Enhancing Tabular Reasoning with Pattern Exploiting Training

Abhilash Reddy Shankarampeta[1*], **Vivek Gupta**[2*†], **Shuo Zhang**[3]
[1]IIT Guwahati; [2]University of Utah; [3]Bloomberg
sareddy53@gmail.com; vgupta@cs.utah.edu; szhang611@bloomberg.net

Abstract

Recent methods based on pre-trained language models have exhibited superior performance over tabular tasks (e.g., tabular NLI), despite showing inherent problems such as not using the right evidence and inconsistent predictions across inputs while reasoning over the tabular data (Gupta et al., 2021). In this work, we utilize Pattern-Exploiting Training (PET) (i.e., strategic MLM) on pre-trained language models to strengthen these tabular reasoning models' pre-existing knowledge and reasoning abilities. Our upgraded model exhibits a superior understanding of knowledge facts and tabular reasoning compared to current baselines. Additionally, we demonstrate that such models are more effective for underlying downstream tasks of tabular inference on INFOTABS. Furthermore, we show our model's robustness against adversarial sets generated through various character and word level perturbations.

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 from INFOTABS (Gupta et al., 2020). The hypotheses **H1, H4** is entailed, **H2, H5** is a neutral and **H3, H6** is a contradiction. Here, the **bold** entries, which correspond to the first column, are the keys, while the corresponding entries in the second column of the same row are their respective values.

1 Introduction

Natural Language Inference (NLI) is the problem of categorizing a hypothesis into entailment, 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., 2019c) have been applied to large datasets like SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), where they have shown performance comparable to that of humans.

However, the existing methods based on language models are 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 inferences (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 human curated datasets (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 comprehensive 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 the current model's reasoning capability, particularly whether they can extract the right knowledge and correctly make rational inferences from that extracted knowledge. We focus on the task of tabular reasoning through table inference on INFOTABS (Gupta et al., 2020). For instance, in table 1, a model must filter out the relevant rows, i.e., extract knowledge, before applying the proper reasoning to categorize H1. Reasoning steps can be complex when involving numerical

[*]Equal Contribution [†]Corresponding Author

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 706–726
November 20–23, 2022. ©2022 Association for Computational Linguistics

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).

It has been proven that LMs pre-trained without explicit supervision on a huge corpus of free web data implicitly incorporate several types of knowledge into their parameters (Peters et al., 2019). For extracting this knowledge from language models (LM), various methods utilize probing (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) strategies. This internalized knowledge cannot be retrieved when fine-turning for a subsequent task. One explanation is that the objectives of pre-training and fine-tuning are vastly different. This variation in training objectives also diminishes the expected performance gains of the task, hence necessitating further pre-training on training data (Xiong et al., 2020; Roberts et al., 2020; Eisenschlos et al., 2020). Therefore, reframing the subsequent task as a joint pre-training objective becomes essential. Hence, we reformulate the tabular NLI, i.e., our downstream task as a cloze-style problem, a.k.a, a mask language modeling (MLM) problem. For fine-tuning, we utilize the efficient Pattern-Exploiting Training (PET) technique (Schick and Schütze, 2021a,b; Tam et al., 2021). PET entails establishing pairs of cloze question patterns and verbalizers that enable subsequent tasks to utilize the knowledge of the pre-trained language models. In addition, PET does not need model upgrades, such as adding more layers or parameters during pre-training.

Compared to direct fine-tuning-based techniques, i.e., training a classifier layer on top of LM, our method improved +8.1 and +25.8 on factual and relational knowledge evaluation tasks, respectively (see table 4). On INFOTABS , a tabular inference dataset, our PET training approach outperforms +1.72 on α_1 (similar to dev), +2.11 on α_2 (adversarial set), and +2.55 on α_3 (zero-shot set), see table 5) the existing baselines. This shows the effectiveness of our approach, especially on adversarial and out-of-domain challenging instances. Furthermore, we evaluate our improved model against instance perturbations to examine its robustness. These perturbations are generated by modifying existing INFOTABS instances, namely by changing names, numbers, places, phrases (paraphrasing), and characters (spelling errors). In addition, we also incorporated counterfactual instances (i.e., negation) to evaluate the model's robustness against pre-trained knowledge overfitting. The improvement in the counterfactual setting demonstrates that our approach benefits the model to ground better with premise table evidence.

Our main contributions are the following:

- We propose a method for generating prompts for determining if current models can infer from knowledge.
- We enhance the model's reasoning via prompt learning, i.e., PET, to extract knowledge from semi-structured tables.
- Our experiments on INFOTABS show that our proposed approach preserves knowledge and improves performance on downstream NLI tasks. The results are robust when assessed on multiple curated adversarial test sets.

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

2 Motivation

Case for Reasoning on Semi-structured Data. Reasoning semi-structured data acquire skills such as arithmetic and commonsense, understanding the text types in the tabular cells, and aggregating information across numerous rows if necessary. For example, to judge the H1 in table 1, the model needs to understand *"duration"* and *"length"* are the same in the context of the table, which is about a music album. Also, numerical reasoning is required to compare *"46:06" minutes"* is less than *"50 minutes"*. 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 factual information.

Implicit Knowledge is Required for Reasoning. For instance, for H3 in table 1, the model needs to first extract the relevant row, i.e., *"Released"* row from the table, then compares the phrase *"end of 1979"* with the "*Released*" row value *"29 March 1979"* implicitly. 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"*, but *"November"* or *"December"* is (implicit commonsense temporal knowledge). While previous works tried to incorporate knowledge via pre-training (Eisenschlos et al., 2020; Neeraja et al., 2021). In this work, we integrate knowledge and reasoning ability simultaneously using Pattern Exploiting Training (Tam et al., 2021). This approach improves the existing knowledge and enhances reasoning compared to existing methods.

Robustness is Critical for Model Evaluation. Tabular reasoning models typically fail on modest input modification, a.k.a. adversarial manipulation of inputs, highlighting the model's poor robustness and generalizability limit (Gupta et al., 2021). Thus, evaluating reasoning models on adversarial sets generated by minimal input perturbation becomes vital. As a result, we propose additional adversarial test sets, such as using character and word level perturbations to evaluate various aspects of model understanding and reasoning over tables. For example, if H1 (table 1) is changed to *"Breakfast in Wales is a pop album with a duration of fewer than 50 minutes."* now the label of hypothesis H1 is changes from **entailment** to **neutral** since we do not know any information of *"Breakfast in Wales"* from table 1. These minor input perturbations can alter the hypothesis' semantic interpretation. Idealistically, a robust model with superior reasoning ability should perform well on these input perturbed adversarial sets, as our technique also demonstrates.

3 Our Approach

In this section we describe our method to **(a)** evaluate pre-trained LM knowledge for tabular reasoning, **(b)** enhance model tabular reasoning capability using PET training, **(c)** and assess model robustness to input perturbations.

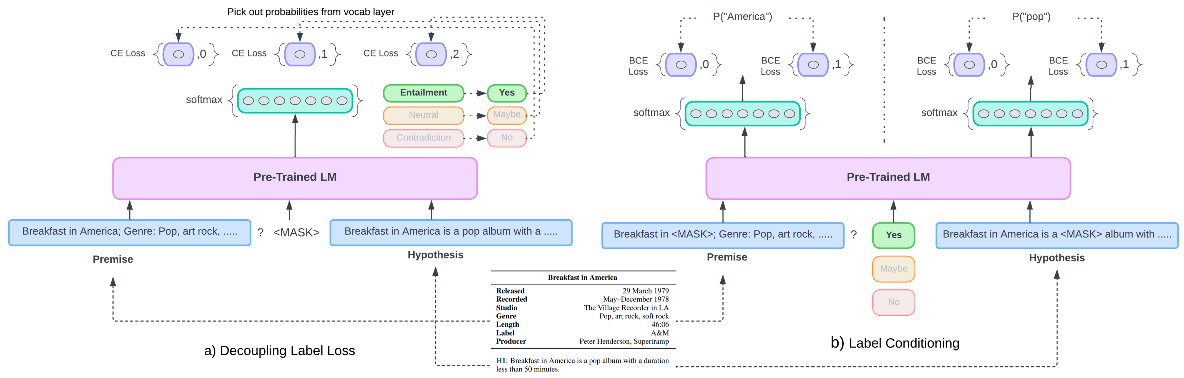
3.1 Evaluation of Pre-training Knowledge

To examine how pre-training affects knowledge-based reasoning for tabular data, we focus on two types of knowledge (a.) factual knowledge (awareness of specific factual knowledge about entities), (b.) and relational knowledge (awareness of possible right relations between two distinct entities). For instance, in the sentence *"Breakfast in America was released on March 29, 1979"*, *"Breakfast in America"* and *"March 29, 1979"* are considered as factual knowledge, while their relationship term, i.e., *"released"* corresponds to relational knowledge.

We evaluate factual and relational knowledge in the language model before and after training for the downstream task like reasoning. In specific, we query the model using "fill-in-the-blank" cloze statements (a.k.a. prompts). As gauging knowledge using prompts is limited by how the prompts are constructed. We use part-of-speech tagging to detect nouns and verbs that are then used to mask names, numbers, and dates. 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). We construct the prompts from both entailed and contradictory hypotheses. For prompts derived from entailed hypotheses, the model must predict the correct masked word, i.e., a term semantically equivalent to the word in the hypothesis. In contrast, for the prompts derived from contradicting hypotheses, the model should predict a semantically different term with the same entity type as the one mentioned in the hypothesis. To study the effect of the premise, we also query the model with the premise. To do this we modify the input as *premise + prompt*.

Prompts for Factual Knowledge Evaluation As most factual knowledge is contained in proper nouns and numbers, we randomly mask proper nouns or numbers in the hypothesis to generate a prompt and query the Language Model to fill the masked tokens. 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 are connected by *"duration"*. We randomly mask either *"Breakfast in America"* or *"46"* to generate prompt *"Duration of Breakfast in America is <mask> minutes"*. Occasionally, a masked term can be a number in numeric form (e.g., 2); however, the model predicted word form ("two"). We solved this issue by converting the predicted word into its numeric form or vice versa. E.g. *"Breakfast in America is produced by <mask> producers"*, where *<mask>* = *two*.

Prompts for Relational Knowledge Evaluation. Similar prompts are leveraged for relational knowledge. For example, to predict *<mask>* = *released* for *"Breakfast in America was <mask> towards the end of 1979"*, the model needs to understand that *"Breakfast in America"* is a music album to predict *"released"* instead of *"eaten"* which is highly probable due the neighbor context term *"Breakfast"*. We also use WordNet (Miller, 1995) to discover syn-



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.

Figure 1: The training uses the two ADAPET components. Here, the blue boxes represent the task inputs (entailed, in this case) a) Decoupling Label Loss: Using the cross entropy loss across all labels, the model must predict the right and wrong labels at the masked-out position. b) Label Conditioning: The model should predict the original token at a randomly masked-out position if the input text has the entail label. Otherwise, not if the label is contradiction or neutral.

onyms for the masked term and see if the predicted word is among them.

3.2 Knowledge Incorporation for Reasoning

The issue of deducing inferences from tabular premises is similar to the typical NLI problem, except that the premises are tables rather than sentences. When evaluating the reasoning skills, we use a variety of representations of the tabular premise (see section 4, appendix A.1). We also study the effect of pretraining on an NLI task 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) to the entire 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 are 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?"* label conditioned mask language modeling is introduced by randomly masking out context tokens. If the label is "entail", during training, the model is obligated to predict the original token; however, if the label is "contradiction" or "neutral", the model is forced to ignore the original token.

Masked Language Modeling. ADAPET randomly masks tokens (RoBERTa style) from the context. Inspired by SpanBERT (Joshi et al., 2020), ERNIE (Sun et al., 2019), we sample and mask the entire words based on pre-defined conditions. In Conditional Whole Word Masking (CWWM), we create a set of words S_w from a given sentence, and the POS of the words in that set must be from {"Adjective", "Adverb", "Noun, "Verb", "Proper Noun", "Adposition", "Numeral", "Coordinating Conjunction", "Subordinating Conjunction" }[1]. We sample words from the set S_w and mask all tokens matching the sampled word concurrently while maintaining the same overall masking rate.

3.3 Robustness with Input Perturbations

We apply a range of character- and word-level perturbations to hypotheses to simulate circumstances where the input is slightly noisy or deviates from the training data distribution. We use TextAttack (Morris et al., 2020), NLP Checklist (Ribeiro et al.,

[1] https://universaldependencies.org/u/pos/

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.

2020), and manual perturbations for generating the adversarial data. These adversarial sets will test the dependence of the model on word overlap, numerical comprehension, and hypothetical assertions. Refer to tables 2 and 9 for examples.

Character-level perturbation employs perturbations such as introducing random characters, switching characters, removing a random character, and substituting a random character in the randomly selected word. This alteration does not impact the label of the hypothesis because it does not alter the sentence's meaning.

Location perturbation modifies the identified locations (countries, cities, and nationalities) in a sentence to another place specified in the location map. The NER model (TextAttack) identifies the location in a given sentence and replaces it with a sampled location from a dictionary. Here, cities are replaced with other cities and similar changes for countries. This perturbation transforms the entail clauses into contradictions but does not affect the original neutral and contradiction labels.

Name perturbation randomly replaces a person's name with the other one from a name list. This perturbation alters the label of every hypothesis into a neutral because the perturbed hypothesis and premise mention different persons.

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.

Perturbing Numbers 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 transforms entailment into a contradiction by negating the given sentence, keeping neutrals intact.

Paraphrasing paraphrases the given sentences without the loss of meaning using manual paraphrasing and Pegasus model[2]. Paraphrasing does not affect the inference label as it does not change the semantic meaning of the hypothesis.

Composition of Perturbations perturbs sentences by applying various distinct perturbations sequentially. E.g., in **num+para+name** we perturbed a sentence *"Supertramp, produced an album that was less than 60 minutes long"*, with premise table 1 to *"Supertramp, produced an album that was less than 40 minutes long"* (number) then *"Supertramp released an album which lasted less than 40 minutes."* (paraphrase) then *"James released an album which lasted less than 40 minutes"* (name).

4 Experiments and Analysis

Dataset. Our experiments we use INFOTABS, a tabular inference dataset introduced by Gupta et al. (2020). The dataset is diverse in terms of the tables domains, categories, and corresponding keys (entity types and forms) it contains, as illustrated in examples table 1. In addition, Gupta et al. (2020) reveals that inference on corresponding hypotheses requires extensive knowledge and commonsense reasoning ability. Given the premise table, hypoth-

[2] https://biturl.top/MzQnMv

esis in the dataset is labeled as either an Entailment (E), Contradiction (C), or Neutral (N).

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 training sets and α_1,α_2 are similar. However, the training and fine-tuning tables are exclusive. Each of the test sets $\alpha_1, \alpha_2, \alpha_3$ has 200 unique tables paired with 9 hypothesis sentences (3E, 3C, 3N), totalling 1800 table-hypothesis pairs. Table 3 depict the statistics of perturbed sets from INFOTABS.

Model. We use the pre-trained RoBERTa-Large (RoBERTa$_L$) (Liu et al., 2019c) language model from HuggingFace (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), RoBERTa$_L$ trained for tabular inference using PET (ADAPET), and finetuning INFOTABS on ADAPET (ADAPET+CLS). Here we define fine-tuning 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 [3] (Nie et al., 2020; Bowman et al., 2015; Nie et al., 2019a). All models are trained on 16538 table-hypothesis pairs (1740 tables) for 10 epochs with a 1e-5 learning rate.

Table Representation. We explored two ways to represent table (a.) *Table as paragraph* uses Better Paragraph Representation for table representation, (b.) and *Distracting Row Removal* prunes tables based on the similarity between hypothesis and tables rows. We investigated the 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

Our experiments answer the following questions:

RQ1: Can the large language model use pre-trained knowledge for reasoning? Does our adaptive training method enhance model reasoning?

RQ2: Does fine-tuning downstream tasks benefit model reasoning? Can our adaptive training benefit model via enhancing its reasoning knowledge?

RQ3: Is our adaptive method-based model robust to input perturbations? Can our method enhance model's semantic-syntactic comprehension?

Models Knowledge Evaluation. To answer RQ1, we evaluate the knowledge in the presence and absence of the premise using the Entail and Contradictory hypotheses, which are taken from the evidence in the premise tables. We do not use Neural statements as they may contain subjective and out-of-table information.

Type	Input	RoBERTa$_L$		ADAPET	
Top 1 Accuracy		w/o	+CLS	w/o	+CLS
Factual	only E	35.5	26.2	34.3	29.2
	prem + E	59.4	29	59.7	44.8
	only C	37.2	24.6	36.9	29.8
	prem + C	54.6	26.5	49.7	39.9
	only E∪C	36.3	25.4	35.5	29.5
	prem + E∪C	57.7	27.8	54.6	42.5
Relational	only E	48.9	27	52.8	35.6
	prem + E	57.7	22.4	58.7	41
	only C	44.7	27.3	47.3	35.6
	prem + C	51.8	24	52.9	34
	only E∪C	46.7	27.2	49.9	35.6
	prem + E∪C	54.6	23.2	55.7	37.3

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

In all the settings (tables 4 and 11) with and without premise, our model outperformed RoBERTa$_L$+CLS. The addition of the premise enhances model performance further. This can be ascribed to additional knowledge in the premise that our PET-trained model can leverage efficiently for reasoning. From table 4, we observe that for all settings, our approach gave $\tilde{1}00\%$ improvement in relational knowledge evaluation compared to RoBERTa$_L$+CLS. Even training a classifier on top of ADAPET outperforms RoBERTa$_L$+CLS. We also evaluated on contradiction hypothesis to assess if the model can rightly identify false claims despite having correct entity types.

There is a significant difference between the Top 1 accuracy of premise+E and premise+C for factual knowledge evaluation as the model should not

[3] https://biturl.top/e6Vney

Splits	Premise	RoBERTa$_L$	ADAPET				ADAPET+CLS			
		+CLS	token	CWWM	+mixNLI	+MNLI	token	CWWM	+mixNLI	+MNLI
Dev	BPR	76.83	77.5	77.67	79.07	78.07	77.66	77.27	**79.63**	78.46
	DRR@4	76.39	76.67	76.97	78.57	77.33	76.88	77.11	**78.64**	77.44
	DRR@8	75.36	77.77	77.63	78.83	77.93	77.81	77.57	**79.42**	78.96
α_1	BPR	75.29	76.87	75.93	77.33	77.47	77.47	78.05	77.96	**78.33**
	DRR@4	75.78	77.5	77.53	**78.6**	78.17	77.18	77.66	78.04	78.13
	DRR@8	75.61	78.3	78	79	78.2	78.03	78.7	78.63	**79.05**
α_2	BPR	66.5	67.93	68.07	**72.4**	69.8	68.48	69.55	72.16	70.09
	DRR@4	67.22	69.33	69	70.23	69.03	68.92	68.29	**70.58**	69.24
	DRR@8	67.11	69.43	69.37	71.87	69.97	69.24	69.81	**72.13**	70.61
α_3	BPR	64.26	63.73	64.6	66.23	64.13	64.98	65.67	**68.4**	66.03
	DRR@4	64.88	67.43	67.5	68.7	67.33	66.02	66	**68.74**	67.37
	DRR@8	67.53	68.07	67.63	**70.2**	68	66.66	67.59	69.2	68.31

Table 5: Reasoning results on INFOTABS comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking.

predict the masked token in the prompt from a contradiction statement, especially in factual prompts. And for relational knowledge, irrespective of the label of the hypothesis, the model should predict the masked token correctly if the model rightly understands the entity types of words in the sentence. 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, thus demonstrating the benefits of our approach. A similar observation was reported with Top 5 accuracy (table 11).

Knowledge Incorporation for Reasoning. To answer RQ2, we experiment with various premise representations of tables as paragraphs (BPR, DRR@4, DRR@8) (see table 5). We observe that Roberta-Large with ADAPET improves performance in all premise representations except for α_3 with BPR compared to RoBERTa$_L$+CLS 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). Results in table 5 are the average accuracy of the models tested on multiple seeds.

With ADAPET, we also improve performance using linearized table (see table 7) compared to Gupta et al. (2020) (+1.04 in α_1, +0.58 in α_2, +0.69 in α_3). ADAPET (token masking, no pre-training) tops RoBERTa$_L$+CLS in every premise representation and test split. +1.72 in α_1, +2.11 in α_2, +2.55 in α_3 with DRR@4. CWWM with ADAPET also outperformed RoBERTa$_L$+CLS. However, the performance of the two masking procedures is comparable for all test sets, even with the classifier setting.

We notice that the DRR@8 representation outperforms the best, especially in α_3 due to removing the irrelevant rows (+4.34 over BPR, +0.64 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 (token masking, no pretraining), there has been an improvement of +2.13 units (no CLS) and +2.54 units (with CLS) with DRR@8 over no pre-training. We also observed that pre-training in more diverse data helps improve performance (Andreas, 2020; Pruksachatkun et al., 2020). Models which are pre-trained on mixNLI[3] outperformed MNLI pre-trained in almost every setting (+0.8 in α_1, +1.9 in α_2, +2.2 in α_3 with no CLS, DRR@8).

Robustness to Input Perturbation. To answer RQ3, we evaluate our model on several challenging input perturbations. The perturb test sets are generated using various character-level, and word-level perturbations are also tested with BPR, DRR@4, and DRR@8 table representations (see table 6). To generate these sets, we applied perturbations on dev, and α_1 sets as the distribution of these sets are similar to the training set. We also human-verified

Perturb	**RoBERTa$_L$**	**ADAPET**				**ADAPET+CLS**			
	+CLS	**token**	**CWWM**	**+mixNLI**	**+MNLI**	**token**	**CWWM**	**+mixNLI**	**+MNLI**
num+para+name	13.04	10.1	7.1	11.7	10.1	11.7	13.81	**16.62**	13.55
number+name	15.72	14.6	9.0	14	13.2	15.6	15.36	**18.94**	15.85
negation+name	19.08	16.1	7.2	**20**	11.6	14.43	12.88	14.37	12.1
num+paraphrase	27.46	59.5	**61.0**	58.4	57.3	52.5	51.49	56.63	54.95
paraphrase+name	30.79	22.6	18.3	28.3	24.9	27.01	27.3	**30.85**	27.71
name	32.7	24.7	19.0	31.1	28	28.9	29.96	**33.44**	30.69
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
number+negation	36.13	42.7	31.8	**53.2**	28.3	37.91	47.32	37.75	24.04
negation+char	39.39	41.4	38.5	**47.6**	40.1	42.9	41.94	42.06	40.85
negation	53.7	58.1	53.3	**64.8**	56.1	57.6	56.83	59.15	53.88
number+char	54.43	58.8	**65.2**	57.1	60.3	55.79	47.9	57.1	59.28
number	56.1	57.8	**62.0**	57.8	57	52.44	51.37	55.79	54.6
character	63.05	62.8	63.3	65.9	64.4	64.05	64.44	66.05	**66.83**
location	67.6	70	**70.2**	67.7	69.1	69.81	66.8	67.4	65.98
paraphrase	70.56	72.3	73.2	**73.8**	73.4	71.6	70.5	72.66	72.3
INFOTABS (α_1)	76.56	78.1	78.9	**80.2**	78.9	78.27	77.66	78.5	78.66

Table 6: Adversarial Reasoning results on perturbed sets with DRR@8 comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training), token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking. Rows in the tables are sorted in ascending order w.r.t RoBERTa$_L$+CLS performance.

our perturbation examples; refer to appendix A.5.

Except for the perturbations involving names, our method ADAPET (no pre-training) outperforms RoBERTa$_L$+CLS. We see the max improvement of ADAPET in the Negation (+4.4); this implies our model can handle counterfactual statements well. We observed that training a classifier head on top of ADAPET performed better with the adversarial sets involving multiple perturbations. In the challenge set with *number+paraphrase* all the ADAPET-based models outperformed RoBERTa$_L$+CLS by 2x times. We observed that using NLI pre-training also helps substantially improve the robustness. With the use of mixNLI and MNLI pre-trained weights, the performance of ADAPET-based models improved substantially compared to those without pre-training, even outperforming RoBERTa$_L$+CLS. From table 6, it is clear that with hypotheses involving multiple perturbations, RoBERTa$_L$+CLS tends to perform more poorly compared to the ADAPET-based model. (For quality analysis of perturbations see appendix A.5). The performance on all perturb sets is much worse than that of the corresponding model on dev, α_1 sets. Improving the performance of these sets is crucial.

What did we learn? Reformulating the NLI task as an MLM problem enabled the inclusion of premise table knowledge into Language Models (LM) for efficient reasoning. Using ADAPET, we have shown that knowledge can be retained and assimilated into reasoning tasks more effectively. ADAPET training also improves the model's ability to reason on downstream tasks. Similar observation is also observed in prior works Xiong et al. (2020); Sun et al. (2019) where MLM is utilized to incorporate external knowledge, although the later require additional table based pre-training. Moreover, Gupta et al. (2021); Lewis et al. (2021) have shown that the LM utilizes spurious patterns to accomplish reasoning tasks. Our perturb sets study informed us that our ADAPET-based method is more robust than direct classification to semantic-syntactic alternations. (see appendix B for further discussions)

5 Related Work

Tabular Reasoning. Many recent papers discussed NLP challenges associated with semi-structured table data such as Tabular NLI (Gupta et al., 2022, 2020; Neeraja et al., 2021), fact verification (Chen et al., 2020a; Zhang et al., 2020a), question answering (Zhu et al., 2021; Zhang and Balog, 2020; 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; Zhang et al., 2020b; Nan et al., 2021; Chen et al., 2021b; Yoran et al., 2021, and others) are some examples. Several studies have offered techniques for encoding Wikipedia tables, such as

TAPAS(Herzig et al., 2020), TaBERT (Yin et al., 2020), TabStruc (Zhang et al., 2020a), TABBIE (Iida et al., 2021), StruBERT (Trabelsi et al., 2022), Table2Vec (Zhang et al., 2019a), 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 tabular data augmentation.

Knowledge Incorporation and Evaluation. A line of works have been proposed to integrate knowledge into the LMs using pretrained entity embeddings (Zhang et al., 2019b; 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). Several methods, including probing classifiers, have been proposed to extract and assess knowledge from LMs (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).

Model 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 (e.g., Ribeiro et al., 2016, 2018a,b; Alzantot et al., 2018; Iyyer et al., 2018; Glockner et al., 2018; Naik et al., 2018; McCoy et al., 2019; Nie et al., 2019b; Liu et al., 2019a). Moradi and Samwald (2021) introduces a textual perturbation infrastructure that incorporates character- and word-level systematic perturbations to imitate real-world noise. Goel et al. (2021) offered a toolbox to evaluate NLP systems on subpopulations, transformations, evaluation sets, and adversarial attacks.

6 Conclusion

In this work, we have validated the effects of factual and relational knowledge in the language model via handcrafted prompts for tabular reasoning. Through prompt learning, i.e., Pattern-Exploiting Training, we extracted knowledge from semi-structured tables and further improved the model's reasoning capabilities. Our intensive experiments on the INFOTABS demonstrate that our approach can conserve knowledge and enhance tabular NLI performance. The conclusions hold up well when tested against carefully crafted adversarial test sets based on character and word-level perturbations.

Method Limitations: Entity tables are the focus of our solution. Its scalability in constructing prompts and other tables with different structures is limited by the idea that manually identified pattern from the specific dataset and template-based prompts. In addition, as not different from other NLP tasks, automatically detecting knowledge patterns and bridging patterns to prompts, especially for semi-structured tables, is under-explored. Furthermore, investigating prompting for sophisticated structured tables such as nested structures (e.g., lists inside tables), hierarchical tables (e.g., table inside a table), and multi-modal tables (pictures within table) will necessitate substantial effort.

Future Directions: We have identified the following 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, masking *Breakfast in America* in H3 (table 1) 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. (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., 2019b), we want to improve and assess the reasons behind the results in table 6.

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.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

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.

Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. 2022.

Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3283, Dublin, Ireland. 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.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. 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.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019b. 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. 2019c. 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.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. 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. 2019a. 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, Yicheng Wang, and Mohit Bansal. 2019b. Analyzing compositionality-sensitivity of nli models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.

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, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, 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.

Mohamed Trabelsi, Zhiyu Chen, Shuo Zhang, Brian D. Davison, and Jeff Heflin. 2022. Strubert: Structure-aware bert for table search and matching. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 442–451, New York, NY, USA. Association for Computing Machinery.

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. 2020a. 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.

Li Zhang, Shuo Zhang, and Krisztian Balog. 2019a. Table2vec: Neural word and entity embeddings for table population and retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1029–1032, New York, NY, USA. Association for Computing Machinery.

Shuo Zhang and Krisztian Balog. 2020. Web table extraction, retrieval, and augmentation: A survey. *ACM Trans. Intell. Syst. Technol.*, 11(2).

Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020b. Summarizing and exploring tabular data in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1537–1540, New York, NY, USA. Association for Computing Machinery.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019b. 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.

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, 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 a 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 semicolon (";") is used to separate rows.

- *Table Pruning:* For a particular hypothesis, not all of the entries in the premise table are essential. Sometimes, the entire table with the hypothesis as input might be longer than the specified input length of the language model. Inspired by 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, so we also consider top-8 rows. We use the top 4 and 8 relevant rows from DRR (DRR@4 and DRR@8, respectively) for evaluation.

A.2 Results with Linearized Table

We experiment with premise as a linearized table and compared our results with Gupta et al. (2020), see table 7. Our proposed approach was able to outperform the baselines in Gupta et al. (2020) by a significant margin.

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

A.3 Reasoning on Entail / Contradict 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, see table 8. We found that DRR@4, DRR@8 representations of premise performs better that BPR because of the less distracting premise.

Splits	RoBERTa$_L$+CLS	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 8: Results on two label classification (Entailment & Contradiction).

A.4 Robustness on Perturbation Set

We evaluate robustness with premise representation. In tables 13 and 14 we show the performance of the model on the adversarial tests which are trained and tested with BPR, DRR@4 representations of premise. We found the results are similar to the results in table 6.

A.5 Qualitative Analysis of Perturbation Sets

On a randomly sampled subset containing 100 examples from each of the perturbation sets, we task a human evaluator to label them and give a score (out of 5) to the grammar of the hypotheses (see table 10). For most cases, i.e., 11 out of 14, we observe a correct of > 80% indicating the correction of our adversarial tests. Furthermore, in half of the cases (7/14), the correctness score was above 95%. Grammar analysis shows that most sentences are highly grammatical, with an average score of 4.5/5.0. In the perturbation *"number+paraphrase"* we only observed 77% of label correctness. This could be due to changing numbers, followed by paraphrasing, which changed some contradiction hypotheses to neutral ones. A similar observation is also observed in *"number+char"* where numbers are modified in character perturbation. We also compare the models' performance on these sampled perturbed sets after human corrections in labels and grammar (see table 12). We observed that the performance on these corrected sets is similar to the generated perturbed sets, as in table 14.

A.6 Models Knowledge Evaluation

We also evaluated the model's knowledge of the top 5 accuracy metric table 11. The results follow a similar pattern on the top 1 accuracy metric.

A.7 Error Analysis

In fig. 7, when compared to fig. 6 there is a substantial improvement in identifying NEUTRAL and CONTRADICTION, but there is also a confusion in identifying ENTAILMENT. Using the NLI-pretrained model improves the detection of ENTAILMENT. A similar observation is also observed with using classifying layer (+CLS) (see figs. 7 and 9).

In fig. 2, we see the greatest inconsistency is with NEUTRAL being misidentified as ENTAILMENT across all models, and this is not that significant with using the classifying layer (+CLS) (see figs. 3 and 5). Although with the classifying layer, there is increased confusion about CONTRADICTION being predicted as ENTAILMENT.

Table 15 shows a subset of the validation set labeled based on the different ways the model must think to put the hypothesis in the correct category. On average, all the ADAPET-based models perform similarly, but the human scores are better than the model we utilize. We observe that for certain reasoning types, such as Negation and Simple Look-up, neither humans nor the model arrives at the correct hypothesis, demonstrating the task's difficulty. For Numerical, Lexical, and Entity type reasoning, our model comes very close to human scores.

In table 16, we observed that the City category on proposed models performs worse probably as a result of the engagement of more numeric and specific hypotheses compared to the other categories, as well as longer average table size. Our models perform extremely well in identifying ENTAILMENT in Food & Drinks category because of their smaller table size on average and hypothesis requiring no external knowledge to reason as compared to CONTRADICTION. Our models also struggle in detecting NEUTRAL and CONTRADICTION in Organization category.

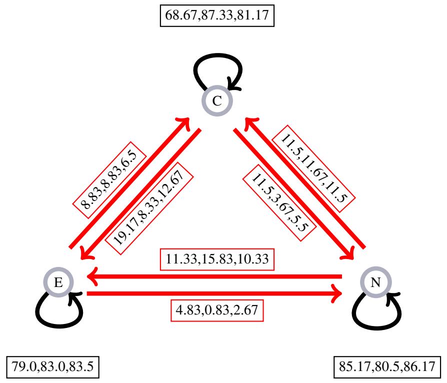


Figure 2: Consistency graph for predictions of ADAPET(token) vs (a) RoBERTa$_L$+CLS (b) ADAPET (CWWM) (c) ADAPET (pretrained mixNLI) in that order respectively.

Perturb	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 9: More examples of various perturbations used to generate the adversarial test sets based on table 1

Perturbation	**Label Correctness(%)**	**Grammar Score**
character	99	4.46
location	79	4.5
name	97	4.5
negation	93	4.36
number	81	4.5
paraphrase	89	4.42
negation+char	88	4.3
negation+name	96	4.5
number+char	77	4.3
number+name	96	4.5
number+negation	80	4.44
num+paraphrase	77	4.48
num+para+name	95	4.42
paraphrase+name	94	4.5

Table 10: Results on Label Correctness (% of our generated labels match with human's predictions) and average Grammar score (out of 5) from human evaluation.

Type	**Input**	**RoBERTa$_L$**		**ADAPET**	
Top 5 Accuracy		**w/o**	**+CLS**	**w/o**	**+CLS**
Factual	only E	50.4	40.6	52.4	46.6
	prem + E	72	45.3	71.5	60.7
	only C	55.2	37.4	56	47.8
	prem + C	74.6	39.3	70.2	56
	only E∪C	52.7	39.1	54.1	47.2
	prem + E∪C	73.3	42.5	70.9	58.5
Relational	only E	64.9	51.6	67.3	57.5
	prem + E	70.8	49.1	72.2	66.3
	only C	64.7	53.1	65.8	57.8
	prem + C	71.1	53.3	72	62
	only E∪C	64.8	52.4	66.5	57.6
	prem + E∪C	70.9	51.3	72.1	64.1

Table 11: Top 5 accuracy of Factual & Relational Knowledge Evaluation on DRR@4.(w/o - no CLS, RoBERTa$_L$+CLS

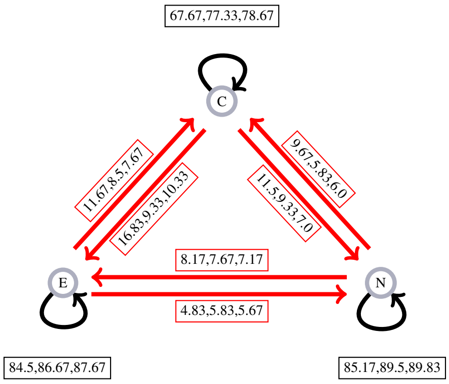


Figure 3: Consistency graph for predictions of ADAPET(token)+CLS vs (a) RoBERTa$_L$+CLS (b) ADAPET (CWWM)+CLS (c) ADAPET (pretrained mixNLI)+CLS in that order respectively.

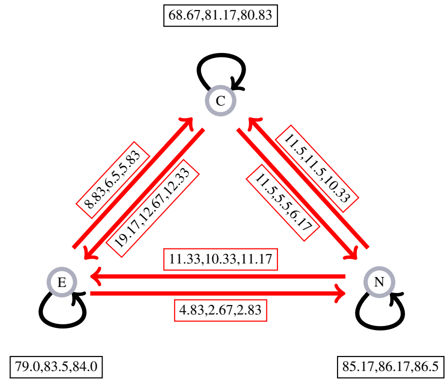


Figure 4: Consistency graph for predictions of ADAPET(token) vs (a) RoBERTa$_L$+CLS (b) ADAPET (pretrained mixNLI) (c) ADAPET (pretrained MNLI) in that order respectively.

Perturb	**RoBERTa$_L$**	**ADAPET**				**ADAPET+CLS**			
	+CLS	**token**	**CWWM**	**+mixNLI**	**+MNLI**	**token**	**CWWM**	**+mixNLI**	**+MNLI**
character	62	**69**	61	64	65	**69**	55	65	53
location	64	**70**	69	66	63	69	68	69	63
name	36	**40**	31	37	**40**	35	41	35	36
negation	43	**65**	63	**65**	59	57	55	55	58
number	62	**69**	69	68	69	68	66	59	54
paraphrase	66	**77**	71	76	**77**	70	68	74	71
negation+char	32	41	42	42	**44**	43	30	4	39
negation+name	15	10	10	**18**	13	16	9	12	12
number+char	5	50	54	55	**60**	49	40	54	50
number+name	22	20	17	24	**26**	23	25	24	21
number+negation	33	58	**54**	51	43	5	47	44	32
num+paraphrase	52	52	58	60	50	**59**	55	54	56
num+para+name	**18**	10	3	8	15	14	15	**18**	10
paraphrase+name	33	**38**	28	35	33	36	34	36	28

Table 12: Adversarial Reasoning results on human corrected perturbation sets with DRR@4 comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking.

Perturb	**RoBERTa$_L$**	**ADAPET**				**ADAPET+CLS**			
	+CLS	**token**	**CWWM**	**+mixNLI**	**+MNLI**	**token**	**CWWM**	**+mixNLI**	**+MNLI**
negation+name	11.74	10.4	10.2	**21.1**	15.6	17.35	14.37	13.89	12.93
num+para+name	14.06	10.6	8.4	**20.7**	12	17.13	16.88	14.83	13.04
number+name	17.26	12.5	10.2	**20.9**	14.8	18.42	18.81	18.42	16.88
paraphrase+name	33	25.8	20.6	**37.6**	31.5	31.2	33.41	32.1	31.3
random	33.33	33.33	33.3	33.33	33.33	33.33	33.33	33.33	33.33
name	34.6	26.5	20.4	**36.4**	33.4	32.41	34.82	33.96	33.2
negation+char	37.71	38.5	40.3	**47.8**	41.3	43.56	40.21	41.25	40.49
number+negation	38.36	30.2	48.7	**54.8**	30.1	37.69	47.26	38.7	26.06
negation	48.9	54.2	57.2	**65.4**	55.3	58.27	55.27	58.45	55.6
number	56.63	**62.3**	55.8	51.9	56	55.43	50.53	53.52	56.1
num+paraphrase	56.98	**62.3**	57.6	49.7	54.5	55.55	49.34	52.26	55.19
number+char	59.11	**66.1**	60.3	45.1	55.6	55.9	49.32	52.46	60.2
character	61.5	64.1	62.5	64.4	66.1	64.9	63.16	**66.61**	65.94
location	68.2	72.4	**72.7**	68.1	70.1	69.08	67.69	66.47	69.48
paraphrase	68.44	72.3	71.8	**72.6**	72.3	72.05	70.33	71.7	**72.66**
dev	76.83	78.1	76.4	**79.8**	79.1	78.72	78.05	79.22	78.55
α_1	75.29	78.1	76.1	77.4	77.4	77.38	77.83	78	**78.38**

Table 13: Adversarial Reasoning results on perturbed sets with BPR comparing RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking. Rows in the tables are sorted in ascending order w.r.t RoBERTa$_L$+CLS performance.

Perturb	**RoBERTa$_L$**	**ADAPET**				**ADAPET+CLS**			
	+CLS	**token**	**CWWM**	**+mixNLI**	**+MNLI**	**token**	**CWWM**	**+mixNLI**	**+MNLI**
number+name	14.17	20	12.9	14.5	18.3	17.78	17.13	**20.8**	16.49
num+para+name	15.08	16.3	8.7	9.5	15.2	15.08	16.88	**17.9**	11.25
negation+name	18.66	17.1	13.9	7.8	11.6	**18.48**	13.23	10.31	10.55
number+negation	28.63	36.9	43.2	41.5	23.1	39.31	**45.86**	37.91	25.78
paraphrase+name	30.9	32.3	22.6	26.7	27.4	32.2	32.36	**32.48**	26.55
name	32.4	32.1	25.7	29.8	30.5	33.56	33.6	**33.7**	30.01
random	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
negation+char	40.38	42.5	41.1	39.7	37.4	**45.4**	40.61	40.49	38.9
negation	46.46	**59.4**	57	56	52	59.03	56.89	58.4	55.7
num+paraphrase	52.56	57.3	59.5	58.4	**59.4**	57.7	51.86	51.13	48.9
number+char	53.34	55.5	63.2	61.6	**64.8**	55.3	49.81	55.85	54.9
number	54.9	59.5	59.1	56.9	**59.8**	55.91	52.09	51.97	51.13
character	56.88	63.7	63.7	**67.1**	63.3	65.16	60.88	65.16	65.27
paraphrase	66.3	72.5	72.9	**73.1**	72.2	69.88	68.44	73.1	72.22
location	69.65	**73**	71.2	70	69.9	69.97	65.825	68.59	68.1
dev	76.39	76.4	77.8	**78.2**	77.2	76.27	78.05	78.16	77.5
α_1	75.78	76.5	78	**79.4**	79.2	76.44	77.66	78.22	78.11

Table 14: Adversarial Reasoning results on perturbed sets with DRR@4 RoBERTa$_L$+CLS, ADAPET, ADAPET+CLS (without pre-training (token, CWWM), with mixNLI, MNLI pre-training). token, CWWM - masking strategies, mixNLI, MNLI pre-training uses RoBERTa style token masking. Rows in the tables are sorted in ascending order w.r.t RoBERTa$_L$+CLS performance.

Reasoning Type	ENTAILMENT					NEUTRAL					CONTRADICTION				
	RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS	
	+CLS	token	+mixNLI	token	+mixNLI	+CLS	token	+mixNLI	token	+mixNLI	+CLS	token	+mixNLI	token	+mixNLI
Numerical (11, 3, 7)	9	9	10	10	8	3	2	3	3	3	6	6	4	6	5
Lexical Reasoning (5, 3, 4)	5	4	4	3	5	2	1	1	1	2	2	3	2	3	3
Subjective/OOT (6, 41, 6)	3	3	3	3	3	37	36	36	37	35	4	4	1	3	5
KCS (31, 21, 24)	25	21	26	20	25	20	20	18	19	18	21	22	18	21	21
Temporal (19, 11, 25)	16	13	15	15	14	7	6	5	6	7	18	20	15	17	17
Multirow (20, 16, 17)	13	12	15	15	13	13	12	11	11	13	15	16	14	15	13
Coref (8, 22, 13)	5	6	5	6	6	19	20	18	20	18	7	10	8	7	8
Quantification (4, 13, 6)	2	2	2	2	2	11	11	12	12	12	2	3	3	3	3
Named Entity (2, 2, 1)	1	2	2	1	2	1	1	2	1	1	1	1	1	1	1
Simple Lookup (3, 0, 1)	2	3	3	2	3	0	0	0	0	0	0	0	0	0	0
Negation (0, 0, 6)	0	0	0	0	0	0	0	0	0	0	4	6	5	5	4
Entity Type (6, 8, 6)	6	5	5	4	6	7	7	7	7	7	6	6	5	6	4

Table 15: Reasoning wise number of correct predictions of DRR@4 on subset of dev set, (a, b, c) are human prediction count.

Categories	ENTAILMENT					NEUTRAL					CONTRADICTION				
	RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS		RoBERTa$_L$	ADAPET		ADAPET+CLS	
	+CLS	token	+mixNLI	token	+mixNLI	+CLS	token	+mixNLI	token	+mixNLI	+CLS	token	+mixNLI	token	+mixNLI
Album	71	79	74	76	81	76	86	88	86	93	60	79	79	74	74
Animal	78	81	89	89	85	70	81	81	85	81	56	70	74	81	78
City	59	63	63	57	69	67	80	65	71	75	53	61	63	65	55
Country	78	75	83	64	78	56	67	64	61	72	56	69	72	58	67
Food&Drinks	96	88	88	88	88	67	75	75	71	79	83	88	79	71	71
Movie	85	75	83	80	80	75	85	70	82	73	62	75	80	73	80
Musician	87	78	84	83	88	86	90	85	89	89	75	83	79	78	78
Organization	83	50	100	75	92	58	75	50	83	75	58	58	58	50	50
Painting	78	81	81	81	85	93	93	93	96	93	78	89	85	78	85
Person	74	73	78	74	78	81	85	80	78	81	67	79	76	77	74
Others	71	69	82	69	80	64	78	69	73	73	49	73	69	67	60

Table 16: Category wise accuracy scores of DRR@4 on dev set

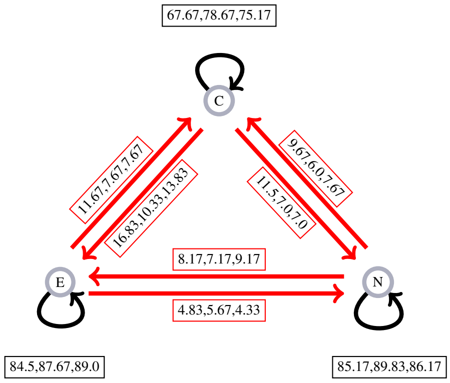


Figure 5: Consistency graph for predictions of ADAPET(token)+CLS vs (a) RoBERTa$_L$+CLS (b) ADAPET (pretrained mixNLI)+CLS (c) ADAPET (pretrained MNLI)+CLS in that order respectively.

B Further Discussion

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; 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 leverage the advantage of pre-trained models 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 .

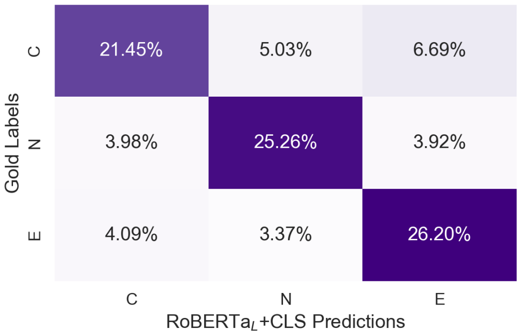


Figure 6: Confusion Matrix: Gold Labels vs predictions of RoBERTa$_L$+CLS.

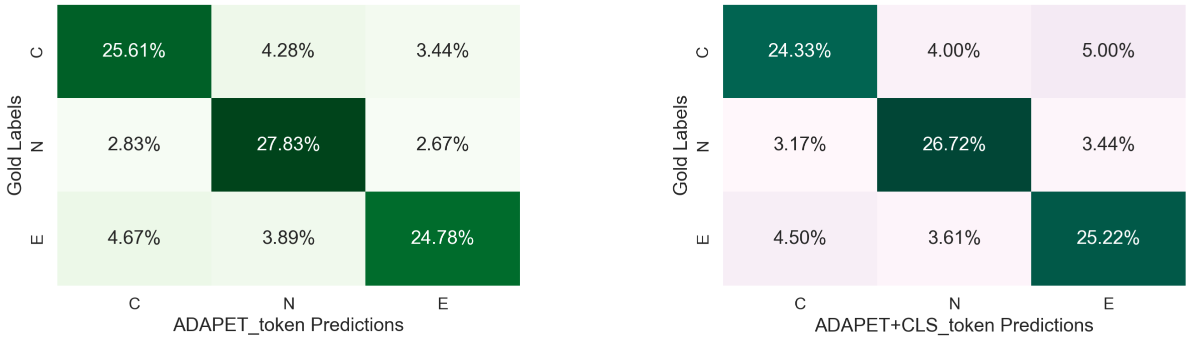


Figure 7: Confusion Matrix: Gold Labels vs predictions of ADAPET(token), ADAPET(token)+CLS.

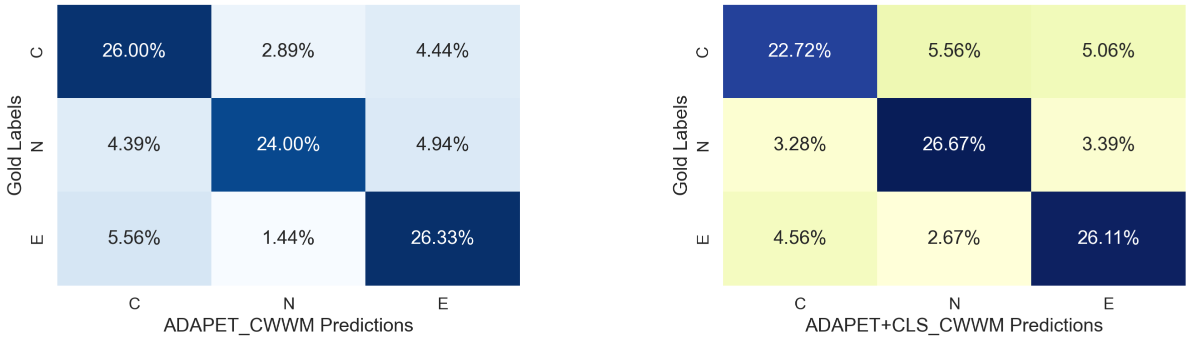


Figure 8: Confusion Matrix: Gold Labels vs predictions of ADAPET(CWWM), ADAPET(CWWM)+CLS.

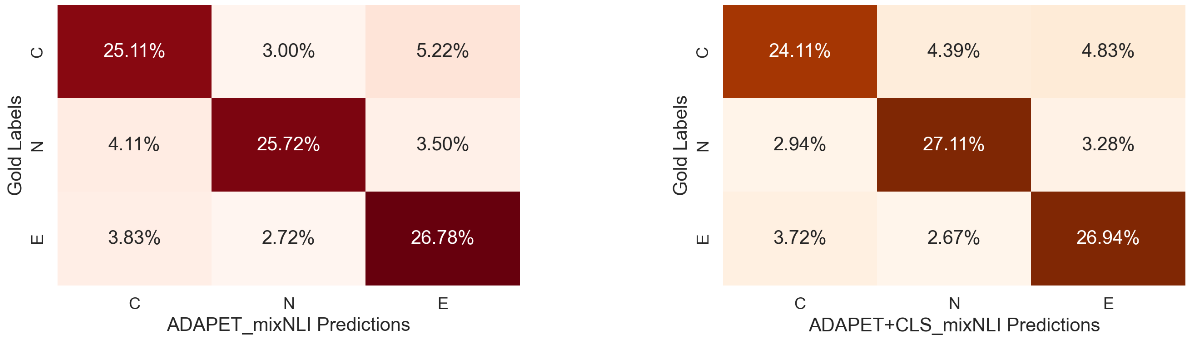


Figure 9: Confusion Matrix: Gold Labels vs predictions of ADAPET (pretrained mixNLI), ADAPET (pretrained mixNLI)+CLS.