

Teaching Probabilistic Logical Reasoning to Transformers

Aliakbar Nafar
Michigan State University
nafarali@msu.edu

Kristen Brent Venable
Florida Institute for Human
and Machine Cognition
bvenable@ihmc.org

Parisa Kordjamshidi
Michigan State University
kordjams@msu.edu

Abstract

In this paper, we evaluate the capability of transformer-based language models in making inferences over uncertain text that includes uncertain rules of reasoning. We cover both Pre-trained Language Models (PLMs) and generative Large Language Models (LLMs). Our evaluation results show that both generations of language models struggle with reasoning over uncertain text. We propose a novel end-to-end fine-tuning approach, Probabilistic Constraint Training (PCT), that utilizes probabilistic logical rules as constraints in the fine-tuning phase without relying on these rules in the inference stage. To assess the effectiveness of PCT, we utilize the related corpora and, additionally, create a new and more challenging benchmark that, unlike the previous ones, uses instance-specific rules. Our study demonstrates that PCT improves the transformer-based language model’s intrinsic reasoning and makes their probabilistic logical reasoning process more explicit and explainable. Furthermore, PCT equips these models to effectively handle novel situations, including higher reasoning depth, new domains, and complex probabilistic structures.

1 Introduction

Language models have demonstrated high performance across a wide range of Natural Language Processing (NLP) tasks (Liu et al., 2019) which in the case of Large Language Models holds even in zero-shot setting (Chen, 2023). However, they struggle to reason over uncertain text involving logical probabilistic rules (Saeed et al., 2021; Jin et al., 2023). This is confirmed by the reported poor results in arithmetic reasoning when using transformers (Mishra et al., 2022) which is required for probabilistic logical reasoning. Additionally, logical probabilistic inference requires coherent step-by-step reasoning. However, PLMs’ evaluation of various question-answering (QA) benchmarks shows they produce contradictory results

that violate the expected steps of reasoning, such as following transitivity or symmetry rules (Asai and Hajishirzi, 2020). This has led to the development of hybrid approaches, where reasoning tasks are outsourced to Neuro-Symbolic engines, bypassing the need for reasoning by transformers (Zhang et al., 2023). To overcome these limitations, we embed probabilistic reasoning into transformers by imposing the rules of logical probabilistic reasoning as constraints during their training phase.

There are only a few research efforts dealing with uncertainty in text. Understanding logical and uncertain rules in natural language form has been investigated in recent research on question answering (Clark et al., 2020; Saeed et al., 2021), and there have been several attempts to teach transformers how to follow these rules (Asai and Hajishirzi, 2020; Faghihi et al., 2023). While incorporating hard logical rules is undoubtedly important and is still being investigated, in the real world, most of the external knowledge and rules involve uncertainty. For example, only a small fraction of the logical rules in DBpedia can be deemed certain (Saeed et al., 2021). Inference over text that includes uncertainty concerning facts, relations, and rules is required in many natural language comprehension tasks. For example, scientific content often utilizes hedges to express the measure of certainty in factual statements (Pei and Jurgens, 2021; National Academies of Sciences et al., 2017).

A related but different challenge is the explainability of the solutions provided by transformer-based language models. Without the capability of providing the underlying components and steps necessary to answer a question, a Language Model’s reasoning remains inexplicable even when it accurately answers a question (Clark et al., 2019). In this paper, we propose a method that forces the transformer to follow coherent reasoning steps to answer the final question, as shown in Table 1, yielding a more explainable model. This feature

RuleBERT	RuleTaker-pro
(Fact 1) David is a cousin of Ann. (Fact 2) Mike is a child of Ann. (Rule 1, 0.90) If A is a spouse of B and C is a child of B, then C is a child of A. (Rule 2, 0.15) If A is a cousin of B, then A is a spouse of B.	(Fact 1) Dave is big. (Fact 2) Erin is sad. (Rule 1) Usually, If someone is big then they are green. (Rule 2) Normally, If someone is green then they are round. (Rule 3) Seldom, If someone is sad then they are round.
(Query) Mike is a child of David.	(Query) Dave is round.
Required Steps of Reasoning to Answer	
Fact 1 (1.00) & Rule 2 (0.15) $\implies$ Fact 3: David is a spouse of Ann. (0.15) (Inferred) Fact 3 (0.15) & Fact 2 (1.00) & Rule 1 (0.90) $\implies$ Fact 4: Mike is a child of David. (0.135) (Inferred) Answer: 0.135	Fact 1 (1.00) & Rule 1 (0.90) $\implies$ Fact 3: Dave is green. (0.90) (Inferred) Fact 3 (0.90) & Rule 2 (0.80) $\implies$ Fact 4: Dave is round. (0.72) (Inferred) Answer: 0.72
Approach: Converting Probabilistic Reasoning Steps to Equality Constraints	
Constraint 1: $P(\text{Fact 1}) * 0.15 = P(\text{Fact 3})$ Constraint 2: $P(\text{Fact 3}) * P(\text{Fact 2}) * 0.90 = P(\text{Fact 4})$	Constraint 1: $P(\text{Fact 1}) * 0.90 = P(\text{Fact 3})$ Constraint 2: $P(\text{Fact 3}) * 0.80 = P(\text{Fact 4})$

Table 1: Left column: an example from RuleBERT with two facts and two rules. Right column: an example from RuleTaker-pro with two facts and three rules. The reasoning steps required to infer the Query and the constraints applied in these steps are shown in the bottom rows.

is an inherent property and a byproduct of our usage of probabilistic logical constraints and Neuro-Symbolic modeling in our approach.

In this paper, to deal with reasoning over uncertain text, we look into a problem setting that involves calculating the probability of a given hypothesis (Query) based on a provided context that includes *linguistic expression* of probabilistic logical rules and facts. The underlying reasoning is probabilistic logical inference. We utilize two QA datasets: RuleBERT (Saeed et al., 2021) and our newly developed RuleTaker-pro, created to include context-specific rules. Table 1 shows examples of our datasets and their required reasoning steps to answer the Query. We convert the reasoning steps to equality constraints (shown in the Approach section of Table 1) and impose these constraints to ensure consistency of the outputs with the rules during the training of PLMs but not inference. Despite the simplicity of the reasoning patterns in our approach, we will show the transferability of learning to more complex structures. In summary, our contributions are as follows:

1) We propose a new approach, Probabilistic Constraint Training (PCT), that explicitly imposes probabilistic reasoning rules during PLM fine-tuning. This approach provides an effective level of abstraction to the models to generalize and transfer reasoning under uncertainty to new domains and to more complex depths of reasoning. 2) We develop a novel evaluation benchmark for probabilistic reasoning over text with context-specific uncertain

rules whose probabilities can not be captured from the training data and must be extracted from the text.¹ 3) We conduct thorough experiments comparing our constraint-based fine-tuning approach with LLMs and show the superiority of our technique.

2 Related work

Previous works mostly looked into the integration of crisp logic (Saha et al., 2020; Tafjord et al., 2021). The earlier work on QA with probabilistic rules in the text is RuleBERT (Saeed et al., 2021), which serves as the baseline for our comparative study. While RuleBERT pioneers this field and introduces Weighted binary cross-entropy loss to incorporate probabilistic learning in transformers, it lacks a mechanism to follow the probabilistic reasoning steps explicitly. Additionally, our experiments revealed that the rules in textual form in this dataset are not properly utilized by the models (see Section 6.1), which prompted us to introduce RuleTaker-pro with context-specific rules.

Reasoning Steps. Explicit elucidation of reasoning steps in QA models has been central in recent literature. Saha et al. improve PLMs’ reasoning by mapping their output to an inference graph, necessitating the model to learn its nodes and edges. While Tafjord et al. utilize T5 to create an inference path, this and similar studies have focused on using non-probabilistic logical rules. In several other related works, the reasoning for QA is approached

¹The code and dataset are available at [GitHub](#).

by generating an output that follows a predefined formal language for theorem proving given the logical rules, which is a very different approach from ours (Wang and Deng, 2020; Polu and Sutskever, 2020; Tafjord et al., 2021). Wu et al. introduce reasoning in LLMs by generating intermediate reasoning steps as an extra output. However, we enable PLMs to incorporate this reasoning in training with no additional output.

Constraints. Our approach’s primary contribution is incorporating probabilistic constraints in the loss function. While various studies incorporate logical constraints into the loss function (Nandwani et al., 2019; Li et al., 2019; Asai and Hajishirzi, 2020; Ribeiro et al., 2019; Faghihi et al., 2023; Guo et al., 2020), no work has explored the application of probabilistic constraints in this context to date.

Neuro-Symbolic Methods. Central to our approach is the implementation of an end-to-end model, ensuring the transferability of our model to various domains without the need to modify the model’s architecture or decision processes. In contrast, numerous studies in this field rely on a **pipeline** approach, often incorporating a **Neuro-Symbolic engine**. Zhang et al. proposes a framework in which Transformers extract the factual knowledge in the text. Consequently, a symbolic engine conducts the reasoning inference. More general approaches use deep neural methods to process the input and use either an existing engine (Manhaeve et al., 2018) such as Problog (De Raedt et al., 2007) or define a language to create a logical structure as an inference engine (Li et al., 2023).

3 Background

3.1 Problem Definition

We focus on the challenge of performing probabilistic logical reasoning within a QA task where a set of facts F , a set of rules R , and a hypothesis h are provided in a **textual** context. While these rules, facts, and hypothesis are provided **only** in their **textual form** as a part of the input to the task, we have their formal information as a part of the metadata. For example, fact $Big(Dave)$ and the rule $Spouse(A, B) \ \& \ Child(C, B) \rightarrow Child(C, A)$ would be given as input in forms: “Dave is big.”, and “If A is a spouse of B and C is a child of B, then C is a child of A.”, respectively. The facts and hypothesis consist of factoids that define properties for an entity “Has_Property(Entity)” or relations between two entities “Relation(Entity1, Entity2)”.

The rules have the form $(p_1, p_2, \dots, p_n) \rightarrow q, Pr$, where p_i represents a premise fact, q is a new inferred fact, and Pr is the probability of the rule. In RuleBERT rules, Pr is not directly mentioned in the rule’s text and must be learned from the data (or extracted from the metadata to be used in the loss during training), while in RuleTaker-pro, Pr is mentioned in the form of adverbs of uncertainty. q ’s probability is computed as the rule probability multiplied by the premise facts probabilities. If the premise facts are mentioned in the context, they would be certain and have a probability of 1.00; otherwise, if they are inferred facts, their probability is derived. The objective is to utilize F and R to infer a probability between 0 and 1 as our task output, which indicates the probability of a given hypothesis h . For example, h =“Sara and John are cousins” obtains a probability of 0.20 by the model.

3.2 Base Model

The backbone of our model is RoBERTa Large, supplemented by two linear layers and a sigmoid activation function applied to its classifier token (CLS). The model takes the textual representations of facts and rules (context) and hypothesis as input formatted as [CLS] text(R)+text(F) [SEP] text(h) [SEP]. Subsequently, it predicts a probability for the given hypothesis.

The LLM models that we use as baselines for comparisons are GPT3.5 and GPT4 (Brown et al., 2020). Due to the high cost of fine-tuning LLMs, we limit our experiments to zero-shot and few-shot. Input comprises a task explanation, text(R)+text(F), and text(h). The explanation instructs the model about the objective and output format, either “True”, “False” (corresponding to a probability greater or less than 0.5), or the hypothesis probability (between 0.0 and 1.0).

3.3 Deep Learning with Logical Constraints

Among the research focused on constraint integration within neural models, we opt for the class of methods that incorporate constraint violation in the loss function during training without altering the model’s architecture (Nandwani et al., 2019; Li et al., 2019; Faghihi et al., 2023). In general, to employ the logical and symbolic constraints in deep models, they must be converted into soft logic for the sake of differentiability. Usually, three main approaches are used for this conversion: Product, Gödel, and Łukasiewicz (Li et al., 2019). For instance, the logical rule, $(p_1, p_2, \dots, p_n) \rightarrow q$, using

the Product surrogate, is written as follows,

$$\min(1, P(q)/[P(p_1) * P(p_2) * \dots * P(p_n)]), \quad (1)$$

where $P(p_i)$ is the probability of the fact p_i . We can express the enforcement of this implication’s truth as follows,

$$|1 - \min(1, P(q)/[P(p_1) * P(p_2) * \dots * P(p_n)])| = 0, \quad (2)$$

where $|\cdot|$ denotes the absolute value. These methods of constraint conversion are defined for logical constraints and **do not directly apply to probabilistic reasoning rules**, which is why we will introduce a novel method of constraint integration for our goal of enforcing probabilistic reasoning.

4 Training with Probabilistic Constraints

We aim to develop a model capable of following probabilistic reasoning steps to infer the probability of a given hypothesis. These reasoning steps for the examples in Table 1 are outlined in the *Required Steps of Reasoning to Answer* row. In each step, a combination of facts and a rule results in a new *intermediate inferred fact* until the final hypothesis is inferred. These steps are formulated as constraints, and our proposed model is trained to adhere to them by incorporating them into the loss function. The *Approach* row of Table 1 shows examples of the reasoning steps’ conversion into constraints in which the probabilities assigned to facts must follow the rule definition. For instance, if Fact 1 and Rule 1 result in a new fact, Fact 1’s probability ($P(\text{Fact 1})$) multiplied by Rule 1’s probability must be equal to the inferred fact’s probability ($P(\text{inferred Fact})$). In the upcoming subsections, we will explain the process of formulating and utilizing constraints.

4.1 Constraint Integration

We formulate the probabilistic reasoning as obeying a set of constraints derived from probabilistic inference calculations, based on an assumed probabilistic network. We distinguish between *Simple* and *Complex* probabilistic reasoning patterns based on their underlying inference network. A probabilistic reasoning pattern is *Simple* if any deducible fact can be drawn from it via only a single reasoning path. The examples provided in Table 1 are *Simple* because “Dave is round.” can be inferred only from Rule 2 and Fact 3 and Fact 3 can only be inferred from Fact 1 and Rule 1. On the other hand, a *Complex* reasoning encompasses at least one fact that can be deduced from two or more

different rules (reasoning paths). By altering the second fact from “Erin is sad” to “Dave is sad”, we create a *Complex* example because it enables inference of “Dave is round” from Fact 2 and Rule 3 as well. Our focus lies primarily on formulating the simple version of probabilistic reasoning for defining constraints. The *Complex* examples are still incorporated in our datasets and used during training and testing. Later, we investigate how our proposed model can also generalize over the *Complex* probabilistic networks.

Given a *Simple* network, our model executes probabilistic inference for the rule $(p_1, p_2, \dots, p_n) \rightarrow q, Pr$ by multiplying the probability of premise facts by the probability of the rules to obtain the probability of the inferred fact. Formally, the model should fulfill the constraint,

$$|P(q) - P(p_1) * P(p_2) * \dots * P(p_n) * Pr| = 0. \quad (3)$$

Our unique definition of constraint constitutes the key novelty of our approach (see Table 1 for Examples of constraints). To satisfy this constraint, the left side of the above equation should approach zero. Note that while this constraint guarantees adherence to the probabilistic rules, it might not ensure the best results on the end task accuracy, and this remains subject to experimentation.

4.2 Training and Inference

Training To generate the constraints for each dataset example, we use the chains of probabilistic reasoning that include the paths of inference for every inferable fact (available in the dataset metadata; see section 5). Examples of these constraints can be found at the bottom of Table 1. We denote the violation from each constraint as C_i , a scalar value that ranges from 0 to 1, that is, the left-hand side of Equation 3. Our training objective centers on minimizing the violation of these constraints. We initiate the process with warm-up iterations on the original QA task to train the model. Following this, we continue the training while adding the constraint violation losses to the primary loss.

There are multiple methods of incorporating constraints into loss. We utilize a training algorithm inspired by (Nandwani et al., 2019), designed for logical constraints, which we alter to apply to our probabilistic constraints. This method keeps the underlying architecture of the model the same, allowing us to transfer this model to other domains. It also assigns Lagrangian Multipliers λ to each rule, which signifies its difficulty during training.

While there are variations and heuristics for including constraint violation in the loss, such as (Li et al., 2019), we found the employed version a more principled way of implementing the optimization objective. We explain our definition of constraints in this method and our unique way of formulating them as a part of the loss, but the details of the rest of the optimization algorithm are not our contribution and, thus, are not discussed here. We refer the reader to see Appendix A.6 for details of the training algorithm. As per the methodology outlined in (Nandwani et al., 2019), we apply the dual formulation of the objective as follows,

$$Loss = TaskLoss + \sum_{i=1}^m \lambda_j * C_i, \quad (4)$$

where “TaskLoss” denotes the primary task loss aiming to minimize the predicted probability error for the hypothesis. The new additional term is the constraint violation loss used in its dual form with Lagrangian multipliers, λ_j , where j is the index of rule j used in constraint violation i (C_i). m is the number of selected constraints. λ_j is adjusted during training and ultimately indicates a rule’s propensity to violation. Consequently, as training progresses, the loss function predominantly impacts the rules with the highest accumulated λ_j .

Inference During inference, the model receives the context that includes textual rules and facts, while the formal rules and constraints that were employed during training, are not available to the model. We expect the model to learn to obey the rules that were utilized in the loss function during training. This ensures the model’s generalizability and transferability across various domains.

5 Dataset Creation

Motivation RuleTaker-pro is created to address some of the shortcomings of the RuleBERT dataset. RuleBERT (Saeed et al., 2021) is built using about 100 rules with fixed probabilities that are applied to many examples in the dataset. The probabilities of these rules are extracted from an external source and remain constant for all examples in the dataset. However, we want a dataset with example-specific rules to make the required reasoning more realistic. For example, the probability of two married people being cousins in the context of one culture is high, while it is close to zero in another or, in the medical domain, the prevalence or mortality of a disease

varies depending on gender or location (Zirra et al., 2023; Menotti et al., 2023).

Rule Generation We developed RuleTaker-pro by modifying RuleTaker’s crisp logical rules $(p_1, p_2, \dots, p_n) \rightarrow q$ (with Pr equal to 1.0) to include probabilities while the rest of the context remains unchanged (examples shown in the right side of Table 1). We leverage a Gaussian random generator to produce probabilities. The mean and variance of the Gaussian generator depend on the depth of reasoning to ensure a balanced dataset with a mean probability of 0.50 and an equal number of answers above and below 0.5 probability. After assigning probabilities to the rules, we use Problog (De Raedt et al., 2007), a probabilistic logical inference tool that facilitates the encoding of probabilistic facts and rules, to compute the probability of the hypothesis. The resulting rules are similar to RuleBERT rules $(p_1, p_2, \dots, p_n) \rightarrow q, Pr$. See Appendix A.2 for details of data creation and distribution, which demonstrates its robustness.

Adverbs of Uncertainty In the context, we include the probability of the rule as an adverb of uncertainty like *Usually*, *Normally*, and *Seldom* with associated probabilities of 0.90, 0.80, and 0.15, respectively. A key difference between RuleTaker-pro and RuleBERT is including instance-specific rules. For example, the rule “If A is a cousin of B, then A is a spouse of B.” from RuleBERT will always have the probability of 0.15 in all the examples. However, in RuleTaker-pro, the same rule may hold different probabilities depending on the adverb assigned to it in different instances. A rule such as “Usually, if someone is big, then they are green.” carries a probability of 0.90 in one context, while “Seldom, if someone is big then they are green.” carries a probability of 0.15 in some other context. Given this difference, the model has to extract the rules from each context and can not use the information learned about the rules from the training data. See Appendix A.1 for more details.

Metadata Metadata about the inference of all facts and their depths are in the dataset and will be used to create constraints which would be used to train our model in PCT during training but **are not directly used during training or inference**. Notably, ambiguity and cycles have already been removed from the RuleTaker dataset for the logical rules and are not an issue in our dataset, as confirmed by our ProbLog solver. In addition, 20% of examples in RuleTaker had a *Complex* inference

architecture, a ratio which we will keep as well.

6 Experiments

In this section, we address four questions using our synthesized RuleTaker-pro and the RuleBERT datasets: **Q1**. How do textual rules affect probabilistic reasoning (6.1)? We will also discuss the baseline results this section. **Q2**. To what extent does the baseline language model improve with PCT concerning probabilistic reasoning and intermediate inferred facts (6.2)? We also include the ablation study to investigate the impact of various losses and datasets on our approach using multiple metrics. **Q3**. Can we transfer the probabilistic reasoning capabilities of the language model when pre-trained with PCT(6.3)? **Q4**. How do LLMs compare to fine-tuned BERT-based models (6.4)?

Evaluation Metrics. We use several performance measures following (Saeed et al., 2021). Binary Accuracy (BA) deems predictions correct if ground truth and predicted probability both fall under or over 0.5. The CA25, CA10, and CA1 require the predicted probability to be in a window of ± 0.25 , ± 0.10 , and ± 0.01 of the ground truth, respectively. (Saeed et al., 2021) applies CA10 and CA1 metrics to dataset splits with isolated rules, while BA is used for all reasoning depths for datasets involving all the rules. For comparison, we use BA for RuleBERT, but we thoroughly evaluate RuleTaker-pro using all relevant criteria. We use an extra metric, CS, to measure soft **Constraint Satisfaction** that deems the constraint (defined in Equation 3) satisfied if the following inequality holds:

$$|P(q) - P(p_1) * P(p_2) * \dots * P(p_n) * Pr| < Threshold. \quad (5)$$

This means that the difference between the predicted and calculated probability of an inferred fact, based on premise facts, must be less than a threshold. This threshold is 0.01 for CS1, 0.10 for CS10, and 0.25 for CS25.

6.1 Q1: Effect of Rules in Textual Format

RuleBERT Firstly, we investigate whether RoBERTa utilizes the text of the rule in the RuleBERT dataset by keeping and removing them from the context in two different experiments. For example, if we remove the textual rules in Table 1, the input will only include Facts 1 and 2. We report the results of these two settings in Table 2, where columns indicate the maximum depth of reasoning in training (M1-M5), and rows correspond to

the reasoning depth of testing (D1-D5). We omit M0 as depth 0 does not use any rules, making it irrelevant to our investigation of PCT. We observe that the accuracy improves across most models and depths when the rules’ text is excluded, suggesting that RoBERTa is **not using** it, and including it may even add unnecessary complexity. Thus, we conjecture that in RuleBERT dataset, RoBERTa can implicitly learn the probabilities of these rules from the facts and hypothesis in training data alone without using the textual rules explicitly.

Roberta With Text of the Rules					
	M1	M2	M3	M4	M5
D1	76.9	79.8	79.9	70.7	64.9
D2	77.5	77.8	76.6	70.4	65.4
D3	78.4	76.9	76.2	78.8	71.6
D4	76.2	73.4	72.4	78.2	73.8
D5	77.1	73.0	69.6	77.5	78.1
Roberta Without Text of the Rules					
D1	76.8	82.0	82.2	83.6	82.1
D2	75.4	78.8	78.2	80.0	78.5
D3	77.9	80.6	80.6	82.8	80.6
D4	75.0	76.2	77.2	79.6	77.0
D5	78.4	75.2	78.7	79.6	76.7
Roberta + PCT					
D1	79.1	81.7	82.4	84.1	81.1
D2	78.5	79.7	77.3	80.9	77.7
D3	79.8	83.4	81.9	86.2	82.2
D4	77.4	81.4	80.2	85.1	81.3
D5	80.1	84.3	84.3	86.1	83.6

Table 2: BA results of RoBERTa fine-tuned on RuleBERT. Columns indicate the maximum depth of reasoning in training (M1-M5), and rows correspond to the reasoning depth of testing (D1-D5). The results are shown for three different training settings: Roberta With Text of the Rules, Roberta Without Text of the Rules and Roberta + PCT.

Our baseline differs from (Saeed et al., 2021). This discrepancy arises from our approach of freezing 22 transformer layers for faster training and more fine-tuned hyper-parameters, which yield superior accuracy at higher depths (We use the same loss function, Weighted binary cross-entropy). Moreover, we also train our models with (Saeed et al., 2021) original setting, and again, the text of the rules did not yield any positive impact on the performance (see Appendix A.4.1 for details).

RuleTaker-pro We compare baseline results for the RuleTaker-pro dataset using Cross-Entropy (CE) and MSE loss functions for CA1 and CA10 metrics in Table 3. Here, the Weighted binary cross-entropy was abandoned due to underperformance on RuleTaker-pro. The models are trained with maximum depths 1, 2, 3, and 5 (max), as these

are the depths provided in the original RuleTaker training data. However, the testing is done on all depths 1 to 5, and their average accuracy is shown according to CA1 and CA10. *CS* also averages over all depths. Though MSE excels in CA10, it underperforms in CA1 and CS1, especially when trained at higher depths. Our investigation into multiple cases indicates that MSE’s low CS1 results from the minor MSE approximation errors at lower depths, magnified at higher depths when multiplied along the chain of probabilities.

Loss	Metric	M1	M2	M3	Mmax
CE	CA1	38.2	38.3	20.4	33.8
	CS1	47.8	35.7	16.2	20.7
MSE	CA1	30.3	32.2	26.1	26.0
	CS1	25.2	14.8	14.4	12.9
CE	CA10	46.4	49.6	49.9	53.2
	CS10	52.2	44.9	35.6	38.2
MSE	CA10	58.1	62.7	66.6	74.8
	CS10	45.1	34.4	32.8	33.3

Table 3: Baseline RoBERTa’s results for the RuleTaker-pro dataset using Cross-Entropy (CE) and MSE loss functions for CA1 and CA10 metrics. The models are trained with maximum depths 1, 2, 3, and 5 (max). The average accuracy of questions of all depths is shown according to CA1 and CA10. *CS* shows constraint satisfaction as an average over all depths.

Given our goal of achieving exact inference probabilities following the path of reasoning, CA1 is a more relevant measure for PCT evaluation. Also, given CE’s higher CA1 and CS1 performance, we will focus mainly on CE and CA1’s results which are detailed in Table 4. Detailed results for all losses and depths for metrics CA1, CA10, BA, MSE, and L1 are available in Appendix A.10.

Unlike RuleBERT, RuleTaker-pro uses example-specific rules, **requiring the text of the rules** to determine the answer. Without rules, the predictions of our model are not better than random guesses. In RuleTaker-pro, we initially generated probabilistic rules by including the probability in the text, such as "With the probability of 15%, if someone is green, then they are sad". However, we also considered using adverbs of uncertainty (Farkas et al., 2010) instead of numbers, changing the rule to "Seldom, if someone is green, then they are sad". Adverbs of uncertainty improved the models in Dev BA by 0.5%-2%, thus we followed this approach in RuleTaker-pro creation (see Appendix A.1).

D/M	RoBERTa			
	M1	M2	M3	Mmax
Total	38.2	38.3	20.4	33.8
D1	56.0	52.7	29.6	43.7
D2	36.4	38.2	20.3	32.8
D3	29.3	31.3	14.9	28.3
D4	27.4	28.5	14.0	27.1
D5	24.9	26.7	14.7	28.2
CS1	47.8	35.7	16.2	20.7
D/M	RoBERTa + PCT			
	M1	M2	M3	Mmax
Total	38.0	39.5	41.1	37.6
D1	53.3	50.8	50.5	46.9
D2	37.4	40.4	42.2	37.0
D3	26.4	32.9	36.0	32.4
D4	26.5	31.9	33.9	31.8
D5	23.3	30.4	33.4	31.4
CS1	44.9	42.6	34.5	35.2

Table 4: Results of RoBERTa fine-tuned on RuleTaker-pro with CE loss, according to CA1 metric. Columns indicate the maximum depth of reasoning in training (M1-Mmax), and rows correspond to the reasoning depth of testing (D1-D5). The bottom section shows the improved results after the incorporation of PCT.

6.2 Q2: Effectiveness of PCT

RuleBERT Table 2 displays the impact of PCT on improving RuleBERT’s accuracy over the baseline results of RoBERTa, especially at deeper depths. These results are for the baseline without the text of the rules (results with the rule’s text yield similar outcomes; see Appendix A.4.2). Using PCT, the CS25 accuracy of intermediate inferred facts increases from an average of 50% to over 90%. Increasing the constraint satisfaction of intermediate inferred facts works synergistically with the accuracy of the model by compelling the model to reason, thus, enhancing it, especially at deeper depths. Appendix A.5 includes more details about inferred intermediate facts.

RuleTaker-pro By deploying PCT in RuleTaker-pro, we observe a similar trend to RuleBERT. As illustrated in Table 4, by incorporating exact probabilities into the constraints, PCT improves the accuracy of CA1 in most models. Another place where PCT shows improved generalization is when it is used to train the models at lower depths, i.e., 2 and 3, and tested at higher depths. This shows that the reasoning learned with PCT is transferred to higher depths. However, at depth 1, due to the limited number of applicable constraints, the change in accuracy is minor. Similar to RuleBERT, we observe a sharp increase of about 50% in the CS in all the models trained with PCT.

Error Analysis. Our findings indicate that im-

provements in constraint consistency are not always proportionate to improvements in accuracy. This discrepancy is prevalent in nearly all tasks involving constraints, as evidenced by related studies (Ribeiro et al., 2019). Notably, to maintain the consistency of outputs, the model might yield incorrect results. Incorporating PCT encouraged the model to output lower probabilities than the baseline model, thus reducing the magnitude of the constraint loss. For instance, in the model trained at depth 3 with PCT, the average output probabilities for all the test dataset questions declined from a baseline of 52% to 45%. When the model is trained with depth 1 with PCT, the constraint satisfaction decreases, likely due to its reduced ability to accurately process questions with a higher reasoning depth. In short, while the best results are achieved when both CS and CA increase, a high CS does not invariably guarantee a corresponding increase in CA. See Appendix A.8 for detailed examples.

6.3 Q3: Transferability Analysis

Experiments in Section 6.2 highlighted the effectiveness of PCT in transferring *reasoning* from a model trained at lower depths to answer questions at higher depths. Here, we evaluate the transferability of PCT from different perspectives.

Transferring Reasoning From Simple to Complex Examples. As highlighted in Section 5, 20% of the inference questions in RuleTaker-pro have the *complex* architecture. Table 5 presents our models’ performance on *simple* and *complex* questions separately, with the models predictably faring better on the former. Employing CE+PCT increases accuracy for both question types, making the difference between them negligible. This suggests that the models can do probabilistic reasoning even in *complex* instances. However, for MSE and MSE+PCT models, the difference between the performance over different question types remains substantial. Using PCT along with cross-entropy loss in the CE+PCT model was more effective in learning probabilistic reasoning because PCT directs the model to output the exact probability values that do not violate the rules. However, the MSE model does not see the same benefit due to cascading errors in the approximated probabilities of the inferred facts, as discussed in Section 6.2. In the case of MSE, adding PCT still improves accuracy.

Domain Transfer. We evaluated the transferability of the probabilistic reasoning and constraint

	CE			CE+PCT		
	M2	M3	Mmax	M2	M3	Mmax
S	39	20	34	41	40	37
C	34	18	32	36	38	36

	MSE			MSE+PCT		
	M2	M3	Mmax	M2	M3	Mmax
S	33	27	27	36	37	35
C	24	19	20	27	28	30

Table 5: RuleTaker-pro results on Simple (S) and Complex (C) examples trained with Cross Entropy and MSE, before after addition of PCT.

satisfaction capabilities to another domain by training our model on RuleTaker-pro with CE+PCT and fine-tuning it on RuleBERT. This transfer direction is selected due to the superior constraint satisfaction of the model that was trained on the RuleTaker-pro dataset. We compare the RuleBERT baseline with two transfer learning approaches: 1) Pre-training RoBERTa on the RuleTaker-pro dataset with a simple CE loss (Augmented Data) to ensure the improvements are not the result of increased data alone, 2) Pre-training RoBERTa on RuleTaker-pro dataset with CE+PCT loss (Transfer Learning of PCT), aiming to understand the specific impact of pre-training with PCT. The findings, detailed in Table 6, show that only lower-depth results improved by data augmentation, while higher depths and overall accuracy improved by transferring from CE+PCT. Transferring from the CE+PCT also increased CS measures for depths 2, 3, and 5 by about +4, +14, and +7, respectively. In contrast, using Augmented Data did not result in any changes in CS measures.

6.4 Q4: LLM Results

To evaluate LLMs, we add instructions and examples (for few-shot settings) to their prompts. The LLM results for RuleTaker-pro for CA1 are shown in Table 7. We observe that even GPT3.5 with few-shot examples and GPT4 fall short of RoBERTa’s accuracy. GPT4 with few-shot is not included in the table since adding few-shot examples to GPT4 or using COT did not improve but hurt our model. A similar outcome is reported on a different dataset (Shi et al., 2022) where incorporation of COT either marginally helped the model or hurt its accuracy at different depths of reasoning for multi-hop spatial reasoning (Yang et al., 2023). We believe that COT can potentially improve the LLM results, but it requires a significant time investment in prompt engineering and

	Baseline RoBERTa		
	M2	M3	M5
D2	77.8	76.6	65.4
D3	76.9	76.2	71.6
D4	73.4	72.4	73.8
D5	73.0	69.6	78.1
	Augmented Data		
	M2	M3	M5
D2	76.8	80.6	83.4
D3	75.9	83.2	81.6
D4	70.4	76.4	74.8
D5	68.0	72.6	67.1
	Transfer Learning of PCT		
	M2	M3	M5
D2	84.8	84.6	72.4
D3	84.9	82.2	72.6
D4	84.4	77.4	73.8
D5	86.0	66.6	81.1

Table 6: Improvements in the binary accuracy (BA) and constraints satisfaction of RuleBERT models in Table 2 after transfer learning from RuleTaker-pro. The results include the Baseline RoBERTa, the Augmented Data model that is trained on RuleBERT and then finetuned on RuleTaker-pro and Transfer Learning of PCT.

example selection. LLMs are undermined even more after we add PCT and improve RoBERTa’s results. The gap in accuracy becomes even wider if we use the CA10 metric, where the CA10 accuracies remain almost the same as in CA1. This indicates that if the LLM cannot predict the exact probability, its prediction will not be even close to the correct answer. The results of LLMs on the RuleBERT dataset are as poor as a random baseline in all settings, even with the addition of COT for GPT4. See Appendix A.9 for the details of prompt instructions, RuleTaker-pro CA10, and RuleBERT results. We further discuss the underperformance of LLMs on these and similar datasets in the same section of the appendix.

	RoBERTa	GPT3.5	GPT3.5*	GPT4
D1	44	28	41	41
D2	33	20	26	27
D3	28	23	25	26
D4	27	18	20	17
D5	28	18	20	21

Table 7: LLM results on RuleTaker-pro for CA1 and CA10 metrics. * indicates using few-shot examples. The chosen RoBERTa model is M5 trained with CE since it performs the best regarding CA1.

7 Conclusion and Future Work

Addressing the problem of reasoning over uncertain rules in textual format, we create a new dataset, RuleTaker-pro, extending the limited resources for studying this problem. We investigate how uncer-

tain rules can be represented in the text and used by the learning models. We propose a novel approach that explicitly uses the rules of probabilistic reasoning as constraints in the loss. This approach improves the performance and reasoning of the backbone language models. Our experiments on LLMs have revealed that they struggle to perform probabilistic reasoning in zero-shot and few-shot scenarios, despite their impressive capabilities in solving other NLP tasks. Our future objective is to develop models that utilize the text of the rules more effectively and transfer their reasoning abilities to more realistic QA domains featuring uncertainty and more advanced structures of probabilistic reasoning. Also, it is worth exploring prompt engineering methods for Large Language Models to ease the use of uncertain text and inference on them.

Acknowledgements

This project is supported by the National Science Foundation (NSF) CAREER award 2028626 and partially supported by the Office of Naval Research (ONR) grant N00014-20-1-2005 and grant N00014-23-1-2417. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation nor the Office of Naval Research. We thank all reviewers for their thoughtful comments and suggestions.

Limitations

One limitation of our work is the fixed structure of the rules in our datasets, which limits the model’s transferability to other domains with more open forms of explaining probabilistic rules. Another limitation is that we take a small step to formalize probabilistic reasoning over text. However, this does not mean the outcome language models are fully capable of language understanding and reasoning. Finally, running our models, based on RoBERTa large while possible, is computationally expensive, limiting their usage with all our different settings. This is exacerbated when it comes to utilizing Large Language Models that, in their current state, are very expensive to use even in zero-shot and few-shot settings.

References

- Akari Asai and Hannaneh Hajishirzi. 2020. [Logic-guided data augmentation and regularization for consistent question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Wenhu Chen. 2023. [Large language models are few\(1\)-shot table reasoners](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. [Transformers as soft reasoners over language](#). *CoRR*, abs/2002.05867.
- Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. 2007. Problog: A probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, page 2468–2473, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Hossein Rajaby Faghihi, Aliakbar Nafar, Chen Zheng, Roshanak Mirzaee, Yue Zhang, Andrzej Uszok, Alexander Wan, Tanawan Premisri, Dan Roth, and Parisa Kordjamshidi. 2023. [Gluecons: A generic benchmark for learning under constraints](#).
- Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. [The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text](#). In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.
- Quan Guo, Hossein Rajaby Faghihi, Yue Zhang, Andrzej Uszok, and Parisa Kordjamshidi. 2020. [Inference-masked loss for deep structured output learning](#). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2754–2761. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models.
- Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Sriku-mar. 2019. [A logic-driven framework for consistency of neural models](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.
- Ziyang Li, Jiani Huang, and Mayur Naik. 2023. [Scallop: A language for neurosymbolic programming](#). *Proc. ACM Program. Lang.*, 7(PLDI).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. [Deepprolog: Neural probabilistic logic programming](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Alessandro Menotti, Paolo Emilio Puddu, Hanna Tolonen, and Anthony Kafatos. 2023. [Cardiovascular mortality in northern and southern european cohorts of the seven countries study at 60-year follow-up](#). *Journal of Cardiovascular Medicine*, 24(2):96–104.
- Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. [NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.
- Yatin Nandwani, Abhishek Pathak, Mausam, and Parag Singla. 2019. [A primal dual formulation for deep learning with constraints](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Engineering National Academies of Sciences, Medicine, Division of Behavioral, Social Sciences, Education, and Committee on the Science of Science Communication: A Research Agenda. 2017. [Communicating Science Effectively: A Research Agenda](#). National Academies Press (US),

- Washington (DC). Copyright 2017 by the National Academy of Sciences. All rights reserved.
- Jiaxin Pei and David Jurgens. 2021. [Measuring sentence-level and aspect-level \(un\)certainly in science communications](#). *CoRR*, abs/2109.14776.
- Stanislas Polu and Ilya Sutskever. 2020. [Generative language modeling for automated theorem proving](#). *CoRR*, abs/2009.03393.
- Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. [Are red roses red? evaluating consistency of question-answering models](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.
- Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. 2021. [RuleBERT: Teaching soft rules to pre-trained language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1460–1476, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal. 2020. [PROVER: Proof generation for interpretable reasoning over rules](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 122–136, Online. Association for Computational Linguistics.
- Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. [Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts](#). *ArXiv*, abs/2204.08292.
- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. [ProofWriter: Generating implications, proofs, and abductive statements over natural language](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.
- Mingzhe Wang and Jia Deng. 2020. [Learning to prove theorems by learning to generate theorems](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 18146–18157. Curran Associates, Inc.
- Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. [Chain of thought prompting elicits knowledge augmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6519–6534, Toronto, Canada. Association for Computational Linguistics.
- Zhun Yang, Adam Ishay, and Joohyung Lee. 2023. [Coupling large language models with logic programming for robust and general reasoning from text](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5186–5219, Toronto, Canada. Association for Computational Linguistics.
- Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, and Eric Xing. 2023. [Improved logical reasoning of language models via differentiable symbolic programming](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3062–3077, Toronto, Canada. Association for Computational Linguistics.
- A. Zirra, S. C. Rao, J. Bestwick, R. Rajalingam, C. Maras, C. Blauwendraat, I. F. Mata, and A. J. Noyce. 2023. [Gender differences in the prevalence of parkinson’s disease](#). *Movement Disorders Clinical Practice*, 10(1):86–93.

A Appendix

A.1 Adverbs in RuleTaker-pro

In creation of RuleTaker-pro, we utilize 8 different adverbs of frequency shown in the Table 8. Using adverbs of frequency improved the Dev binary accuracy consistently in all depths. The results are shown in Table 9.

A.2 RuleTaker-pro Generation Algorithm

In order to make a balanced dataset with an equal number of labels, we generate a random probability for each rule based on a Gaussian random generator. Then the adverb with the closest probability to the generated probability is chosen. The rule probability generations are generated so that half of the answers are above and half are below 0.50.

The algorithm to change a logical context to a probabilistic one is shown in Algorithm 1. “FIND_ADVERB” function gets a random probability from 0 to 100 as input and returns an adverb to it based on the closest probability of an adverb in Table 8. In the procedure “ADD_PROBABILITIES”, a logical context and question are given as input. Then, in line 6, it is randomly decided whether or not the final answer to this instance should be above or below 0.50 to ensure balance in the final results of the dataset. In the rest of the algorithm, until the pre-selected above or below 0.50 probability for the answer is achieved, random probabilities would be assigned to the rules in the context. The random function that assigns these probabilities is a Gaussian function with a mean of 40 and std of 60. The random probabilities are added with the value h , initially set to $depth * 10$, and it increases or decreases slightly to help achieve the desired answer after reaching failure. h is created based on the depth of the dataset group to create a balanced average of answer probabilities. A real example of the created dataset is shown and analyzed in section A.8.

Adverbs	always	usually	normally	often	sometimes	occasionally	seldom	never
Probability	1.00	0.90	0.80	0.65	0.50	0.30	0.15	0.0

Table 8: The adverb of uncertainty and their respective probabilities that we link to them.

CE	M1	M2	M3	Mmax
With adverbs	96.24	94.97	93.12	89.71
With probabilities	95.78	93.71	92.52	88.01

Table 9: Dev BA for models M1 to Mmax trained with CE loss.

Algorithm 1 Assigning Gaussian-based probabilities to logical rules to create a probabilistic dataset while ensuring that the resulting dataset is balanced with heuristics.

```

1: function FIND_ADVERB( $x$ )
2:   Determine adverb and its associated probability based on the range of  $x$  in Table 8
3:   return adverb
4: end function

5: procedure ADD_PROBABILITIES( $c, q, d$ ) ▷  $c$ 
   is context,  $q$  is question and  $d$  is the depth of
   the dataset group (not the instance)
6:    $Above0.50 \leftarrow \text{RANDOM}(False, True)$ 
7:    $h \leftarrow 10 * depth$ 
8:   while not Answer is  $Above0.50$  do
9:      $new\_c = c$ 
10:    for each rule in  $context$  do
11:       $p_i = \text{RANDOMGAUSS}(40,60)+h$ 
12:       $adverb = \text{FIND\_ADVERB}(p_i)$ 
13:      add  $adverb$  to  $new\_c$ 
14:     $Answer \leftarrow \text{PROBLOG}(new\_c, q)$ 
15:    if  $Above0.50$  then
16:       $h = h + 5$ 
17:    else
18:       $h = h - 5$ 
19: end procedure

```

Statistics about the splits, their unique context and questions, and their balanced average answer produced by our algorithm are shown in Table 10.

Split	D	Rows	Queries	MA
Train	1	13549	807	0.49
Train	2	16145	810	0.48
Train	3	19960	812	0.48
Train	5	23805	812	0.50
Dev	1	1946	551	0.50
Dev	2	2290	586	0.48
Dev	3	2837	629	0.48
Dev	5	3412	694	0.50
Test	1	3930	690	0.49
Test	2	4592	718	0.48
Test	3	5687	765	0.48
Test	5	6829	789	0.50

Table 10: RuleTaker-pro Dataset Statistics. Split determines the split of the dataset which could be train, dev or test. D determines the depth of the question. Rows shows the total rows and Queries shows the number of unique queries. MA is the Mean of all Answers which should be around 0.50 for a balanced dataset.

RuleTaker-pro depth distribution for all depths and the number of True and False labels are shown in Table 11.

A.3 ProbLog

ProbLog is a tool that allows us to encode probabilistic facts and rules. Then it will calculate any queries in the context of the defined facts and rules, which is exactly what we need for RuleTaker-pro. For example, Table 1’s right column would be shown in Problog pseudo code in the Figure 1a.

A more complicated example would occur when there is more than one way to reach an inferred intermediate fact. Imagine that the second fact in the example of Table 1’s right column is “David is

	M1	M2	M3	Mmax
D0 T	10626	9590	7441	2616
D0 F	10719	9485	7650	2720
D1 T	6422	4613	4438	3802
D1 F	6452	4465	4272	3692
D2 T	0	3441	2930	2442
D2 F	0	3469	2949	2520
D3 T	0	0	2597	2118
D3 F	0	0	2642	2026
D4 T	0	0	0	1852
D4 F	0	0	0	1858
D5 T	0	0	0	1761
D5 F	0	0	0	1734

Table 11: RuleTaker-pro depth distribution for all depths and the number of True and False labels. M* shows the distribution for the training set of max depth *. T and F stand for True and False labels which would indicate a probability of final answer being higher or lower than 0.50.

Input:
Dave_is_big .
Erin_is_sad .
0.90:: Green :- Big .
0.80:: Round :- Green .
0.15:: Round :- Sad .
query(Dave_is_round).
query(Erin_is_round).
Output:
[(Dave_is_round, 0.72),
(Erin_is_round, 0.15)]

(a) Encoding of the Table 1’s right column example in ProbLog pseudo code.

Input:
Dave_is_big .
Dave_is_sad .
0.90:: Green :- Big .
0.80:: Round :- Green .
0.15:: Round :- Sad .
query(Dave_is_round).
Output:
[(Dave_is_round, 0.762)]

(b) Encoding of the Table 1’s right column example in ProbLog pseudo code if the second fact is replaced with “David is sad.”

sad.”. In that case, the probability that “David is round” would be 0.762 as shown in Figure 1b.

A.4 RuleBERT’s Additional Results

A.4.1 RuleBERT’s Original Setting

The original RuleBERT baseline from (Saeed et al., 2021) is shown in Table 12. We also train our models with their settings, both with and without including the text of the rules. These new results are shown in Table 13. The text of the rules is still not useful for the models.

	M1	M2	M3	M4	M5
D1	86.0	88.4	88.7	88.9	88.9
D2	65.5	73.0	75.1	75.0	72.0
D3	58.1	63.6	68.4	69.0	65.6
D4	46.8	54.7	62.6	66.6	62.7
D5	35.6	49.6	70.3	78.5	74.4

Table 12: RuleBERT baseline results trained and tested on different depths (Saeed et al., 2021).

	RoBERTa With Rules’ Text				
D/M	M1	M2	M3	M4	M5
D1	76	91	87	91	93
D2	76	87	79	83	83
D3	67	85	76	76	73
D4	66	82	69	63	51
D5	53	75	54	34	28
	RoBERTa Without Rules’ Text				
D1	88	90	88	92	89
D2	87	88	77	78	74
D3	84	85	73	72	67
D4	82	80	65	60	51
D5	80	68	44	29	21

Table 13: M shows the maximum depth of the training data, and D shows the depth of the test data. Here RoBERTa is trained with the original (Saeed et al., 2021)’s training setting and parameters.

A.4.2 PCT With Rules’ Text

Since the text of the rules decreases the accuracy of our models, we removed it in our original PCT result, but if we do include the text, PCT would still improve the accuracies as shown in Table 14.

A.5 CA25 Accuracy of Intermediate Inferred Facts

CA25 Intermediate Inferred Facts for M5 is depicted in Figure 2. The model is trained for 6 epochs to show the accuracy over time. PCT accuracy remains consistently over 0.90 while the baseline models accuracy fluctuates and remains below 0.60.

A.6 PCT Algorithm Pseudo-Code

The PCT algorithm pseudo-code is shown in Algorithm 2. Lines 2-4 apply the taskloss, and lines 5-13 apply constraints loss and update the λ_j . The rate at which λ_j is updated depends on PCT variable (α) decayed at each iteration’s end.

	RoBERTa				
	M1	M2	M3	M4	M5
D1	76.9	79.8	79.9	70.7	64.9
D2	77.5	77.8	76.6	70.4	65.4
D3	78.4	76.9	76.2	78.8	71.6
D4	76.2	73.4	72.4	78.2	73.8
D5	77.1	73.0	69.6	77.5	78.1
	RoBERTa + PCT				
D1	78.3	83.1	77.5	77.9	67.7
D2	78.9	79.7	76.6	78.0	68.9
D3	79.1	80.8	81.3	81.3	78.9
D4	77.7	77.0	79.0	80.8	82.6
D5	77.8	74.1	84.1	82.7	88.6

Table 14: BA results of RoBERTa fine-tuned on RuleBERT when the rule’s text is included with and without PCT. Columns indicate the maximum depth of reasoning in training (M1-M5), and rows correspond to the reasoning depth of testing (D1-D5).

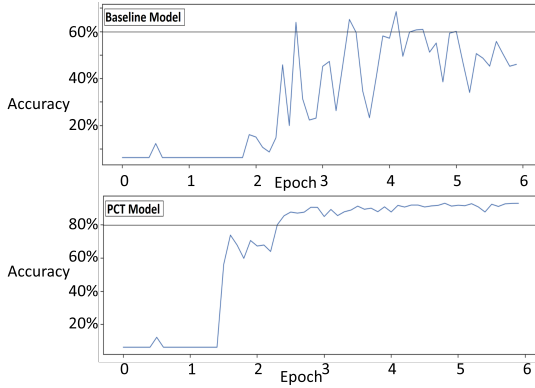


Figure 2: The CS25 of intermediate inferred facts over 6 Epochs of training for M5.

A.7 Training Parameres

A.7.1 RuleTaker-pro

To train RuleTaker-pro, we use RoBERTa Large for four epochs with a learning rate of $1e-5$. When we use PCT, the alpha (PCT variable) varies from 1.0 to 0.001 depending on the depth of the training dataset with higher depths training with smaller alphas.

A.7.2 RuleBERT

To train RuleBERT, we also use RoBERTa Large for four epochs, but we freeze the first 22 layers of the transformer. The learning rate varies between numbers $1e-6$ for higher depth datasets with more examples and $2e-6$ for lower depth datasets. When using PCT, the alpha is 0.01 for lower depths (1-3) and 0.001 for higher depths (4-5). In Table 15, the

Algorithm 2 PCT algorithm

```

1: for each batch in data do
2:   Apply model on batch to get the logits
3:   Calculate Taskloss (CE/MSE/L1loss)
4:   Backward propagate the loss
5:   if Not warm-up iteration then
6:     Get the next constraints batch
7:     Apply model on constraints batch
8:      $cl \leftarrow 0$   $\triangleright$  initialize constraints loss
9:     for each constraint do
10:       $l \leftarrow abs(q - p_1 \times p_2 \dots \times p_n \times Pr)$ 
11:       $cl \leftarrow cl + l \times \lambda_j$ 
12:       $\lambda_j \leftarrow \alpha \times l$ 
13:   Backward propagate the cl
14:   Take optimizer step, and Reset gradients
15:   decay  $\alpha$ 

```

effect of alpha on the PCT Dev BA is shown. As shown, a higher alpha will help the model reach higher accuracy earlier. However, the best result is achieved with an alpha of 0.01.

A.8 Error Analysis Examples

We analyze an example shown in Figure 3 that benefited from PCT. Initially, the base model predicted 0.50 for the final answer, which was incorrect, as the answer should have been 0.85. After training the model using PCT, the model correctly predicted 0.85. This demonstrates the potential of the PCT model for incorporating additional constraints in the inference process. However, it should be noted that this is an ideal case that may not always be reproduced in practice. The PCT model can be adapted to alter the probability of the depth2 fact to satisfy the constraint if needed. In other scenarios, the model may keep the 0.50 prediction for depth 3 and change the prediction for depth 2. In this case, the model satisfies the constraint, yet the final prediction is incorrect. In the worst case, the model may predict 0.0 for all elements and still satisfy the constraint.

It has been observed that the predicted probabilities of the PCT models are lower on average than those of the baseline models. This is due to the fact that lower predicted probabilities make it easier to satisfy the constraints, and thus, even models that improve overall accuracy tend to have lower average predicted probabilities.

Depth3	Epoch1	Epoch2	Epoch3	Epoch4	Epoch5	Epoch6
Baseline	49	70	77.95	75.85	70.925	72.62
PCT with $\alpha = 0.1$	49	79.15	78.42	76.9	77	64.51
PCT with $\alpha = 0.01$	49	79.32	80.87	79.32	78.17	78.57
PCT with $\alpha = 0.001$	49	70.90	78.55	80.85	78.55	78.75

Table 15: Accuracy obtained using PCT during training with different hyper-parameter (α) for depth 3 of reasoning for 6 epochs on RuleBERT dataset. Normally we train our models for 4 epochs, but here we use 6 epochs to observe the learning process better.

Context: The cow is round. Always, If something is nice and round then it does not visit the lion. The mouse visits the cow. The rabbit does not see the cow. The lion is round. The rabbit is big. The cow likes the rabbit. The lion likes the rabbit. Always, If something is big and it does not see the rabbit then it visits the mouse. The mouse is green. Usually, If something visits the lion then it visits the mouse. Always, if something is green, then it visits the lion. Always, if the rabbit is big, then the rabbit is green. Hypothesis: The rabbit visits the mouse. (Depth 3), P3 = 85% Required Intermediate Facts: The rabbit visits the lion. (Depth 2), P2 = 100% The rabbit is green. (Depth 1), P1 = 100% The rabbit is big. (Depth 0), P0 = 100%
Base model: The rabbit visits the mouse. (Depth 3), P3 = 50% The rabbit visits the lion. (Depth 2), P2 = 100% Constraint: P2*85% $\neq$ P3 (violated Constraint and Incorrect Answer)
PCT Model The Ideal Case: The rabbit visits the mouse. (Depth 3), P3 = 85% The rabbit visits the lion. (Depth 2), P2 = 100% Constraint: P2*85% = P3 (Satisfied Constraint and Correct Answer)
PCT Model The Problem Case: The rabbit visits the mouse. (Depth 3), P = 50 % The rabbit visits the lion. (Depth 2), P = 59 % Constraint: P2*85% = P3 (Satisfied Constraint and Incorrect Answer)
PCT Model The Worst Case: The rabbit visits the mouse. (Depth 3), P = 0 % The rabbit visits the lion. (Depth 2), P = 0 % Constraint: P2*85% = P3 = 0 (Satisfied Constraint and Incorrect Answer)

Figure 3: In the given example, the fact “The rabbit visits the lion.” can be inferred from the context with a probability of 1.00 at depth 2. Both the base model and the PCT model accurately predicted the probability of this fact. However, only the PCT model took into account the additional bold rule in the text, which led to an 0.85 probability for the hypothesis.

A.9 LLM Prompt Instructions and Additional Results

To effectively evaluate LLMs like, we adjust our approach with our datasets to make them suitable for zero-shot and in-context settings for generative

models. These adaptations involved adding a text explaining the task before the context. For RuleBERT, we use the following explanation, “*Answer the following logical probabilistic question with only one word, True or False.*” and add the probability of the rules to their text. For RuleTaker-pro, we use “*Answer the following logical probabilistic question in the format .##, which is the probability of the question asked rounded to 2 decimals, for example, .13%.*” After this text, we provide the context and pose the hypothesis as a question.

To test RuleBERT in LLMs, we included the probability of the rules in the text; Otherwise, the model has no way of extracting them. The results are shown in Table 16.

Model	GPT3.5	GPT3.5*	GPT4
Depth1	19%	43%	29%
Depth2	58%	53%	46%
Depth3	58%	58%	60%
Depth4	51%	56%	46%
Depth5	56%	43%	58%

Table 16: RuleBERT BA results are show for GPT3.5 and GPT4. * indicates few-shot setting.

RuleTake-pro results for CA1 and CA10 are compared in Table 17. The only model that improves with regard to an increase in the threshold of the final answer is RoBERTa. This suggests that if the LLM can now predict the final answer, it would not predict anything close to it.

As we mentioned previously CoT did not improve our models. Normally, one would expect CoT to always improve the results by adding a little extra reasoning and explanation for the LLM, but here it does not do the same for the following reasons: 1)The reasoning here is very complex and may require a combination of up to 5 rules and five facts, which are explained with long text to infer the final answer 2) In addition to the previous reason, the text of the rules is very large, and the context includes random rules mixed with useful rules that

	CA1			
	RoBERTa	GPT3.5	GPT3.5*	GPT4
D1	44	28	41	41
D2	33	20	26	27
D3	28	23	25	26
D4	27	18	20	17
D5	28	18	20	21
	CA10			
D1	56	33	45	43
D2	52	28	36	37
D3	51	25	34	34
D4	52	21	33	29
D5	53	22	31	29

Table 17: LLM results on RuleTaker-pro for CA1 and CA10 metrics. * indicates using few-shot examples. The chosen RoBERTa model is M5 trained with CE since it performs the best regarding CA1.

are not needed to answer the final question. 3)The task requires math and number extraction, which LLMs historically struggle with.

(Shi et al., 2022) is another dataset who suffers from the first two reasons and as a results dose not improve with COT. Given these complexities, the provided instructions and examples in COT that we initially tried ended up actually hurting the model counterintuitively. For example, in some cases, the model would try to imitate the exact solutions in the few-shot examples to answer the questions and hallucinate the probabilities that don’t exist in the text in its reasoning. Given these complexities, in these datasets, CoT is not just a baseline model and needs significant careful prompt engineering to be useful.

A.10 Additional RuleTaker-pro Results

In Table 18, The binary results for RuleTaker-pro trained with MSE and CE is shown.

Additional detailed baseline and PCT results of RuleTaker-pro are shown in Tables 19 and 20.

	CE Loss				MSE Loss			
BA	M1	M2	M3	Mmax	M1	M2	M3	Mmax
Total	76.93	82.65	88.74	91.05	76.19	84.84	87.73	91.39
D1	97.19	94.85	92.18	93.39	97.28	95.92	92.64	94.33
D2	75.58	89.11	91.26	91.26	74.41	90.91	91.88	91.74
D3	68.19	77.35	89.42	91.00	42.88	81.93	88.59	90.34
D4	65.16	71.35	84.93	88.70	38.04	74.38	81.82	89.17
D5	58.61	65.05	80.96	88.31	57.70	66.96	76.43	88.21
MSE	M1	M2	M3	Mmax	M1	M2	M3	Mmax
Total	0.1574	0.1278	0.0965	0.0716	0.4693	0.6585	0.6298	0.0716
D1	0.0866	0.0996	0.1076	0.0983	0.1992	0.0173	0.0190	0.0983
D2	0.1939	0.1261	0.1065	0.0876	0.1902	0.0257	0.0247	0.0876
D3	0.2352	0.1826	0.1149	0.0818	0.1915	0.0698	0.0313	0.0818
D4	0.2511	0.2003	0.1281	0.0710	0.1910	0.0982	0.0423	0.0797
D5	0.3082	0.2436	0.1428	0.710.	0.1963	0.1237	0.0618	0.0710
L1	M1	M2	M3	Mmax	M1	M2	M3	Mmax
Total	0.2505	0.2216	0.1903	0.1664	0.3628	0.1055	0.0798	0.1664
D1	0.2004	0.2138	0.2236	0.2175	0.3693	0.0525	0.0581	0.2175
D2	0.3118	0.2434	0.2243	0.2076	0.3638	0.0770	0.0786	0.2076
D3	0.3528	0.3010	0.2316	0.1972	0.3642	0.1570	0.1032	0.1972
D4	0.3672	0.3182	0.2443	0.1960	0.3659	0.2090	0.1326	0.1960
D5	0.4136	0.3519	0.2495	0.1761	0.3737	0.2480	0.1627	0.1761

Table 18: The Binary accuracy, MSE and L1 of the baseline model trained and tested on the RuleTaker-pro dataset at different depths.

	CE (CA1)				MSE (CA1)			
D/M	M1	M2	M3	Mmax	M1	M2	M3	Mmax
Total	38.21	38.34	20.45	33.89	30.39	32.26	26.17	26.04
D1	56.03	52.71	29.63	43.77	50.46	49.48	38.15	37.26
D2	36.40	38.28	20.31	32.87	26.49	31.13	25.52	28.43
D3	29.30	31.30	14.98	28.39	18.81	22.06	18.98	19.90
D4	27.49	28.53	14.03	27.11	18.54	21.37	17.79	17.32
D5	24.97	26.78	14.70	28.29	19.83	21.14	19.33	15.50
CS1	47.88	35.79	16.22	20.78	25.24	14.88	14.47	12.97
	CE (CA10)				MSE (CA10)			
D/M	M1	M2	M4	Mmax	M1	M2	M3	Mmax
Total	46.45	49.69	49.95	53.25	58.15	62.75	66.67	74.80
D1	61.56	59.55	53.41	56.17	91.76	84.45	80.94	82.81
D2	45.76	52.08	52.42	51.59	53.60	69.97	77.25	77.32
D3	38.88	44.87	48.37	51.45	42.88	51.20	61.44	71.60
D4	37.75	42.45	47.36	51.97	38.04	46.51	51.50	69.39
D5	33.63	38.67	43.80	53.07	32.62	37.05	43.30	63.64
CS10	52.24	44.97	35.67	38.25	45.13	34.49	32.86	33.34

Table 19: The accuracy of the **baseline** models trained and tested on the RuleTaker-pro dataset. The rows show different test depths (depths 1 to 5). Total indicates the weighted average accuracy of all depths, and CS* shows the constraint satisfaction at the indicated thresholds. The best results for each depth are in bold.

	CE+PCT (CA1)				MSE+PCT (CA1)			
D/M	M1	M2	M3	Mmax	M1	M2	M3	Mmax
Total	38.0	39.5	41.1	37.6	37.4	34.7	36.4	34.3
D1	53.3	50.8	50.5	46.9	56.50	49.8	52.6	37.6
D2	37.4	40.4	42.2	37.0	35.99	34.2	38.1	33.8
D3	26.4	32.9	36.0	32.4	25.9	25.8	26.5	32.6
D4	26.5	31.9	33.9	31.8	24.1	25.5	24.9	31.6
D5	23.3	30.4	33.4	31.4	22.0	24.0	24.0	33.1
CS1	44.9	42.6	34.5	35.2	20.5	19.3	15.4	13.0
	CE+PCT (CA10)				MSE+PCT (CA10)			
D/M	M1	M2	M3	Mmax	M1	M2	M3	Mmax
Total	46.6	50.8	52.5	52.9	58.9	63.3	67.2	68.7
D1	59.7	57.8	50.5	57.9	92.4	82.7	83.5	70.9
D2	47.78	51.4	42.2	51.8	57.7	73.2	76.1	68.2
D3	39.2	47.4	50.5	50.0	41.5	52.2	60.4	68.6
D4	36.2	47.0	50.1	50.4	36.1	47.3	51.4	70.0
D5	35.1	47.0	48.6	49.8	34.0	38.1	44.6	63.8
CS10	49.7	47.3	45.6	46.8	49.6	36.0	34.9	33.7

Table 20: RuleTaker-pro results trained with **PCT**. The rows show different test depths (depths 1 to 5). Total indicates the weighted average accuracy of all depths, and CS* shows the constraint satisfaction at the indicated thresholds.