Logical Reasoning over Natural Language as Knowledge Representation: A Survey

Anonymous EMNLP submission

Abstract

Logical reasoning is central to human cognition and intelligence. Past research of logical reasoning within AI uses formal language as knowledge representation (and symbolic reasoners). However, reasoning with formal language has proved challenging (e.g., brittleness and knowledge-acquisition bottleneck). This paper provides a comprehensive overview on a new paradigm of logical reasoning, which uses natural language as knowledge representation (and pretrained language models as reasoners), including philosophical definition and categorization of logical reasoning, advantages of the new paradigm, benchmarks and methods, challenges of the new paradigm, desirable tasks & methods in the future, and relation to related NLP fields. This new paradigm is promising since it not only alleviates many challenges of formal representation but also has advantages over end-to-end neural methods.

1 Introduction

An argument consists of premise(s) and a conclusion. Logical reasoning is a form of thinking in which premises and relations between premises are used in a rigorous manner to infer conclusions that are entailed (or implied) by the premises and the relations (Nunes, 2012). It consists of three reasoning types, namely deductive reasoning, inductive reasoning, and abductive reasoning (Flach and Kakas, 2000) (more illustration on the categorization can be found in §2). It is important since the ability to reach logical conclusions on the basis of prior information is recognized as central to human cognition and intelligence (Goel et al., 2017).

The past research of logical reasoning within AI uses formal language (e.g., first-order logic) as knowledge representation and symbolic reasoners (Muggleton and Raedt, 1994). This paradigm has resulted in impressive applications such as expert systems (Metaxiotis et al., 2002). However, building and reasoning over formal language have proved challenging (Musen and Van der Lei, 1988), with representative disadvantages of brittleness and knowledge-acquisition bottleneck.

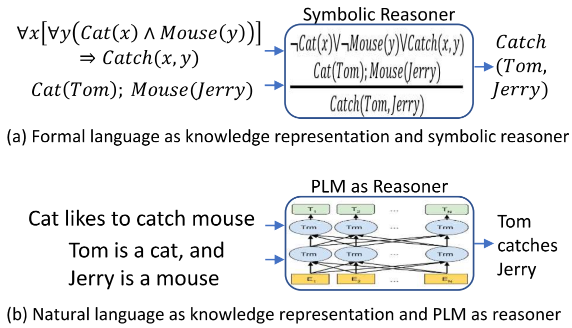

Figure 1: Comparison between the previous paradigm which uses formal representation and symbolic reasoner, and the new paradigm which uses natural language as knowledge representation and PLM as reasoner.

Since the rapidly developed NLP techniques, natural language has been explored as a new knowledge representation, and pretrained language model (PLM) has been used as a new corresponding reasoner for deductive reasoning (Clark et al., 2020), abductive reasoning (Bhagavatula et al., 2020), and inductive reasoning (Yang et al., 2022b). Therefore, all three reasoning types of logical reasoning have been investigated with natural language as knowledge representation. Recent research also shows that PLMs can be finetuned or prompted to have a good level of ability for each of the reasoning types.

In this paper, we summarize the three previously separately investigated reasoning types together with a new overall concept, logical reasoning over natural language (LRNL) as knowledge representation, and provide the first survey of LRNL. Illustrated in Figure 1, LRNL means a new paradigm for logical reasoning that uses new knowledge representation (natural language) and new reasoner (PLM). Recent methods in this area are generally modular-based: multiple PLMs each as one module playing a different function, combined together to perform complex tasks. They make one step of reasoning with one inference of PLM. For complex problems, they usually have access to a knowledge base which stores relevant textual knowledge to be retrieved as premises to support the reasoning process to reach a conclusion, which might be used as a new premise for next step's reasoning. By iteratively repeating this process, a final conclusion may be made. Although looks similar to expert systems, we discuss how LRNL is possible to overcome many main challenges of the previous paradigm such as brittleness and knowledge-acquisition bottleneck in §3.1.

In addition to the comparison with formal language, in §3.2 we discuss that LRNL could be viewed as a new type of neural-symbolic (NeSy) method, which has unique advantages over existing NeSy methods. We also discuss how LRNL, as a NeSy method, has advantages over existing end-to-end neural methods (e.g., explainability, controllability, less catastrophic forgetting) in §3.3. These advantages make a LRNL system possible to deal with many challenging problems today.

In the remaining sections of this survey, we review papers on LRNL (including deductive reasoning §4, inductive reasoning §5, and abductive reasoning §6), list challenges (§7), and possible future directions (§8). Our main focus is to understand language model's logical reasoning ability through the three subtypes of logical reasoning to provide finer analysis and avoid ambiguity. Therefore we focus on papers that specialized on one (or more) of the three subtypes of logical reasoning (instead of only "reasoning"). We find that these papers generally use smaller, typically finetuned PLMs, and recent methods are generally modular-based. We do not intentionally cover out-of-box large language model (LLM) prompting techniques such as chain-of-thought (Wei et al., 2022) unless they have a specification. For each reasoning type, we summarize existing task formulations, datasets, and methods under each task. In §A.1 we also discuss the relation of LRNL to related NLP fields, which could help to form a clear shape of LRNL in NLP.

2 Definition and Categorization of Logical Reasoning

There are many subjects related to logical reasoning, including philosophy, logic, and AI. Among them, the definition and categorization aspects of logical reasoning are handled by philosophy research. However, debate exists in philosophy research on the categorization of logical reasoning.

One group believes that every argument can be classified as deduction argument, inductive argument, or fallacy (Salmon, 1989). Without considering fallacy, given that an argument consists of premises and a conclusion, when the premises can provide conclusively support to the conclusion (which means that if the premises of the argument were all true, it would be impossible for the conclusion of the argument to be false), this argument is a deductive argument. Conversely, when the premises can not provide conclusively support to the conclusion, the argument is inductive.

The other group has the same definition of deductive reasoning, but they believe that further categorization of non-deductive reasoning is necessary. Without considering fallacy, they believe in a trichotomy of deductive, inductive, and abductive reasoning (Peirce, 1974). However, even for the second group, the definition and difference between inductive and abductive reasoning are also controversy (Flach and Kakas, 2000).

Nevertheless, Console and Saitta (2000) argue that from the utility perspective of AI, a distinction between inductive and abductive reasoning is possible: both inductive and abductive reasoning provide explanations about the world but their explanations differ in the degree of generality. For instance, an inductive hypothesis allows the validity of properties, observed on a set of individuals, to be generalized to other individuals not in the observations, whereas an abductive one allows unobserved properties to be applied to observed individuals. More details about the difference and an example can be found in §A.2.

Considering that inductive and abductive reasoning can be distinctive enough when formulated in NLP, in this paper, we adopt the second group, particularly Console and Saitta (2000)'s view of definition and categorization of logical reasoning. We survey deductive, inductive, and abductive reasoning in NLP separately in the following sections.

3 Advantages of LRNL

3.1 Advantages over Formal Language

Building and reasoning over formal language have proved challenging (Musen and Van der Lei, 1988; Cropper et al., 2022), with disadvantages such as (1) brittleness (expert system fails when its knowledge base does not contain complete knowledge for a problem), (2) knowledge-acquisition bottleneck (human experts are needed to encode their knowledge with formal representation), (3) inability to handle raw data such as natural language, (4) sensitivity to label errors, and (5) failure to recognize different symbols with similar meanings.

Nevertheless, the new paradigm of logical reasoning, LRNL, has systematic strengths over these challenges. Specifically, PLMs contain knowledge themselves (Davison et al., 2019), which makes it possible for them to provide good answers even when some required explicit knowledge is not present in a knowledge base (Talmor et al., 2020) (less brittle), and be less affected by input errors (Meng et al., 2021); with natural language as knowledge representation, such a system can naturally handle raw input, and is possible to utilize the enormous web corpora to automatically construct a rule base using information extraction (Ji, 2018) or inductive reasoning (Yang et al., 2022b) (less affected by knowledge-acquisition bottleneck); using embeddings for concepts (Mikolov et al., 2013), it semantically "understands" the meaning of symbols and therefore robust for paraphrasing.

3.2 Advantages over Existing NeSy Systems

LRNL could be seen as a new type of NeSy in addition to the existing 6 types summarized by Kautz (2022), as its goal and design of methodology are typically symbolic (logical reasoning with knowledge bases), while avoiding any symbolic representation, using (currently pure) neural methods. Therefore LRNL can avoid many bottlenecks of the other NeSy methods caused by symbolic representation, such as symbolic knowledge acquisition and scalability (Wang and Yang, 2022).

3.3 Advantages over E2E Neural Methods

As a NeSy method, LRNL systematically has some advantages over end-to-end neural methods, such as interpretability (Cambria et al., 2023) (since its stepwise reasoning nature), more controllability (LRNL reasons following a given knowledge base), and less catastrophic forgetting (LRNL uses an explicit knowledge base to store knowledge).

4 Deductive Reasoning

4.1 Existing Task Formulations

Existing tasks for deductive reasoning can be summarized as hypothesis classification, proof generation, proof generation with incomplete information, and implication enumeration. Datasets for the tasks are summarized in Table 1. "Proof generation" tab with ✗ means it is for hypothesis classification task.

Dataset	Human written	Realistic	Multi-step	Theory included	Theory sufficient	Proof generation	Size
D*	✗	✗	✓	✓	✓	✗	500k
ParaRules	✓	✗	✓	✓	✓	✗	40k
Birds-electricity	✓	✓	✓	✓	✓	✗	5k
Leap-of-thought	✗	✓	✗	✓	✗	✗	33k
PARARULE-Plus	✗	✗	✓	✓	✓	✗	400k
FOLIO	✓	✓	✓	✓	✓	✗	1,435
D*(CWA)	✗	✗	✓	✓	✓	✓	500k
D*(OWA)	✗	✗	✓	✓	✗	✓	500k
EntailmentBank	✓	✓	✓	✓	✓	✓	1,840
ENWN	✓	✓	✓	✓	✓	✓	100

Table 1: Summary of deductive reasoning datasets: D*, ParaRules, and birds-electricity (Clark et al., 2020);leap-of-thought (Talmor et al., 2020); PARARULE-Plus (Bao et al., 2022);FOLIO (Han et al., 2022);D*(CWA) and D*(OWA) (Tafjord et al., 2021);EntailmentBank (Dalvi et al., 2021);ENWN (Sprague et al., 2022).

Hypothesis Classification Each data example for hypothesis classification task is a tuple $(theory, hypothesis, correctness)$, where $theory$ typically has the form $(fact^*, rule^*)$, $hypothesis$ is a question, and $correctness$ can be $True$ or $False$ (or $Unknown$). This task requires to predict the $correctness$ for the $hypothesis$ given the $theory$.

Proof Generation The proof generation task has the same setting as the hypothesis classification task, except that in addition to predicting a $correctness$, the proof generation task also requires to provide a $proof$ given $theory$ to explain the $correctness$. The $proof$ is a directed tree $(\mathcal{N}, \mathcal{E})$ with nodes $n \in \mathcal{N}$ and edges $e \in \mathcal{E}$. Each node is an item of knowledge in $theory$ (usually a $fact$ or a $rule$), or a generated intermediate reasoning conclusion, or the $hypothesis$ itself; Each edge points from a premise node to a conclusion node to form a deductive argument, which typically needs one-step inference (not multi-step).

Proof Generation with Incomplete Information This task is the same as the proof generation task, except that $theory$ lacks one $node$ to form a complete $proof$. Specifically, given $theory$, it requires to predict the $correctness$ of $hypothesis$ with a $proof$, as well as recovering the missing $node$.

Implication Enumeration Given a $theory$, this task requires to enumerate implications of the $theory$, using deductive reasoning.

4.2 Methods

4.2.1 Hypothesis Classification

There are mainly three categories of methods for the hypothesis classification task regarding a multi-task aspect. The first category of methods only conducts the classification task itself; Methods from the second category can predict $correctness$ as well as generate a $proof$. However, the $correctness$ is not necessarily consistent with the predicted $proof$. The third category is similar to the second, except that $correctness$ always follows $proof$.

Until now, methods from the first category directly use transformer-based PLMs (Vaswani et al., 2017), with the target of analyzing and benchmarking their performance in different settings (datasets). Specifically, Clark et al. (2020) find that finetuned RoBERTa-large (Liu et al., 2019) can achieve 95%+ accuracy on the test set of D* and ParaRules datasets; Talmor et al. (2020) further demonstrate that LMs can be trained to reliably perform deductive reasoning using both implicit, pre-trained knowledge and explicit natural language statements ($theory$) to make predictions; Han et al. (2022) evaluate finetuned medium-sized language models and few-shot prompting on LLMs on the FOLIO dataset. However, they find that LLM with few-shot prompting only performs slightly better than random results.

The second category methods typically infer PLMs only once, and then utilize the final layer embeddings or generations to obtain $correctness$ and $proof$. Specifically, PRover (Saha et al., 2020) and multiPRover (Saha et al., 2021) use the [CLS] token to predict $correctness$, and leverage the final layer embeddings of knowledge items in $theory$ to generate $proof$; All-At-Once ProofWriter (Tafjord et al., 2021) and EntailmentWriter (Dalvi et al., 2021) generate $correctness$ and linearized $proof$ at the same time.

The third category methods create a $proof$ first, and then predict $correctness$ from the $proof$. §4.2.2 illustrates these methods in detail.

4.2.2 Proof Generation

Current methods for the proof generation task roughly consist of three stages. In each stage, one key new technique is considered and developed. In stage 1, PLMs are used for forming $proof$ in one inference step. In stage 2, modular-based, stepwise frameworks are developed to create $proof$ (each module is usually implemented with a single PLM). In stage 3, a verifier is added as a new module to make sure that each reasoning step reflects the belief of PLMs. We will introduce the motivation and typical method for each stage.

Methods for stage 1 typically utilize the last layer embeddings (Saha et al., 2020, 2021) or generations (Tafjord et al., 2021; Dalvi et al., 2021) to create $proof$. Methods utilizing embedding typically (1) obtain an averaged embedding for each knowledge item in $theory$, and (2) pass each embedding to a node classifier, and each embedding pairs to an edge classifier to predict nodes and edges for $proof$. Constraints are usually used to enforce the structure of $proof$. Generation methods directly generate linearized $correctness$ and full $proof$ given linearized $theory$ and $hypothesis$.

The motivations of stage 2 methods are generally concerned with end-to-end methods, which is considered to lack interpretability (Liang et al., 2021; Qu et al., 2022; Sanyal et al., 2022b; Bostrom et al., 2022), suffer from compositional generalization problems (Liang et al., 2021; Creswell et al., 2022), have limited input size (Ribeiro et al., 2022), are not casual (Creswell et al., 2022), and lack constraints on the validity of each inference step (Hong et al., 2022).

Methods in stage 2 can be summarized as having two components, an inference module and a reasoning controller. The inference module can be a deduction module (Tafjord et al., 2021; Ribeiro et al., 2022; Creswell et al., 2022; Sanyal et al., 2022b; Bostrom et al., 2022), an abduction module (Liang et al., 2021; Qu et al., 2022), or both (Hong et al., 2022; Sprague et al., 2022). The deduction module performs deductive reasoning, and reasons forwardly from $theory$ to $hypothesis$ to construct $proof$; the abduction module performs abductive reasoning, and reasons backwardly from $hypothesis$ to $theory$ to construct $proof$. The reasoning controller in general performs a search process that each step it searches through the $theory$ and generated intermediate conclusions space to select (retrieve) premises for next step inference. The search processes include exhaustive search (Tafjord et al., 2021; Liang et al., 2021) or heuristic search (Qu et al., 2022; Ribeiro et al., 2022; Creswell et al., 2022; Sanyal et al., 2022b; Bostrom et al., 2022; Hong et al., 2022; Sprague et al., 2022). The reasoning controller usually can also stop the search process if it detects the goal.

Method	Generation based	Inference w/ hypothesis	Stepwise	Proof direction	Heuristic search	Verifier	Human-authored realistic proof	Stage
PRover (Saha et al., 2020)	✗	✓	✗	N/A	N/A	✗	✗	1
multiPRover (Saha et al., 2021)	✗	✓	✗	N/A	N/A	✗	✗	1
EntailmentWriter (Dalvi et al., 2021)	✓	✓	✗	N/A	N/A	✗	✓	1
ProofWriter (Tafjord et al., 2021)	✓	✗	✓	$\rightarrow$	✗	✗	✗	2
EVR (Liang et al., 2021)	✓	✗	✓	$\leftarrow$	✗	✗	✗	2
IBR (Qu et al., 2022)	✗	✓	✓	$\leftarrow$	✓	✗	✗	2
IRGR (Ribeiro et al., 2022)	✓	✓	✓	$\rightarrow$	✓	✗	✓	2
SI (Creswell et al., 2022)	✓	✗	✓	$\rightarrow$	✓	✗	✗	2
FaiRR (Sanyal et al., 2022b)	✓	✗	✓	$\rightarrow$	✓	✗	✗	2
MetGen (Hong et al., 2022)	✓	✗	✓	Both	✓	✗	✓	2
SCSearch (Bostrom et al., 2022)	✓	✗	✓	$\rightarrow$	✓	✗	✓	2
ADGV (Sprague et al., 2022)	✓	✗	✓	Both	✓	✓	✓	3
NLProofS (Yang et al., 2022a)	✓	✓	✓	$\rightarrow$	✓	✓	✓	3
Entailer (Tafjord et al., 2022)	✓	✓	✓	$\leftarrow$	✓	✓	✓	3
Teachme (Dalvi et al., 2022)	✓	✓	✓	$\leftarrow$	✓	✓	✗	3

Table 2: Methods for Proof Generation task. "Generation based" means whether $proof$ is created by generative inference model, otherwise is by utilizing embeddings to classify nodes and edges of $proof$. "Inference w/ $hypothesis$" means whether $hypothesis$ is provided during inference. $\rightarrow$ and $\leftarrow$ denote forward/backward stepwise proof generation. "Heuristic seach" with ✗ means exhaustive search. "Human-authored realistic proof" means whether the dataset adopted uses human-authored $proof$, whose contents are consistent with the real world.

Motivation of stage 3 methods is similar, basically that stage 2 methods lack explicit verifiers to avoid hallucinating invalid steps (Yang et al., 2022a), and to ensure that the inference processes reflect PLM's own beliefs (Tafjord et al., 2022).

Methods in stage 3 can be summarized as utilizing explicit verifier(s) (implemented with a PLM) to check the validity of each inference step. One way is to add a new module (additional to the inference module and reasoning controller in stage 2), working as a "fact checker" to verify the generated inference step (Yang et al., 2022a; Tafjord et al., 2022); The other one, called round-trip consistency, is only suitable for methods that use both deduction and abduction modules, where deduction and abduction modules work as the verifier for each other (Sprague et al., 2022).

In addition to the general 3 stages, a new aspect is attended to, which is whether teachable by humans. Build based on Entailer (Tafjord et al., 2022), TeachMe (Dalvi et al., 2022) shows that user corrections can help override erroneous model beliefs, and that a system can gradually improve by accumulating user corrections. Compared to Entailer, it adds an interaction module and a dynamic memory module to obtain and store human corrections.

We summarize and analyze the experiment results of proof generation task in §A.4.

4.2.3 Proof with Incomplete Information

ADGV (Sprague et al., 2022) is the only method focusing on this task. It uses both deduction and abduction modules, and the reasoning controller performs heuristic search. The abduction module is used to recover the missing premise.

4.2.4 Implication Enumeration

Tafjord et al. (2021) is the only paper mentioned this task. They compare the performance of "All-At-Once" and "Iterative" ProofWriter on this task. They find that "All-At-Once" performs worse, mainly because it struggles with problems that are more complex than training examples.

4.3 Robustness of PLM as Reasoner

The previously introduced methods only focus on solving the deductive reasoning tasks, while it is unclear whether PLMs can be used as robust deductive reasoners. To investigate the problem, Gaskell et al. (2022) create a more challenging synthetic dataset on hypothesis classification task in terms of complexity, and test PLM's performance on it. They find that with large and complex enough training examples, transformers can perform well on the dataset. In addition, they find that transformers exhibit some degree of generalization and scale-invariance ability; Richardson and Sabharwal (2022) propose an adversarial attack method for synthetic datasets on the hypothesis classification task. They find that transformers are often fooled if the query literally appears within the body of a rule, and transformers struggle to correctly bind variables on either side of a rule; Sanyal et al. (2022a) proposed a synthetic deductive reasoning dataset to

Dataset	Human written	Human labeled	Realistic	Rule provided	Not restricted rule types	Generation	Size
property-norm	✗	✗	✓	✗	✗	✗	23k
DEERLET	✗	✓	✓	✓	✓	✗	846
DEER	✓	✓	✓	✓	✓	✓	1.2k

Table 3: Summary of inductive reasoning datasets: property-norm (Misra et al., 2022), DEERLET and DEER (Yang et al., 2022b). "Not restricted rule types" means whether the data is not restricted in a specific topic (e.g., taxonomic).

evaluate the robustness of language models to minimal logical edits in the inputs and different logical equivalence conditions, and find that PLMs are not robust to their proposed logical perturbations.

5 Inductive Reasoning

5.1 Existing Task Formulations

Existing tasks for inductive reasoning can be summarized as rule classification and rule generation tasks. Datasets for the tasks are summarized in Table 3. "Generation" tab with ✗ means it is for the rule classification task.

Rule Classification Given a generated $rule$ and $facts$ where the $rule$ is generated from, the task is to classify whether the $rule$ can be accepted. The current evaluation aspects are from requirements of both inductive reasoning and natural language.

Rule Generation Given multiple manually selected $facts$ with similar patterns, the task is to induce a $rule$ that (1) can entail the $facts$, and (2) is more general than all of the $facts$. Here "more general" means larger information coverage scope. More detailed illustrations can be found in §A.5.

5.2 Methods

5.2.1 Rule Classification

Misra et al. (2022) analyze language model's ability to generalize novel property knowledge (has sesamoid bones) from concept(s) (robins) to others (sparrows, canaries). As illustrated in §A.5, they analyze the language models' ability to classify a new fact (but not a rule) as correct or not, given facts. It could be seen that the correctness of a rule is implicitly predicted by testing multiple facts entailed by the rule.

Yang et al. (2022b) propose three requirements of rule confirmation from philosophy literature ($rule$ and $facts$ should not be in conflict; $rule$ should reflect reality; $rule$ should generalize over $facts$) on inductive reasoning and one requirement of rule confirmation from NLP requirement ($rule$ should not be trivial or incomplete).

Dataset	Human written	Realistic	Multi-step	Theory included	Generation	Size
αNLI	✓	✓	✗	✗	✗	22k
αNLG	✓	✓	✗	✗	✓	76k
AbductionRules	✗	✗	✗	✓	✓	114k
D*-Ab	✗	✗	✓	✓	✓	14k

Table 4: Summary of abductive reasoning datasets: αNLI and αNLG (Bhagavatula et al., 2020), AbductionRules (Young et al., 2022), and D*-Ab (Tafjord et al., 2021). "Realistic" means whether the data is consistent with the real world. "Multi-step" means whether multiple reasoning steps are needed to get the result.

5.2.2 Rule Generation

Yang et al. (2022b) assume that the inductive reasoning task is so difficult that a proper system should contain a rule populator and (multiple) rule verifiers that filter bad rules from different aspects. Accordingly, they propose a framework named chain-of-language-models (CoLM). Specifically, one LM generates $rules$ given $facts$ and a rule template, the other four LMs filter generated rules mainly from inductive reasoning requirements that were selected from philosophy literature.

6 Abductive Reasoning

6.1 Existing Task Formulations

Existing tasks for abductive reasoning can be summarized as explanation classification, and explanation generation w/o and w/ theory. Datasets for the tasks are summarized in Table 4. In the table, the "generation" tab and "theory included" tab can be used to determine the task it is used for.

Explanation Classification Given observation O_1 at time t_1, observation O_2 at time t_2 ($t_2 > t_1$), a plausible hypothesis h^+ and a implausible hypothesis h^- that explain O_1 and O_2, this task is to select the most plausible hypothesis from h^+ and h^-. O_1 and O_2 each contains a single sentence.

Explanation Generation without Theory Given observation O_1 at time t_1, observation O_2 at time t_2 ($t_2 > t_1$), this task is to generate a valid hypothesis h^+ given O_1 and O_2. O_1 and O_2 each is described in a single sentence.

Explanation Generation with Theory Given a theory C and a possible observation O not provable from C, the task is to generate a new hypothetical fact h such that $C \cup \{h\} \models O$. Here C contains multiple facts and rules, where each fact or rule contains a single sentence. O is in single sentence.

6.2 Methods

6.2.1 Explanation Classification

Methods for this task generally introduce knowledge in various ways to improve performance. Specifically, Mitra et al. (2019) explore ways to incorporate additional unstructured textual knowledge retrieved from a story corpus through prompt; Paul and Frank (2020) encode and incorporate knowledge from COMET's generation (Bosselut et al., 2019) directly into transformer's internal attention; Lourie et al. (2021) and Paul and Frank (2021) incorporate knowledge by multi-task training; Du et al. (2021) incorporate knowledge with an additional pre-training stage using $\mathcal{ARI}$ independent story corpora;

In addition to knowledge integration, many different aspects of explanation classification tasks are also investigated. Specifically, Bhagavatula et al. (2020) rewrite the objective using Bayes Rule and formulate a set of probabilistic models that make various independence assumptions on the new objective. They find that the most sophisticated probabilistic model works the best; Zhu et al. (2020) frame this task as a ranking task to also measure the plausibility of hypothesis in addition to discriminating it; Paul and Frank (2021) conduct this task in an unsupervised setting by pretraining on a counterfactual reasoning dataset, which is related to abductive reasoning. Kadikis et al. (2022) propose a method to select suitable PLMs for this task. It is based on the cosine similarity of $embed(O_1, O_2)$ and $embed(h_i)$ for each PLM without finetuning. Zhao et al. (2023) assume that different h are mutually exclusive, and improve performance by incorporating an additional loss item as regularization to enforce an unbalanced probability prediction over different h.

6.2.2 Explanation Generation without Theory

In general, methods for this task either incorporate knowledge or improve the decoding method to be more suitable for this task.

For knowledge integration, Bhagavatula et al. (2020) utilize textual knowledge generated from COMET and investigate two ways of knowledge integration — via texts or via embeddings, and find that the embedding-based method is more effective; Ji et al. (2020) leverage structural knowledge from ConceptNet (Speer et al., 2017) for this task.

For improving decoding method, Qin et al. (2020) are motivated by the fact that the target h^+ to generate happens before O_2. They accordingly propose an unsupervised decoding algorithm that can incorporate both past and future contexts.

6.2.3 Explanation Generation with Theory

Tafjord et al. (2021) explore the ability of a finetuned T5-11B (Raffel et al., 2020) on $P(h|C, O)$. Their results indicate that finetuned T5-11B can reach a high test accuracy of 93% on D*-Ab.

7 Challenges of LRNL

Computationally Efficient Reasoner Many tasks in logical reasoning over formal language have very high algorithmic complexity (Muggleton et al., 2012). Thanks to the low computational cost of each deduction step over formal language, such complex tasks could be possible. However, each deduction step in LRNL typically costs one inference of a LLM, which makes tasks with high algorithmic complexity nearly prohibitive.

Robust Deductive Reasoner Symbolic deductive reasoners are not restricted to train data distributions, while neural deductive reasoners are restricted to their training data (Gontier et al., 2020; Richardson and Sabharwal, 2022); In addition, neural deductive reasoners are also vulnerable to adversarial attacks (Gaskell et al., 2022), while symbolic reasoners are robust to the attacks. The lack of robustness can lead to restricted application domains and incorrect deductive inferences.

Reliable Rule Generation Currently, the rule generation method in inductive reasoning relies on out-of-box LLMs, since a finetuned rule generation model could be restricted in a domain. The annotation of an inductive reasoning dataset should only be done by experts and is very time consuming (Yang et al., 2022b). Given the two restrictions, how to improve the quality of generated rules given related facts could be a challenging open problem.

Reliable Explanation Generation Abduction is a form of non-monotonic reasoning (Paul, 1993), and potentially has a large search space of conclusions given premises. Therefore, how to generate more (all) reasonable explanations can be challenging (Bhagavatula et al., 2020).

Reliable Verifier on Reasoning Steps Many state-of-the-art methods on deductive reasoning (Yang et al., 2022a; Tafjord et al., 2022;

Sprague et al., 2022) and inductive reasoning (Yang et al., 2022b) use verifier to check the correctness of generated reasoning results. However, the current verifiers only reflect the internal beliefs of PLMs. It is doubtful whether PLMs have obtained the knowledge for verification.

Better Automatic Evaluation Metrics It is generally difficult to automatically evaluate generative reasoning implications, especially with realistic and not synthetic datasets. The difficulty mainly lies in that the same semantic meaning can be expressed with diversified forms, and that different conclusions might be all acceptable (especially in abductive and inductive reasoning). This may lead to biased evaluation when using automatic metrics.

Building Larger Benchmarks For complicated reasoning tasks especially in realistic and natural language settings, usually experts are needed for annotation, and the process is very time consuming (Dalvi et al., 2021; Sprague et al., 2022; Yang et al., 2022b). Therefore it can be challenging to construct significantly larger benchmarks.

Understanding the Internal Mechanism of PLMs for Reasoning Until now research works only focus on investigating whether the input/output behaviors of PLMs can be used to simulate a reasoner (Clark et al., 2020) or complete reasoning tasks. However, it is still a challenging open research question to understand the internal mechanism of PLMs for reasoning.

More Impacts on (NLP) Applications As illustrated in §3, overall LRNL can be seen as a new type of neuro-symbolic method, which takes the advantages from both the symbolic and subsymbolic aspects, and can systematically alleviate many main challenges of both symbolic and subsymbolic methods. These characteristics make a LRNL system possible (but might still challenging) to deal with many (NLP) applications such as medical diagnosis and legal NLP tasks, since many medical and legal problems could be seen as pure logical reasoning problems with very large rule base (e.g., medical knowledge and laws).

8 Possible Future Directions

Probabilistic Inference In reality, pure deductive reasoning has not always been used. When people include "likely" in their expressions, uncertainty is introduced, which makes the reasoning process probabilistic; In addition, inductive reasoning and abductive reasoning are by default non-monotonic reasoning. This uncertainty aspect has not been focused in current research. It is probably beneficial to learn from how symbolic reasoning handles uncertainty (Halpern, 2017).

Reasoning with Incomplete Information The current proof generation task requires all necessary premises provided to create a proof tree. Only one work (Sprague et al., 2022) focuses on proof generation with the incomplete information task. However, the task they adopt only overlooks one premise, while in reality more might be missing.

Inductive Reasoning on Web Corpora Currently, the dataset for rule generation tasks in inductive reasoning provides manually selected facts (Yang et al., 2022b). However, to best leverage a system's ability in handling natural language, it should be able to work on raw web corpora to induce rules, which leads to a more challenging task of inductive reasoning on web corpora.

Abductive Reasoning with (Long) Theory Many tasks such as medical diagnosis conduct abductive reasoning with a long theory (e.g., medical knowledge). However, current abductive reasoning research only covers abductive commonsense reasoning (Bhagavatula et al., 2020) without given theory, or only given short, synthetic, not realistic knowledge as theory (Tafjord et al., 2021).

Interactions between Reasoning Types Multiple reasoning types can be used together for complex tasks. Existing works only utilize deductive reasoning with abductive reasoning to create a proof tree (Hong et al., 2022; Sprague et al., 2022). However, many other collaborations are possible, such as using inductive reasoning to collect a (large) rule base, which is to be used as the theory base for deductive reasoning.

9 Conclusion

In this paper, we propose a new concept, logical reasoning over natural language as knowledge representation (LRNL), and provide a detailed and up-to-date review of LRNL. Moreover, we have introduced the philosophical foundations, advantages of LRNL, benchmarks and methods, challenges, desirable tasks & methods, and the relation of LRNL to related NLP fields (§A.1).

10 Limitations

The scope of this survey paper does not cover out-of-box LLMs' prompting techniques such as chain-of-thought (Wei et al., 2022) for reasoning. However, chain-of-thought methods do not specifically focus on any specific reasoning type of logical reasoning – including deductive reasoning, inductive reasoning, and abduction reasoning. Instead, they focus on mathematical reasoning and commonsense reasoning. We also discuss the difference of mathematical and commonsense reasoning from logical reasoning in §A.1. More discussions on the difference between chain-of-thought and papers reviewed in this paper can be found in §A.1.1.

11 Ethics Statement

This article follows the ACL Code of Ethics. To our best knowledge, there are no foreseeable potential risks to use the datasets and methods in this paper.

References

A. Aamodt and E. Plaza. 1994. Case-based reasoning:foundational issues,methodological variations,and system approaches. *AI communications*.

Qiming Bao, Alex Yuxuan Peng, Tim Hartill, Neset Tan, Zhenyun Deng, Michael Witbrock, and Jiamou Liu. 2022. Multi-step deductive reasoning over natural language: An empirical study on out-of-distribution generalisation. The 2nd International Joint Conference on Learning and Reasoning and 16th International Workshop on Neural-Symbolic Learning and Reasoning (IJCLR-NeSy 2022).

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. *CoRR*, abs/2201.06028.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3829–3839. European Language Resources Association.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Inf. Process. Manag.*, 60(1):103111.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3882–3890. ijcai.org.

Luca Console and Lorenza Saitta. 2000. On the relations between abductive and inductive explanation. *Abduction and Induction: Essays on their Relation and Integration*, pages 133–151.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *CoRR*, abs/2205.09712.

Andrew Cropper, Sebastijan Dumancic, Richard Evans, and Stephen H. Muggleton. 2022. Inductive logic programming at 30. *Mach. Learn.*, 111(1):147–172.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7358–7370. Association for Computational Linguistics.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems. *CoRR*, abs/2204.13074.

Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, and Andrew McCallum. 2022. Knowledge base question answering by case-based reasoning over subgraphs. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 4777–4793. PMLR.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9594–9611. Association for Computational Linguistics.

Joe Davison, Joshua Feldman, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1173–1178. Association for Computational Linguistics.

Li Du, Xiao Ding, Ting Liu, and Bing Qin. 2021. Learning event graph knowledge for abductive reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5181–5190. Association for Computational Linguistics.

Peter A Flach and Antonis C Kakas. 2000. Abductive and inductive reasoning: background and issues. In *Abduction and induction*, pages 1–27. Springer.

Alexander Gaskell, Yishu Miao, Francesca Toni, and Lucia Specia. 2022. Logically consistent adversarial attacks for soft theorem provers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4129–4135. ijcai.org.

Vinod Goel, Gorka Navarrete, Ira A Noveck, and Jérôme Prado. 2017. The reasoning brain: The interplay between cognitive neuroscience and theories of reasoning.

Ben Goertzel, Nil Geisweiller, Lucio Coelho, Predrag Janičić, and Cassio Pennachin. 2011. *Real-World Reasoning: Toward Scalable, Uncertain Spatiotemporal, Contextual and Causal Inference*, volume 2. Springer Science & Business Media.

Nicolas Gontier, Koustuv Sinha, Siva Reddy, and Christopher Pal. 2020. Measuring systematic generalization in neural proof generation with transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Joseph Y Halpern. 2017. *Reasoning about uncertainty*. MIT press.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq R. Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. FOLIO: natural language reasoning with first-order logic. *CoRR*, abs/2209.00840.

Ruixin Hong, Hongming Zhang, Xintong Yu, and Changshui Zhang. 2022. METGEN: A module-based entailment tree generation framework for answer explanation. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1887–1905. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *CoRR*, abs/2212.10403.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 725–736. Association for Computational Linguistics.

Heng Ji. 2018. Information extraction. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems, Second Edition*. Springer.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3441–3460. Association for Computational Linguistics.

Emils Kadikis, Vaibhav Srivastav, and Roman Klinger. 2022. Embarrassingly simple performance prediction for abductive natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 6031–6037. Association for Computational Linguistics.

Henry Kautz. 2022. The third ai summer: Aaai robert s. engelmore memorial lecture. *AI Magazine*, 43(1):93–104.

Janet L Kolodner. 1997. Educational implications of analogy: A view from case-based reasoning. *American psychologist*, 52(1):57.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Trans. Assoc. Comput. Linguistics*, 3:585–597.

Zhengzhong Liang, Steven Bethard, and Mihai Surdeanu. 2021. Explainable multi-hop verbal reasoning through internal monologue. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1225–1250. Association for Computational Linguistics.

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.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13480–13488. AAAI Press.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10367–10378. Association for Computational Linguistics.

Kostas S. Metaxiotis, Dimitris Askounis, and John E. Psarras. 2002. Expert systems in production planning and scheduling: A state-of-the-art survey. *J. Intell. Manuf.*, 13(4):253–260.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6097–6109. Association for Computational Linguistics.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. 2022. A property induction framework for neural language models. *CoRR*, abs/2205.06910.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *CoRR*, abs/1909.08855.

Stephen H. Muggleton and Luc De Raedt. 1994. Inductive logic programming: Theory and methods. *J. Log. Program.*, 19/20:629–679.

Stephen H. Muggleton, Luc De Raedt, David Poole, Ivan Bratko, Peter A. Flach, Katsumi Inoue, and Ashwin Srinivasan. 2012. ILP turns 20 - biography and future challenges. *Mach. Learn.*, 86(1):3–23.

Mark A Musen and Johan Van der Lei. 1988. Of brittleness and bottlenecks: Challenges in the creation of pattern-recognition and expert-system models. In *Machine Intelligence and Pattern Recognition*, volume 7, pages 335–352. Elsevier.

Terezinha Nunes. 2012. *Logical Reasoning and Learning*, pages 2066–2069. Springer US, Boston, MA.

Debjit Paul and Anette Frank. 2020. Social commonsense reasoning with multi-head knowledge attention. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2969–2980. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2021. Generating hypothetical events for abductive inference. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021, Online, August 5-6, 2021*, pages 67–77. Association for Computational Linguistics.

Gabriele Paul. 1993. Approaches to abductive reasoning: an overview. *Artif. Intell. Rev.*, 7(2):109–152.

Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *CoRR*, abs/2212.09597.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 794–805. Association for Computational Linguistics.

Hanhao Qu, Yu Cao, Jun Gao, Liang Ding, and Ruifeng Xu. 2022. Interpretable proof generation via iterative backward reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2968–2981. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Danilo Neves Ribeiro, Shen Wang, Xiaofei Ma, Rui Dong, Xiaokai Wei, Henghui Zhu, Xinchi Chen, Peng Xu, Zhiheng Huang, Andrew O. Arnold, and Dan Roth. 2022. Entailment tree explanations via iterative retrieval-generation reasoner. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 465–475. Association for Computational Linguistics.

Kyle Richardson and Ashish Sabharwal. 2022. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11209–11219. AAAI Press.

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 2020, Online, November 16-20, 2020*, pages 122–136. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, and Mohit Bansal. 2021. multiprover: Generating multiple proofs for improved interpretability in rule reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3662–3677. Association for Computational Linguistics.

Merrilee H Salmon. 1989. Introduction to logic and critical thinking.

Soumya Sanyal, Zeyi Liao, and Xiang Ren. 2022a. Robustlr: Evaluating robustness to logical perturbation in deductive reasoning. *CoRR*, abs/2205.12598.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022b. Fairr: Faithful and robust deductive reasoning over natural language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1075–1093. Association for Computational Linguistics.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1466–1476. The Association for Computational Linguistics.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4505–4514. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Zayne Sprague, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction with incomplete information. *CoRR*, abs/2211.00614.

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, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3621–3634. Association for Computational Linguistics.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. *CoRR*, abs/2210.12217.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenguan Wang and Yi Yang. 2022. Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing. *arXiv preprint arXiv:2210.15889*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Kaiyu Yang, Jia Deng, and Danqi Chen. 2022a. Generating natural language proofs with verifier-guided search. *CoRR*, abs/2205.12443.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022b. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.

Zonglin Yang, Xinya Du, Erik Cambria, and Claire Cardie. 2023. End-to-end case-based reasoning for commonsense knowledge base completion. In *EACL*.

Zonglin Yang, Xinya Du, Alexander M. Rush, and Claire Cardie. 2020. Improving event duration prediction via time-aware pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3370–3378. Association for Computational Linguistics.

Nathan Young, Qiming Bao, Joshua Bensemann, and Michael Witbrock. 2022. Abductionrules: Training transformers to explain unexpected inputs. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 218–227. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*.

Wenting Zhao, Justin T Chiu, Claire Cardie, and Alexander M Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. *arXiv preprint arXiv:2305.14618*.

Yunchang Zhu, Liang Pang, Yanyan Lan, and Xueqi Cheng. 2020. L2r^2: Leveraging ranking for abductive reasoning. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1961–1964. ACM.

A Appendix

A.1 Relation to Related (NLP) Fields

In this section, we first introduce related NLP fields to general logical reasoning, then introduce fields that are only related to deductive reasoning, inductive reasoning, or abductive reasoning. We hope that this section could be helpful to form a clear shape of LRNL in NLP.

A.1.1 Logical Reasoning

Neuro-Symbolic Computing Neural-symbolic computing (NeSy) is a hybrid of symbolism and connectionism to exploit advantages from both sides (Wang and Yang, 2022; Cambria et al., 2022). The knowledge representation of its symbolic part basically is a knowledge graph or propositional logic or first-order logic (Wang and Yang, 2022). LRNL could be seen as a new type of NeSy in addition to the existing 6 types summarized by Kautz (2022), as its goal and design of methodology are typically symbolic (logical reasoning with knowledge bases), while avoiding any symbolic representation, using (currently pure) neural methods.

Natural Language Inference Natural language inference (NLI) is generally considered as the semantic concepts of entailment and contradiction (Bowman et al., 2015). Here logical reasoning tasks can be viewed as special types of NLI focusing on particular reasoning aspects.

Question Answering The form of LRNL looks similar to question answering (QA), however, QA is conducting one-step logical reasoning only when the context provides enough information to answer the question (deductive reasoning), or the answer is a generalization of an argument in context or question (inductive reasoning), or the answer is to provide explanations to the question (abductive reasoning).

Commonsense Reasoning Commonsense reasoning (CR) and logical reasoning (LR) are similar in that they both involve "knowledge" and "reasoning". Compared to LR, CR focuses more on the "knowledge" aspect. Some typical tasks include whether a system has commonsense knowledge (Bosselut et al., 2019; Yang et al., 2020), and whether a system's answer is commonsense-knowledge-aware (Bisk et al., 2020); LR focuses more on the "reasoning" aspect, e.g., whether a system's i/o behaviors follow reasoning requirements (Clark et al., 2020).

Chain of Thoughts Chain of thoughts (COT) (Wei et al., 2022) is a prompting technique that can elicit the step-by-step reasoning ability of LLMs without finetuning.

COT can potentially be used for each of the three sub-reasoning types of logical reasoning. In fact, for a given (commonsense reasoning) question, some reasoning steps of COT could be deductive, and others can be inductive or abductive. Since the purpose of this paper is to provide a finer analysis on logical reasoning, we do not intentionally cover prompting techniques such as COT.

It is also argued by several modular-based deductive reasoning methods that COT's reasoning is not casual (Creswell et al., 2022), limited by input size (Ribeiro et al., 2022), and contains unrelated or incorrect steps (Hong et al., 2022; Tafjord et al., 2022).

Overall, it could be interesting to use COT-related methods specifically for deductive, inductive, or abductive reasoning (as opposed to modular-based methods), and it is a less-explored research direction.

A.1.2 Deductive Reasoning

Multi-hop Reasoning Compared to proof generation, many multi-hop reasoning tasks (Yang et al., 2018; Jiang et al., 2020; Min et al., 2019; Sinha et al., 2019) are much simpler, often being single-branched (Qu et al., 2022), consisting of only 2-3 supporting facts, and are more coarse-grained, involving large chunks of texts such as passages instead of simple, short sentences (Yang et al., 2022a).

Nevertheless, some multi-hop reasoning datasets can be considerd as conducting deductive reasoning. For instance, for each data in CLUTRR (Sinha et al., 2019) dataset, a set of facts that can make conclusive support to the target kinship relation is included in background information as input for each target relation, hence from the philosophical definition (Salmon, 1989), it requires to perform deductive reasoning.

Mathematical Reasoning In many mathematical reasoning tasks such as math word problem solving (Koncel-Kedziorski et al., 2015) and geometry problem solving (Seo et al., 2015), the conclusion can be conclusively entailed by the premise. Therefore these tasks belong to deductive reasoning. We do not review math-related papers because we want to focus solely on the challenge of deductive reasoning while mathematical reasoning involves numbers in the text, which introduces additional challenges.

A.1.3 Inductive Reasoning

Information Extraction Information Extraction (IE) is a task of extracting pre-specified types of facts from written texts or speech transcripts, and converting them into structured representations (Ji, 2018). The rule generation task here also extracts rules from facts represented in written texts. The difference is that IE pursues extracting the exact information from existing texts, while inductive reasoning aspires to induce more general rules from existing texts, where the information in rules goes beyond what is exactly stated in the texts.

Case-based Reasoning Case-based Reasoning (CBR) is a classic AI subject, whose methods share a general methodology of four steps: retrieve, reuse, revise, and retain (Aamodt and Plaza, 1994). Recently there has been research works devoting to bridge the research of CBR and NLP, by using NLP techniques for CBR challenges (Yang et al., 2023) and improving NLP tasks with CBR methodologies (Das et al., 2021, 2022; Yang et al., 2023). CBR could be seen as a type of analogical reasoning (Kolodner, 1997), and analogical reasoning belongs to inductive reasoning (Salmon, 1989). However, CBR is a different inductive reasoning type than the "generalization" process (from facts to rules) described in Flach and Kakas (2000), but more on the general description on inductive reasoning (Salmon, 1989) that premises cannot provide conclusive support to the conclusion.

A.1.4 Abductive Reasoning

Casual Reasoning In logic research, causal reasoning aims at an epistemological problem of establishing precise causal relationships between causes and effects. It is generally considered a form of inductive reasoning (Goertzel et al., 2011), since inductive reasoning is to derive rules that lead from one to another. When the focus is to derive possible causes from effects, the problem belongs to abductive reasoning (Goertzel et al., 2011).

A.2 More Details About the Difference Between Inductive Reasoning and Abductive Reasoning

We adopt Console and Saitta (2000)'s view on the difference between inductive and abductive reasoning: both inductive and abductive reasoning provide explanations about the world but their explanations differ in the degree of generality.

For instance, an inductive hypothesis allows the validity of properties, observed on a set of individuals, to be generalized to other individuals not in the observations, whereas an abductive one allows unobserved properties to be applied to observed individuals.

The distinction between inductive and abductive hypotheses strictly parallels the dichotomy *extension* vs. *intension*, or *generality* vs. *informativeness*. In other words, an inductive hypothesis extends or generalizes to unobserved individuals, while an abductive one provides more specific information (e.g., unobserved properties) about existing specific individuals.

For example, if a white ball is found in a bag, inductive reasoning might lead to the conclusion that "all balls in this bag are white", while abductive reasoning might lead to the conclusion that "someone put the white ball into this bag".

In this example, inductive hypothesis generalizes the property of existing individual (a found white ball) to unobserved individuals (other not seen balls in the bag), while abductive hypothesis provides more specific information about the current individual (who brought this ball to the bag).

To summarize in simple words, in common situations, pure inductive reasoning is to only provide (usually sample to population) generalizations, while pure abductive reasoning is to only provide specific explanations.

In reality, some hypotheses can be both inductive and abductive. Console and Saitta (2000) labels non-deductive hypotheses as inductive, inductive/abductive or abductive.

A.3 Related Surveys on Reasoning

Huang and Chang (2022); Qiao et al. (2022) mainly reviews the prompting techniques for LLMs, but do not focus on papers that specialized on logical reasoning (the coverage of the two fields are quite different).

Yu et al. (2023) is a concurrent work of ours and reviews papers related to reasoning. However, it does not focus on logical reasoning, particularly the three subtypes of logical reasoning. The advantage of our survey is that we provide a finer analysis on logical reasoning (including more detailed definition and categorization of logical reasoning from philosophy literature, comparison with the classic AI paradigm on logical reasoning, and organizing the survey based on the three subtypes of logical reasoning).

A.4 Experiments Summarization

In this section, we summarize the experiment results of important and literature-abundant task.

Until now there has been only one or two papers working on inductive reasoning. Methods for abductive reasoning generally leverage different resources (such as multi-task, additional knowledge resources, and ancillary loss) and lack an progressive relationship between each other, therefore are less comparable. Currently the $ProofGeneration$ task in deductive reasoning are the most literature-abundant, and methods for this task have progressive relationships with each other. Therefore here we mainly summarize results and analyze for the $ProofGeneration$ task.

Table 5 shows the summarized experiment results. We select the most widely used tasks to display their performance. Among the task, the setting of ParaRules is trained on D3 (D* dataset with depth 3) and test on ParaRules test set; the setting of Birds-Electricity is trained on D5 (D* dataset with depth 5) and test on bird-electricity set; setting for EntailmentBank is the task 3 which uses full corpus as input (so that many distractors exist in input); setting for OBQA and QuaRTz are zero-shot setting while model pretrained on another dataset (EntailmentBank).

Among the methods, Creswell et al. (2022) and Bostrom et al. (2022) design unique metrics using EntailmentBank dataset, and Sprague et al. (2022) focus on a unique task (proof generation task with incomplete information), therefore we do not list their experiments results in the table.

Overall methods for proof generation tasks tend to use different datasets for evaluation, making them less comparable.

A.5 Meaning of "More General" Required by Inductive Reasoning

This section is collected from Yang et al. (2022b)'s appendix, to help illustrate inductive reasoning.

Given an argument consisting of a premise and a conclusion, if the conclusion involves new information that is not covered by the premise and can not be conclusively entailed by the premise, the argument is an inductive argument (Salmon, 1989).

When the conclusion has a larger scope of information coverage than the premise, and can entail the premise, it can be said that the conclusion is "more general" to the premise (Yang et al., 2022b). In this case, we termed the premise as a "fact", and the conclusion as a "rule"; When the conclusion contains new pieces of information and cannot entail the premise, as defined by Salmon (1989), the argument is still an inductive argument. But in this case, we termed the premise as a "fact", and the

Methods	ParaRules	Birds-Electricity	EntailmentBank (Task 3)							OBQA	QuaRTz
	Full Accuracy (FA)	Full Accuracy (FA)	Leaves F1	Leaves All-Cor.	Steps F1	Steps All-Cor.	Intermediates F1	Intermediates All-Cor.	Overall All-Correct	Accuracy	Accuracy
PRover	95.1	80.5	-	-	-	-	-	-	-	-	-
multiPRover	94.5	81.8	-	-	-	-	-	-	-	-	-
EntailmentWriter	-	-	39.7	3.8	7.8	2.9	36.4	13.2	2.9	-	-
ProofWriter	98.5	97.0	-	-	-	-	-	-	-	-	-
EVR	-	63.1	-	-	-	-	-	-	-	-	-
IBR	95.7	93.5	-	-	-	-	-	-	-	-	-
IRGR	-	-	45.6	12.1	16.3	11.8	38.8	36.5	11.8	-	-
Selection-Inference	-	-	-	-	-	-	-	-	-	-	-
FaiRR	98.6	-	-	-	-	-	-	-	-	-	-
MetGen	-	-	34.8	8.7	9.8	8.6	36.7	20.4	8.6	-	-
SCSearch	-	-	-	-	-	-	-	-	-	-	-
ADGV	-	-	-	-	-	-	-	-	-	-	-
NLProofS	-	-	43.2	8.2	11.2	6.9	42.9	17.3	6.9	-	-
Entailer	-	-	-	-	-	-	-	-	-	76.8	74.3
Teachme	-	-	-	-	-	-	-	-	-	77.0	75.9

Table 5: Proof Generation Task Results.

conclusion as another “fact”.

For instance, if facts that are about cats and dogs are good accompaniment of humans, then some examples of a “more general” rule can be (1) mammals are good accompaniment of humans, or (2) domesticated animals are good accompaniment of humans, or (3) animals with four legs are good accompaniment of human.

In these examples, the rules cover a larger scope than the facts (e.g., mammals compared to cats; domesticated animals compared to cats), and therefore the rules are “more general” than the facts.

“More general” means not only about finding higher taxonomic rank, but can be in unlimited forms. For instance, if the fact is about the Sun rises and falls every day, then some examples of a “more general” rule can be (1) the Earth is the king of the universe or (2) the Earth is rotating itself.

Both rule examples are “more general” than the given fact, since the rule can entail not only the given fact, but also other not mentioned facts such as the observable movements of the other stars in the Milky Way.