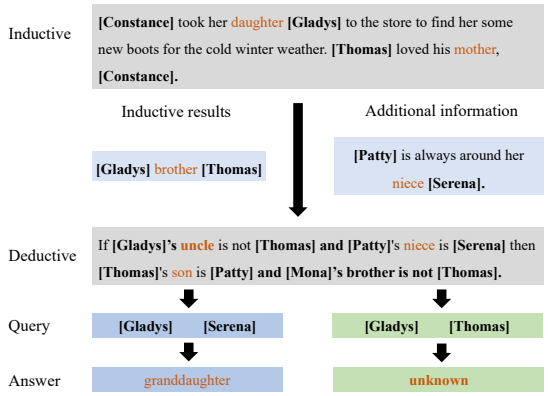
$\mathcal{MTR}$: A Dataset Fusing Inductive, Deductive, and Defeasible Reasoning

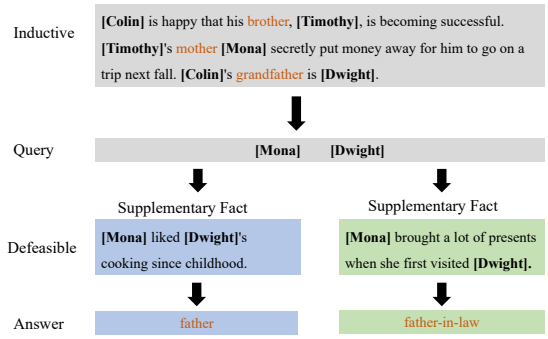
Yitian Li[1,2], **Jidong Tian**[1,2], **Wenqing Chen**[3], **Caoyun Fan**[1,2],
Hao He[1,2‡] and **Yaohui Jin**[1,2‡]
[1]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]State Key Lab of Advanced Optical Communication System and Network,
Shanghai Jiao Tong University
[3]School of Software Engineering, Sun Yat-sen University
{yitian_li, frank92,} @sjtu.edu.cn
chenwq95@mail.sysu.edu.cn
{ fcy3649, hehao, jinyh}@sjtu.edu.cn

Abstract

A long-standing difficulty in AI is the introduction of human-like reasoning in machine reading comprehension. Since algorithmic models can already perform as well as humans on simple quality assurance tasks thanks to the development of deep learning techniques, more difficult reasoning datasets have been presented. However, these datasets mainly focus on a single type of reasoning. There are still significant gaps in the studies when compared to the complex reasoning used in daily life because we can mix and match different types of reasoning unconsciously. In this work, we introduce a brand-new dataset, named $\mathcal{MTR}$. There are two subsets of it: (1)the first is mainly used to explore mixed reasoning abilities and combines deductive and inductive reasoning; (2)the second integrates inductive and defeasible reasoning for detecting non-monotonic reasoning ability. It consists of more than 30k instances, requiring models to infer relations between characters in short stories. Compared with the corresponding single reasoning datasets, $\mathcal{MTR}$ serves as a more challenging one, highlighting the gap in language models' ability to handle sophisticated inference.



(a) An example of $\mathcal{D}$-$\mathcal{MTR}$.



(b) An example of $\mathcal{F}$-$\mathcal{MTR}$.

Figure 1: Examples of different logical combinations in $\mathcal{MTR}$. Named entities are represented by words in bold and in parenthesis, whereas relationships are represented by words in orange.

1 Introduction

Natural language understanding (NLU) has long pursued the goal of working like a human that can perceive information and conducts logical reasoning over knowledge (Du et al., 2022). Deep neural networks (DNNs) have achieved great success recently (Devlin et al., 2019) and have excelled at information perception tasks such as text classification and sentiment analysis (Lee-Thorp et al., 2022; Yang et al., 2019; Schick and Schütze, 2021). However, logical reasoning, which needs to confront a novel environment or complex task, exposes the weakness of DNNs' tendency to make decisions by non-generalizable shortcuts (Du et al., 2022). Although an array of existing datasets are available for exploring different reasoning capabilities of neural networks, such as CLUTRR (Sinha et al., 2019), RuleTakers (Clark et al., 2020) ,and WIQA (Tandon et al., 2019), most of them primarily highlight monotonic logic (Choi, 2022) and a single form of reasoning. For instance, RuleTaker (Clark et al., 2020) and LogicNLI (Tian et al., 2021) only in-

‡ Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10078–10089

clude deductive reasoning, while CLUTRR (Sinha et al., 2019) is exclusively related to inductive reasoning. As a non-monotonic reasoning dataset, δ-NLI (Rudinger et al., 2020) only contains instances with the simplified form of reasoning. These settings bring two problems: (1) The establishment of these datasets does not conform to our daily reasoning habits. (2) They also impact the effectiveness of the evaluation of models. To solve the problems, we explore combining different reasoning forms. We first refer to the theory of the human reasoning process, which shows that induction and deduction are two major monotonic forms to make logical reasoning. Based on the theory, we integrate induction and deduction to generate cases. Besides, most of our day-to-day reasoning is always accompanied by non-monotonic reasoning (Choi, 2022). Psychological theoretical research also shows that human reasoning lacks clarity and does not distinguish between various forms of reasoning in a straightforward manner (Johnson-Laird, 2010b). Different forms of reasoning are interrelated and support each other (Liu et al., 2020; Li et al., 2022). We also explore introducing non-monotonic logic into the dataset.

Especially, inspired by CLUTRR (Sinha et al., 2019) for inductive reasoning diagnosis on relationships, we introduce a new dataset for multiple reasoning combinations, $\mathcal{MTR}$[1](**M**ulti-**T**ype **R**easoning Dataset). $\mathcal{MTR}$ is the semi-automatic extension to CLUTRR and includes two parts, $\mathcal{D}$-$\mathcal{MTR}$ and $\mathcal{F}$-$\mathcal{MTR}$. **Part I** ($\mathcal{D}$-$\mathcal{MTR}$) includes various deductive rules and combines deductive reasoning with inductive reasoning. As a result, $\mathcal{D}$-$\mathcal{MTR}$ involves negation logic in relationship understanding, which prevents the reasoning process from using shortcuts. In practice, we have introduced an additional relation, "unknown", to represent the situation where the relationship cannot be inferred through the given text reasoning. Examples are provided in Figure 1(a). **Part II** ($\mathcal{F}$-$\mathcal{MTR}$) makes the process of inductive reasoning defeasible. We first construct new inductive reasoning stories. These stories have the characteristic that when only given an inductive reasoning story, two compatible relationships between family members can be inferred. As the example in Figure 1(b), both *"father"* and *"father-in-law"* are reasonable without supplementary facts. However, with the additional new fact, this inference tends towards one of the answers. For example, if we subsequently learn that *"someone cooks for him since he was a child"*, the choice of *"father-in-law"* is greatly abandoned.

[1]The dataset will soon be available.

Dataset	Deductive	Inductive	Defeasible
RuleTaker	✓		
LogicNLI	✓		
LogiQA	✓		
ProofWriter	✓		
CLUTRR		✓	
HotpotQA		✓	
QuaRTz		✓	
δ-NLI			✓
$\mathcal{MTR}$(Ours)	✓	✓	✓

Table 1: Comparison with existing reading comprehension datasets and our $\mathcal{MTR}$.

We also experiment on $\mathcal{MTR}$, with several state-of-the-art neural models developed for NLU. Results show that models' performance on $\mathcal{MTR}$ is significantly reduced compared with the one on CLUTRR. This phenomenon is evident that state-of-the-art neural models still lack logical reasoning capabilities in logic-entangling scenarios. Further analysis on $\mathcal{D} - \mathcal{MTR}$ shows that similar inference rules can significantly interfere with models' hybrid inference and models trained on non-inferential order data have better anti-interference ability. During non-monotonic reasoning tests on $\mathcal{F} - \mathcal{MTR}$, neural models cannot benefit from the supplementary facts before answering a defeasible inference query.

2 Background and Related Work

2.1 Reasoning Datasets

Many datasets have been proposed to test the reasoning ability of NLU systems. RuleTaker (Clark et al., 2020) is a dataset known as deductive reasoning. Many neural methods have been developed for this dataset and achieved results when only dealing with single deductive reasoning. ProofWriter (Tafjord et al., 2021) and LogicNLI (Tian et al., 2021) also focus on deductive reasoning but enrich in logical forms. The dataset LogiQA (Liu et al., 2020) also includes multiple types of deductive reasoning. In contrast, many datasets like HotpotQA (Yang et al., 2018), QuaRTz (Tafjord et al., 2019), CLUTRR (Sinha et al., 2019), etc., deal with inductive reasoning over textual inputs. δ-NLI (Rudinger et al., 2020) is a dataset for defeasible inference in natural lan-

guage. However, few datasets include the combination of various reasoning types, which make them difficult to evaluate the logical reasoning ability of models comprehensively.

2.2 Reasoning Definition

Traditional logic have two branches: deduction and induction (Gillies, 1994). These classical logic forms can ensure the certainty of reasoning both syntactically and semantically. But in real-world situations, a clash of knowledge frequently appears (Allaway et al., 2022), introducing uncertainty into the daily reasoning process. Nonmonotonic reasoning is therefore recommended as a crucial artificial intelligence thinking technique (Strasser and Antonelli, 2019; Ginsberg, 1987). Defeasible reasoning is a type of nonmonotonic logic, where logical conclusions are not monotonically true.

Deductive Reasoning is described as applying broad concepts to specific situations (Johnson-Laird, 2010a; Sanyal et al., 2022). Deductive reasoning relies on making logical premises and basing a conclusion around those premises. Starting with a rule, the deduction task is then applied to a real-world scenario. For instance, we can conclude that *"Socrates is mortal"* based on the tenets *"All men are mortal"* and *"Socrates is a man"* (Johnson-Laird, 1999; Heit and Rotello, 2010).

Inductive Reasoning is described as drawing conclusions by going from the specific to the general. It includes the process of making predictions about novel cases based on past experience or observations (Hayes et al., 2010; Lavrac and Dzeroski, 1994). The form of induction task begins with facts about individual cases and then generalizes to a general rule, such as deducing from the facts that *"Swallows can fly"* and *"Orioles can fly"* the consequence that *"All birds can fly"* (Heit, 2000; Hayes and Heit, 2018). So, if given *"Tweety is a bird"*, we can entail that *"Tweety can fly"*.

Defeasible Reasoning is the mode of reasoning where conclusions are modified with additional information (Pollock, 1987). It has been studied by both philosophers and computer scientists (Koons, 2005). The conclusion is not logically sound and could be refuted by fresh information, such as the clarification that *"Tweety is a penguin"* provided in the case above (Lascarides and Asher, 1991).

Statistics	$\mathcal{D}$-$\mathcal{MTR}$	$\mathcal{F}$-$\mathcal{MTR}$
#Instances	30k	2k
Avg. Length	76.6	43.2
Max. Length	189	121
#Hop	≤ 5	≤ 5

Table 2: Statistics of $\mathcal{D}$-$\mathcal{MTR}$ and $\mathcal{F}$-$\mathcal{MTR}$.

3 Dataset

3.1 Overview

To evaluate the models' ability under complex mixed-reasoning scenarios, we create a new dataset $\mathcal{MTR}$, including multi-type reasoning that requires kinship inferring. $\mathcal{MTR}$ is the extension to the existing natural language inductive dataset, CLUTRR (Sinha et al., 2019), and includes the mixed reasoning types of induction, deduction, and defeasibility. As a result, $\mathcal{MTR}$ is a dataset with the expansion of relation types and complexity. Specifically, we add two kinds of logic (deduction and defeasibility) to construct two sub-datasets, $\mathcal{D}$-$\mathcal{MTR}$ and $\mathcal{F}$-$\mathcal{MTR}$. The detailed statistics are summarized in Table 2.

3.2 $\mathcal{D}$-$\mathcal{MTR}$

$\mathcal{D}$-$\mathcal{MTR}$ is the subset that inferring is accomplished by the combination of induction and deduction. Taking Type ① in Figure 2 as an example, we can perceive the relationship chains that "[*son*, Sharon, Christopher], [*aunt*, Christopher, Diana], [*daughter*,Debra,Diana]" from the story. Based on inductive reasoning, we can infer that "[*mother*, Sharon, Debra]". Next, we need to choose the appropriate deduction rule that *"If [Sharon]'s mother is [Debra] then [Debra]'s Sister is [Lois]"*. After combining inductive and deductive reasoning, we can judge the final relationship between [Sharon] and [Lois].

We adopt a semi-automatic method to generate $\mathcal{D}$-$\mathcal{MTR}$ with three steps: 1) logic generation, 2) logic correction and 3) natural language generation. As for the logic generation, we adopt an automatic method to generate each logic expression to ensure the validity of deductive reasoning. Specially, we define a set of logical templates T in advance. It concentrates on diverse first-order logical forms (including conjunction $\wedge$, disjunction $\vee$, negation $\neg$, and implication $\rightarrow$). Then we conduct the knowledge base (KB) that contains all single rules, such as [*grandfather*, X, Y] $\vdash$ [[*father*, X, Z], [*father*, Z, Y]]. In this type of knowledge graph, we present the vector computation method and the accompa-

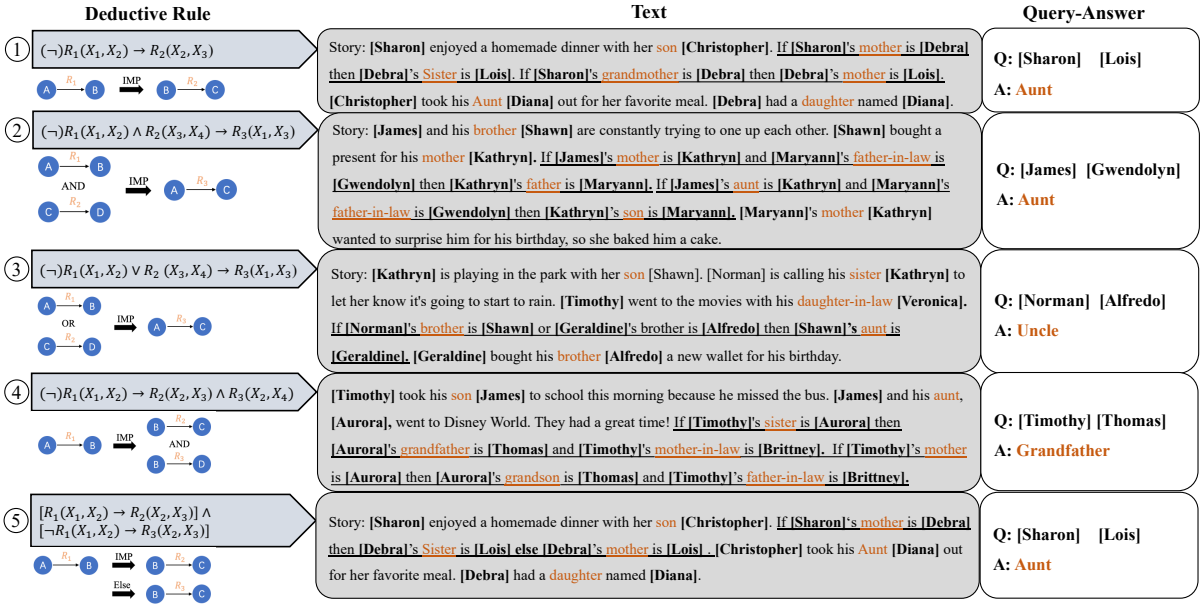


Figure 2: Examples of each type of deductive logical reasoning in $\mathcal{D}$-$\mathcal{MTR}$. Circles with different letters indicate the different entities. Underlined sentences indicate corresponding rules and partially displayed noise rules.

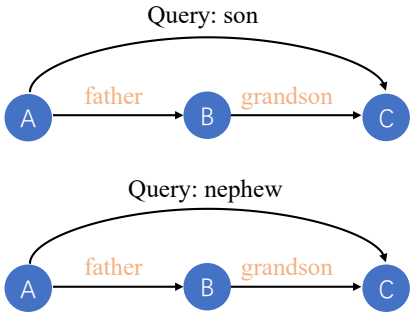
nying subtraction operation. The above example can be changed to the following form: [*father*, Z, Y] = [[*grandfather*, X, Z] − [*father*, Z, Y]]. We design relation checking and final label generation using the relation knowledge base. With regard to the creation of natural language, we first use a rule-based approach to produce the initial language expressions and then perform manual adjustments. Grammatical errors and semantic ambiguities are corrected manually. Figure 2 shows the statistics and representative examples of the reasoning types in our dataset. For example, although Type② and Type③ are only distinguished conjunction(∧) and disjunction(∨), there is a big difference in deductive reasoning and requires a strong logical reasoning ability of the models.

We incorporated different methods formulate the deductive rules to stop the model from processing reasoning through erroneous statistical correlations. On the one hand, we introduce two negations in deductive reasoning. By introducing negation, the model is prevented from drawing relational conclusions from erroneous correlations. (1) Negation words are used to introduce the first type of negation, which is logical negation, such as the fact *"Colin's grandfather is not Dwight."*. (2) We formulate relation contradiction as contradictory statements without negation cues. Relation contradiction events are not identifiable as negations on their own, but demonstrate reversed semantic or pragmatic meaning when paired with their affirmative counterparts (e.g., the fact that *"Colin's grandfather is Dwight."* vs. *"Colin's father is Dwight."*). These negated or contradictory statements shift the relation implications of the original premise in non-trivial ways.

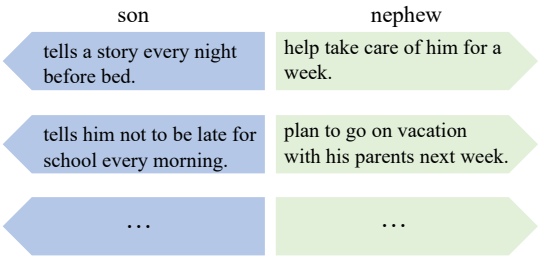
On the other hand, we don't just make up our deductive principles at random. The five rules of deductive logic are all artificially designed. For instance, in Type⑤, we make the reasoning more challenging by allowing both positive and negative derivations existing simultaneously. The model can only carry out further reasoning after determining whether or not the conditions are true. To avoid spurious reasoning brought by a single rule, we add homologous interference rules to each form of reasoning. In practice, we have also introduced an additional relation, "unknown", to represent the situation where the relationship between the two cannot be inferred from given facts or rules. This innovative relation keeps the option of expanding to more unknown relationships while also increasing uncertainty during model inference.

3.3 $\mathcal{F}$-$\mathcal{MTR}$

$\mathcal{F}$-$\mathcal{MTR}$ is designed to evaluate the non-monotonic reasoning ability, specifically combining defeasible and inductive inference in natural language. As shown in Figure 1(b), we find that the results of the same inductive reasoning text are



(a) A defeasible pair where the identical line of reasoning leads to a different conclusion.



(b) Some supplementary facts are used to update the defeasible reasoning

Figure 3: Examples of the defeasible pair and supplementary facts collected for $\mathcal{F}$-$\mathcal{MTR}$.

not unique. Both *"father"* and *"father-in-law"* are reasonable. The likelihood of a particular choice changes when the supplementary fact is supplied, either strengthening or weakening it. In a word, neural models can benefit from the supplementary fact before answering a defeasible inference query.

To generate $\mathcal{F}$-$\mathcal{MTR}$, we first construct defeasible pairs D. The defeasible pairs must satisfy the following conditions, all of which can be derived from the same relationship path. Figure 3(a) shows an example of defeasible pair. Both [*son*, A, C] and [*nephew*, A, C] can be deduced through the same relational path ([[*father*, A, B],[*grandson*, B, C]]). Then, for each defeasible pair, we build supplemental facts $\mathcal{U}$ that is used for update reasoning. We employ three post-graduate students to collect supplemental facts. There are at least four supplementary facts provided for each option to be chosen at random. Specifically, given the premise of a text, the model's conclusion derivation is not unique. Given a supplementary fact $u \in \mathcal{U}$, the model may determine whether a specific fact is less likely to be true or more likely to be true. Given the inference instance in Figure 3(a), the model is unable to distinguish between the defensible pair. When we are given additional information, such as *"tells a story every night before bed."*(see in Figure 3(b)), the model would infer that "son" is most likely true.

We conduct hop extension on inductive inferences to produce the final dataset with defeasible reasoning. The new relational paths are selected from the defeasible pairs D. It can guarantee that the outcomes of the new reasoning is ambiguous. Then, we select one supplementary fact $u \in \mathcal{U}$ at random for each of the different defeasible pairs to make one of the inference directions stronger or weaker.

4 Experiments

4.1 Baselines

We conduct experiments on several natural language understanding systems to systematically measure their reasoning ability. Bidirectional LSTMs (Hochreiter and Schmidhuber, 1997; Graves, 2012; Cho et al., 2014) (with and without attention) are always used to reason on unstructured text. Relation Networks (RN) (Santoro et al., 2017) and Compositional Memory Attention Network (MAC) (Hudson and Manning, 2018) are recently proposed methods, which outperform other systems when dealing with relational reasoning. Pre-trained models also give the current state-of-the-art results on machine reading. In particular, we measure the reasoning ability of BERT (Devlin et al., 2018), as well as a trainable LSTM encoder on top of the pre-trained BERT embeddings. In our task, both BERT and BERT-LSTM (a one-layer LSTM encoder is added on top of pre-trained BERT embeddings) are 12-layered frozen and encode the sentences into 768-dimensional vectors.

4.2 Experimental Setup

The final dataset $\mathcal{D}$-$\mathcal{MTR}$ contains 30K questions split into [25K|5k] questions in the [train|test] folds. During the experiments, we performed sequential and random operations on the data set, specifically referring to arranging the sentences in the input text according to logical reasoning order and illogical order(using "Sequential" and "Random" for abbreviation). We also compare how these models perform on the single inductive reasoning dataset CLUTRR and test in the same way. The final dataset $\mathcal{F}$-$\mathcal{MTR}$ contains 2K questions. We assess the accuracy of the $\mathcal{F}$-$\mathcal{MTR}$ with and without supplementary facts. We consider a model to be correct if it predicts one of the answers without

Data(Accuracy)		Models					
Sequence	Type	BiLSTM-Attention	BiLSTM-Mean	RN	MAC	BERT	BERT-LSTM
Sequential	①	$0.13_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.12_{\pm 0.03}$	$0.08_{\pm 0.02}$	$0.14_{\pm 0.01}$	$0.09_{\pm 0.03}$
	②	$0.23_{\pm 0.02}$	$0.21_{\pm 0.03}$	$0.10_{\pm 0.02}$	$0.24_{\pm 0.01}$	$0.13_{\pm 0.02}$	$0.23_{\pm 0.02}$
	③	$0.13_{\pm 0.02}$	$0.09_{\pm 0.02}$	$0.08_{\pm 0.02}$	$0.10_{\pm 0.03}$	$0.12_{\pm 0.02}$	$0.15_{\pm 0.01}$
	④	$0.46_{\pm 0.02}$	$0.38_{\pm 0.03}$	$0.14_{\pm 0.03}$	$0.37_{\pm 0.05}$	$0.10_{\pm 0.03}$	$0.40_{\pm 0.04}$
	⑤	$0.73_{\pm 0.05}$	$0.73_{\pm 0.06}$	$0.60_{\pm 0.05}$	$0.63_{\pm 0.05}$	$0.17_{\pm 0.02}$	$0.65_{\pm 0.05}$
	Average	$0.34_{\pm 0.03}$	$0.30_{\pm 0.03}$	$0.22_{\pm 0.03}$	$0.29_{\pm 0.03}$	$0.13_{\pm 0.03}$	$0.33_{\pm 0.02}$
Random	①	$0.26_{\pm 0.06}$	$0.23_{\pm 0.06}$	$0.09_{\pm 0.03}$	$0.28_{\pm 0.02}$	$0.11_{\pm 0.01}$	$0.26_{\pm 0.03}$
	②	$0.17_{\pm 0.02}$	$0.16_{\pm 0.02}$	$0.10_{\pm 0.03}$	$0.16_{\pm 0.02}$	$0.10_{\pm 0.02}$	$0.18_{\pm 0.02}$
	③	$0.10_{\pm 0.04}$	$0.10_{\pm 0.04}$	$0.08_{\pm 0.02}$	$0.09_{\pm 0.01}$	$0.10_{\pm 0.02}$	$0.09_{\pm 0.03}$
	④	$0.36_{\pm 0.03}$	$0.31_{\pm 0.03}$	$0.10_{\pm 0.02}$	$0.32_{\pm 0.03}$	$0.08_{\pm 0.01}$	$0.32_{\pm 0.04}$
	⑤	$0.64_{\pm 0.06}$	$0.62_{\pm 0.06}$	$0.49_{\pm 0.03}$	$0.55_{\pm 0.04}$	$0.15_{\pm 0.03}$	$0.62_{\pm 0.05}$
	Average	$0.31_{\pm 0.04}$	$0.28_{\pm 0.04}$	$0.17_{\pm 0.03}$	$0.28_{\pm 0.02}$	$0.11_{\pm 0.02}$	$0.29_{\pm 0.03}$
CLUTRR	Average	$0.61_{\pm 0.08}$	$0.59_{\pm 0.08}$	$0.54_{\pm 0.07}$	$0.61_{\pm 0.06}$	$0.30_{\pm 0.07}$	$0.56_{\pm 0.05}$

Table 3: Results on $\mathcal{D}$-$\mathcal{MTR}$ and CLUTRR. "Sequential" means that we train on $\mathcal{D}$-$\mathcal{MTR}$ in the logical order of inference input and test on the sequential test set. On the contrary, "Random" means that we train on $\mathcal{D}$-$\mathcal{MTR}$ in the random logical order of inference input and test on the random test set.

providing any additional information. We adopt a similar setting as Sinha et al. (2019) during training. Specially, all models were trained for 40 epochs with Adam optimizer with a learning rate of 1e-3. We train our models with a batch size of 8. All experiments were run 5 times with random data classification.

4.3 Main Results

Total Accuracy. Table 3 illustrates the performance of different models on $\mathcal{D}$-$\mathcal{MTR}$ and CLUTRR. Overall, we find that the BiLSTM-Attention baseline outperforms other models across most testing scenarios (0.34 on "Sequential", 0.31 on "Random", and 0.61 on CLUTRR). In contrast, BERT does not build reasoning on relational texts without fine-tuning. Under the same models, the performance of $\mathcal{D}$-$\mathcal{MTR}$ (both "Sequential" and "Random") is much lower than that of CLUTRR. It can demonstrate that dealing with multiple reasoning types is significantly more difficult than dealing with single logic. Next, we compare the results of $\mathcal{D}$-$\mathcal{MTR}$ between "Sequential" and "Random". Relational reasoning is more difficult for the models to conduct on non-reasoning order text. Models struggle with Type② and Type③, indicating a deficit in handling conjunction $\wedge$ and disjunction $\vee$. The performance on Type⑤ is all considerably above average. The results on Type① are worse than our expectation. Compared with other types, Type① does not contain complex combinations of first-order logic in deductive reasoning, but it causes great difficulties to models. In the follow-up, we will undertake a more detailed analysis.

Accuracy	(w/o) Supplementary Facts	
Models	w	o
BiLSTM-Attention	0.04	0.09
BiLSTM-Mean	0.03	0.08
RN	0.02	0.04
MAC	0.03	0.08
BERT	0.05	0.03
BERT-LSTM	0.08	0.07

Table 4: Results on $\mathcal{F}$-$\mathcal{MTR}$. We train on the $\mathcal{D}$-$\mathcal{MTR}$ sequentially and test on $\mathcal{F}$-$\mathcal{MTR}$ with and without supplementary facts. "w" indicates that the input texts are supplemented with additional information. "o" indicates that the input texts lack extra facts.

Table 4 illustrates the performance of different models on $\mathcal{F}$-$\mathcal{MTR}$. We consider a model to be correct if it predicts one of the answers on $\mathcal{F}$-$\mathcal{MTR}$ with no supplementary facts associated to the input texts. As a result, we can consider it a single inductive reasoning problem. Results show that models trained on $\mathcal{D}$-$\mathcal{MTR}$(both "Sequential" and "Random") do not have the ability to transfer to $\mathcal{F}$-$\mathcal{MTR}$(o). This demonstrates that simply adding deductive reasoning to the data set does not increase the model's inductive reasoning abilities, but rather interferes with them.

When we compare the accuracy with and with-

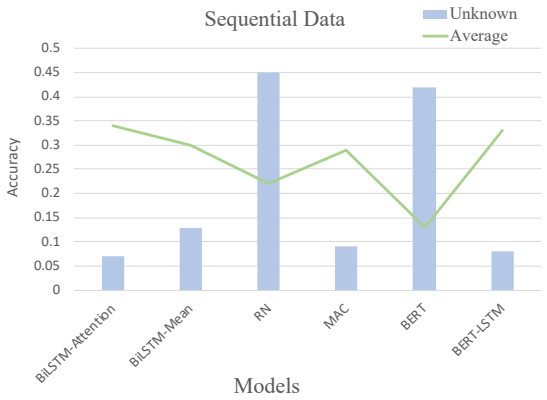
out the supplemental facts, we find that models can be loosely classified into two groups. The first category includes models lacking BERT, such as BiLSTM-Attention, BiLSTM-Mean, RN, and MAC, whose performance on $\mathcal{F}$-$\mathcal{MTR}$ with supplemental facts is roughly half that without. These models do not deal with defeasible reasoning. Supplementary facts used to aid defeasible reasoning actually hinder rather than help model reasoning. However, supplementary facts can help with the rest's models. Although there is still a performance discrepancy when compared to $\mathcal{D}$-$\mathcal{MTR}$, it can be seen that supplementary facts enhance the defeasible inference. This phenomenon supports the notion that pre-trained language models contain a plethora of information (Petroni et al., 2019). This knowledge assists the models in distinguishing differences between defeasible pairs.

4.4 Further Analysis

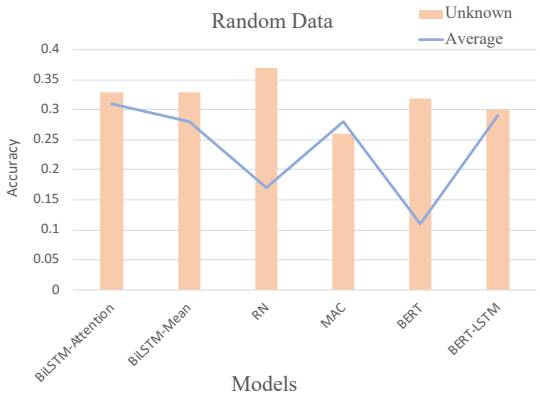
In this section, we provide further analysis of our designed dataset. On the one hand, we present additional insight as to why the virtual label "unknown" is introduced into $\mathcal{D}$-$\mathcal{MTR}$. On the other hand, more experiments are conducted to investigate the unexpected deductive rule (Type①).

Analysis of "unknown". As shown in Figure 1(a), the "unknown" label indicates that the relationship between the two cannot be deduced from the known material. Situations that are unknown or cannot be reasoned about are common in everyday life. Therefore, it is critical to complement the space of the relation. We can summarize two effects of the "unknown": 1) "unknown" provides more accurate relation information for model training, thereby effectively suppressing the impacts of spurious correlations caused by dataset bias; 2) "unknown" makes the diagnostic scenarios more complete and complex, so it can better distinguish the relation reasoning abilities of different models.

As shown in Figure 5, we examine the models' accuracy on the new label "unknown" and compared it to the overall dataset average. On "Sequential" data (Figure 5(a)), most models are unable to reason about this new label and perform much worse than the average. Judging the conclusion of the "unknown" requires the models to exclude all inferable relations. "RN" and "BERT" do poorly on the full dataset, but deduce most of the answers as "unknown"(BERT: from 0.13 on average up to



(a) Performance of different models when trained on "Sequential" $\mathcal{D}$-$\mathcal{MTR}$.

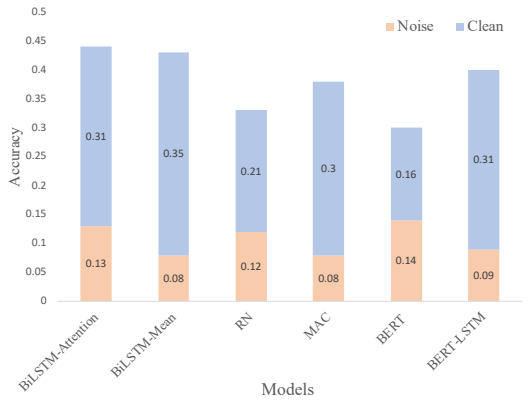


(b) Performance of different models when trained on "Random" $\mathcal{D}$-$\mathcal{MTR}$.

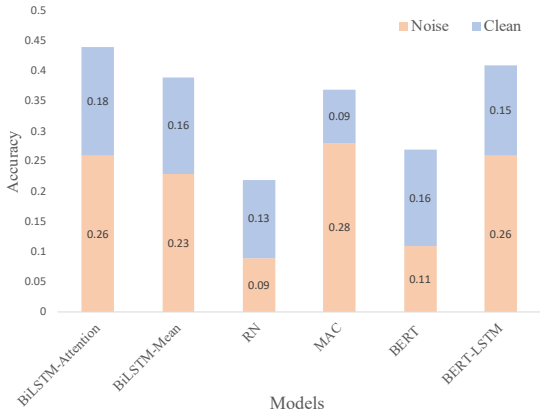
Figure 4: The different accuracy of "unknown". The bar in the figure represents the average accuracy rate, and the line represents the accuracy rate of the "unknown" label.

0.42 on "unknown"). This demonstrates that the two models have not fully developed their reasoning abilities. On "Random" data(Figure 5(b)), the performance of the models on the new labels is significantly improved (BiLSTM-Attention from 0.07 up to 0.33, MAC from 0.09 up to 0.26) and remains close to the average. Comparing "Sequential" data with "Random" data, the former calls for a more robust level of model reasoning. Tests demonstrate that models trained on more complicated datasets perform better on new labels. This is implicit evidence to support that the "unknown" demands more precise reasoning abilities from the model.

Analysis of Type①. To further understand the results on Type① in Table 3, we perform the analysis of why models produce surprising results. The form of Type① is defined as "$(\neg)R_1(X_1, X_2) \rightarrow R_2(X_2, X_3)$". It only includes negation($\neg$) and



(a) Results on "Sequential" $\mathcal{D}$-$\mathcal{MTR}$.



(b) Results on "Random" $\mathcal{D}$-$\mathcal{MTR}$.

Figure 5: Further results on Type ①. In both "Sequential" and "Random" $\mathcal{D}$-$\mathcal{MTR}$, we test on two different Type ① test datasets. "Noise" represents the result on Type①, where we retain deductive inference rules that have the same form but are unrelated to reasoning. "Clean" represents the improvement of accuracy after removing the interference rules.

implication($\rightarrow$) but is almost the hardest type to handle in "Sequential". We discover that models tend to mistakenly select the deductive principles that are seemingly similar. When there are just two classes of first-order logic, noise is considerably more deceiving. By contrasting rules that account for noise versus those that do not, we shall illustrate the propositions(shown in Figure 5).

After using the data without noise rules, the accuracy is improved, as indicated by the blue "Clean" portion. On Type① dataset free of noise and models without retraining, all models exhibit noticeable performance gains. In partial, models with more inference capability also perform better after eliminating noise(such as BERT-LSTM and BiLSTM-Attention improve accuracy by 0.31 after removing

[Josephine] and her father **[Cruz]** like to spend the holidays together.

[Cruz] will often invite his grandson **[Norman]** to join them.

Supplementary fact: **[Josephine]** said to visit **[Norman]**'s father next time.

Query: **[Josephine]** **[Norman]**

Model (w) : father

Model (o) : son

Defeasible Pair: son / nephew

Figure 6: Case study of BiLSTM-Attention on $\mathcal{F}$-$\mathcal{MTR}$. Model(w) and Model(o) are the model's prediction results with and without additional supplemental facts, respectively. The word in the gray background has the model's attention.

interference rules in "Sequential" $\mathcal{D}$-$\mathcal{MTR}$). Comparing the results in Figure 5(a) and Figure 5(b), we discovered that the performance improvement of the models trained on "Random" data is significantly less than that trained on "sequential" data.

4.5 Case Study

To further understand the defeasible inference process of the model, we perform a case study on $\mathcal{F}$-$\mathcal{MTR}$. A comparison between the prediction made by BiLSTM-Attention with and without additional facts is shown in the situation in Figure 6. Two possible relations can be deduced from the existing text in the absence of the supplementary fact. BiLSTM-Attention can successfully predict one of the correct relations (son).

When we provide the new fact *"[Josephine] said to visit [Norman]'s father next time."*, the model provides the wrong prediction "father". This means that the model cannot capture the fact that [Josephine] is not the "father" of [Norman], which is even affected. However, it is easy for us to rule out the "son" relationship between [Josephine] and [Norman] from the provided fact. We examine the model's inference process and discover that the model focuses on the erroneous relationship "father" in the fact. This shows that the model does not capture the information of the entire sentence, but focuses on the part. It also demonstrates how far behind humans in terms of complicated reasoning state-of-the-art neural models perform.

5 Conclusion

In this paper, we propose $\mathcal{MTR}$, a large-scale logical reasoning dataset including deductive, in-

ductive, and defeasible reasoning. It is a more complex relational inference dataset with a mixture of various inferences. In addition to testing the reasoning capacities of state-of-the-art neural models, our dataset helps to re-examine some deficiencies in the research of logical artificial intelligence in the era of deep learning NLP. The results demonstrate that even the most advanced machine readers lag well below human ability.

Limitations

There are two limitations: (1) Although $\mathcal{MTR}$ include three types of reasoning types (deductive, inductive ,and defeasible reasoning), we only focus on relation reasoning task. For other tasks, it is also necessary to construct more datasets with the fusion of multiple reasoning types. (2) Our primary focus remains monotonic reasoning, however, the combined reach of deduction and induction is only the tip of the iceberg of human reasoning (Choi, 2022). This also inspires us to focus on more non-monotonic reasoning and more logical combinations.

Acknowledgements

This work was supported by the National Key Research and Development Program (2022YFC3303102) and Shanghai Pujiang Program (21PJ1407300).

References

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen R. McKeown, Doug Downey, and Yejin Choi. 2022. Penguins don't fly: Reasoning about generics through instantiations and exceptions. *CoRR*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *IJCAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Yilun Du, Shuang Li, Joshua B. Tenenbaum, and Igor Mordatch. 2022. Learning iterative reasoning through energy minimization. In *ICML*.

Donald Gillies. 1994. A rapprochement between deductive and inductive logic. *Bull. IGPL*.

Matthew L Ginsberg. 1987. Readings in nonmonotonic reasoning.

Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Brett K Hayes and Evan Heit. 2018. Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Brett K Hayes, Evan Heit, and Haruka Swendsen. 2010. Inductive reasoning. *Cognitive science*.

Evan Heit. 2000. Properties of inductive reasoning. *Psychonomic Bulletin & Review*.

Evan Heit and Caren M Rotello. 2010. Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*.

Phil Johnson-Laird. 2010a. Deductive reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Philip N Johnson-Laird. 1999. Deductive reasoning. *Annual review of psychology*.

Philip N Johnson-Laird. 2010b. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

Robert Koons. 2005. Defeasible reasoning.

Alex Lascarides and Nicholas Asher. 1991. Discourse relations and defeasible knowledge. In *29th Annual Meeting of the Association for Computational Linguistics*.

Nada Lavrac and Saso Dzeroski. 1994. Inductive logic programming. In *WLP*.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. 2022. Fnet: Mixing tokens with fourier transforms. In *NAACL*.

Yitian Li, Jidong Tian, Wenqing Chen, Caoyun Fan, Hao He, and Yaohui Jin. 2022. To what extent do natural language understanding datasets correlate to logical reasoning? A method for diagnosing logical reasoning. In *COLING*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *IJCAI*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *EMNLP-IJCNLP*.

John L Pollock. 1987. Defeasible reasoning. *Cognitive science*, 11(4):481–518.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *EMNLP*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. Fairr: Faithful and robust deductive reasoning over natural language. In *ACL*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *EMNLP-IJCNLP*.

Christian Strasser and G. Aldo Antonelli. 2019. Nonmonotonic logic. In *The Stanford Encyclopedia of Philosophy*.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *ACL/IJCNLP*.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. Quartz: An open-domain dataset of qualitative relationship questions. In *EMNLP-IJCNLP*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *EMNLP-IJCNLP*.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *EMNLP*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NIPS*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section Limitation

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 and 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All sources are open sources.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3 and Section 4.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
All artifacts in this work are general and public.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 5.

C ☑ Did you run computational experiments?

Section 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We only consider public language models in this work.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
All packages are common used in NLP, such as Transformers and Pytorch.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.