLI-pretrained model improves the model's reasoning and generalization across all three test sets and the adversarial sets used to assess robustness. We also observed that pre-training in more diverse data helps in the performance. Models which are pre-trained on mixNLI[2] outperformed MNLI pre-trained in almost every setting (+1.3 in α_1, +1.47 in α_2, +2.1 in α_3 with DRR@8)

4.1.3 Robustness on Adversarial Data

To answer RQ3, we evaluate our model on several challenging adversarial sets. The adversarial test sets generated using various character-level and word-level perturbations are also tested with BPR, DRR@4, and DRR@8 table representations (see table 7). To generate these sets, we applied perturbations on dev, and α_1 sets as the distribution of

Test Splits	Premise	KI	ADAPET			ADAPET+CLS		
			w/o	+mixNLI	+MNLI	w/o	+mixNLI	+MNLI
Dev	BPR	76.83	78.1	**79.8**	79.1	78.72	79.22	78.55
	DRR@4	76.39	76.4	**78.2**	77.2	76.27	78.16	77.5
	DRR@8	76.42	77.3	77.7	77.3	77.55	**79.38**	78.83
α_1	BPR	75.29	78.1	77.4	77.4	77.38	78	**78.38**
	DRR@4	75.78	76.5	**79.4**	79.2	76.44	78.22	78.11
	DRR@8	76.56	78.1	**80.2**	78.9	78.27	78.5	78.66
α_2	BPR	66.5	67.9	**72.9**	70.6	67.5	72.33	70.61
	DRR@4	67.22	67.9	70.3	68.8	68.56	**70.89**	69.72
	DRR@8	68.11	69.1	72.5	71.2	69.38	**72.67**	69.88
α_3	BPR	64.26	63.7	66.3	64.7	64.88	**68.44**	65.11
	DRR@4	64.88	66.4	68.8	67.2	65.57	**69.44**	67.11
	DRR@8	68.66	66.9	**70.9**	68.7	67.44	70.05	68

Table 6: Reasoning results on InfoTabs comparing KI, ADAPET, ADAPET+CLS (without pre-training (w/o), with mixNLI, MNLI pre-training). Note that results for BPR, DRR@4 with KI are from (Neeraja et al., 2021)

these sets are similar to the training set.

We see the max improvement of ADAPET in the Negation (+4.4); this implies our model counterfactual statements. In the challenge set with *number+paraphrase* all the ADAPET-based models outperformed Knowledge InfoTabS by 2x times. We observed that training a classifier head on top of ADAPET performed better with the adversarial sets involving multiple perturbations. We also observed that using NLI pre-training also helps significantly improve the robustness.

Except for the perturbations involving names, our method ADAPET (no pre-training) outperforms Knowledge InfoTabS (KI). With the use of mixNLI and MNLI pre-trained weights, the performance of ADAPET based models improved significantly compared to those without pre-training, even outperforming Knowledge InfoTabs. From table 7 it is also clear that with hypotheses involving multiple perturbations, KI tends to perform poorly while the ADAPET based model outperformed.

Some of these test sets are challenging as we can see their accuracy is below random guessing. Mainly when Compositional Perturbations are used, it affects the performance of all models. Performance on these sets is far behind the corresponding model's performance on dev, α_1 sets. Improving the performance of these sets is vital (Section 5).

5 Discussion

What did we learn? Using ADAPET, we have shown that we can preserve and incorporate knowledge. ADAPET training also enhances model reasoning ability and hence downstream performance on tabular inference tasks. Reformulating the NLI task as an MLM task helped incorporate the knowledge from premise tables into the Language Models (LM). Similar observation is also observed in prior works Xiong et al. (2020); Sun et al. (2019) where MLM is utilized to incorporate external knowledge. Many studies Gupta et al. (2021); Lewis et al. (2021) have also shown that the LM leverage spurious patterns to solve reasoning task. From our adversarial sets and analysis, we have learned that our approach is much more robust than Knowledge InfoTabS to the various character and word level perturbations.

Why table as a paragraph? A massive data corpus is used to pre-train the large language models. In contrast to semi-structured data, the bulk of pre-training data is unstructured. These models should, of course, perform better on unstructured data and struggle with semi-structured data. Tables in InfoTabS (Gupta et al., 2020) are semi-structured in nature. These tables do not explicitly state the relationship between the keys and values, and they can also have variable schemas. The album's overall duration is 46:06 minutes, according to the row with key Length and value 46:06. It is difficult to comprehend implicitly that "Length" refers to time length in minutes. Because of the absence of implicit information, a simple table linearization will not be sufficient. Gupta et al. (2020); Neeraja et al. (2021) experimented with various forms of table representations. They found that representing tables as paragraphs gave better results and can also leverage the advantage of pre-trained models

Perturb	**KI**	**ADAPET**			**ADAPET+CLS**		
		w/o	**+mixNLI**	**+MNLI**	**w/o**	**+mixNLI**	**+MNLI**
num+para+name	13.04	10.1	11.7	10.1	11.7	**16.62**	13.55
number+name	15.72	14.6	14	13.2	15.6	**18.94**	15.85
negation+name	19.08	16.1	**20**	11.6	14.43	14.37	12.1
num+paraphrase	27.46	**59.5**	58.4	57.3	52.5	56.63	54.95
paraphrase+name	30.79	22.6	28.3	24.9	27.01	**30.85**	27.71
name	32.7	24.7	31.1	28	28.9	**33.44**	30.69
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33
number+negation	36.13	42.7	**53.2**	28.3	37.91	37.75	24.04
negation+char	39.39	41.4	**47.6**	40.1	42.9	42.06	40.85
negation	53.7	58.1	**64.8**	56.1	57.6	59.15	53.88
number+char	54.43	58.8	57.1	**60.3**	55.79	57.1	59.28
number	56.1	**57.8**	**57.8**	57	52.44	55.79	54.6
character	63.05	62.8	65.9	64.4	64.05	66.05	**66.83**
location	67.6	**70**	67.7	69.1	69.81	67.4	65.98
paraphrase	70.56	72.3	**73.8**	73.4	71.6	72.66	72.3
dev	76.83	78.1	**79.8**	79.1	78.72	79.22	78.55
α_1	76.56	78.1	**80.2**	78.9	78.27	78.5	78.66

Table 7: Adversial Reasoning results on DRR@8 InfoTabs comparing KI (Neeraja et al., 2021), ADAPET, ADAPET+CLS (without pre-training (w/o), with mixNLI, MNLI pre-training). Rows in the tables are sorted in ascending order w.r.t KI performance.

datasets like MNLI for even better performance.

Why NLI task as cloze-style questions? While Gururangan et al. (2018) showed MLM pre-training with unlabeled target data could further improve the performance on downstream tasks. Chiang (2021) also showed that using MLM pre-training makes models robust to lexicon-level spurious features. Wei et al. (2021) presented a methodology for analysis that connects the pre-training and downstream tasks to an underlying latent variable generative text model. They observed that prompt tuning achieves downstream assurances with less stringent non-degeneracy constraints than head tuning. By reformulating the NLI task as cloze style questions, we can use label conditioned MLM with prompt tuning, which resulted in a better performance on tabular reasoning on InfoTabs.

Future Directions Based on our observations and discussions, we have identified the following potential future directions:

(a.) *Designing better prompts for knowledge evaluation*: Our current prompts treat entail and contradictory statements as the same while evaluating knowledge. In the presence of the premise, table 1 masking *Breakfast in America* in H3 and using that as an input model will predict Breakfast in America even though the hypothesis is a contradiction. We want to work on developing prompts label conditioned evaluation based on existing work on prompt engineering. (Shin et al., 2020; Liu et al., 2021).

(b.) *Improving Robustness:* While our models' performance on the challenging adversarial test sets is lower than benchmarks on InfoTabs, we do not know its reason. The created test sets may be challenging because they focus on phenomena that existing models cannot capture or exploit blind spots in a model's training set. Following the ideas of Inoculation by Fine-Tuning (Liu et al., 2019a) we want to improve and assess the reasons behind the results in table 7.

Furthermore, it will be interesting to extend the adaptive training approach to tabular reasoning tasks such as TabFact (Chen et al., 2020a), WikiTableQuestions (Pasupat and Liang, 2015), and others. Ablation and significant studies should be examined to understand the causes for performance gain with adaptive training.

6 Related Work

Tabular Reasoning: There has been a slew of papers published recently that investigate a variety of natural language processing problems using semi-structured table data. Tabular NLI (Gupta et al., 2020; Neeraja et al., 2021), fact verification (Chen et al., 2020a; Zhang et al., 2020), question answering (Pasupat and Liang, 2015; Krishnamurthy et al., 2017; Abbas et al., 2016; Sun et al., 2016; Chen et al., 2020b; Oguz et al., 2020; Lin et al., 2020; Zayats et al., 2021; Chen et al., 2021a, and others), and text generation from tables (Parikh et al., 2020;

Nan et al., 2021; Chen et al., 2021b; Yoran et al., 2021, and others) are some examples.

Several recent studies have also presented solutions for encoding Wikipedia relational tables, parallel to InfoTabs (Gupta et al., 2020) data set. Examples of such works are TAPAS(Herzig et al., 2020), TaBERT (Yin et al., 2020), TabStruc (Zhang et al., 2020), TABBIE (Iida et al., 2021), TabGCN (Pramanick and Bhattacharya, 2021) and RCI (Glass et al., 2021), amongst others. Works suchs as (Yu et al., 2018, 2021; Eisenschlos et al., 2020; Neeraja et al., 2021; Müller et al., 2021, and others) investigate the enhancement of tabular inference by pre-training.

Knowledge Evaluation: Many methods have been proposed to extract and evaluate knowledge from LMs. Some of the methods include probing classifiers (Hewitt and Liang, 2019; Voita and Titov, 2020; Hou et al., 2022, and others), attention visualization (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), and prompting (Petroni et al., 2019; Shin et al., 2020; Jiang et al., 2020). Many works have been published to study and create the prompts (Shin et al., 2020; Liu et al., 2021; Miller, 1995; Qin and Eisner, 2021, and others).

Knowledge Incorporation: Although the pretrained language models benefit from the large-scale corpus, they are limited by implicit knowledge representation. Various works have been proposed to integrate knowledge into the LMs using pretrained entity embeddings (Zhang et al., 2019; Peters et al., 2019, and others), external memory (Logan et al., 2019; Khandelwal et al., 2020; Lu et al., 2021), unstructured text (Xiong et al., 2020; Sun et al., 2019).

Robustness: Many works proposed ways to evaluate robustness to noise, fairness, consistency, explanation, error analysis, and adversarial perturbations to test the model's robustness and reliability. Ribeiro et al. (2020) provides a framework for testing NLP models inspired by software engineering. Moradi and Samwald (2021) introduces a perturbation architecture for textual inputs that includes a variety of character- and word-level systematic perturbations to replicate the many sorts of noise that an NLP system may encounter in real-world use scenarios. Goel et al. (2021) proposed a toolkit to identify challenges with evaluating NLP systems and assess them on sub-populations, transformations, evaluation sets, and adversarial attacks.

7 Conclusion

We introduced simple and effective prompts to evaluate knowledge and investigated finetuning and prompt-based methods for improving reasoning on tabular data. We also studied the effect of various table representations (BPR, DRR). Furthermore, we also constructed challenging test sets to assess the model's robustness to names, numbers, location, and others. With the help of label conditioned masking and reformulating NLI as cloze task questions, our approach outperforms previous baselines on the downstream tabular inference task. In conclusion, our proposed approach demonstrates the capability of extracting information from tables and reasoning over them.

Acknowledgement

We thank members of the Utah NLP group for their valuable insights and suggestions at various project stages and reviewers for their helpful comments. Additionally, we appreciate the inputs provided by Vivek Srikumar and Ellen Riloff. Vivek Gupta acknowledges support from Bloomberg's Data Science Ph.D. Fellowship.

References

Faheem Abbas, Muhammad Kamran Malik, Muhammad Umair Rashid, and Rizwan Zafar. 2016. Wikiqa — a question answering system on wikipedia using freebase, dbpedia and infobox. In *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, pages 185–193.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Qianglong Chen, Feng Ji, Xiangji Zeng, Feng-Lin Li, Ji Zhang, Haiqing Chen, and Yin Zhang. 2021a. KACE: Generating knowledge aware contrastive explanations for natural language inference. 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 2516–2527, Online. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021b. Open question answering over tables and text. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. 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, Online. Association for Computational Linguistics.

Ting-Rui Chiang. 2021. On a benefit of mask language modeling: Robustness to simplicity bias. *CoRR*, abs/2110.05301.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Michael Glass, Mustafa Canim, Alfio Gliozzo, Saneem Chemmengath, Vishwajeet Kumar, Rishav Chakravarti, Avi Sil, Feifei Pan, Samarth Bharadwaj, and Nicolas Rodolfo Fauceglia. 2021. Capturing row and column semantics in transformer based question answering over tables. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1212–1224, Online. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Vivek Gupta, Riyaz A. Bhat, Atreya Ghosal, Manish Srivastava, Maneesh Singh, and Vivek Srikumar. 2021. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *CoRR*, abs/2108.00578.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, 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, Online. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Yifan Hou, Guoji Fu, and Mrinmaya Sachan. 2022. Understanding knowledge integration in language models with graph convolutions. *CoRR*, abs/2202.00964.

Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. 2021. TABBIE: Pretrained representations of tabular data. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3446–3456, Online. Association for Computational Linguistics.

Nupur Jain, Vivek Gupta, Anshul Rai, and Gaurav Kumar. 2021. TabPert : An effective platform for tabular perturbation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 350–360, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. 2020. Bridging textual and tabular data for cross-domain text-to-SQL semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4870–4888, Online. Association for Computational Linguistics.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019a. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Yinquan Lu, Haonan Lu, Guirong Fu, and Qun Liu. 2021. KELM: knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. *CoRR*, abs/2109.04223.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Thomas Müller, Julian Eisenschlos, and Syrine Krichene. 2021. TAPAS at SemEval-2021 task 9: Reasoning over tables with intermediate pre-training. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 423–430, Online. Association for Computational Linguistics.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 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, Online. Association for Computational Linguistics.

J. Neeraja, Vivek Gupta, and Vivek Srikumar. 2021. Incorporating external knowledge to enhance tabular reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association* *for Computational Linguistics: Human Language Technologies*, pages 2799–2809, Online. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Sejr Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2020. Unified open-domain question answering with structured and unstructured knowledge. *CoRR*, abs/2012.14610.

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, Online. Association for Computational Linguistics.

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, Beijing, China. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. 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, Hong Kong, China. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Aniket Pramanick and Indrajit Bhattacharya. 2021. Joint learning of representations for web-tables, entities and types using graph convolutional network. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1197–1206, Online. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 771–782, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Colin Wei, Sang Michael Xie, and Tengyu Ma. 2021. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning. *CoRR*, abs/2106.09226.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Alignment over heterogeneous embeddings for question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2681–2691, Minneapolis, Minnesota. Association for Computational Linguistics.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4514–4525, Online. 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Ori Yoran, Alon Talmor, and Jonathan Berant. 2021. Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills. *CoRR*, abs/2107.07261.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 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, Brussels, Belgium. Association for Computational Linguistics.

Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. 2021. Representations for question answering from documents with tables and text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2895–2906, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.

A Appendix

A.1 Table Representation

We explored two ways to represent table as follows:

- *Premise as a paragraph:* Instead of using a universal template like " The key of $title$ is $value$. Following (Neeraja et al., 2021), we use Better Paragraph Representation (BPR) templates based on table categories and keys associated with entity types. In reference to *Breakfast in America* (table 1), the row "**Released**: *29 March 1979*", is transformed into "The *released* of *Breakfast in America* is *29 March 1979*." using universal template. "*Breakfast in America* was *released* on *29 March 1979*." using BPR.

- *Premise as a Linearized Table:* In accordance with (Chen et al., 2020a), we describe tables as a series of "key : value" tokens. A comma is used to separate multiple values for the same key from one another, while a semi-colon is used to separate rows.

- *Table Pruning:* For a particular hypothesis, not all of the entries in the premise table are essential. Sometimes, the length entire table with the hypothesis as input might be higher than the specified input length of the language model. Inspired from (Neeraja et al., 2021), we used alignment methods used in (Yadav et al., 2019, 2020) to remove distracting rows (DRR). By choosing the top 4 rows, we observed that some vital rows are missing for some examples, making the model detect them as neutral, especially in out-of-domain test sets like α_3. For evaluation, we use the top 4 and top 8 relevant rows from DRR (DRR@4 and DRR@8, respectively).

A.2 Results with premise as a Linearized Table

We experimented with premise as a linearized table and compared our results with Gupta et al. (2020)

Test Splits	Gupta et al. (2020)	Ours
Dev	**77.61**	76.7
α_1	75.06	**76.1**
α_2	69.02	**69.6**
α_3	64.61	**65.3**

Table 8: Results on Linearized Table comparing Gupta et al. (2020) and our approach (ADAPET)

A.3 Reasoning only on Entailed and Contradictory hypothesis

We also study the classification of Entailed and Contradictory hypotheses when the model is trained and tested on the data without any Neutral hypotheses.

Test Splits	KI	ADAPET		
	DRR@4	BPR	DRR@4	DRR@8
Dev	81.5	83.5	**84.3**	82.8
α_1	80.25	83.8	**84.3**	**84.3**
α_2	64.66	65.9	66.9	**67.7**
α_3	76	75.1	**78.5**	77.4

Table 9: Results on Entail vs Contradiction.

A.4 Robustness on Adversarial Data

We also evaluate robustness with premise representation. In tables 11 and 12 we show the performance of the model on the adversarial tests which are trained and testes with BPR, DRR@4 representations of premise.

Perturbation	Original text	Perturbed text
neg+char	The genres of the album are pop and rock.	The gejnres of the alzum are not pbp and rock.
neg+name	Peter Henderson's album was recorded in 1979.	John Doe's album was not recorded in 1979.
num+char	The album was recorded in 1979.	The album was recqorded in the last hplf of 459.
num+name	Peter Henderson's album was recorded in 1979.	John Doe's album was recorded in 731.
num+neg	The album was released on 29 March 1978.	The album was not released on 29 March 346.
num+para	The album was recorded in 1979.	In the second part of 1278, the album was recorded.
para+name	Peter Henderson produces only rock albums.	Only rock albums are produced by John Doe.
num+para+name	Peter Henderson's album was recorded in 1979.	The album by John Doe was recorded in 3147.

Table 10: More examples of various perturbations used to generate the adversarial test sets based on table 1

Perturb	**KI**	**ADAPET**			**ADAPET+CLS**		
		w/o	**+mixNLI**	**+MNLI**	**w/o**	**+mixNLI**	**+MNLI**
negation+name	11.74	10.4	**21.1**	15.6	17.35	13.89	12.93
num+para+name	14.06	10.6	**20.7**	12	17.13	14.83	13.04
number+name	17.26	12.5	**20.9**	14.8	18.42	18.42	16.88
paraphrase+name	33	25.8	**37.6**	31.5	31.2	32.1	31.3
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33
name	34.6	26.5	**36.4**	33.4	32.41	33.96	33.2
negation+char	37.71	38.5	**47.8**	41.3	43.56	41.25	40.49
number+negation	38.36	30.2	**54.8**	30.1	37.69	38.7	26.06
negation	48.9	54.2	**65.4**	55.3	58.27	58.45	55.6
number	56.63	**62.3**	51.9	56	55.43	53.52	56.1
num+paraphrase	56.98	**62.3**	49.7	54.5	55.55	52.26	55.19
number+char	59.11	**66.1**	45.1	55.6	55.9	52.46	60.2
character	61.5	64.1	64.4	66.1	64.9	**66.61**	65.94
location	68.2	**72.4**	68.1	70.1	69.08	66.47	69.48
paraphrase	68.44	72.3	**72.6**	72.3	72.05	71.7	**72.66**
dev	76.42	77.3	77.7	77.3	77.55	**79.38**	78.83
α_1	75.29	78.1	77.4	77.4	77.38	78	**78.38**

Table 11: Adversial Reasoning results on BPR InfoTabs comparing KI (Neeraja et al., 2021), ADAPET, ADAPET+CLS (without pre-training (w/o), with mixNLI, MNLI pre-training). Rows in the tables are sorted in ascending order w.r.t KI performance.

Perturb	**KI**	**ADAPET**			**ADAPET+CLS**		
		w/o	**+mixNLI**	**+MNLI**	**w/o**	**+mixNLI**	**+MNLI**
number+name	14.17	20	14.5	18.3	17.78	**20.8**	16.49
num+para+name	15.08	16.3	9.5	15.2	15.08	**17.9**	11.25
negation+name	18.66	17.1	7.8	11.6	**18.48**	10.31	10.55
number+negation	28.63	36.9	**41.5**	23.1	39.31	37.91	25.78
paraphrase+name	30.9	32.3	26.7	27.4	32.2	**32.48**	26.55
name	32.4	32.1	29.8	30.5	33.56	**33.7**	30.01
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33
negation+char	40.38	42.5	39.7	37.4	**45.4**	40.49	38.9
negation	46.46	**59.4**	56	52	59.03	58.4	55.7
num+paraphrase	52.56	57.3	58.4	**59.4**	57.7	51.13	48.9
number+char	53.34	55.5	61.6	**64.8**	55.3	55.85	54.9
number	54.9	59.5	56.9	**59.8**	55.91	51.97	51.13
character	56.88	63.7	**67.1**	63.3	65.16	65.16	65.27
paraphrase	66.3	72.5	**73.1**	72.2	69.88	73.1	72.22
location	69.65	**73**	70	69.9	69.97	68.59	68.1
dev	76.39	76.4	**78.2**	77.2	76.27	78.16	77.5
α_1	75.78	76.5	**79.4**	79.2	76.44	78.22	78.11

Table 12: Adversial Reasoning results on DRR@4 InfoTabs comparing KI (Neeraja et al., 2021), ADAPET, ADAPET+CLS (without pre-training (w/o), with mixNLI, MNLI pre-training). Rows in the tables are sorted in ascending order w.r.t KI performance.