Inference on Tables as Semi-structured Data

Vivek Gupta
School of Computing, University of Utah

October 26, 2022

1 Introduction

This dissertation examines reasoning and inference over semi-structured tabular text, specifically entity-centric InfoBox tables (cf. Figure 1). Semi-structured tabular text is widespread and involves comprehension of the meaning of text fragments and implicit relationships between them. The dissertation argues that semi-structured data can serve as a crucial testing ground for increasing the understanding of how Natural Language Processing (NLP) models to reason about information.

Breakfast in America		Relevance
Released[4]	29 March 1979[4]	H3
Recorded[3,4]	May-December 1978[3,4]	H2, H3
Studio	The Village Recorder in Los Angeles[3]	
Genre	Pop, Art Rock, Soft Rock	
Length[2]	46:06[2]	H1
Label	A&M	
Producer[1]	Peter Henderson, Supertramp[1]	H1

H1: Supertramp produced[1] an album that was less than an hour long[2].
H2: Most of Breakfast in America was recorded[3] in the last month of 1978[3].
H3: Breakfast in America was released[4] the same month recording[4] ended.

Figure 1: A semi-structured premise (the table 'Breakfast in America') example from (Gupta et al., 2020). Hypotheses H1 are entailed by it, H2 is neither entailed nor contradictory, and H3 is a contradiction. The 'Relevance' column shows the hypotheses that use the corresponding row for reasoning. The colored text (and superscripts) in the table and hypothesis highlights relevant token level alignment.

To explore this, we first introduce a new dataset called INFOTABS (Gupta et al., 2020) (refer §2.1), which is used to investigate the task of Natural Language Inference (NLI). NLI is the process of determining whether a human-written textual hypothesis is true (ENTAIL), false (CONTRADICT), or cannot be determined, i.e., possibly true/false (NEUTRAL), based on premises that are extracted from Wikipedia info-boxes. INFOTABS's semi-structured, multi-domain and heterogeneous nature of the tabular data admits complex, multi-faceted reasoning. More precisely, this dissertation is devoted to the following hypothesis:

Thesis Statement: *Reasoning and inference over semi-structured tabular text, more precisely entity-centric InfoBox tables, is a critical component of Natural Language Understanding. The task poses numerous real challenges, including effective table representation, successful knowledge addition, model robustness to perturbation, and requisite evidence extraction. A fusion of tools and techniques involving several areas of Natural Language Processing is utilized to address these issues.*

While working with tabular data (c.f. §2), the following challenges and questions were addressed:

1. *How do models designed for the raw text adapt to the tabular data?(§2.1)* The infobox table does not explicitly state the relationship between the keys and values (Gupta et al., 2020). Additionally, only a portion of the table is necessary for the model to predict; the remainder acts as a distraction (Neeraja et al., 2021; Gupta et al., 2022b). In Figure 1 All keys except *"Length"* and *"Producer"* are unrelated to hypothesis H1. Distracting rows frequently produce erroneous correlations, leading to correct predictions for the wrong reasons.

2. *How do we represent and incorporate knowledge into a tabular model?(§2.2)* Tables often lack the necessary context to comprehend the meaning of a text fragment (such as a key) and its relationship to other elements (such as value and other keys). For example, in Figure 1, we need to interpret the key *'Length'* n the context of music albums for the given table. Furthermore, due to inadequate training, data models trained on tables are often feeble in implicit lexical knowledge (Neeraja et al., 2021). This affects interpreting the meaning of words such as "*less than*" in H1 (c.f. Figure 1).

3. *How to ensure the model is using the correct evidence-based reasoning?(§2.3 and §2.4)* Recent studies show that deep learning systems are brittle and memorize spurious patterns such as annotation artefacts, often amplify societal biases (Bolukbasi et al.; Zhao et al., 2017; Poliak et al., 2018; Niven and Kao, 2019). To investigate this issue in the context of tables, we developed several systematic logical probes (Gupta et al., 2022a) to evaluate models' reasoning abilities in terms of correct evidence selection (Gupta et al., 2022b), robustness to input perturbation particularly hypothesis, and reasoning ability over counterfactual information.

Any modeling strategy needs to address these problems in the absence of large labeled corpora. Our study in §2.2 tackles some of these challenges and demonstrates the effectiveness of table-specific pre-processing techniques. Furthermore, to address the issue of the presence of distraction information as well as handle correct evidence selection, we causally link the information extraction with the tabular inference problem and introduce the task of *"Trustworthy Tabular Reasoning*" (§2.4). NLP models also extract relevant rows as reasoning evidence in addition to the primary inference task. For the future work (c.f. §3), we propose to address the following questions:

1. Is complete supervision a necessity for evidence extraction in *Trustworthy Tabular Reasoning*? Human annotation of relevant information is a time-consuming and expensive process. In many real-world scenarios, particularly industrial ones, the available resources, i.e., annotation budget and time, are constrained. Thus, it is critical to investigate resource-efficient techniques (e.g., active learning for structured spaces §3.1) for evidence extraction.

2. Can current NLP models reason about dynamic tables, especially ones that contain information that evolves (temporal alterations)? Numerous data pieces inside an entity evolve and change throughout time (c.f. §3.2). For instance, a country's population or the mayor of a city may change regularly. Robust models must consider this and ensure that they can reason well across temporal dimensions.

2 Current Work

2.1 INFOTABS: A Dataset for Tabular Inference[1]

To study semi-structured inference, we created a new dataset called INFOTABS[2], comprising of human-written textual hypotheses based on premises that are extracted from Wikipedia info-boxes. INFOTABS consists of $23,738$ premise-hypothesis pairs, whose premises are based on Wikipedia infoboxes. Figure 1 shows an example table from the dataset with three hypotheses. The dataset contains $2,540$ distinct infoboxes representing a variety of domains. All hypotheses were written and labeled by MTurk workers. The tables have a *title* and two columns, as shown in the example. Since each row takes the form of a key-value pair, we will refer to the elements in the left column as the *keys*, and the right column provides the corresponding *values*.

INFOTABS incorporates several diverse kinds of reasoning adapted from the Glue (Wang et al., 2018) and SuperGlue (Wang et al., 2019) benchmarks, as shown in Figure 2 which are typically missing in previous NLI datasets. For example, in Figure 1, consider the hypothesis sentence H1. In order to determine whether the hypothesis entails the premise, one needs to look up multiple rows (*'Length'* and *'Producer'*), conclude that *'Length'* in Album terms denotes the total length of the album's songs (i.e. Album Singles), and *'46:06'* where the album length is in minutes rather than an hour (using common sense). In addition to the regular training and development sets, to differentiate models' true learning ability from learning spurious correlated patterns in the data (artifacts), we created three challenge test sets. The α_1 set represents a standard test set that is topically and lexically similar to the training data. In the α_2 set, hypotheses are designed to be

[1] This work was published as Gupta et al. (2020) [2] INFOTABS website: https://infotabs.github.io

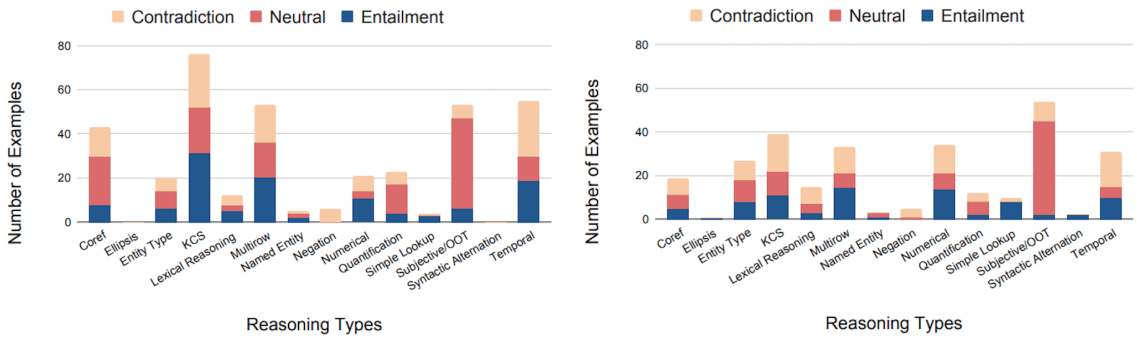

Figure 2: Numbers of examples per reasoning types for a random subset of Development and α_3 data splits of INFOTABS. OOT and KCS are short forms of out-of-table and Knowledge and Common Sense, respectively.

lexically adversarial, and the α_3 tables are drawn from topics not present in the training set; refer to Table 1 for data splits details. For our analysis, we will use all three test sets.

Data split	# tables	# pairs
Train	1740	16538
Dev	200	1800
α_1 test	200	1800
α_2 test	200	1800
α_3 test	200	1800

Table 1: Number of tables and premise-hypothesis pairs for each data split

Model	Dev	α_1	α_2	α_3
Human	**79.78**	**84.04**	**83.88**	**79.33**
Hypothesis Only	60.51	60.48	48.26	48.89
RoBERTa$_{\text{LARGE}}$	75.55	74.88	65.55	64.94
5xCV	$73.59_{(2.3)}$	$72.41_{(1.4)}$	$63.02_{(1.9)}$	$61.82_{(1.4)}$

Table 2: Results of the *Table as a Paragraph* strategy on INFOTABS subsets with RoBERTa$_L$ model, hypothesis-only baseline and majority human agreement. The last row represents the average performances (and standard deviations as subscripts) using models obtained via five-fold cross validation.

We use a universal template[3] to represent the table as a paragraph i.e. *Table as a Paragraph*, with each row representing a sentence. We found that existing state-of-the-art models, e.g., RoBERTa-Large for NLI, underperform on our dataset compared to the majority human agreement performance, suggesting that reasoning about tables can pose a difficult modeling challenge. The penultimate row of Table 2.1 presents the performance of the model trained on the entire training data, while the last row presents the performance of the 5xCV models. The results demonstrate that model performance is reasonably stable to variations in the training set. Table 2.1 also shows the hypothesis-only baseline (Poliak et al., 2018; Gururangan et al., 2018) and human agreement on the labels. [4]. To study the stability of the models to variations in the training data, we performed a 5-fold cross-validation (5xCV). An average cross-validation accuracy of 73.53% with a standard deviation of 2.73% was observed on the training set, which is close to the performance on the α_1 test set (74.88%). In addition, we also evaluated performance on the development and test sets.

2.2 External Knowledge Integration for Tabular Reasoning[5]

We study the below questions in regards to the tabular reasoning problem. How will these models designed for raw text adapt to tabular data? How do we represent data and incorporate knowledge into these models? Can better pre-processing of tabular information enhance table comprehension? In the absence of large labeled corpora, any modeling strategy needs to explicitly address these problems. We examined these challenges and effective pre-processing approaches for addressing them that work well on tabular data [6]:

- **Better table representation**: The table does not explicitly state the relationship between the keys and values. §2.1 suggested the use of a universal template which leads to most sentences being ungrammatical, e.g., "*The recorded of Breakfast in America is 29 March 1998.*". To address this issue, we propose using entity specific templates **(BTR)** by using value entity types DATE or MONEY or CARDINAL or BOOL. The final sentence now become grammatically correct,

[3] The *row-key* of *table-title* are *row-values*. [4] Preliminary experiments on the development set showed that RoBERTa$_L$ outperformed other pre-trained embeddings. We found that BERT$_B$, RoBERTa$_B$, BERT$_L$, ALBERT$_B$ and ALBERT$_L$ reached development set accuracies of 63.0%, 67.23%, 69.34%, 70.44% and 70.88%, respectively. While we have not replicated our experiments on these other models due to prohibitively high computational costs, we expect the conclusions to carry over to these other models as well. [5] This work was published as Neeraja et al. (2021) [6] Knowledge-INFOTABS website: https://knowledge-infotabs.github.io/

e.g., "*Breakfast in America was recorded on March 29th, 1998.*". Furthermore, we also add category-specific information, e.g., "*Breakfast in America is an album.*".

- **Missing implicit lexical knowledge:** This affects interpreting meaning of words like *'less than'*, and *'most of'* in H1 and H2 respectively. Limited training data affects the interpretation of *synonyms*, *antonyns*, *hypernyms*, *Hyponyns*, and *Co-hyponyms* words such as fewer, over, more than, less than, over, under, negations, and others. We find that pre-training on a large Natural Language Inference dataset helps expose the model to diverse lexical constructions and the representation in also now more tuned to the NLI task. So firstly, we intermediately pre-train with MNLI data **(KG implicit)** and then subsequently fine tune on the tabular inference INFOTABS dataset.

- **Presence of distracting information**: Only select rows are relevant for a given hypothesis. For example, the key *'Recorded'* is relevant for the hypothesis H2 and H3 but irrelevant for the hypothesis H3. Furthermore, due to BERT tokenization limit, useful rows in longer tables might be cropped. To handle this we first propose a simple preprocessing solution, Distracting Row Removal **(DRR)**, where we select only rows relevant to the hypothesis. For this, we adopt the Alignment based retrieval algorithm with fastText vectors as detailed in Yadav et al. (2019, 2020). For example, we prune the table with only rows *'Length'* and *'Producer'* for hypothesis H1. We also explore through the lens of evidence extraction and propose better unsupervision and supervision approches, as describe later in §2.4.

- **Missing domain knowledge about keys**: We need to interpret the meaning of the table key in the correct context for this table. In the case of H1, we need to interpret *'Length'* in the album context. For example, here, the length must be interpreted as in *album* context: "*The such of total playtime of all the songs in the record album.*" rather as "*The dimension of the larger side of a portrait.*" as in *painting* context. We append explicit information **(KG explicit)** to enrich the keys. Explicit knowledge helps improve the model's ability to disambiguate the meaning of the keys. We utilize BERT on wordnet examples to get key embeddings, then use BERT to get key embeddings from the premise, and finally select the best match, based on similarity score, and add its definition to the premise.

Premise	Dev	α_1	α_2	α_3
Human	**79.78**	**84.04**	**83.88**	**79.33**
Para	75.55	74.88	65.55	64.94
BTR	76.42	75.29	66.50	64.26
+KG implicit	**79.57**	78.27	71.87	66.77
+DRR	78.77	78.13	70.90	68.98
+KG explicit	79.44	**78.42**	**71.97**	**70.03**

Table 3: Accuracy with the proposed modifications on the Dev and test sets. Here, + represents the change with respect to the previous row. Here the Human and Para results are taken from §2.1

Premise	Dev	α_1	α_2	α_3
Para	75.55	74.88	65.55	64.94
DRR	76.39	75.78	67.22	64.88
KG explicit	77.16	75.38	67.88	65.50
KG implicit	**79.06**	**78.44**	**71.66**	**67.55**

Table 4: Ablation results with individual modifications.

Our proposed effective preprocessing approaches lead to substantial improvements in prediction quality, especially on adversarial α_2 and α_3 test sets as shown in Table 3[7] and 4. The proposed solutions are also applicable to question answering and generation problems with both tabular and textual inputs.

2.3 Systematic Probing for Evidence-Based Tabular Reasoning [8]

Given the surprisingly high accuracies in Table 2.1, especially on the α_1 test dataset, can we conclude that the RoBERTa-based model reasons effectively about the evidence in the tabular input to make its inference? That is, does it arrive at its answer via a sound logical process that takes into account all available evidence along with common sense knowledge? Merely achieving high accuracy is not

[7] Reported numbers are the average over three random seed runs with a standard deviation of 0.33 (+KG explicit), 0.46 (+DRR), 0.61 (+KG implicit), 0.86 (BPR), over all sets. All improvements are statistically significant with $p < 0.05$, except α_1 for BPR representation w.r.t to Para (Original). [8] This work will appear as Gupta et al. (2022a)

sufficient evidence of reasoning: the model may arrive at the right answer for the wrong reasons leading to improper and inadequate generalization over unseen data.

Although "Reasoning" is a multi-faceted phenomenon, and fully characterizing it is almost impossible. However, one can probe for the *absence* of evidence-grounded reasoning i.e. "reasoning failures" via model responses to carefully constructed inputs and their variants. For e.g. there are certain pieces of information in the premise (irrelevant to the hypothesis) when changed, should not impact the outcome, thus making the outcome *invariant* to these changes. For example, deleting irrelevant rows from the premise should not change the model's predicted label. Contrary to this is the relevant information ("evidence") in the premise. Changing these pieces of information should vary the outcome in a predictable manner, making the model *covariant* with these changes. For example, deleting relevant evidence rows should change the model's predicted label to NEUTRAL[9]. Overall, the guiding premise for this (in-/co-)variants perturbation work is:

> *Any "evidence-based reasoning" system should demonstrate expected, predictable behavior in response to controlled changes to its inputs.*

Directly checking for such property there would require a lot of labeled data—a big practical impediment. Fortunately, in the case of tabular semi-strcutured data, the (in-/co-)variants associated with these dimensions allow controlled and semi-automatic edits to the inputs leading to predictable variation of the expected output. This insight underlies the design of probes using which we examine the robustness of the reasoning employed by a model performing tabular inference. We instantiate the above strategy along three specific dimensions and introduce specific probes[10], described below using the running example in Figure 1.

(a.) Avoiding Annotation Artifacts A model should not rely on spurious lexical correlations. In general, it should not be able to infer the label using only the hypothesis. Lexical differences in closely related hypotheses should produce predictable changes in the inferred label. Two possible scenarios arise with a hypothesis alteration, without a change in the premise table, either (a) leads to a change in the label (i.e., the label covaries with the variation in the hypothesis) i.e. label flipping, or (b) does not induce a label change (i.e., the label is invariant to the variation in the hypothesis) i.e. label preserving. For example, in the hypothesis H1 of Figure 1 if the token "less than" is replaced with "more than", the model prediction should change from ENTAIL to CONTRADICT.

To create such probe, we identify a set of reasoning categories and characterize the relationship between a tabular premise and a hypothesis. We use a subset of (a) **Named Entities**: such as *Person, Location, Organisation*; (b) **Nominal modifiers**: nominal phrases or clauses; (c) **Negation**: markers such as *no, not*; (d) **Numerical Values**: numeric expressions representing *weights, percentages, areas*; (e) **Temporal Values**: Date and Time; (f) and **Quantifiers**: like *most, many, every*. to perform controlled changes in the hypotheses.

Although we can easily track these expressions in a hypothesis using tools like entity recognizers and parsers, it is non-trivial to automatically modify them with a predictable change on the hypothesis label. For example, some label changes can only be controlled if the target expression in the hypothesis is correctly aligned with the facts in the premise. Hence, we follow the following strategy: (a) We avoid perturbations involving the NEUTRAL label altogether, as they often need changes in the premise (table) as well. (b) We generate all label-preserving and some label-flipping transformations automatically by leveraging the syntactic structure of a hypothesis and the monotonicity properties of function words like prepositions. (c) We annotate the CONTRADICT to ENTAIL label-flipping perturbations manually.

From the analysis of artifact probe, we show that the model heavily relies on correlations between a hypothesis' sentence structure and its label [11]. Thus, models should be systematically evaluated on adversarial sets like α_2 for robustness and sensitivity. This observation is concordant with multiple studies that probe deep learning models on adversarial examples in a variety of non-tabular tasks such as question answering, sentiment analysis, document classification, natural language inference, etc. (e.g. Ribeiro et al., 2020; Richardson et al., 2020; Goel et al., 2021; Lewis et al., 2021; Tarunesh et al., 2021).

[9] This strategy has been either explicitly or implicitly also employed for recent non-tabular work work (Ribeiro et al., 2020; Gardner et al., 2020). [10] INFOTABS Probing website: `https://tabprobe.github.io` [11] For complete results refer to the publication (Gupta et al., 2022a).

(b.) Evidence Selection A model should use the correct evidence in the premise for determining the hypothesis label. For example, ascertaining that the hypothesis H1 is entailed requires the *Length* and *Producer* rows of Figure 1.

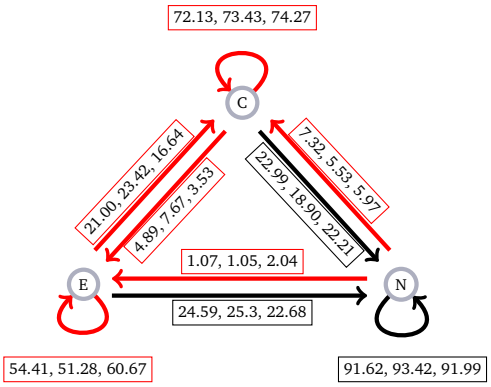

Table 5: Changes in model predictions after deletion of relevant rows. Directed edges are labeled with transition percentages from the source node label to the target node label. The number triple corresponds to α_1, α_2 and α_3 test sets respectively and for each source node, adds up to 100% over the outgoing edges. Red lines represent invalid transitions while **black** lines represent valid transitions.

Datasets	α_1	α_2	α_3	Avg.
ENTAIL	5.14	6.97	6.09	6.07
NEUTRAL	3.9	3.54	5.01	4.15
CONTRADICT	5.94	5.09	6.91	5.98
Avg.	4.99	5.2	6.01	-

Table 6: Percentage of invalid transitions after deletion of irrelevant rows. For an ideal model, all these numbers should be zero.

Dataset	α_1	α_2	α_3	Avg.
ENTAIL	75.41	74.70	77.31	75.80
NEUTRAL	8.39	6.58	8.01	7.66
CONTRADICT	77.02	81.10	77.80	78.64
Avg.	53.60	54.14	54.35	

Table 7: Percentage of invalid transitions following deletion of relevant rows. For an ideal model, all these numbers should be zero.

To better understand the model's ability to select evidence in the premise, we use two kinds of controlled edits: (a) automatic edits without any information about relevant rows, and, (b) semi-automatic edits using knowledge of relevant rows via manual annotation. We define four types of table modifications that are agnostic to the relevance of rows to a hypothesis: (a) *row deletion*, i.e. deleting information, (b) *row insertion*, i.e. inserting new information, (c) *row-value update*, i.e., changing existing information, and (d) *row permutation*, i.e., reordering rows. Each modification allows certain desired (valid) changes to model predictions.[12] We examine below the case of row deletion in detail and refer the reader to the Gupta et al. (2022a) for the others[13].

Row deletion should lead to the following desired effects: (a) If the deleted row is relevant to the hypothesis (e.g., *Length* for H1), the model prediction should change to NEUTRAL. (b) If the deleted row is irrelevant (e.g., *Producer* for H1), the model should retain its original prediction. NEUTRAL predictions should remain unaffected by row deletion. Figure 5 shows the label transitions i.e. model response when relevant rows are deleted.. The fact that even after the deletion of relevant rows, ENTAIL and CONTRADICT predictions don't change to NEUTRAL a large percentage of times (mostly the original label remains unchanged and at other times, it changes incorrectly), indicates that the model is likely utilizing spurious statistical patterns in the data for making the prediction. We summarize the combined invalid transitions for each label for relevant and irreleavant row deletion in Table 6 and Table 7. The large percentage of invalid transitions after relevant row deletion in the ENTAIL and CONTRADICT cases indicates a rather high utilization of spurious statistical patterns by the model to arrive at its answers.

Overall from evidence-selection probing[14], we found the model does not look at correct evidence for correct reasoning and rather leverages spurious patterns and statistical correlations to make predictions. A recent study by Lewis et al. (2021) on non-tabular question-answering shows that models indeed leverage spurious patterns to answer a large fraction (60-70%) of questions.

(c.) Robustness to Counterfactual Changes A model's prediction should be *grounded* in the provided information even if it contradicts the real world, i.e., to counterfactual information. For example, if the month and year of the *Released* date changed to "December" and 1978 respectively, then the model should change the label of H3 in Figure 1 to ENTAIL from CONTRADICT. Since this information about release date contradicts the real world, the model cannot rely on its pre-trained

[12] In performing these modifications, we ensure that the modified table does not become inconsistent or self-contradicting.
[13] For more details examples refer to the publication (Gupta et al., 2022a). [14] For complete results refer to the publication (Gupta et al., 2022a)

knowledge, say from Wikipedia. For the model to predict the label correctly, it needs to reason with the information in the table as the primary evidence. Although the importance of pre-trained knowledge cannot be overlooked, it must not be at the expense of primary evidence.

For this probe, we limit ourselves to modifying only the ENTAIL and CONTRADICT examples. We omit the NEUTRAL cases because the majority of them in INFOTABS involve out-of-table information; producing counterfactuals for them is much harder and involves the laborious creation of new rows with the right information. The task of creating counterfactual tables presents two challenges. First, the modified tables should not be self-contradictory. Second, we need to determine the labels of the associated hypotheses after the table is modified. We use the evidence selection data (probe 1) to gather all premise-hypothesis pairs that share relevant keys. Counterfactual tables are generated by swapping the values of relevant keys from one table to another. In addition, we also generated counterfactuals by swapping the table title and associated expressions in the hypotheses with the title of another table, resulting in a counterfactual table-hypothesis pair, as in the row swapping strategy. This strategy also preserves the hypothesis label similar to row swapping. The above approaches are *Label Preserving* as they do not alter the ENTAIL labels. Counterfactual pairs with flipped labels i.e. *Label Flipped* are also important for filtering out the contribution of artifacts or other spurious correlations that originate from a hypothesis (probe 2). So, in addition, we also created counterfactual table-hypothesis pairs where the original labels are flipped. These counterfactual cases are, however, non-trivial to generate automatically, and are therefore created manually. Additionally, we also developed a interactive annotation platform (Jain et al., 2021) for generating counterfactual tabular instances.

From counterfactual probes [15], we found that the model relies on knowledge of pre-trained language models than on tabular evidence as the primary source of knowledge for making predictions. This is in addition to the spurious patterns or hypothesis artifacts leveraged by the model. Similar observations are made by Clark and Etzioni (2016); Jia and Liang (2017); Kaushik et al. (2020); Huang et al. (2020); Gardner et al. (2020); Tu et al. (2020); Liu et al. (2021); Zhang et al. (2021); Wang et al. (2021) for unstructured text. We also perform an inoculation study (Liu et al., 2019), in which we found that fine-tuning on challenge sets improves model performance on challenge sets but degrades on the original α_1, α_2, and α_3 test sets. That is, changes in the data distribution during training have a negative impact on model performance. This adds weight to the argument that the model relies excessively on data artifacts.

2.4 Evidence Extraction for Trustworthy Tabular Reasoning[16]

As evident from §2.3, existing NLI systems optimized solely for label prediction cannot be trusted. It is not sufficient for a model to be merely *Right* but also *Right for the Right Reasons*. In particular, at least identifying the relevant elements of inputs as the *'Right Reasons'* is essential for trustworthy reasoning[17]. We address this issue by introducing the task of *Trustworthy Tabular Inference*, where the goal is to extract relevant rows as evidence and predict inference labels [18].

To illustrate this task, consider an example from the INFOTABS dataset in Figure 1, which shows a premise table and three hypotheses. The figure also marks the rows needed to make decisions about each hypothesis, and also indicates the relevant tokens for each hypothesis. For trustworthy tabular reasoning, in addition to predicting the label ENTAIL for H1, CONTRADICT for H2 and NEUTRAL for H3, the model should also identify the evidence rows—namely, the rows *Producer* and *Length* for hypothesis H1, *Recorded* for hypothesis H2, *Released* and *Recorded* for hypothesis H3.

Trustworthy Tabular Inference is a table reasoning problem that seeks not just the NLI label, but also relevant evidence from the input table that supports the label prediction. We use T^R, a *subset* of T, to denote the relevant rows or evidence. Then, the task is defined as follows.

$$f(\mathrm{T}, \mathrm{H}) \rightarrow \{\mathrm{T}^R, \mathrm{y}\} \tag{1}$$

In our example table, this task will also indicate the evidence rows T^R of *Producer* and *Length* for hypothesis *H1*, *Recorded* for hypothesis *H2*, and *Released* and *Recorded* for hypothesis *H3*.

[15] For complete results refer to the publication (Gupta et al., 2022a). [16] This work will appear as Gupta et al. (2022b)
[17] We argue that a reasoning system can be deemed trustworthy only if it exposes how its decisions are made, thus admitting verification of the reasons for its decisions. [18] Trustworthy Tabular Inference website: `https://tabevidence.github.io`

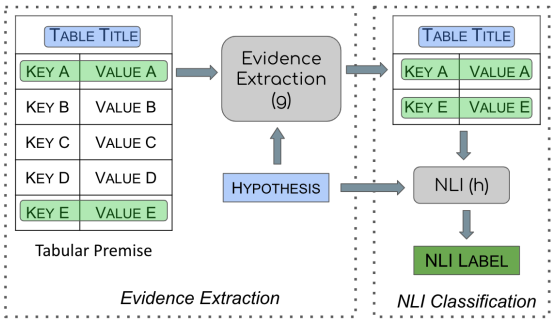

Table 8: High level flowchart showing our approach for evidence extraction based trustworthy tabular inference.

Sampling (Ratio)	α_1	α_2	α_3
Random Negative ($1\times$)	69.42	71.94	54.12
Hard Negative ($1\times$)	80.88	84.37	68.28
No Sampling ($6\times$)	83.76	85.41	71.26
Hard Negative ($3\times$)	**84.49**	**86.58**	**72.61**
Human Oracle	**88.62**	**89.23**	**88.56**

Table 9: F1-scores of supervised evidence extractors on the INFOTABS using various strategy for negative example strategies. Here, Hard Negative are sampled as most similar irrelevant premise table rows using the unsupervised DRR §2.2 approach.

While the notion of evidence is well-defined for the ENTAIL and CONTRADICT labels, the NEUTRAL label requires explanation. To decide on the NEUTRAL label, one must first search for relevant rows (if any), i.e., identify evidence in the premise tables. In fact, this is a causally correct sequential approach. Indeed, INFOTABS has multiple neutral hypotheses that are partly entailed by the table; if any part of a hypothesis contradicts the table, then the inference label should be CONTRADICT. For example, in our example table, the premise table indicates that the album was recorded in 1978, emphasizing the importance of the *Recorded* row for the hypothesis H2. For NEUTRAL examples, we refer to any such pertinent rows as evidence.

Trustworthy inference has an intrinsic sequential causal structure: extract evidence first, then predict the inference label using the extracted evidence data, knowledge/common sense, and perhaps formal reasoning (Herzig et al., 2021; Paranjape et al., 2020). Therefore, we propose a two-stage sequential prediction approach for the task, as shown in Table 8, comprising of an evidence extraction stage, followed by an inference stage. In the evidence extraction stage, the model extracts the necessary information needed for the second stage. In the inference stage, the NLI model uses only the extracted evidence as the premise for the label prediction task.

Notation. The function f in Eq. 1 can be rewritten with functions g and h, $f(.) = g(.), h \circ g(.)$, as

$$f(\mathrm{T}, \mathrm{H}) = \{g(\mathrm{T}, \mathrm{H}) \ , h\left(g(\mathrm{T}, \mathrm{H}), \mathrm{H}\right)\} \tag{2}$$

Here, g extracts the evidence rows T^R subset of T, and h uses the extracted evidence T^R and the hypothesis H to predict the inference label y, as

$$g(\mathrm{T}, \mathrm{H}) \rightarrow \mathrm{T}^R \quad ; \quad h(\mathrm{T}^R, \mathrm{H}) \rightarrow \mathrm{y} \tag{3}$$

To obtain f, we need to define the functions g and h, and a flexible representation of a semi-structured table T. To represent a table T, we use the **B**etter **T**able **R**epresentation (BTR) heuristic from §2.2.

Category	**Evidence Extraction Train Set**	**Evidence Extraction Test Set**	α_1	α_2	α_3
Baselines	WMD	WMD	70.38	62.55	61.33
	No Extraction	No Extraction	74.88	65.55	64.94
	DRR	DRR	75.78	67.22	64.88
Unsupervised	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	DRR (Re-Rank + Top-2$_{(\tau=1)}$)	74.66	67.38	65.83
Supervised	Oracle	Supervised ($3\times$ HN)	77.34	71.15	68.92
Human	Oracle	Oracle	**78.83**	**71.61**	**71.55**
Human NLI	-	-	**84.04**	**83.88**	**79.33**

Table 10: Tabular NLI performance with the extracted relevant rows as the premise. Here, DRR is the distracting row removal approach as discussed in §2.2, WMD represent the word mover distance based approach of INFOTABS and HN represent Hard Negative Sampling. Re-Rank represent reranking using exact key/values matches. Here, τ represent the absolute threshold hyperparameter for varying Top-2 extraction.

For the evidence extraction phase, we explore both unsupervised and supervised evidence extraction approaches over INFOTABS. Our best unsupervised evidence extraction method i.e. SimCSE outperforms a previously developed baseline DRR §2.2 by 4.3%, 2.5% and 5.4% absolute score on

the three test sets, for details refer to Gupta et al. (2022b). For supervised evidence extraction, we annotate the INFOTABS training set (17K table-hypothesis pairs with 1740 unique tables) with relevant rows following the methodology of Gupta et al. (2022a), and then train a RoBERTa$_{\text{LARGE}}$ classifier. The supervised model improves the evidence extraction performance by 8.7%, 10.8%, and 4.2% absolute scores on the three test sets over the unsupervised approach, refer to Table 9. Finally, for the full inference task, refer to table 10, we demonstrate that our two-stage approach with best extraction, outperforms the earlier baseline by 1.6%, 3.8%, and 4.2% on the three test sets, refer to Table 9. Overall, we demonstrated that employing extracted evidence-based inference benefits both interpretability and performance on the downstream tabular inference task.

3 Future Work

3.1 Active Learning over Semi-structure Data

We observe that training supervision for evidence extraction is crucial for Trustworthy Tabular Inference, yet annotating the entire training set is cost-intensive and time-consuming. Furthermore, adding new domain instances with a complete supervision approach incurs high costs in terms of time, computation, and money before any deployable model delivery. In many real-world contexts, particularly industrial ones, both the budget and time for annotation are resource-constrained. Therefore, we ask a natural question: *Is complete supervision necessary for the task of Trustworthy Tabular Inference?*. More specifically, we ask: *Is it conceivable, with resource-efficient minimal annotations, to develop an effective model for evidence extraction and thus Trustworthy Tabular Inference?*.

Random Sampling. One approach to minimise supervision is sampling random instances for the annotation process. However, this strategy has several limitations; as explained (a) there is no guarantee that sampled data will offer theoretically optimal final performance; due to the random nature of the sampling, extraction performance could be non-deterministic, with performance fluctuating between random samples, (b) the approach ignores category and row-key imbalances, which can result in the development of a model incapable of reasoning consistently across all reasoning types, (c) the approach suffers from the problem of instance duplication with similar reasoning, which is a prevalent issue with crowd-sourcing data. E.g., in the INFOTABS dataset, consider the following two instances from *Movie* category in Table 11. Both hypotheses H1 and H2 require similar reasoning and relevant row extraction despite being from two different tables., and (d) the approach disregards the available knowledge derived from unsupervised evidence extraction model predictions. Furthermore, information acquired through inference label prediction without extraction, i.e., from complete table inference, might be used to select challenging examples.

Index	Table Title	Hypothesis	Relevant Rows Keys
H1	Despicable Me	Despicable Me was a box office flop.	Box Office, Budget
H2	Problem Child	Problem Child was a box office disappointment.	Box Office, Budget
H3	Mulan	Mulan lost producers money.	Box Office, Budget

Table 11: Example of Instances from INFOTABS with similar reasoning and same relevant rows-keys.

Optimal Strategy. Thus, a random sampling-based approach will not be the optimal one. An ideal optimal strategy ensures that the instances are diverse, i.e., avoid repetition and have good coverage of categories, row-keys, and reasoning types. An ideal strategy should prefer the selection of more complex instances. E.g., in Table 11, hypothesis H3 should be chosen above hypotheses H1 and H2. H3 is complicated since it requires an additional step of complex second-order reasoning, i.e., "*common sense: producers invest money in movies*" $\rightarrow$ "*table fact: the movie lost money/fails if Movie Budget $>$ it is Box Office Collection*" implying the relevant rows as *'Box Office', 'Budget'*. To get better sampling for resource constraint annotation, we plan to utilize ideas from guided constraint semi-supervision (Chang et al., 2007; Ganchev et al., 2010). We plan to explore the following two directions: (a) To ensure sample diversity, we intend to first replace the TITLE OF TABLE with TABLE CATEGORY and the NAMED ENTITIES with the appropriate NAMED ENTITIES TAGS in the hypothesis sentences[19], and then to utilize an appropriate contextual model to obtain their

[19] replacement ensure that model weigh row-key and sentence intent over varying individual entities which capturing hypothesis similarity.

semantically rich representation. Then, we plan to use Determinantal Point Processes[20] (Kulesza and Taskar, 2010, 2011, 2012) over these vector representations for optimal diverse sampling. (b) The second strategy focuses on leveraging existing prediction knowledge derived from unsupervised or semi-supervised evidence extraction techniques. We plan to improve global table-level instance selection by using quantifiable properties such as mutual entropy, mutual information, prediction confidence, and global sparsity of multiple row-level predictions. Alternatively, one can also utilize strategies from active learning (Settles, 2009) over structured spaces with the available unsupervised or partially supervised model for the next annotation instance selection. Finally, we seek a technique that employs the inference model across the entire table, i.e., no extraction for the sample selection task. If time permits, we intend to explore the online learning context of the sample selection problem.

3.2 Dynamic Temporal Evidence-Based Tabular Reasoning

Numerous components of information about an entity evolve throughout time. E.g., in figure 3 from the Wikipedia infobox of "Joe Biden" one can observe temporal variation in entity information across several relational aspects (keys) such as, e.g., official position, marital status, political affiliation, and others.

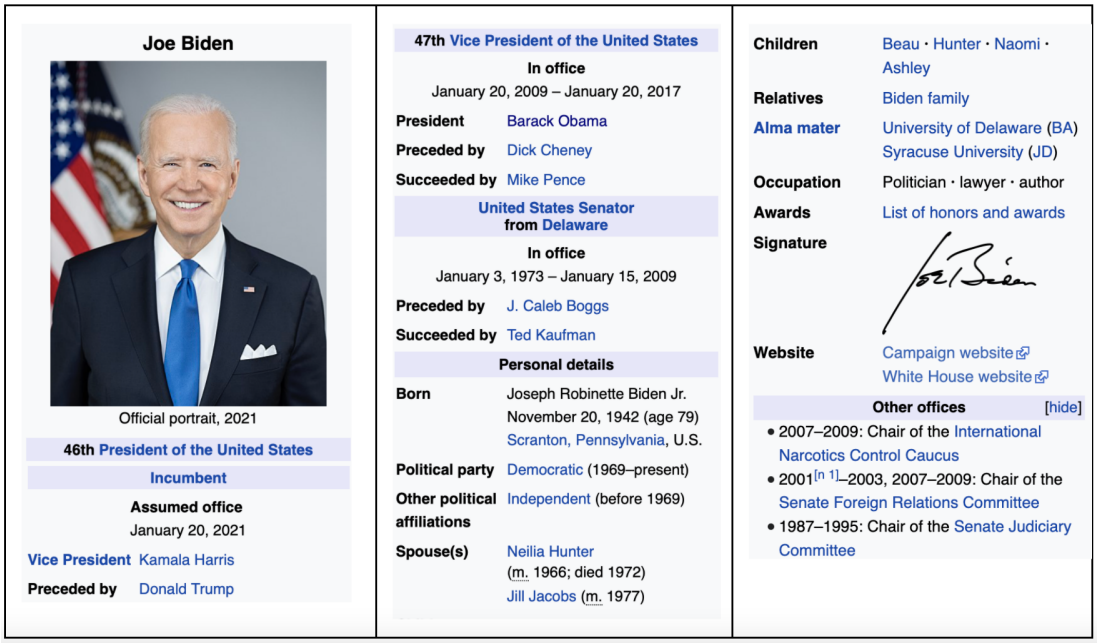

Figure 3: Wikipedia InfoBox of "Joe Biden", the Current President of the United States of America.

Using the information present in Figure 3 one can easily reason over several interesting temporal questions as shown in Table 12[21].

S.no	Question	Answer
1	What was Joe Biden's political affiliation in 1953?	Independent
2	What position did Joe Biden hold in year 2012?	Vice President of USA
3	How many positions did Joe Biden's hold in 2009?	Four
4	How long has Joe Biden chaired the Senate Foreign Relations Committee?	Six

Table 12: Dynamic Temporal Reasoning over Semi-structured InfoBox tables.

Question Types. To handle such temporal variation and challenging questions, an NLP model should be capable of reasoning across time (i.e., temporal alteration). We propose to construct a novel dataset of table-based question answers revolving around temporal questions. The dataset

[20] DPPs provide efficient and precise algorithms for sampling, marginalization, conditioning, and other types of inference. DPPs are used in real-world applications. These include locating diverse sets of high-quality search results, creating informative summaries from diverse sentences extracted from documents, modeling non-overlapping human poses in images or videos, and automatically creating timelines of important news stories. [21] more detailed examples https://bit.ly/3LHMFrx

would contain an infobox table and multiple questions revolving around time. The questions will be broadly of the following two types: (a) the question contain temporal aspect and answer seek relational keys. Q1. is an example of this type. (b) the question contains relational keys and the answer seeks time. Q4 is an example of this type.

Temporal Reasoning Types. We also plan to include varying levels of reasoning difficulty in the questions as follows: (a) mention times in questions that is explicitly mention in the entity tables. (b) mention time in the questions, which can be inferred implicitly from the tables. (c) questions involve reasoning across multiple relational keys across time which vary temporally. (d) mentions relational keys explicitly or implicitly in the question with an answer involving temporal reasoning. (e) mentions temporal relationships such as 'within', 'between', 'before', 'after', etc. and seeks either a relational key or value as the answer. and (f) any other complex temporal reasoning questions.

Data Construction. For dataset construction, we first plan to re-purpose/recast existing temporal questions answering datasets such as (a) Time-Sensitive-QA (Chen et al., 2021b), TORQUE (Ning et al., 2020) entity-specific reading comprehension dataset with time-sensitive questions over paragraph taken from wikipedia pages, (b) TempQA-WD (Neelam et al., 2022), CRONQUESTIONS (Saxena et al., 2021), TempQuestions (Jia et al., 2018a) question answering datasets over knowledge graph embedding with temporal link. In all cases, we plan to replace the Wikipedia paragraph or knowledge graph with the Wikipedia infobox table. We also plan to explore questions form open domain (Zhang and Choi, 2021) and cloze-form (Dhingra et al., 2022) or event-centric (Ning et al., 2018; Wen et al., 2021; Chen et al., 2021a) temporal questing answering datasets. Additionally, we intend to use the Amazon Mechanical Turk platform to crowdsource difficult template queries that are not addressed by recasting. We will also utilize crowdsourcing for manual and automated paraphrasing (Zhao et al., 2021) to further rephrase these templates for additional lexical variation.

Benchmark Models. The purpose of the proposed dataset is to study any model's temporal reasoning and understanding ability in a grounded context of the form of succinct semi-structured data such as entity tables. We intend to investigate temporaly tuned language models trained on knowledge-based question answering datasets like CRONKBQA (Saxena et al., 2021), TEQUILA (Jia et al., 2018b), EXAQT (Jia et al., 2021), OTR-QA(Shang et al., 2021), TempoQR(Mavromatis et al., 2021), and others. If time permits, we also plan to explore methods that incorporate temporal aspects while masked language model pre-training (Dhingra et al., 2022; Logan IV et al., 2021), rather than during fine-tuning on the downstream NLI task.

Other Research Questions. We anticipate that our proposed approach will spark future research on other related questions, including (a) Table Retrieval Based Temporal Question Answering: This task is a natural open-domain extension of the proposed work, where correct tables need to be retrieved before applying temporal reasoning (b) Dynamic temporal variation across varied temporal tables: this setting explore an entity-table whose information is gradually varied across time. E.g. using the infobox table of *Salt Lake City* of *year 2019* and *year 2022* to answer challenging questions as shown in Table 13[22].

S.no	Question	Answer
1	What is the water area of Salt Lake City?	1.3 sq mi (3.3 km2) (Y-2019); 0.47 sq mi (1.22 km2) (Y-2022)
2	Between 2010 and 2020, how much did SLC's population grow?	13283
3	Change is population rank if SLC from 2010 to 2020?	+6
4	What is the city area of SLC in year 2020?	110.81 sq-mi (286.99 km^2)

Table 13: Dynamic Temporal Reasoning over temporally varied Semi-structured InfoBox tables.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven

[22] more detailed examples: https://bit.ly/3qZ5DlA

learning. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 280–287, Prague, Czech Republic. Association for Computational Linguistics.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021a. Event-centric natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021b. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Peter Clark and Oren Etzioni. 2016. My Computer Is an Honor Student — but How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine*, 37(1):5–12.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Kuzman Ganchev, Jo227;o Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(67):2001–2049.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. 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 Shrivastava, Maneesh Singh, and Vivek Srikumar. 2022a. Is my model using the right evidence? systematic probes for examining evidence-based tabular reasoning. *Transactions of the Association for Computational Linguistics*, 10.

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. 2022b. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 2022 Conference of the Association for Computational Linguistics*. 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, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

William Huang, Haokun Liu, and Samuel R. Bowman. 2020. Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 82–87, 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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. TEQUILA: Temporal Question Answering over Knowledge Bases. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1807–1810, New York, NY, USA. ACM.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 792–802.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Alex Kulesza and Ben Taskar. 2010. Structured determinantal point processes. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In *ICML*, pages 1193–1200.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083*.

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.

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

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert L Logan IV, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2021. Fruit: Faithfully reflecting updated information in text. *arXiv preprint arXiv:2112.08634*.

Costas Mavromatis, Prasanna Lakkur Subramanyam, Vassilis N. Ioannidis, Soji Adeshina, Phillip R. Howard, Tetiana Grinberg, Nagib Hakim, and George Karypis. 2021. Tempoqr: Temporal question reasoning over knowledge graphs.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, et al. 2022. A benchmark for generalizable and interpretable temporal question answering over knowledge bases. *arXiv preprint arXiv:2201.05793*.

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.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. 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.

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.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8713–8721.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. 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 6663–6676, Online. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Chao Shang, Peng Qi, Guangtao Wang, Jing Huang, Youzheng Wu, and Bowen Zhou. 2021. Open temporal relation extraction for question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. 2021. Trusting RoBERTa over BERT: Insights from checklisting the natural language inference task. *arXiv preprint arXiv:2107.07229. Version 1*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2021. Logic-driven context extension and data augmentation for logical reasoning of text. *arXiv preprint arXiv:2105.03659. Version 1.*

Haoyang Wen, Yanru Qu, Heng Ji, Qiang Ning, Jiawei Han, Avi Sil, Hanghang Tong, and Dan Roth. 2021. Event time extraction and propagation via graph attention networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 62–73, Online. Association for Computational Linguistics.

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.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double perturbation: On the robustness of robustness and counterfactual bias evaluation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3899–3916, Online. Association for Computational Linguistics.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Mengjie Zhao, Fei Mi, Yasheng Wang, Minglei Li, Xin Jiang, Qun Liu, and Hinrich Schütze. 2021. Lmturk: Few-shot learners as crowdsourcing workers. *arXiv preprint arXiv:2112.07522*.