Natural Language Reasoning, A Survey

FEI YU, The Chinese University of Hong Kong, Shenzhen, China
HONGBO ZHANG, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
PRAYAG TIWARI, School of Information Technology, Halmstad University, Halmstad, Sweden
BENYOU WANG, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China

This survey article proposes a clearer view of **Natural Language Reasoning (NLR)** in the field of **Natural Language Processing (NLP)**, both conceptually and practically. Conceptually, we provide a distinct definition for NLR in NLP, based on both philosophy and NLP scenarios; discuss what types of tasks require reasoning; and introduce a taxonomy of reasoning. Practically, we conduct a comprehensive literature review on NLR in NLP, mainly covering classical logical reasoning, **Natural Language Inference (NLI)**, multi-hop question answering, and commonsense reasoning. The article also identifies and views backward reasoning, a powerful paradigm for multi-step reasoning, and introduces defeasible reasoning as one of the most important future directions in NLR research. We focus on single-modality unstructured natural language text, excluding neuro-symbolic research and mathematical reasoning.[1]

CCS Concepts: • **Computing methodologies** → **Natural language processing**; *Machine learning;*

Additional Key Words and Phrases: Natural language reasoning, pre-trained language models

ACM Reference Format:
Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural Language Reasoning, A Survey. *ACM Comput. Surv.* 56, 12, Article 304 (October 2024), 39 pages. https://doi.org/10.1145/3664194

1 INTRODUCTION

NLP has shown significant advancements in recent years, particularly with the introduction of transformers and **Pre-trained Language Models (PLMs)**. However, their abilities[2] to perform NLR are still far from satisfactory. Reasoning, the process of making inferences based on existing knowledge, is a fundamental aspect of human intelligence and is essential for complex tasks such as decision-making. Building an artificial intelligence system capable of reasoning is both

[1] https://github.com/FreedomIntelligence/ReasoningNLP
[2] In this survey, we refer to transformer-based PLMs.

This work is supported by the Shenzhen Science and Technology Program (JCYJ20220818103001002), Shenzhen Doctoral Startup Funding (RCBS20221008093330065), and Tianyuan Fund for Mathematics of National Natural Science Foundation of China (NSFC) (12326608).
Authors' Contact Information: Fei Yu, The Chinese University of Hong Kong, Shenzhen, China; e-mail: 222043013@link.cuhk.edu.cn; Hongbo Zhang, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China; e-mail: hongboz183@gmail.com; Prayag Tiwari, School of Information Technology, Halmstad University, Halmstad, Halland, Sweden; e-mail: prayag.tiwari@ieee.org; Benyou Wang (Corresponding author), School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China; e-mail: wangbenyou@cuhk.edu.cn.

ACM 0360-0300/2024/10-ART304
https://doi.org/10.1145/3664194

the ultimate goal of the research community and the necessary way to improve the performance of complex applications. Compared to reason with formal language, reasoning with natural language expressions provides a more natural human–computer interaction interface and opens the door to research on defeasible reasoning, such as abduction and induction, which are incapable of formal-based symbolic methods.

PLMs such as BERT [34] and GPT [121] have been the essential components in NLP research since they occurred. Pre-trained on large-scale text corpora, PLMs are capable of natural language understanding. Recent progresses suggest that PLMs also have the potential to solve reasoning problems [25, 149, 154, 170]. Specifically, PLMs can perform soft deductive reasoning over natural language statements [25], reason with implicit knowledge memorized in their parameters [154], and perform multi-step reasoning step by step just with a few demonstrations or instructions when the model size is large enough via **Chain-of-Thought (CoT)** prompting [81, 170]. Recently, ChatGPT and GPT4 also have demonstrated impressive reasoning capabilities to the community [4, 16].

However, while reasoning has attracted increasing attention recently [25, 27, 28, 81, 113, 151, 170], there is still lacking a distinct definition of reasoning and the term "reasoning" is sometimes of mistaken usage, which may affect the communication and development toward reasoning in the NLP community. For example, while it belongs to "commonsense reasoning," few people might deem that telling about a shared lived experiences [10], e.g., "name something that you might forget in a hotel room," is reasoning. Another example is that sometimes "NLI" is introduced as a task of natural language understanding [12], but other times as a task of reasoning [25]. By now, not all of the tasks named with "reasoning" are thought of as reasoning (e.g., commonsense reasoning), and not all of the tasks named "without reasoning" are thought of as non-reasoning (e.g., NLI and multi-hop question answering). This raises a question: what is reasoning actually, and how can we identify reasoning tasks if their names are not very indicative? Although many research works [25, 59, 181, 187] refer to a definition of reasoning from philosophy and logic, the definition cannot capture the reasoning in NLP well enough. For example, while reasoning is philosophically defined as "using evidence and logic to arrive at conclusions" [59], it fails to clarify whether implicit commonsense knowledge can be evidence and what types of conclusion are reasoning products—e.g., how about named-entity disambiguation?

To promote the research on reasoning in NLP, we make an attempt to propose a clearer view of NLP reasoning, both conceptually and practically. Conceptually, we propose a definition for NLP reasoning based on both philosophy and NLP scenarios, discuss what types of tasks require reasoning, and introduce a taxonomy of reasoning. Practically, we provide a comprehensive literature review on NLR in NLP based on our clarified definition, mainly covering classical logical reasoning, NLI, multi-hop question answering, and commonsense reasoning. Reviewing papers of all sizes of PLMs, we capture general methodologies that can be applied to different model sizes: end-to-end reasoning, forward reasoning, and backward reasoning. Finally, we discuss some limitations and future directions of reasoning.

This article distinguishes itself from previous surveys by providing a distinct definition of NLR from the perspective of NLP, advancing beyond the philosophical and logical definitions [59] or the simple listing of representative reasoning applications and tasks [116]. By integrating relevant reasoning theories into the NLP domain, we underscore the subtleties and shared aspects among NLR tasks. Grounded in our definition, we present an examination of the rationale behind the classification of certain NLP tasks as reasoning and the identification of tasks inaccurately labeled as reasoning. In methodology, our review extends beyond the scope of **Large Language Models (LLMs)** [59, 116] to include smaller language models, thereby revealing the under-explored paradigm of backward reasoning. In terms of applications and tasks, our survey specifically focuses on NLR, diverging from the wider scopes of earlier surveys [59, 116]. This results in a broader

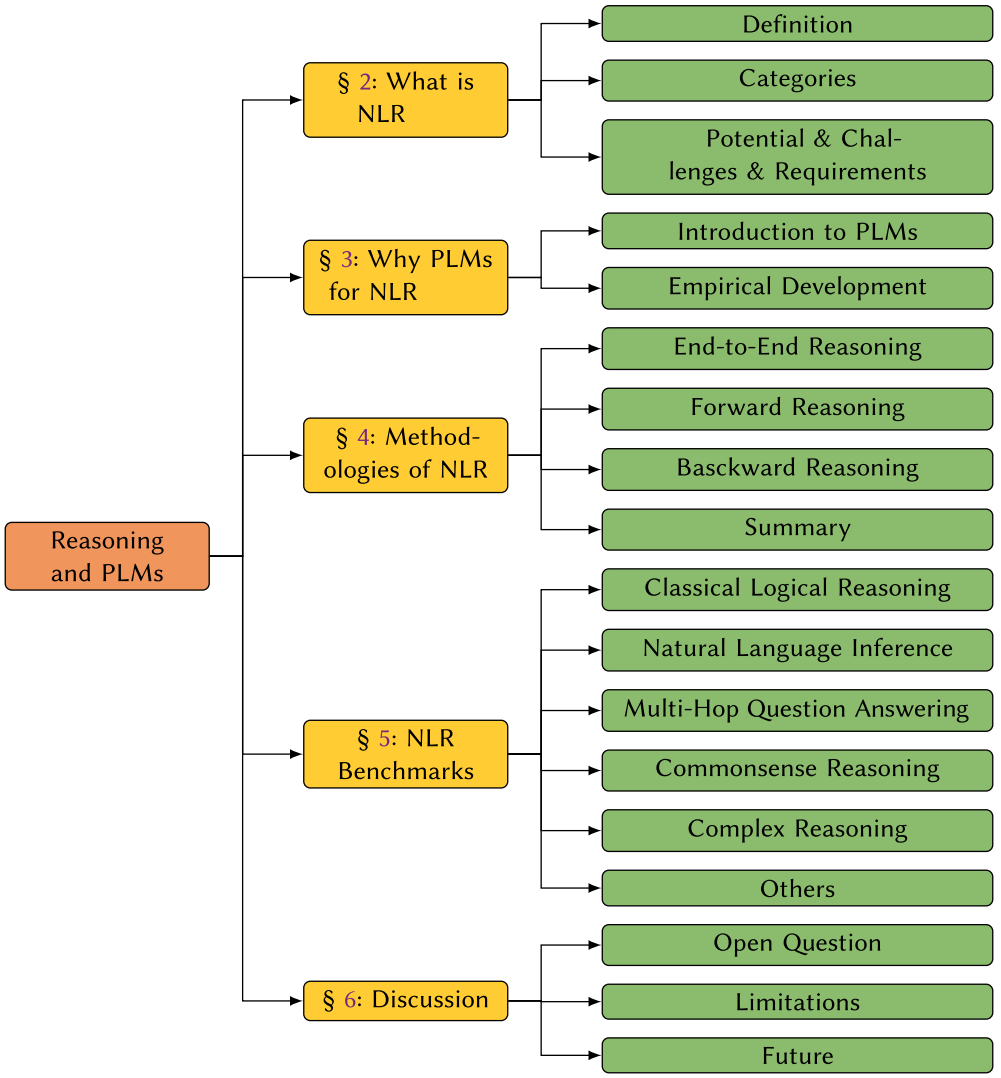

Fig. 1. Architecture of this survey.

spectrum of NLR applications and tasks, such as multi-hop question answering and defeasible reasoning. Additionally, we offer a comprehensive collection of benchmarks for NLR.

In this article, we focus on the single-modality unstructured natural language text (without knowledge triples, tables, and intermediate formal language) and NLR (rather than symbolic reasoning and mathematical reasoning).[3] Concretely, we conduct a review of related works that utilize transformer-based PLMs, with a deliberate exclusion of neuro-symbolic techniques. We sorted the collected papers and categorized the methodologies of NLR in NLP. We identify the progress and trends in recent years in this domain. The article is organized into five sections (as shown in Figure 1). Timeline of important works is shown in Figure 2.

[3]Although recently it is popular to solve mathematical reasoning problems such as math word problems using NLP methods, we do not cover them in this article since mathematical reasoning is very different to NLR in nature as math is precise and formal.

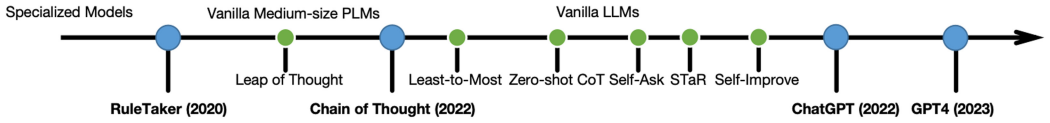

Fig. 2. Timeline of important works.

We collected more than 200 papers related to reasoning or PLMs in recent years. We searched keywords such as inference, reasoning, infer, reason, multi-step, and multi-hop on the top conferences, including ACL, EMNLP, NAACL, ICML, ICLR, and NeurIPS, from 2019 to 2022. We also found some related works from the collected papers.

In conclusion, the main contributions of this survey are:

(1) To our best knowledge, we are the first to provide a distinct definition for NLR in NLP and discuss to what degree some popular benchmarks are related to reasoning.
(2) To our best knowledge, we are the first to conduct a comprehensive review on PLM-based NLR, covering diverse NLR benchmarks, and providing a comprehensive taxonomy of methodology. We also cover backward reasoning, which is neglected but has potential.
(3) We introduce defeasible reasoning, which we believe is one of the future directions with the most potential; compare differences between deductive reasoning and defeasible reasoning; discuss how they can affect NLP solutions; and review current methods.

2 WHAT IS NLR

There is still lacking a distinct definition of NLR in NLP, which affects the development and communication of NLR in the NLP community. To promote understanding, analysis, and communication, we aim to suggest distinct definitions of terms and concepts for NLR in NLP. To realize this goal, we take a look into two relevant areas that have studied reasoning for a long time—philosophy and logic—and transfer the relevant reasoning theory into NLP. First, we propose a definition for NLR in NLP that satisfies the concerns of the NLP community (Section 2.1). Then, we provide categories of NLR and introduce how the differences between them can affect NLP solutions (Section 2.2). Finally, we introduce the potentials, challenges, and requirements to achieve NLR (Section 2.3).

2.1 Definition

Reasoning in NLP has been focused on in recent years, while philosophy has studied reasoning for thousands of years, and logic is seen as the art of correct reasoning, which studies the concepts of inference, systematizes its categories, and develops principles of good reasoning, including formal logic and informal logic [9, 46, 65]. In this section, we first include reasoning theory from philosophy and logic and derive it into NLP reasoning. Then, we review some NLR topics in NLP. Finally, we propose a definition for reasoning in NLP, which combines the definition in philosophy and logic and the concerns of the NLP community.

2.1.1 Definition from Philosophy and Logic. Here we introduce two descriptions and three definitions of reasoning from philosophy and logic: task-based description (Description 2.1), negation-based description (Description 2.2), logic-based definition (Definition 2.1), assertion-based definition (Definition 2.2), and action-based definition (Definition 2.3). The former two descriptions can tell us "what reasoning can do" and "what isn't reasoning," while the latter three provide us different definitions of "what is reasoning." However, the definition from logic (Definition 2.1) restricts reasoning to a subset within the coverage of formal logic. To reach a more generalized definition,

Table 1. Comparison and Combination of Descriptions about Reasoning from Philosophy and NLP

	What Is Reasoning	What Isn't Reasoning
Philosophy	infer a new assertion from a set of assertions infer an action from goals and knowledge	sensation, perception, and feeling direct recourse to sense perceptions or immediate experience
NLP	more than understanding, slow thinking e.g., multi-hop QA, commonsense reasoning	memorize, look up, match information e.g., text summarization, style transfer
Combination	a dynamic process to integrate multiple knowledge to get new conclusions, rather than direct recourse to memorized or provided first-hand information	

we adopt the latter two definitions from philosophy, which are two different classes named theoretical reasoning and practical reasoning, respectively, as the basis for defining NLR in NLP.

Description 2.1 (Task Based). Reasoning is an essential mental activity when conducting conscious tasks with complex computations such as problem-solving, decision-making, persuasion, and explaining [3, 41, 48, 73].

Description 2.2 (Exclusion Based). Reasoning is a dynamic process to get some knowledge *without* direct recourse to sense perceptions or immediate experience, which *is opposed to* sensation, perception, and feeling [3, 14, 165].

Definition 2.1 (Logic-based Reasoning). Reasoning is to discover valid conclusions by applying logic [3, 14, 41, 94, 165].

Definition 2.2 (Assertion-based Reasoning/Theoretical Reasoning). Reasoning is to infer conclusions from a set of premises, consisting of one or more inference steps, where premises and conclusions are assertions that claim something is true or false about the world [3, 9, 14, 130, 165].

Definition 2.3 (Action-based Reasoning/Practical Reasoning). Practical reasoning is to infer actions from goals and knowledge, which is oriented to deciding whether an action is practically reasonable [9, 165].

2.1.2 Definition in NLP We Suggest. According to Definition 2.2, Definition 2.3, and negation-based Description 2.2, we can know "what is reasoning" and "what isn't reasoning" from the perspective of philosophy. There are also some descriptions toward the two questions in NLP. We compare and combine them in Table 1. We also review typical NLR datasets in NLP to observe and capture what the NLP community is concerned about.

From our observations, in NLP, NLR also combines multiple knowledge to derive conclusions. The unique characteristics are (1) knowledge sources and (2) conclusion types. First, common knowledge sources are knowledge bases, context, and PLMs, where the former two can explicitly provide encyclopedic knowledge and contextual knowledge, while the last is implicit knowledge sources. Second, in addition to assertions and actions, it is also popular to infer relations, e.g., causes and effects, of events. We demonstrate examples of these three conclusion types in Table 2.

Correspondingly, we first propose the formal definition of NLP reasoning in Definition 2.4. Then, to facilitate a more intuitive understanding, we elaborate "what isn't reasoning in NLP" and "what NLP reasoning can do" in Description 2.3 and Description 2.4. It should be emphasized that conclusions are new (or unknown) assertions, events, or actions, which distinguishes reasoning from other knowledge-intensive tasks that may also require multiple knowledge. Specifically, we demonstrate some examples of knowledge-intensive datasets and explain why they are not reasoning in Table 3.

Definition 2.4 (NLP Reasoning). NLR is a process to integrate multiple knowledge (e.g., encyclopedic knowledge and commonsense knowledge) to derive some new conclusions about the (realistic or hypothetical) world. Knowledge can be from both explicit and implicit sources. Conclusions are assertions or events assumed to be true in a world, or practical actions.

Table 2. Three Types of Conclusion in Reasoning, Where "Assertion" and "Event" Assume Something True or Likely to Be True in the World

	Premise	Conclusion
Assertion	Cat is animal. Animal can breathe.	Cat can breathe.
Event	John was shot. There are people around. Doctor can save life.	John will be sent to see a doctor.
Action	Marry is in the living room. Marry feels it is hot. Remote control for air conditioner is in the bedroom.	Go to the bedroom, take the remote control, come back, and turn on the air conditioner

Table 3. Examples to Explain What Is Not Reasoning

	CoNLL	CommonGen	Natural Questions
Task	entity linking	generate a sentence describing a daily scenario using the given concepts (constrained text generation)	open-domain QA
Input example	They performed Kashmir, written by Page and Plant.	dog, frisbee, catch, throw	Question: what color was john wilkes booth's hair? Context: ... He stood 5 feet 8 inches tall, *had jet-black hair* ...
Output example	Kashmir -> Kashmir (song); Page -> Jimmy Page; Plant -> Robert Plant	A *dog* leaps to *catch* a *thrown frisbee*.	jet-black
Why not reasoning	Align known entities without producing new assertions, events, or actions	New text, but neither claim true assertions or events nor generate actions	Claim "john wilkes booth's hair is jet-black," but the knowledge is directly given in the context, without demand on knowledge integration

Description 2.3 (NLP Exclusion Based). NLR is to derive new assertions, events, or actions *without* direct recourse to models' memorization, knowledge base storage, and the provided context.

Description 2.4 (NLP Task Based). Reasoning is an important *method* to arrive at the required answers or solutions. This approach is necessary when what we need is neither directly provided by context nor memorized by models and stored by knowledge bases, but reachable by integrating available information.

2.1.3 Key Concepts. We first introduce the key concepts: proposition and inference. Similarly, we derive the definitions from philosophy and logic to NLP. Then, we further clarify the definition of reasoning in NLP.

Definition of key concepts. In logic, the proposition is the basic operation unit in reasoning, and inference is a sub-process of a complete reasoning process. Concretely, while reasoning is performed with statements (as premises and conclusions), the real operation units are the semantics behind sentences, i.e., propositions [65]. Inference is a single step in reasoning [9, 13, 52, 130, 165], and each reasoning can be made of one or more inference steps (Definition 2.2). We put the two key concepts into NLP in Definition 2.5 and Definition 2.6.

Definition 2.5 (NLP Proposition). A proposition is the semantic meaning or information content of a statement rather than its superficial linguistics.[4]

Definition 2.6 (NLP Inference). Inference is a single step that produces a single (intermediate) conclusion from some premises.

[4]We exclude lexical inference from our scope in this article, concentrating on semantics-level reasoning.

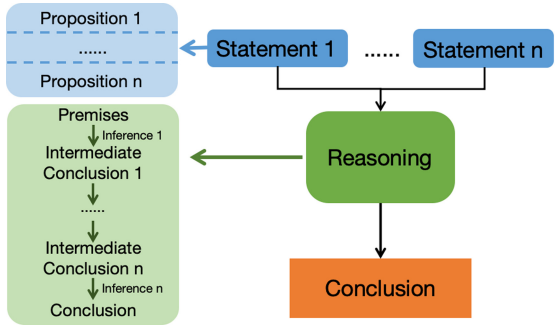

Fig. 3. Reasoning process. The premises can be either explicit or implicit knowledge, e.g., PLMs' memory.

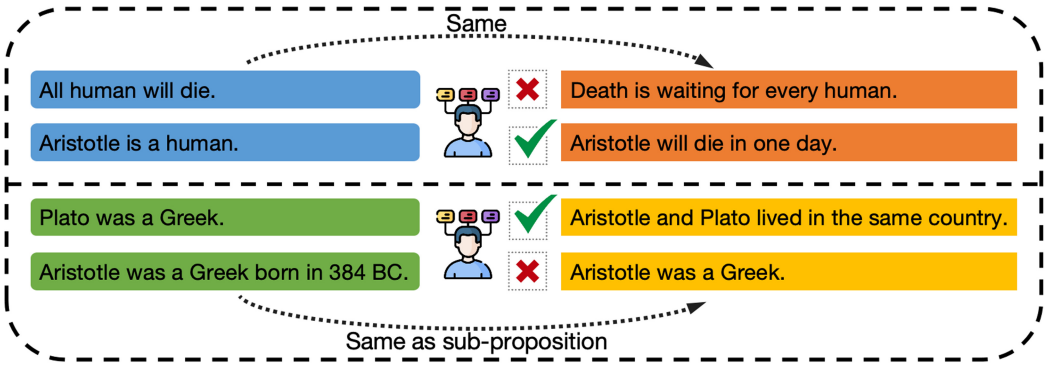

Fig. 4. Examples to show the key idea of "semantic difference," where check mark denotes reasoning while cross denotes not reasoning.

Further clarify the definition of NLP reasoning. We leverage these concepts to clarify the definition of NLP reasoning: what we mean by "integrate multiple information to derive new conclusions" is that (1) a single sentence conveying multiple semantics can provide multiple premises and (2) there must yield new semantics in inference and reasoning; i.e., conclusions are semantically different to all premises. We detail two examples to demonstrate this key idea (Figure 4) and illustrate the definition of reasoning in Figure 3.

2.2 Categories of Inference

While knowledge has been well categorized in NLP (e.g., explicit world knowledge and implicit commonsense knowledge), we find that there is still a lack of reasonable taxonomy for inference. Therefore, we borrow the categories from philosophy and discuss the differences between classes to NLP and how they can affect the solutions.

Inference can be divided into (mainly) deductive, inductive, and abductive [43, 109] or divided into monotonic and defeasible. Actually, the deduction is monotonic inference, while induction and abduction are sub-classes of defeasible inference. Since "monotonic" and "defeasible" can capture the difference between deductive and non-deductive inference, we combine the two taxonomies into one: deductive inference and defeasible inference.

2.2.1 Deduction, Induction, and Abduction. According to Aristotle and Peirce, there are three major inferences: deduction, induction, and abduction [43, 109]. This taxonomy is the most familiar

Table 4. A Simple Example to Show the Difference between Deduction, Abduction, and Induction, Where Text in Black Is the Given Knowledge While Text in Red Is the Inferred Knowledge

Fact1: Aristotle is a human Rule: All humans will die Fact2: Aristotle will die		
Deduction	**Abduction**	**Induction**
(Fact1 + Rule ⟶ Fact2)	(Fact1 + Rule ⟵ Fact2)	(Fact1 + Fact2 ⟶ Rule)

"Fact" denotes specific knowledge while "rule" denotes general principle.

one to the NLP community, adopted and studied by several works [7, 25, 50, 106, 145, 172, 181, 187]. The definitions are shown below.

Definition 2.7 (Deduction). A deductive inference is to infer valid knowledge (conclusion) from the given knowledge (premises).

Definition 2.8 (Induction). An inductive inference is to infer probable knowledge, which describes a more general rule,[5] extrapolated from the given knowledge.

Definition 2.9 (Abduction). An abductive inference is to infer probable knowledge, as the best explanation (i.e., cause), for the given knowledge (i.e., phenomena).

See an example in Table 4. Abduction involves hypothesizing the most plausible explanation for given observations, wherein one observation acts as the conclusion and the rest as premises in its reversed process. This reversal leads to deduction, with the explanation hypothesized during abduction serving as the (missing) premise necessary to complete the logical deduction. Deduction involves applying a general rule to derive more specific knowledge, moving from the general to the specific. Conversely, induction aims to identify a general principle shared among specific facts, moving from the specific to the general.

However, among these three classes, research on abduction and induction is much more underexplored than deduction, while the widely studied deduction is only a very small set of human daily reasoning.

2.2.2 Deductive Inference and Defeasible Inference. Our main goal is to promote research on non-deductive reasoning and highlight the differences and challenges. Therefore, we turn to monotonic inference and defeasible inference, which can better capture the features of deductive and non-deductive inference, respectively.

Key difference. The key difference between monotonic inference and defeasible inference from philosophy is that the former derives valid conclusions[6] while the latter only produces probable conclusions. Since the conclusions of deductive inference are truth preserving, the future added knowledge will not affect their validity, and thus the set of knowledge is incremental, i.e., monotonic. By contrast, the conclusions of non-deductive inference (e.g., induction and abduction) may be wrong, and the newly added knowledge may retract the conclusion, i.e., defeasible. For example, one may inductively infer "birds can fly" with the premises "parrots can fly" and "eagles can fly." However, when one discovers the new knowledge "ostrich cannot fly," the conclusion will be retracted.

[5]A rule encapsulates a shared principle relevant to a range of facts. For example, the principle "A man will die" applies universally to individuals such as "Aristotle" and "Plato," illustrating a commonality despite the specificity of each case.
[6]"Valid" means when the premises are true, it is impossible for the conclusion to be false.

Table 5. The Characteristics of the Deductive Inference and Defeasible Inference

	Deductive Inference	Defeasible Inference
Conclusion	true	probably true
Inference relation	support	strengthen, weaken, rebut
Quality of inference	valid or invalid	weak to strong
Required knowledge	bounded	unbounded

Different characteristics. This difference toward conclusions between deductive inference and defeasible inference leads to many different characteristics, including inference relations between premises and conclusions, the quality of inference, and the requirement of knowledge. Concretely, there is only one inference relation between premises and each conclusion in deductive inference, i.e., support, and the inference is either valid or invalid. Therefore, we can derive a valid conclusion just with several supporting premises. By contrast, knowledge can strengthen, weaken, and even rebut (the probability of) the conclusion in defeasible inference, and the quality of inference varies from weak to strong. Therefore, it is better to collect more comprehensive information to arrive at a more probable conclusion. We compare the characteristics of the deductive inference and defeasible inference in Table 5.

Affects on NLP. These characteristics affect relevant knowledge acquisition, reasoning path structure, and the importance of interpretability in NLP. First, while collecting the supporting knowledge toward the valid conclusion is enough for deductive reasoning, it is better to collect both supportive and opposing knowledge to compare the confidence of different conclusions for defeasible reasoning. Then, there has been increasing attention on reasoning path generation in NLP [133, 151, 170]. However, due to more types of inference relation, the structure of reasoning paths for defeasible reasoning is more complex than deductive reasoning and thus becomes more challenging to generate. Finally, it is more important and sometimes even crucial for NLP models to perform interpretable defeasible reasoning. This is because people with different background knowledge can infer very different and even opposite conclusions by themselves, and thus it is much more difficult to clarify the conclusion without explicit premises and reasoning procedures.

2.3 Potentials, Challenges, and Requirements of NLR

Potentials. Compared to reasoning with precise formal language, natural language provides a better human–computer interaction interface. Besides, natural language opens a door to play with defeasible reasoning, where formal language fails.

Challenges. First, natural language suffers from ambiguity and variety, since there are polysemy, synonymy, and diverse structures. Therefore, while triples and formal languages are precise, statements and propositions are many-to-many in natural language, which poses a challenge on natural language understanding. Second, supervised data of inference is difficult to obtain, which may prevent it from large-scale training. Moreover, the step of reasoning is diverse at the instance level; i.e., different questions may require different inference steps to answer, and it is important to generalize to the unseen steps.

Requirements. Based on the definition (Definition 2.4), the key components in NLP to achieve reasoning are (1) (multiple) knowledge and (2) an algorithm capable of understanding and inference. Correspondingly, there are three stages: knowledge acquisition, knowledge understanding, and inference. First, it requires collecting the relevant knowledge required for reasoning (knowledge acquisition). Then, the algorithm requires to capture propositions underlying the given knowledge

(knowledge understanding). In addition to the general semantics, it should also capture the logical semantics such as negation, conjunction, and disjunction. Subsequently, beginning from these propositions, the algorithm requires integrating some knowledge to infer a new conclusion with one or more steps to reach the final answer (inference). Though knowledge acquisition and understanding are also necessary for reasoning, the two topics are big enough to write another survey, and thus we just focus on inference in this article.

3 WHY PLMS FOR NLR

3.1 Introduction to PLMs

PLMs are based on transformer architecture [163], which is built with many attention modules and pre-trained on massive amounts of text data via unsupervised learning techniques such as predicting masked tokens [34] or generating the next tokens [121]. Since BERT [34] occurred, pretraining-then-finetuning became a common paradigm, which transfers the general abilities of PLMs learned in the pretraining stage to downstream tasks with further task-specific fine-tuning. Since LLMs have been found to be few-shot learners [15], in-context learning has become a new popular paradigm, which can predict a new sample with only a few demonstrations without fine-tuning parameters. Recently, the zero-shot prompting paradigm has also become more popular in LLMs [81].

Types of PLMs. According to the architecture, PLMs can be divided into encoder-only (e.g., BERT [34]), decoder-only (e.g., GPT [121]), and encoder-decoder (e.g., T5 [122]). According to the directivity, PLMs can be divided into bidirectional (encoder-only) and causal (decoder-only and encoder-decoder); while bidirectional PLMs are commonly used for discriminative tasks, causal PLMs can model general tasks but are more capable of generative tasks. According to the model size, there are medium-size PLMs and LLMs, where LLMs are much larger than the former (e.g., 13B parameters).

Advantages of PLMs for NLR. We conclude with four advantages of PLMs for NLR:

- **Ability of natural language understanding**. Transformers represent words and sentences in a context-dependent manner as continuous vectors in a high-dimensional space dealing with ambiguity and uncertainty in nature. After large-scale pretraining, PLMs can learn a powerful understanding capability, which helps them capture and understand knowledge mentioned in the text.
- **Ability to learn implicit knowledge into parameters**. It has been found that PLMs can capture some implicit knowledge that is not explicitly mentioned, such as commonsense knowledge, into their parameters. This is important since it is impossible to explicitly enumerate and provide commonsense knowledge for reasoning.
- **Ability of in-context learning**. LLMs such as GPT-3 exhibit the impressive ability to perform tasks only with some demonstrations without further fine-tuning, which is valuable to alleviate data sparsity problems.
- **Emergent abilities**. Recently, it was found that LLMs have some emergent abilities that only occur when the model size is big enough [169], and LLMs can perform much more complex tasks as their size increases. Moreover, it has been demonstrated that performing multi-step reasoning in a few-shot or zero-shot manner is one of the emergent abilities [170].

3.2 Empirical Development

Recent progress also shows the potential to leverage PLMs on NLR, which exhibits their learning and generalization abilities of reasoning skills with both explicit and implicit knowledge.

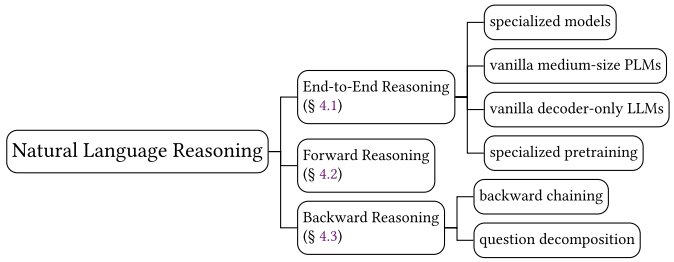

Fig. 5. Taxonomy of NLR.

By fine-tuning on the specific dataset, [25] first demonstrated that PLMs can perform deductive reasoning over explicitly provided natural language statements, which can zero-shot transfer to different domains. Moreover, [154] showed that PLMs can combine memorized implicit taxonomic and world knowledge with explicitly provided knowledge for the deduction. In addition to deduction, PLMs can also learn to perform defeasible reasoning [129, 181, 187].

While LLMs with in-context learning were once thought to be incapable of multi-step reasoning, it has been found that their capabilities of reasoning can be unlocked by generating forward reasoning paths before the final answer [170], which is called CoT prompting. With this prompting, the performance of many multi-step reasoning tasks in Big-Bench Hard can surpass the average human rater. Furthermore, LLMs can perform multi-step reasoning not only with few-shot exemplars; [81] also found that they can automatically produce intermediate steps with a simple "Let's think step by step" prompting in a zero-shot manner. Surprisingly, LLMs can even learn from their self-generated reasoning paths [61, 191]. Moreover, GP4 outperformed a majority of people on several realistic examinations such as the Uniform Bar Exam that also require some reasoning.

In addition, to forward reasoning paths, question decomposition, a backward reasoning method, is also effective in multi-hop question answering, which is beneficial to both medium-size PLMs [102, 108] and LLMs [108, 113].

Moreover, while neural-based methods are blamed for black-box prediction, [27, 138] demonstrated that PLMs can produce faithful reasoning paths and make predictions based on them.

In conclusion, PLMs can learn to perform multi-step reasoning from supervised data or few-shot demonstrations. Their capabilities of natural language understanding, generalization, and leveraging implicit knowledge make them promising to deal with arbitrary natural language, commonsense knowledge, and defeasible reasoning.

4 METHODOLOGIES OF NLR

In this section, we introduce three types of NLR approaches: end-to-end reasoning (Section 4.1), forward reasoning, and backward reasoning. The overall taxonomy is shown in Figure 5.

The key difference among these three categories lies in the reasoning path. Concretely, "end-to-end reasoning" only predicts the final answers without any intermediate text as all the intermediate steps are performed implicitly, while the latter two approaches can produce explicit reasoning paths, containing one or more steps with the intermediate conclusions, showing the process of (possibly multi-step) reasoning that links premises to the conclusion.[7]

Presenting the reasoning path for each prediction can improve the interpretability of a system. Especially, a strict reasoning path can also explicitly expose the supporting knowledge of each

[7]There are also some research works on producing natural language explanations instead of reasoning procedures, but we just focus on reasoning paths in this survey.

Table 6. Comparison of End-to-end Reasoning, Forward Reasoning, and Backward Reasoning

	Direction	Pros	Cons
End-to-End Reasoning	–	most efficient	black box bad generalization
Forward Reasoning	bottom-up	interpretability open-ended	huge search space only effective in LLMs
Backward Reasoning	top-down	interpretability efficient	goal specific

step. Moreover, producing reasoning paths has been demonstrated to be beneficial to the final performance of multi-step reasoning [81, 108, 113, 149, 170]. There are two directions of reasoning.

Two Directions of Reasoning. Multi-step reasoning can be performed by either forward [28, 138, 150, 170] or backward reasoning [77, 86, 102, 113, 151]. Forward reasoning is a bottom-up procedure, which starts from the existing knowledge and repeatedly makes inferences to obtain new knowledge until the problem is solved. The other, backward reasoning, is a top-down procedure, which starts from the problem and repeatedly breaks down into sub-problems until all of them can be solved by the existing knowledge. While backward reasoning targets the specified problems, forward reasoning can freely uncover new knowledge implicated by the existing knowledge without preassigned problems. Accordingly, the search space of forward reasoning is much larger than backward reasoning when solving a specific problem, facing the combinatorial explosion as the step of inference goes. When it comes to theorem proving, which is a verification problem, where the reasoning path is named "proof," forward reasoning and backward reasoning are often called "forward chaining" and "backward chaining," respectively.

We compare these three methods in Table 6 and demonstrate an example in Figure 6. The following subsections will further introduce and discuss the comparison.

4.1 End-to-end Reasoning

End-to-end reasoning is a complete black-box prediction that only outputs the final answers without any explanation, intermediate conclusion, or reasoning path, whether it is a single-step or multi-step reasoning problem. In this process, PLMs are required to recall and integrate knowledge encoded within their parameters implicitly to derive the final answer. There are mainly three kinds of models used to perform end-to-end reasoning: specialized models built upon medium-size PLMs, vanilla medium-size PLMs, and decoder-only LLMs. Besides, there is also some research on specialized pretraining methods.

4.1.1 Training Specialized Models. To perform end-to-end reasoning, models need to aggregate multiple knowledge and reason over them. Correspondingly, there are specialized models improving the capability of multiple evidence aggregation [36, 97, 196, 198] or reasoning [40, 87, 119, 183, 200]. Previous research often incorporated some task-specific inductive biases via architectural designs. For example, graph neural networks are popularly used to leverage edges (e.g., entity–entity relations) to promote information aggregation and integration between nodes (e.g., entity information) [40, 119, 183]. However, these designs only specialize in either specific tasks or datasets. By contrast, ReasonFormer [201] proposed a variant architecture of transformer for general reasoning, with different modules responsible for different predefined fundamental reasoning capabilities. This kind of model can improve performance on specific tasks or datasets. Nevertheless, all of these designs rely heavily on handcrafts, introducing strong prior assumptions, which may hurt the generalization ability to other tasks.

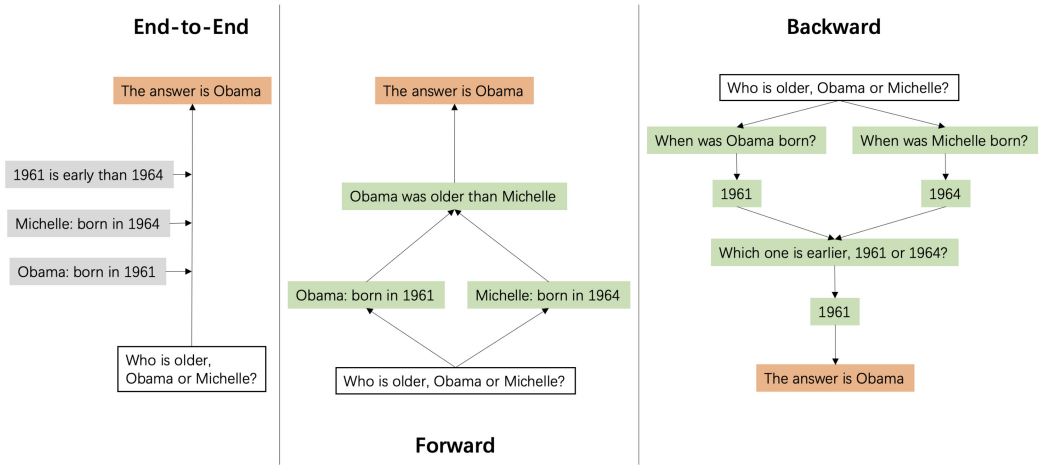

Fig. 6. An example to demonstrate the reasoning procedure of end-to-end reasoning, forward reasoning, and backward reasoning. White colors show the question, green colors the intermediate text, orange colors the answer, and gray colors the implicit knowledge inside the PLM's parameters.

4.1.2 Fine-tuning Vanilla Medium-size PLMs. Medium-size PLMs lack the ability to perform zero-shot reasoning such as theorem proving, argument completion, commonsense reasoning, and abduction without training [6, 25, 187, 205]. Recently, it was found that transformers can be good soft deductive reasoners after in-domain training [6, 25]. By contrast, it is more challenging to perform defeasible reasoning [129, 181, 187].

Deductive reasoning. Both bidirectional and causal PLMs have demonstrated learning ability for deductive reasoning. [25] first found that BERT and RoBERTa (bidirectional PLMs) can perform theorem proving over synthetic natural language facts and rules after training. When it comes to causal PLMs, [6] demonstrated that GPT2 can learn to reason over deductively valid arguments and is able to generalize from simple core schemes to some unseen composite schemes. However, there are two challenging problems in this paradigm: data sparsity and spurious correlations.

Due to data sparsity, many researchers resort to synthetic data, which is far away from the realistic setting [6, 25, 150]. Moreover, researchers demonstrated that training RoBERTa on synthetic data fails to generalize to linguistic variations on theorem proving and commonsense reasoning [25, 205], which indicates they learn less of the general logical structure underlying the linguistic variations. While training on high-quality data [50] can alleviate the spurious correlation problem [50], such data is difficult to annotate on a large scale. Although automatic data collection can obtain large-scale examples, it is restricted to limit reasoning types dependent on the designed heuristic methods [11].

On the other hand, PLMs are found to learn spurious correlations on multi-hop reasoning, theorem proving, and commonsense reasoning [101, 194, 205]. In other words, fine-tuning on specific tasks and datasets may lead models to overfit to the specific spurious correlations underlying them. There are several researchers trying to reduce artifacts in the dataset such as by adding adversarial data [70] and carefully constructing the new dataset [55, 160]. However, it is difficult to construct data without any artifact, and there may be some statistical features inherent in the problem that cannot be avoided in principle [194]. Another line to alleviate shortcuts is increasing attention on reasoning path generation, which may encourage models to perform actual reasoning (Section 4.2).

Defeasible reasoning. The capability of defeasible reasoning seems to be more challenging for vanilla medium-size PLMs to learn. Specifically, [129] demonstrated that the performance of BART-large and T5-large on a defeasible reasoning task, i.e., generate a statement to update the strength of a probable conclusion, is far from satisfaction. There is a similar observation on inductive reasoning [181]. Besides, it is hard to generalize the ability of abduction learned in a synthetic dataset to unseen domains [187].[8] While data sparsity is also a challenge for defeasible reasoning, how to better enable PLMs capable of defeasible reasoning remains a problem.

4.1.3 Few-shot Decoder-only LLMs. Few-shot prompting using decoder-only LLMs without finetuning can alleviate data sparsity and also prevents models from overfitting to specific tasks or datasets. However, there remains the question of whether models can be better capable of reasoning as the model size increases.

Although the performance on reasoning problems improves as the model size increases [28, 50, 129, 181], it is still unclear how much progress can be attributed to the improvement in reasoning capability. [28] demonstrated that (deductive) reasoning problems are much more challenging and the scaling laws (of the Gopher family) work much slower than other tasks in BigBench and vanilla LLMs struggle with multi-step reasoning problems. [113] found that while LLMs memorize more factual knowledge as the model size increases, it seems their ability to implicitly integrate knowledge for deduction does not improve.

Surprisingly, more reasoning capabilities of LLMs can be elicited by CoT prompting, as introduced in Section 4.2.

4.1.4 Specialized Pretraining. To improve the reasoning capability of PLMs, there is some research on introducing inductive biases of reasoning with continual pretraining [33, 71, 111, 139]. There are type-specific inductive biases [71, 111] and type-agnostic inductive biases [33, 139]. For example, [33] incorporated the general inductive bias of reasoning over multiple long evidence texts, while [71] mainly designed for relational reasoning. Inductive biases are introduced with reasoning-related data and training strategies. For example, [139] collected reasoning-related text that involves logical inference keywords and let models to self-supervised predict these keywords. While the pretraining improves performance on multi-hop reasoning and logical reasoning problems, especially in the low-resource setting [33, 71, 139], all of them only worked on encoder-only PLMs, i.e., BERT and RoBERT.

Recently, [111] proposed a new line that leverages programs such as SQL to pretrain PLMs with synthesized (program, execution result) pairs. The results are inspiring that PLMs, including medium-size, large-size, encoder-only, and encoder-decoder, can attain significant improvement in multi-hop reasoning and logical reasoning.

However, it is important to ask whether it is still beneficial to incorporate inductive biases into LLMs or whether simply increasing the model size and pretraining on more data is enough to improve reasoning capability. In other words, can LLMs learn reasoning well enough just by the current general pretraining? Maybe LLMs have already learned powerful reasoning capabilities that just need to be elicited via smart prompting such as CoT [170].

4.2 Forward Reasoning

Forward reasoning repeatedly composes existing knowledge to derive new knowledge until reaching the answers. There are two kinds of benefits to producing a forward reasoning path: trustworthiness [27, 31, 138] and performance improvement [28, 149, 170].

[8]By contrast, the ability of deduction learned in the synthetic dataset can be generalized to other domains [25].

Trustworthiness. Showing how multiple knowledge interacts and contributes to new conclusions can contribute to the system's interpretability. Furthermore, when the prediction is based on the reasoning procedure, it can alleviate the widespread shortcut problem. To exhibit the structure of reasoning, involving the required knowledge and their inference relation, reasoning paths are often represented as directed graphs or trees [31, 64, 96, 133]. Typically, each node represents one piece of knowledge and the edge represents the inference relation between knowledge. For example, a single inference linking two premises to one conclusion can be represented as two nodes linking to their shared parent node.

Deductive reasoning. There is only one inference relation in deductive reasoning, i.e., support. To construct such an interpretable reasoning path, it needs to find the relevant knowledge as premises and infer the conclusions (inference). Since inference is to produce new knowledge with the given premises, it is usually implemented by vanilla generative PLMs [126, 138]. Instead of explicitly selecting or retrieving the required knowledge [133], some works put the context into the input and modeled both knowledge selection and inference as unified generation [150, 180]. However, it may generate hallucinations and invalid inferences. To alleviate this problem, [180] leveraged an additional verifier to score the validity. In addition to just improving the validity of the knowledge node and inference relation edge, some researchers proposed performing faithful reasoning, which forces the prediction to rely on reasoning paths. This is mainly realized by designing decoupled modular frameworks to avoid shortcuts to irrelevant context [27, 56, 138]. For example, [27] iteratively performed knowledge selection and inference alternately in a step-by-step manner, where each inference step only conditions the currently selected knowledge to infer the conclusion without seeing the question and the previous steps. Both supervised modular frameworks based on medium-size PLMs [56, 138] and in-context learning modular frameworks based on LLMs [27] have been explored to perform faithful reasoning. In addition to faithfulness, such step-decoupling behaviors also bring other effects. On the one hand, it is easier to provide supervised training data or in-context exemplars. The supervised framework can leverage one-step supervision [56, 138] to train the system, which alleviates the data sparse problem in multi-step reasoning, while the in-context learning framework can demonstrate representative one-step examples that avoid the challenge of selecting the appropriate exemplars for multi-step reasoning [27, 28]. On the other hand, it brings error propagation. However, all of these works consider the simplest setting, where all the required knowledge is either explicitly provided in context or retrievable from knowledge bases.

Defeasible reasoning. There are more types of inference relations in defeasible reasoning, i.e., strengthen, weaken (the probability of the conclusion), and rebut. Since it is difficult to collect all the supporting premises, research on this line mainly concerns the label of inference relations between statements. In other words, there exists implicit reasoning; i.e., some premises are not explicitly provided. Similar to deductive reasoning, reasoning paths can be generated by one-shot generation [64, 96] or the faithful modular framework [64]. Different from deductive reasoning, it is more challenging to generate defeasible reasoning paths that even fine-tuned LLMs (T5-11B) find difficult.

Challenge. However, there remains a problem with the evaluation of the constructed reasoning path. Specifically, there may be multiple reasoning paths for each problem, which poses challenges on data annotation [31] and automatic evaluation [27]. Annotating all possible reasoning paths for evaluation is impractical, especially for those long-step problems facing combinatorial explosion. And it is also challenging to automatically evaluate the validity of reasoning paths without annotated data.

Performance improvement. Reasoning path can also be used to improve the answer performance on multi-step deductive reasoning, including the in-domain performance of LLMs and the

generalization ability of PLMs. For this purpose, it is not necessary to involve all the required knowledge in the reasoning path or keep the validity of inferences as what we are concerned about are the final results rather than reasoning paths.

Improvement on in-domain performance. First, reasoning paths can improve the in-domain performance by providing an enriching context [149, 170] or supervision signal [23, 189]. Recently, [170] demonstrated that the LLMs' performance of several reasoning tasks such as commonsense reasoning (both deductive and defeasible) can be significantly improved by generating a reasoning path before the final answers, which is called CoT prompting. Before this, while LLMs are successful in classical NLP tasks, they fail in reasoning, especially multi-step reasoning tasks. This finding boosted a series of research on this line [27, 28, 54, 81, 98, 143, 144, 149, 168, 184, 189, 197]. Especially, [81] showed that even a simple zero-shot prompting "let's think step by step" can activate LLMs to perform commonsense reasoning and attained impressive performance. Furthermore, [168] found that the final performance on commonsense reasoning can be greatly further improved by just voting the results on multiple reasoning paths. Besides, in addition to performing reasoning on downstream tasks via few-shot prompting without changing the parameters, supervised fine-tuning LLMs on CoT annotations can further improve their reasoning capability [23, 189]. In addition to commonsense reasoning, the performance of classical logical reasoning and multi-step reasoning are also improved significantly by generating CoT [28, 149]. However, classical logical reasoning is much more challenging than other typical tasks [28]. Instead of one-shot CoT generation, [28] proposed a more inspiring framework (SI) for theorem proving (a task of classical deductive reasoning) based on modules with different prompting, which outperforms 40x larger LLMs with CoT. Moreover, [61, 191] found that LLMs can self-improve their reasoning capabilities by fine-tuning their self-generated reasoning paths. However, such abilities are only effective in LLMs; i.e., the model scale should be large enough, which is also seen as an emergent ability of LLMs that can be elicited by few-shot [170] and even zero-shot prompting [81]. There are some research works transferring the CoT reasoning capability of LLMs to smaller models via knowledge distillation [54, 98, 144].

Improvement on generalization. Moreover, it can improve the generalization ability of PLMs. It has been observed that constructing the proof graph for the goal hypothesis can improve the zero-shot generalization ability of medium-size PLMs to the unseen step of reasoning [133, 135, 150] and to the unseen domain [56, 150] on the theorem proving task, which is likely because it forces models to perform reasoning rather than exploit shortcuts. Also, turning one-shot construction into a stepwise procedure has a better generalization to the unseen steps of reasoning and to cross-datasets [56, 150].

Challenge. However, the search space of forward reasoning suffers from a combinatorial explosion as the number of reasoning steps increases. In addition to performing a single-step inference, planning is also very important to multi-step reasoning, especially to deep steps. It has been observed that while LLMs are capable of a single inference, they still struggle to plan on deep reasoning steps [28, 143]. Yet this topic is under-explored with few research works [27, 180].

In addition to deductive reasoning, leveraging reasoning paths to improve performance on defeasible reasoning is still under-explored.

4.3 Backward Reasoning

Backward reasoning repeatedly breaks down problems into sub-problems and solves them until reaching the answers. Similar to forward reasoning, it can be used to produce trustworthy reasoning paths explicitly represented with knowledge and inference relations [56, 120, 151] or improve the final performance without strict structures [74, 77, 102]. It faces a smaller search space and thus is more efficient than forward reasoning. There are two popular backward reasoning

methods: backward chaining and question decomposition. While the former is a proof-finding strategy, the latter is a general strategy available for general problems. Research works mentioned in this section are mainly about deductive reasoning.

4.3.1 Backward Chaining. Backward chaining is the preferable approach for proof-finding by humans. Beginning from the goal, it repeatedly performs abductive reasoning to derive the potential premises as sub-goals until all the sub-goals can be proved or disproved by the existing knowledge. According to the source of the premises, or sub-goals, there are two kinds of abduction: predict part of premises (others are the existing knowledge) and predict all premises. The first one is to predict the unknown required premise for a conclusion with the existing knowledge, which can be realized by vanilla generative PLMs, either medium size [56] or large size [74]. The other one is to predict all the premises from scratch without relying on the existing explicit knowledge, which can be realized by LLMs [151]. While the former kind of abduction is easier to perform, the latter can solve the scenario where all the required premises do not exist in the knowledge base. Compared to forward chaining, backward chaining has a smaller search space and thus is more efficient [74]. Moreover, [74] proposed a backward chaining modular framework with LLMs as modules, which attains better performance than the existing forward chaining frameworks. Another direction is to perform forward chaining (deduction) and backward chaining (abduction) simultaneously [56]. In addition to proof-finding, backward chaining can also be generalized to more general problems. For example, [151] applied it to a multi-choice question-answering problem by combining the question and each answer choice into a verifiable hypothesis.

However, research works on this line are more under-explored than forward chaining.

4.3.2 Question Decomposition. Question decomposition is a backward reasoning method to improve performance on multi-hop questions that require integrating multiple pieces of knowledge and inferring over them to obtain the answers. It decomposes each question into several simpler sub-questions and answers these sub-questions to derive the final answers. In analogy to forward reasoning, solving a single-hop sub-question is to query a single piece of knowledge, and combining sub-answers to form the final answer is inference. Decomposing a question into sub-questions is an abductive step. In other words, while question decomposition introduces abduction steps, it removes the requirement of multi-step knowledge selection/retrieval.

Multi-hop questions are difficult to answer because they have a long tail distribution and it is challenging to find the relevant multiple pieces of knowledge. Especially, it might be very challenging to find the required knowledge for implicit multi-hop questions, whose superficial text and semantics can be very different from the required knowledge. By contrast, it is easier to query a piece of knowledge and answer each decomposed single-hop sub-question. For example, [108] demonstrated that both medium-size PLMs and LLMs can significantly improve the performance on multi-hop questions with human-decomposed questions. It was also found to be effective in mathematical reasoning and symbolic reasoning [204]. Besides, previous research has also shown that question decomposition is effective with both medium-size PLMs [203] and LLMs [113] on multi-hop questions. Research of this line has a longer history than LLM-only CoT methods.

Decomposition of explicit and implicit multi-hop question. According to the difficulty of decomposition, multi-hop questions can be divided into explicit multi-hop questions and implicit multi-hop questions. Explicit multi-hop questions are those which can be decomposed simply based on their superficial text (syntactical pattern). For example, the question "where was Obama's wife born?" can be decomposed into "who is Obama's wife?" and "where was #1 born?"[9] based on the

[9]"#1" denotes the answer of the first sub-question.

superficial text of the original question. Implicit multi-hop questions, however, are more difficult to decompose since their sub-questions are not syntactically consistent with the questions. For example, the question "can we directly live in the space?" needs to be decomposed into "what do we need to stay alive?" and "are there #1 in the space?," where the key predicate in the first sub-question "need" is not explicitly mentioned in the original question. While explicit multi-hop questions can be decomposed based on their superficial text and syntactical structures via extraction and editing [102], decomposing implicit multi-hop questions is much more difficult. A key challenge is that it lacks large-scale annotated data, which is labor intensive to obtain especially as the number of hops increases. StrategyQA [45] is an implicit multi-hop question dataset annotated with sub-questions and the corresponding knowledge pieces, but its size is small (2.7k). To alleviate the data sparsity problem, there is some research on weak supervision data [110, 203]. Recently, in-context learning provides a new solution [108, 113] to this problem, which requires only a small set of demonstrations.

Framework with respect to sequential and tree structure. There are different structures of decomposition based on the dependencies among the parent question and sub-questions, involving a sequential structure and tree structure. In a sequential structure, each sub-question is linearly dependent on the answer (e.g., a bridge entity) of the antecedent sub-question, and the answer of the last sub-question is the multi-hop question's answer. For example, the answer "Michelle" of the first sub-question "who is Obama's wife?" makes up of its subsequent sub-question "where was #1 born?," whose answer "1964" is also the final answer of the multi-hop question "where was Obama's wife born?" By contrast, in a tree structure, sub-questions are independent of each other with their answers equally contributing to the final answer. For example, the question "who can swim better, elephant or dolphin?" consists of "can elephant swim?" and "can dolphin swim?," and the final answer is derived by composing the corresponding sub-answer "elephant can't swim" and "dolphin can swim." There are three kinds of decomposition-based frameworks: module-based decomposition [102], decompose-then-recompose [108, 110], and generate-then-answer [77, 113]. The first framework designs different modules responsible for different reasoning types, which separate and model the sequential and tree structure independently [102]. The decompose-then-recompose framework first decomposes the multi-hop question into all its comprised sub-questions and recomposes their sub-answers to derive the final answer [108, 110]. However, it ignores the dependencies among sub-questions (sequential structure). By contrast, the last one, generate-then-answer, is sequential in nature, which iteratively generates and answers a single-hop sub-question [77, 113]. It considers the question dependencies in the sequential structure and is compatible with the tree structure but is less efficient than decompose-then-recompose since it can't solve sub-questions of the tree structure in parallel.

However, it is still challenging to solve multi-hop questions with very long hops. Due to the combinatorial explosion, it becomes increasingly difficult to annotate decomposition supervision data and provide representative demonstrations for in-context learning. Also, there are likely to exist multiple decomposition paths when there are long hops, which also puts a challenge on planning. The following are the potential directions we suggest:

— **Hierarchical decomposition**. Instead of directly decomposing the multi-hop question into the simplest single-hop sub-questions, it might be easier for models to perform hierarchical decomposition, i.e., repeatedly decompose multi-hop questions into simpler multi-hop questions until there are only single-hop questions. Moreover, it is also more practical for researchers to annotate supervision data or select appropriate exemplars of in-context learning for layer-by-layer decomposition.

— **Knowledge-aware planning**. When there exist multiple decomposition paths, it is critical to plan for a decomposition way to answerable sub-questions. For this purpose, it is important to be aware of what existing knowledge there is.

4.4 Summary

Reasoning requires models to integrate multiple knowledge and reason over them. Early research mostly improved reasoning performance via architectural designs and only constructed forward reasoning paths for interpretability or faithfulness. Specialized models were designed to improve evidence aggregation, reasoning capability, or faithfulness, but they are constrained to specific tasks, datasets, or reasoning types that hurt the generalization. Since transformers have been found to be soft deductive reasoners after in-domain fine-tuning, vanilla PLMs have been more popular to perform reasoning. However, data sparsity and spurious correlation problems make it difficult for medium-size PLMs to learn the general logical structure of diverse reasoning types. There are also some research works incorporating inductive biases via specialized pretraining, but it is unclear whether this is still of worth as the model size and the number of pretraining data increase. Recently, it was found that an emergent ability comes as PLMs are large enough: generating a reasoning path before the final answer can significantly improve the multi-step reasoning performance, which boosts much research in this line. In addition to the forward reasoning direction, the other reasoning direction is backward, which is more efficient than forward reasoning due to the smaller search space. While forward reasoning can expose arbitrary new knowledge entailed by the existing knowledge, backward reasoning just targets the specific goal or the problem solution. A typical approach of backward reasoning is question decomposition, which can improve performance on multi-hop questions for both medium-size PLMs and LLMs. While there is much research on deductive reasoning, defeasible reasoning is much more challenging for PLMs and is still under-explored.

5 NLR BENCHMARKS

In this section, we review some typical and popular downstream benchmarks thought to require NLR and discuss to what extent they are actually related to reasoning. Although there might be more downstream benchmarks with respect to NLR, here we mainly focus on four of the most popular and familiar to the community: classical logical reasoning, NLI, multi-hop question answering, and commonsense reasoning. We list the corresponding datasets and benchmarks and briefly introduce the development. Besides, we present some datasets collected from realistic examinations or explicitly designed to challenge LLMs, which we name "complex reasoning." In addition to well-known reasoning benchmarks, we also introduce some other tasks that require performing NLR. A figure of the taxonomy is shown in Figure 7.

5.1 Classical Logical Reasoning

Some datasets explicitly target classical reasoning types in philosophy and logic, e.g., deduction, abduction, and induction, following the definitions in the two areas. Thus, we call them "classical logical reasoning tasks." A key characteristic of this topic is that tasks are mostly artificial to study reasoning. There are both deductive reasoning and defeasible reasoning.

5.1.1 Deductive Reasoning. Classical deductive reasoning tasks are defined formally based on formal logic, such as propositional logic and first-order logic. There are mainly three types of task: inference [6, 106, 172], theorem proving [5, 25, 50, 150], and reasoning path generation [106]. The inference task is to reason the conclusion given the premises in a single step, while theorem proving is to predict whether the given proposition is true or false with the given knowledge bases,

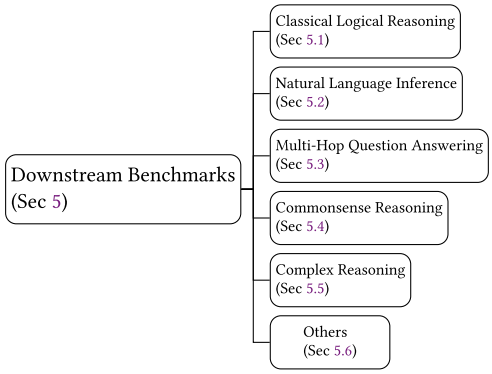

Fig. 7. NLR benchmarks in NLP.

Table 7. Datasets of Classical Deductive Reasoning, Where bAbI-15 Means "the 15th Task in bAbI Tasks"

Dataset	Size	Data Source	Task	Remark
bAbI-15 [172]	–	synthetic	inference	basic deduction
RuleTaker† [25]/**ProofWriter**† [150]	500k	synthetic	theorem proving	the first natural language theorem proving
PARARULE-Plus [5]	400k	synthetic	theorem proving	addresses the depth imbalance issue on ParaRules
AAC [6]	710k	synthetic	inference	based on 8 syllogistic argument schemes
NLSat [127]	406k	synthetic	inference	natural language satisfiability problem
LogicInference [106]	200k	synthetic	inference reasoning path generation	–
FOLIO [50]	1.4k	expert-written	theorem proving	more diverse patterns
LogiGLUE [95]	–	both	hybrid	a collection of many tasks

† denotes there are ground reasoning paths.

which usually requires multiple steps. Obviously, inference is the fundamental task that forms the basic capability of multi-step reasoning tasks such as theorem proving, while reasoning path generation is an interpretable task that can be complementary to multi-step reasoning. However, except for FOLIO [50], all the existing explicit deductive reasoning datasets are synthesized. We list the classical deductive reasoning datasets in Table 7.

Proof-finding and faithful reasoning. Since [25] has proposed a theorem proving dataset and showed that vanilla medium-size PLMs can be soft theorem provers, a series of research works emerged on this task to study NLR, with both vanilla medium-size PLMs [133, 135, 138, 180, 194] and LLMs [27, 28, 74, 143, 150]. However, while the performance of transformers on theorem proving is promising, [194] found that there are some statistical features inherently existing in the problem, which may hinder models from generalization. In addition to just classifying the final label [25, 194], it has been demonstrated that producing proofs can bring better generalization ability to unseen proof depth and out-of-domain data [133, 150] and contribute to interpretability. There are several research works on proof generation or proof-finding, either forward [27, 28, 138, 150] or backward [74, 86, 120], where backward chaining is more efficient than forward chaining on proof-finding intrinsically [74]. To alleviate the combinatorial explosion problem in the search space of the forward chaining, some researchers proposed planning [27, 180]. Moreover, faithful reasoning is also an interesting topic in this problem, where the procedure of reasoning is strictly designed to guarantee that models actually perform reasoning to derive the answer rather than rely on shortcuts [27, 138]. However, while the performance is promising, even approaching perfect sometimes, all research mentioned above is based on synthetic datasets. Moreover, recently, the new expert-written dataset FOLIO [50] showed that when it comes to more diverse natural language, the performance degrades severely. By contrast, the entailment tree generation dataset

Table 8. Datasets of Classical Defeasible Reasoning, Where bAbI-16 Means "the 16th Task in bAbI Tasks"

Dataset	Reasoning	Size	Source	Task	Remark
bAbI-16 [172]	induction	–	synthetic	extraction	induce-then-deduce
CLUTRR [145]	induction	–	synthetic	extractive QA	induce-then-deduce
DEER [181]	induction	1.2k	Wikipedia	generation	rule prediction
AbductionRules [187]	abduction	–	synthetic	generation	abduce from knowledge database
ART [7]	abduction	17.8k	ROCStories [103]	2-choice/generation	abduce from two premises
defeasibleNLI [129]	others	43.8k	other datasets	classification/generation	concern the change of strength

EntailmentBank [31] is often used to study the proof generation and faithful reasoning as with theorem proving [27, 56, 126, 151, 180]. The target hypotheses in this dataset are collected from realistic examinations and proofs are annotated by humans, which is a better alternative for studies on proof generation.

There are also some benchmarks to diagnose models' capabilities on logical semantics understanding [125, 134, 137].

5.1.2 Defeasible Reasoning. Two typical defeasible reasoning types are abduction [7, 187] and induction [172, 181]. There is also another type of defeasible reasoning [129]. Datasets are shown in Table 8. Compared to classical deductive reasoning, research works on defeasible reasoning are still under-explored. Experiments suggested that there remains a large space to improve [181].

Inductive reasoning. Induction produces a more general principle from the given knowledge that can express or explain it. Early datasets require first inducing rules and then applying them to perform deduction, without inducing explicit rules [145, 172]. Recently, a new dataset DEER [181] studies rule prediction, where the task is to induce natural language rules from natural language facts.

Abductive reasoning. Abduction is to predict the best explanation for the observations. According to the mode of reversed reasoning, abduction can provide explanations that constitute premises of whether it is deductive reasoning [187] or defeasible reasoning [7]. Based on the explained objects (i.e., input), abduction may target a small set of premises [7] or a knowledge corpus [187].

Others. In addition to abduction and induction, defeasibleNLI [129] focuses on whether a premise can weaken or strengthen a probable conclusion. There are research works on defeasible inference graphs to improve both human reasoning [96] and machine reasoning performance [97].

5.2 Natural Language Inference

NLI, also known as recognizing textual entailment, is a typical task in NLP. It is a three-way classification task labeled as entailment, contradiction, and neutral, to identify whether the given premise entails a hypothesis. An entailment is described as a conclusion that a person would typically infer from the premise or the implication described by the premise [12, 29].

While NLI is regarded as a natural language understanding [12] or NLR [25] problem, we find it involves examples of both understanding and reasoning problems. Specifically, we identify that there are mainly three types of premise-hypothesis entailment problems: paraphrasing, compound semantics understanding, and reasoning with implicit premises. For the first type, the hypothesis is a paraphrase of the premise. For the second type, the premise is a compound proposition entailing the hypothesis. For the last type, some unstated premises are needed to link the provided premise to the hypothesis. We demonstrate samples from the popular dataset SNLI [12] for each type respectively in Table 9.

Table 9. Examples from SNLI [12] of Three Types of Entailment, Where CSU Indicates "Compound Semantics Understanding"

	Premise	Hypothesis
Paraphrase	Two doctors perform surgery on patient	Doctors are performing surgery
CSU	Two women are embracing while holding to-go packages	Two women are holding packages (Two women are embracing)
Reasoning	A soccer game with multiple males playing (Soccer is a sport)	Some men are playing a sport

The blue-colored sentence is the implicit premise, while the orange-colored sentence is the other semantics of the premise.

Table 10. Datasets of NLI

Dataset	Domain	Size	P Source	H Source	Remark
SNLI [12]/**e-SNLI**† [18]	generic	570k	realistic	human-authored	the first large-scale NLI dataset
MultiNLI [173]	generic	433k	realistic	human-authored	cover more styles and topics
ANLI [104]	generic	162k	realistic	human-authored	collected via adversarial human-and-model-in-the-loop
OCNLI [58]	generic	56k	realistic	human-authored	a large-scale Chinese dataset
XNLI [26]	generic	7.5k	–	–	cross-lingual, based on MultiNLI
SciTail [79]	science	27k	realistic	realistic	the first NLI dataset with entirely realistic data
SciNLI [131]	science	107k	realistic	realistic	–

"P" denotes "Premise" while "H" denotes "Hypothesis". † means that e-SNLI provides explanations for examples of SNLI.

There are several popular generic datasets listed in Table 10, where datasets with realistic hypotheses have fewer hypothesis-only biases than those with human-authored hypotheses. Concretely, it has been found that there are significant biases in human-authored hypotheses [12, 173], with which models can even predict the label without premise [112, 161, 175].

Several datasets and benchmarks of NLI are just understanding problems, such as those presented specifically to probe and improve the model capabilities of coreference resolution, paraphrase, and compound semantics understanding [57, 134, 136, 178, 179]. A representative example is HANS [99], which is a popular controlled evaluation set designed with syntactic heuristics. Also, datasets that are converted from other tasks into NLI style are irrelevant to reasoning when they are not the reasoning problems originally [19, 186]. In addition to the datasets listed in Table 10, NLI is also a popular task in general NLU benchmarks [75, 166, 167, 177], such as OCNLI [58] in CLUE [177] and MultiNLI [173] in GLUE [167].

Interestingly, it was shown that crowdworkers sometimes annotated different labels to the same premise–hypothesis pair [12, 21]. We think this phenomenon can be attributed to the existence of defeasible reasoning, where people with different background knowledge can derive different conclusions.

5.3 Multi-hop Question Answering

Multi-hop question answering studies answering the complex questions that require reasoning over evidence scattered in different contexts;[10] thus, it is also called multi-hop reading comprehension, where candidate contexts are either explicitly provided involving some distractors [55, 160, 171, 182] (distractor setting) or can be retrieved from external knowledge bases such as Wikipedia [45, 115, 182] and WorldTree [76] (retrieval setting). The term "hop" here indicates the number of contexts required to reason rather than the number of inference steps, which describes the behavior moving among different contexts.

[10]There are not only NLR questions but also other types such as numerical comparisons [45, 182].

Table 11. Datasets of Multi-hop Question Answering

Dataset	Domain	Size	CS	QS	AT	Rationale
COMPLEXWEBQUESTIONS [152]	generic	34k	Web	human-rephrased	span	×
BREAK [174]	generic	83k	Wikipedia	human-composed	span	decomposition
WikiHop [171]	generic	51k	Wikipedia	synthetic	option	×
MedHop [171]	medicine	2.5k	Medline	synthetic	option	×
HotpotQA [182]	generic	112k	Wikipedia	semi-synthetic	span yes/no	sentences
R4C [67]	generic	4.6k	Wikipedia	semi-synthetic	span yes/no	triples
BeerQA [115]	generic	530	Wikipedia	human-authored	span yes/no	×
2WikiMultiHopQA [55]	generic	192k	Wikipedia	synthetic	span	sentences triples
MuSiQue [160]	generic	25k	Wikipedia	human-composed	span	paragraphs decomposition★
QASC [76]/**eQASC**† [69]	science	9.9k	WorldTree	human-authored	option	sentences reasoning path [69]★
StrategyQA [45]	generic	2.7k	Wikipedia	human-authored	yes/no	paragraphs decomposition★

† indicates it annotates the rationale for this dataset. "CS" denotes "Context Source", "QS" denotes "Question Source", and "AT" denotes "Answer Type". In CS, the distractor setting is colored blue, while the retrieval setting is colored orange, and black means both. For rationale, ★ means "reasoning path", otherwise "supporting evidence set". "decomposition" indicates the ground annotations of decomposed sub-questions.

Datasets and Benchmarks. We list some typical datasets in Table 11. A key challenge on dataset construction is that it is very label intensive to annotate large-scale multi-hop questions especially when there is a combinatorial explosion as the number of hops increases. Many datasets are synthetic or semi-synthetic [55, 67, 171, 182], where questions are mainly deductive; i.e., the answers are necessarily true with the given contexts. There are two types of rationale: supporting text set [45, 55, 76, 160, 182] and reasoning path including both forward [67, 69] and backward [45, 160]. In addition to these traditional datasets, a recent dataset, COMMAQA [78], proposes to leverage existing systems to cooperate to solve the multi-hop questions rather than train a single giant model for the complex task.

Multi-hop question construction. There are mainly two lines on multi-hop question construction: improve data quality and increase data number. First, it has been found that there are artifacts in HotpotQA that can be leveraged to answer questions without performing multi-hop reasoning [20, 70, 101, 159]. To deal with this problem, one way is to leverage adversarial data [55], and another way is to construct new datasets of high-quality multi-hop questions with carefully designed data collection strategies [55, 160]. Second, as multi-hop questions are difficult to annotate, there are some research works on automatic data generation [39, 107, 188].

Reasoning. After deriving the relevant contexts, it requires aggregating multiple pieces of evidence and reasoning over them. First, there are some specialized models designed for better evidence aggregation [63, 198]. Second, reasoning is usually performed via end-to-end answering [63, 68, 84, 198, 199] or backward decomposition [45, 77, 102, 108, 110, 113]. In this topic, question decomposition (i.e., backward reasoning) is more popular than forward reasoning.

5.4 Commonsense Reasoning

Commonsense reasoning deals with implicit commonsense knowledge, where commonsense knowledge is necessarily required to solve the problem. Such knowledge may be obvious to people but non-trivial to machines since they are difficult to retrieve from the web due to reporting bias—e.g., "when people are hungry, they would like to eat something,"

Table 12. Datasets of "What" Commonsense Reasoning

Dataset	Other Knowledge	Knowledge Source	Size	Task	Rationale
OpenBookQA [100]	science	WorldTree	6k	multi-choice QA	science facts
OpenCSR [87]	science	WorldTree, ARC corpus	20k	free-form QA	×
CREAK [105]	entity	Wikipedia	13k	claim verification	explanation

Table 13. Datasets of "What if"/"Why" Commonsense Reasoning, Where † Denotes There Are Supporting Facts or Reasoning Paths

Dataset	Size	Direction	Context Source	Task	Remark
ROCStories [103]	50k	temporal	human-authored	2-choice QA	–
SWAG [192]	113k	temporal	ActivityNet, LSMDC	multi-choice QA	–
HellaSwag [193]	20k	temporal	ActivityNet, WikiHow	multi-choice QA	an upgraded SWAG
COPA [128]	1k	both	human-authored	2-choice QA	–
Social-IQA [142]	38k	both	human-authored	multi-choice QA	social situations
e-CARE† [37]	21k	both	human-authored	2-choice QA	–
WIQA [158]	40k	forward	ProPara [157]	multi-choice QA	about nature processes
TIMETRAVEL [117]	29k	forward	ROCStories [103]	generation	counterfactual reasoning
ART [7]	20k	backward	ROCStories [103]	2-choice/generation	abductive commonsense reasoning
TellMeWhy [82]	30k	backward	ROCStories [103]	free-form QA	each annotated 3 possible answers
WikiWhy† [53]	9k	backward	human-edited Wikipedia	free-form QA	about Wikipedia entities / events

For direction, "both" indicates there are both forward and backward causal reasoning.

However, although it is named as "commonsense reasoning," not all the datasets are reasoning as defined (Section 2.1), such as querying shared living experiences [10], identifying pragmatic implications [141], and so on [44, 88].

5.4.1 Datasets and Benchmarks. According to the conclusion type, there are mainly three types of reasoning problems in commonsense reasoning: "what" (i.e., assertions or events), "what if/why" (e.g., causal and temporal relations between events), and "how" (i.e., actions).

What. This type of problem is similar to multi-hop question answering, where the problems require combining multiple pieces of knowledge of which some are from external knowledge sources. The key difference is that it requires some commonsense knowledge, which is not explicitly provided, in commonsense reasoning. In other words, the problems require integrating explicit knowledge, such as science [87, 100], with some commonsense knowledge. We list some datasets in Table 12.

What if and why. This type of problem often reasons for causal and temporal relations between events. There are two causal relations, causes and effects, which can be seen as backward causal reasoning and forward causal reasoning, respectively. Take the causality of events as an example: forward causal reasoning asks, "what events are likely to happen next?," while backward causal reasoning asks, "what may cause this event?" in a scenario described by the context, i.e., querying the plausible previous or subsequent events, respectively. Besides, there are some problems that require considering another scenario in addition to the context, which can be seen as constrained causal reasoning. For example, TIMETRAVEL [7, 117] is a counterfactual story-rewriting dataset, where the original story is also given. See relevant datasets and benchmarks in Table 13.

How. This type of problem is mainly about "how to do it." It is more complex and also involves problem-solving and decision-making. See some examples in Table 14.

Others. Some datasets involve multiple types of reasoning. We list some typical datasets in Table 15.

Table 14. Datasets of "How" Commonsense Reasoning

Dataset	Size	Context Source	Option Source	Task	Remark
WikiHow Goal-Step [195]	1489k	WikiHow	automatically generated	multi-choice	goals, steps, and temporal ordering
PIQA [8]	21k	human-authored	human-authored	2-choice	physical causal reasoning

Table 15. Datasets and Benchmarks with Multiple Types of Commonsense Reasoning

	Size	Context Source	Question Source	Task	Remark
CSQA [153] **CoS-E**† [123]/**ECQA**† [1]	12k	–	semi-synthetic	multi-choice QA	ConceptNet concepts [146] explanation [1, 123], commonsense facts [1]
CSQA2 [155]	14k	–	human-authored	boolen QA	data construction via gamification
CosmosQA [62]	35k	blog [17]	human-authored	multi-choice QA	reading comprehension on blogs
Moral Stories [38]	12k	human-authored	–	classification/generation	situated reasoning with social norms

† indicates it annotates the rationale for the dataset.

Table 16. Complex Reasoning Datasets with the Realistic Data from Examinations or Tests, Where "RC" Denotes "Reading Comprehension"

Dataset	Size	Domain	Source	Task
AR-LSAT [202]	2k	law	law school admission test	multi-choice QA
HEAD-QA [164]	6.7k	healthcare	specialized healthcare examination	multi-choice QA
AI2-ARC [24]/**EntailmentBank**† [31]	7.7k	science	grade-school standardized test	multi-choice QA
ReClor [190]/**MetaLogic**† [64]	6k	generic	standardized graduate admission examination	RC + multi-choice QA
LogiQA [92]	8k	generic	national civil servant examination of China	RC + multi-choice QA
ConTRoL [90]	8k	generic	competitive selection and recruitment test	passage-level NLI

† indicates "it annotates reasoning paths for some examples in this dataset".

5.4.2 Reasoning. Since commonsense knowledge is essential to this topic, much research has focused on commonsense knowledge [42, 66, 80, 85, 91, 124, 140, 146]. As for the reasoning system, there are mainly two types of methods: graph based [40, 87, 183, 196] and vanilla PLMs [72, 80, 93, 118, 156, 189], where graph-based methods are designed to aggregate knowledge from commonsense knowledge bases, while vanilla PLMs are used as implicit knowledge bases themselves.

5.5 Complex Reasoning

There are some datasets collected from realistic examinations or tests, which may require domain-specific knowledge and multiple types of reasoning skills (Table 16).

To better diagnose the ability of LLMs, two few-shot prompting benchmarks called MMLU [51] and Big-Bench [147] are proposed, where tasks are much more challenging and even believed to be beyond the capabilities of current language models, in which some require performing reasoning. Among tasks in Big-Bench, [149] identified 23 challenging tasks, named as Big-Bench Hard, in which LLMs failed to surpass the average human rater, and many of which required performing multi-step reasoning. However, when equipped with CoT prompting, GPT3 can outperform human performance on a majority of these hard tasks.

5.6 Others

In addition to the above-mentioned datasets and benchmarks, there are also some other tasks requiring NLR scattered throughout the NLP domain, involving dialog [132], reading comprehension [89], and so on [49]. Note that reasoning is an important method to arrive at the required answers or solutions, which is of more frequent usage in complex problems. In other words, reasoning can occur in many other domains to solve challenging problems that require multiple knowledge to derive conclusions. While there might be more reasoning tasks or datasets, we list just some of them in Table 17.

Table 17. Some other NLP Benchmarks Requiring Natural Language Reasoning

Dataset	Size	Reasoning	Context Source	Task
ShARC [132]	32k	deductive	government document	conversation + boolean QA
ROPES [89]	14k	deductive	science textbook, Wikipedia	RC + extractive QA
ARC [49]	2k	abductive	news comment	2-choice

6 DISCUSSION

In this section, we propose some open questions, introduce some limitations, and suggest some future directions for reasoning. Among these, we also discuss the limitations of ChatGPT and GPT4.

6.1 Open Questions

We propose some open questions toward the reasoning capabilities of LLMs. There are many mysteries in their emergent reasoning capabilities.

- **Why is CoT prompting effective?** Why can LLMs just produce reasoning paths, which can even be wrong, before the final answer brings such significant improvement? And why is CoT prompting only effective for LLMs? What happens to LLMs when working with CoT, which fails to work well in medium-size PLMs?
- **Where do these reasoning capabilities of LLMs come from?** Why can LLMs emerge with reasoning capabilities just as the model size increases? Where does the magic "Let's think step by step" come from? How can they learn these capabilities? While the mechanism of another LLM's magic, in-context learning, has been studied [2, 30, 176], the mechanism of reasoning capabilities remain mysterious and under-explored [114].
- **Do even larger models reason better?** Can LLMs continue improving and learning more competitive reasoning capabilities from the general pretraining as the model size increases? How far could LLMs go just by increasing the model size? Do we still need to build more datasets and design algorithms specifically for reasoning?

6.2 Limitations

We introduce limitations of the current research and those intrinsic in PLMs.

First, there are gaps in defeasible reasoning and reasoning path evaluation:

- **Research gap on defeasible reasoning**. While defeasible reasoning is widely used in our daily life, this topic is still under-explored in NLP. [4] found that it is more challenging for ChatGPT to perform abductive reasoning and inductive reasoning than deduction, among which induction is the much more difficult one.
- **Lack of effective ways to evaluate reasoning paths**. It is still challenging to automatically evaluate generated reasoning paths without ground truth. Evaluating reasoning paths might become increasingly important to build explainable and reliable AI systems, especially when more people contact and use ChatGPT-like products nowadays.

Second, there are also limitations intrinsic to PLMs:

- **Soft deduction can produce invalid conclusions**. Transformers can only predict conclusions with probability, irrespective of whether the conclusion of deductive reasoning is necessarily true in nature, which might prevent it from precise reasoning. This characteristic can result in a sub-optimal solution to deductive problems (including arithmetic reasoning and symbolic reasoning). For example, while ChatGPT is impressive on reasoning tasks, it still fails to achieve perfect performance on the simplest one-step deductive inference task [4].

— **Biases on content**. PLMs make their prediction based on context. While LLMs have made huge progress in reasoning, [32] found that LLMs are biased by content like humans when performing deduction. For example, they perform worse in abstract or counterfactual situations than realistic ones. Such biases will hinder them from actual reasoning and lead to wrong answers, degrading downstream performance. More severely, it might cause harmful societal influences due to some social biases such as gender, which also exist in GPT4 [16].

6.3 Future

We suggest some potential research directions at both the holistic and technical levels in the future.

At the holistic level, first, reasoning should be generalized to more complex settings (longer steps and defeasible reasoning) and more diverse knowledge mediums (languages and modalities). Second, more attention should be placed on interpretability and faithfulness. We introduce these directions as the following:

— **Generalization to longer steps**. The multi-step performance degrades as PLMs encounter samples that require more reasoning steps than those in training data or few-shot exemplars. Although there is research on decoupled one-step inference, which can alleviate the challenge of the OOD problem, it still struggles with planning. How to better generalize to longer steps is an important problem for complex reasoning tasks, which are also challenging to ChatGPT [4].

— **More research on defeasible reasoning**. PLMs are currently the path with the most potential to defeasible reasoning due to advantages we have introduced in Section 3. According to philosophy, non-deductive reasoning is much more common than deductive reasoning in our daily lives and practical scenes. It is worth more effort to explore PLMs on defeasible reasoning since there is a lack of effective methods to deal with defeasible reasoning, while deductive reasoning can be solved by developed symbolic engines,[11] e.g., Prolog coding for first-order logic. Moreover, it might benefit scientific research a lot if AI can induce general rules from specific facts.

— **Reasoning over non-English languages**. In addition to reasoning over English statements, it is important to perform reasoning with other languages, which is much more challenging due to more severe data sparsity problems.

— **Reasoning with multi-modality**. Other types of modalities can also contribute to reasoning, such as tables [22] and images [35, 83, 148, 185]. GPT4 can process images, which might push forward visual reasoning.

— **Interpretability and faithful reasoning**. Transparent and reliable reasoning paths become increasingly important when it generalizes to longer steps and defeasible reasoning. First, when there are many steps, it takes more time and effort for people to check the quality of reasoning. Therefore, unfaithful reasoning might introduce difficulty in people's judgement and decision-making. Second, when it comes to defeasible reasoning, exposing interpretable reasoning paths is much more important and sometimes necessary for people to be convinced. In this case, different people with different background knowledge can derive different and even opposed conclusions, and thus it is crucial to illustrate the evidence collected to reason.

At the technical level, we suggest several directions to improve reasoning capabilities and performance of multi-step reasoning and defeasible reasoning as follows:

[11] Arithmetic reasoning and symbolic reasoning, which are popular recently, are also deductive and can be solved by calculator or code.

— **More prompts to elicit reasoning capabilities from LLMs**. Few-shot CoT and zero-shot CoT prompting are inspiring, and CoT annotations have been used to improve LLMs' reasoning capabilities [23]. It is both interesting and important to find whether there are other prompts that can activate LLMs to perform reasoning or are beneficial to improving reasoning capabilities, especially on complex reasoning.
— **Self-improvement of LLMs**. Data annotations for reasoning paths, especially for long-step and defeasible reasoning, are difficult to obtain. Interestingly, in our case studies, we found that ChatGPT can provide more comprehensive answers than the ground annotations in some existing datasets such as EntailmentBank [31] and WikiWhy [53]. Recent research has demonstrated that LLMs can learn from their self-generated reasoning paths to improve reasoning capabilities [61, 191], which provides the potential to alleviate the data challenge.
— **More exploration on backward reasoning**. Backward reasoning can benefit both medium-size PLMs and LLMs, while CoT prompting only benefits LLMs. Moreover, it is more efficient than forward reasoning with a smaller search space, which can bring more benefits as the depth of reasoning increases. To solve more complex reasoning problems, it is worth conducting more exploration in this direction.
— **More research on planning**. Planning is important to perform longer-step reasoning since the search space will become bigger as the depth increases.
— **Exploration on self-correction**. Since the conclusion of defeasible reasoning can be retracted by newly added evidence, it might be important for PLMs to update and self-correct their previously derived conclusions as the reasoning proceeds. Recent research suggests that LLMs lack the capability to recognize errors they have produced without external cues [60, 162]. However, these models can update and rectify their mistakes when given additional context or external feedback [47], which could be useful in defeasible reasoning scenarios.

REFERENCES

[1] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New dataset and models. 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'21), (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3050–3065. https://doi.org/10.18653/v1/2021.acl-long.238

[2] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? Investigations with linear models. *CoRR* abs/2211.15661 (2022). https://doi.org/10.48550/arXiv.2211.15661 arXiv:2211.15661

[3] Peter Adam Angeles. 1981. *Dictionary of Philosophy*. Barnes & Noble Books.

[4] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *CoRR* abs/2302.04023 (2023). https://doi.org/10.48550/arXiv.2302.04023 arXiv:2302.04023

[5] 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. In *The 2nd International Joint Conference on Learning and Reasoning and 16th International Workshop on Neural-Symbolic Learning and Reasoning (IJCLR-NeSy'22)*.

[6] Gregor Betz, Christian Voigt, and Kyle Richardson. 2021. Critical thinking for language models. In *IWCS*. Association for Computational Linguistics, 63–75.

[7] 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 *ICLR*. OpenReview.net.

[8] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *The 34th AAAI Conference on Artificial Intelligence (AAAI'20), The 32nd Innovative Applications of Artificial Intelligence Conference (IAAI'20), The 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'20)*. AAAI Press, 7432–7439. https://ojs.aaai.org/index.php/AAAI/article/view/6239

[9] Simon Blackburn. 2008. *The Oxford Dictionary of Philosophy*. Oxford University Press.
[10] Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In *EMNLP*. Association for Computational Linguistics, 1122–1136.
[11] Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, and Greg Durrett. 2021. Flexible generation of natural language deductions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6266–6278. https://doi.org/10.18653/v1/2021.emnlp-main.506
[12] 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'15)*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). Association for Computational Linguistics, 632–642. https://doi.org/10.18653/v1/d15-1075
[13] The Editors of Encyclopaedia Britannica. 2017. Inference. *Encyclopedia Britannica*. https://www.britannica.com/topic/inference-reason
[14] The Editors of Encyclopaedia Britannica. 2020. Reason. *Encyclopedia Britannica*. https://www.britannica.com/topic/reason
[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020 (NeurIPS'20)*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
[16] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR* abs/2303.12712 (2023). https://doi.org/10.48550/arXiv.2303.12712 arXiv:2303.12712
[17] Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 spinn3r dataset. In *3rd Annual Conference on Weblogs and Social Media (ICWSM'09)*.
[18] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018 (NeurIPS'18)*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 9560–9572. https://proceedings.neurips.cc/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html
[19] Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP'21) (Findings of ACL, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Vol. ACL/IJCNLP 2021)*, 3354–3361. https://doi.org/10.18653/v1/2021.findings-acl.297
[20] Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *NAACL-HLT*. Association for Computational Linguistics, 4026–4032.
[21] Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 8772–8779. https://doi.org/10.18653/v1/2020.acl-main.774
[22] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations (ICLR'20)*. OpenReview.net. https://openreview.net/forum?id=rkeJRhNYDH
[23] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR* abs/2210.11416 (2022). https://doi.org/10.48550/arXiv.2210.11416 arXiv:2210.11416
[24] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR* abs/1803.05457 (2018). arXiv:1803.05457 http://arxiv.org/abs/1803.05457

[25] Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *IJCAI*. ijcai.org, 3882–3890.
[26] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2475–2485. https://doi.org/10.18653/v1/d18-1269
[27] Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *CoRR* abs/2208.14271 (2022).
[28] Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *CoRR* abs/2205.09712 (2022).
[29] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00509ED1V01Y201305HLT023
[30] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. *CoRR* abs/2212.10559 (2022). https://doi.org/10.48550/arXiv.2212.10559 arXiv:2212.10559
[31] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *EMNLP*. Association for Computational Linguistics, 7358–7370.
[32] Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *CoRR* abs/2207.07051 (2022).
[33] Xiang Deng, Yu Su, Alyssa Lees, You Wu, Cong Yu, and Huan Sun. 2021. ReasonBERT: Pre-trained to reason with distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6112–6127. https://doi.org/10.18653/v1/2021.emnlp-main.494
[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19), Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423
[35] Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, and Zhifang Sui. 2022. Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 932–946. https://doi.org/10.18653/v1/2022.acl-long.66
[36] Li Du, Xiao Ding, Ting Liu, and Bing Qin. 2021. Learning event graph knowledge for abductive reasoning. In *ACL/IJCNLP*. Association for Computational Linguistics, 5181–5190.
[37] Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: A new dataset for exploring explainable causal reasoning. In *ACL*. Association for Computational Linguistics, 432–446.
[38] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *EMNLP*. Association for Computational Linguistics, 698–718.
[39] Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang. 2022. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *ACL*. Association for Computational Linguistics, 6896–6906.
[40] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *EMNLP*. Association for Computational Linguistics, 1295–1309.
[41] Maurice A. Finocchiaro. 1984. Informal logic and the theory of reasoning. *Informal Logic* 6, 2 (1984).
[42] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to reason about social and moral norms. In *EMNLP*. Association for Computational Linguistics, 653–670.
[43] Ahti-Veikko Pietarinen Francesco Bellucci. 2022. Peirce's logic. In *The Internet Encyclopedia of Philosophy, ISSN 2161-0002* (2022). https://iep.utm.edu/peir-log/
[44] Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *ACL*. Association for Computational Linguistics, 3108–3127.

[45] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics* 9 (2021), 346–361.
[46] Alvin I. Goldman. 1986. *Epistemology and Cognition*. Harvard University Press.
[47] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: Large language models can self-correct with tool-interactive critiquing. *CoRR* abs/2305.11738 (2023). https://doi.org/10.48550/ARXIV.2305.11738 arXiv:2305.11738
[48] Trudy Govier. 1989. Critical thinking as argument analysis. *Argumentation* 3, 2 (1989), 115–126.
[49] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *NAACL-HLT*. Association for Computational Linguistics, 1930–1940.
[50] 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 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 (2022).
[51] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations (ICLR'21)*. OpenReview.net. https://openreview.net/forum?id=d7KBjmI3GmQ
[52] Jaakko J. Hintikka. 2022. Logic. In *Encyclopedia Britannica*. https://www.britannica.com/topic/logic
[53] Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2022. WikiWhy: Answering and explaining cause-and-effect questions. *CoRR* abs/2210.12152 (2022). https://doi.org/10.48550/arXiv.2210.12152 arXiv:2210.12152
[54] Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *CoRR* abs/2212.10071 (2022). https://doi.org/10.48550/arXiv.2212.10071 arXiv:2212.10071
[55] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *COLING*. International Committee on Computational Linguistics, 6609–6625.
[56] 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'22)*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 1887–1905. https://doi.org/10.18653/v1/2022.findings-naacl.145
[57] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 9106–9118. https://doi.org/10.18653/v1/2020.emnlp-main.732
[58] Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S. Moss. 2020. OCNLI: Original Chinese natural language inference. In *Findings of the Association for Computational Linguistics (EMNLP'20) (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 3512–3526. https://doi.org/10.18653/V1/2020.FINDINGS-EMNLP.314
[59] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *CoRR* abs/2212.10403 (2022). https://doi.org/10.48550/arXiv.2212.10403 arXiv:2212.10403
[60] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *CoRR* abs/2310.01798 (2023). https://doi.org/10.48550/ARXIV.2310.01798 arXiv:2310.01798
[61] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *CoRR* abs/2210.11610 (2022). https://doi.org/10.48550/arXiv.2210.11610 arXiv:2210.11610
[62] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *EMNLP/IJCNLP*. Association for Computational Linguistics, 2391–2401.
[63] Yongjie Huang and Meng Yang. 2021. Breadth first reasoning graph for multi-hop question answering. In *NAACL-HLT*. Association for Computational Linguistics, 5810–5821.
[64] Yinya Huang, Hongming Zhang, Ruixin Hong, Xiaodan Liang, Changshui Zhang, and Dong Yu. 2022. MetaLogic: Logical reasoning explanations with fine-grained structure. *CoRR* abs/2210.12487 (2022).
[65] Patrick J. Hurley. 2014. *A Concise Introduction to Logic*. Cengage Learning.

[66] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-) Atomic 2020: On symbolic and neural commonsense knowledge graphs. In *35th AAAI Conference on Artificial Intelligence (AAAI'21), 33rd Conference on Innovative Applications of Artificial Intelligence (IAAI'21), The 11th Symposium on Educational Advances in Artificial Intelligence (EAAI'21)*. AAAI Press, 6384–6392. https://ojs.aaai.org/index.php/AAAI/article/view/16792

[67] Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *ACL*. Association for Computational Linguistics, 6740–6750.

[68] Naoya Inoue, Harsh Trivedi, Steven Sinha, Niranjan Balasubramanian, and Kentaro Inui. 2021. Summarize-then-answer: Generating concise explanations for multi-hop reading comprehension. In *EMNLP*. Association for Computational Linguistics, 6064–6080.

[69] Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *EMNLP*. Association for Computational Linguistics, 137–150.

[70] Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *ACL*. Association for Computational Linguistics, 2726–2736.

[71] Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. MERIt: Meta-path guided contrastive learning for logical reasoning. In *Findings of the Association for Computational Linguistics (ACL'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3496–3509. https://doi.org/10.18653/v1/2022.findings-acl.276

[72] Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *CoRR* abs/2205.11822 (2022). https://doi.org/10.48550/arXiv.2205.11822 arXiv:2205.11822

[73] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.

[74] Seyed Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2022. LAMBADA: Backward chaining for automated reasoning in natural language. *CoRR* abs/2212.13894 (2022). https://doi.org/10.48550/arXiv.2212.13894 arXiv:2212.13894

[75] Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabadi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A suite of language understanding challenges for persian. *Trans. Assoc. Comput. Linguistics* 9 (2021), 1147–1162. https://doi.org/10.1162/TACL_A_00419

[76] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *AAAI*. AAAI Press, 8082–8090.

[77] Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 1264–1279. https://doi.org/10.18653/v1/2021.naacl-main.99

[78] Tushar Khot, Kyle Richardson, Daniel Khashabi, and Ashish Sabharwal. 2022. Hey AI, can you solve complex tasks by talking to agents? In *Findings of the Association for Computational Linguistics (ACL'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 1808–1823. https://doi.org/10.18653/V1/2022.FINDINGS-ACL.142

[79] Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTaiL: A textual entailment dataset from science question answering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18), the 30th innovative Applications of Artificial Intelligence (IAAI'18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'18)*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 5189–5197. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17368

[80] Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *ACL*. Association for Computational Linguistics, 4831–4836.

[81] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *CoRR* abs/2205.11916 (2022).

[82] Yash Kumar Lal, Nathanael Chambers, Raymond J. Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP'21) (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 596–610. https://doi.org/10.18653/v1/2021.findings-acl.53

[83] Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. DVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue. 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'21) (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5651–5665. https://doi.org/10.18653/v1/2021.acl-long.439

[84] Kyungjae Lee, Seung-won Hwang, Sang-eun Han, and Dohyeon Lee. 2021. Robustifying multi-hop QA through pseudo-evidentiality training. In *ACL/IJCNLP*. Association for Computational Linguistics, 6110–6119.

[85] Douglas B. Lenat. 1995. CYC: A large-scale investment in knowledge infrastructure. *Commun. ACM* 38, 11 (1995), 32–38. https://doi.org/10.1145/219717.219745

[86] Zhengzhong Liang, Steven Bethard, and Mihai Surdeanu. 2021. Explainable multi-hop verbal reasoning through internal monologue. In *NAACL-HLT*. Association for Computational Linguistics, 1225–1250.

[87] Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William W. Cohen. 2021. Differentiable open-ended commonsense reasoning. In *NAACL-HLT*. Association for Computational Linguistics, 4611–4625.

[88] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics (EMNLP'20) (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 1823–1840. https://doi.org/10.18653/v1/2020.findings-emnlp.165

[89] Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering (MRQA@EMNLP'19)*, Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen (Eds.). Association for Computational Linguistics, 58–62. https://doi.org/10.18653/v1/D19-5808

[90] Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context—Investigating contextual reasoning over long texts. In *35th AAAI Conference on Artificial Intelligence (AAAI'21), 33rd Conference on Innovative Applications of Artificial Intelligence (IAAI'21), The 11th Symposium on Educational Advances in Artificial Intelligence (EAAI'21)*. AAAI Press, 13388–13396. https://ojs.aaai.org/index.php/AAAI/article/view/17580

[91] Hugo Liu and Push Singh. 2004. ConceptNet–a practical commonsense reasoning tool-kit. *BT Technol. J.* 22, 4 (2004), 211–226.

[92] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI'20)*, Christian Bessiere (Ed.). ijcai.org, 3622–3628. https://doi.org/10.24963/ijcai.2020/501

[93] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *ACL*. Association for Computational Linguistics, 3154–3169.

[94] John Locke. 1847. *An Essay Concerning Human Understanding*. Kay & Troutman.

[95] Man Luo, Shrinidhi Kumbhar, Ming Shen, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2023. Towards LogiGLUE: A brief survey and A benchmark for analyzing logical reasoning capabilities of language models. *CoRR* abs/2310.00836 (2023). https://doi.org/10.48550/ARXIV.2310.00836 arXiv:2310.00836

[96] Aman Madaan, Dheeraj Rajagopal, Niket Tandon, Yiming Yang, and Eduard H. Hovy. 2021. Could you give me a hint? Generating inference graphs for defeasible reasoning. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP'21) (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5138–5147. https://doi.org/10.18653/v1/2021.findings-acl.456

[97] Aman Madaan, Niket Tandon, Dheeraj Rajagopal, Peter Clark, Yiming Yang, and Eduard H. Hovy. 2021. Think about it! Improving defeasible reasoning by first modeling the question scenario. In *EMNLP*. Association for Computational Linguistics, 6291–6310.

[98] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *CoRR* abs/2212.08410 (2022). https://doi.org/10.48550/arXiv.2212.08410 arXiv:2212.08410

[99] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL'19), Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 3428–3448. https://doi.org/10.18653/v1/p19-1334

[100] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2381–2391. https://doi.org/10.18653/v1/d18-1260

[101] Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *ACL*. Association for Computational Linguistics, 4249–4257.

[102] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *ACL*. Association for Computational Linguistics, 6097–6109.

[103] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'16)*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, 839–849. https://doi.org/10.18653/v1/n16-1098

[104] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL'20)*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 4885–4901. https://doi.org/10.18653/V1/2020.ACL-MAIN.441

[105] Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A dataset for commonsense reasoning over entity knowledge. In *NeurIPS Datasets and Benchmarks*.

[106] Santiago Ontañón, Joshua Ainslie, Vaclav Cvicek, and Zachary Fisher. 2022. LogicInference: A new dataset for teaching logical inference to seq2seq models. *CoRR* abs/2203.15099 (2022). https://doi.org/10.48550/arXiv.2203.15099 arXiv:2203.15099

[107] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Unsupervised multi-hop question answering by question generation. In *NAACL-HLT*. Association for Computational Linguistics, 5866–5880.

[108] Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Is a question decomposition unit all we need? *CoRR* abs/2205.12538 (2022). https://doi.org/10.48550/arXiv.2205.12538 arXiv:2205.12538

[109] Charles Sanders Peirce. 1992. *Reasoning and the Logic of Things: The Cambridge Conferences Lectures of 1898*. Harvard University Press.

[110] Ethan Perez, Patrick S. H. Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8864–8880. https://doi.org/10.18653/v1/2020.emnlp-main.713

[111] Xinyu Pi, Qian Liu, Bei Chen, Morteza Ziyadi, Zeqi Lin, Yan Gao, Qiang Fu, Jian-Guang Lou, and Weizhu Chen. 2022. Reasoning like program executors. *CoRR* abs/2201.11473 (2022). arXiv:2201.11473 https://arxiv.org/abs/2201.11473

[112] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (SEM@NAACL-HLT'18)*, Malvina Nissim, Jonathan Berant, and Alessandro Lenci (Eds.). Association for Computational Linguistics, 180–191. https://doi.org/10.18653/v1/s18-2023

[113] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *CoRR* abs/2210.03350 (2022).

[114] Ben Prystawski and Noah D. Goodman. 2023. Why think step-by-step? Reasoning emerges from the locality of experience. *CoRR* abs/2304.03843 (2023). https://doi.org/10.48550/arXiv.2304.03843 arXiv:2304.03843

[115] Peng Qi, Haejun Lee, Tg Sido, and Christopher D. Manning. 2021. Answering open-domain questions of varying reasoning steps from text. In *EMNLP*. Association for Computational Linguistics, 3599–3614.

[116] 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 (2022). https://doi.org/10.48550/arXiv.2212.09597 arXiv:2212.09597

[117] Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *EMNLP/IJCNLP*. Association for Computational Linguistics, 5042–5052.

[118] 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 *EMNLP*. Association for Computational Linguistics, 794–805.

[119] Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *ACL*. Association for Computational Linguistics, 6140–6150.

[120] Hanhao Qu, Yu Cao, Jun Gao, Liang Ding, and Ruifeng Xu. 2022. Interpretable proof generation via iterative backward reasoning. In *NAACL-HLT*. Association for Computational Linguistics, 2968–2981.

[121] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. OpenAI.

[122] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. http://jmlr.org/papers/v21/20-074.html

[123] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! Leveraging language models for commonsense reasoning. In *ACL*. Association for Computational Linguistics, 4932–4942.

[124] Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18), Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 463–473. https://doi.org/10.18653/v1/P18-1043

[125] Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22)*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 8729–8755. https://aclanthology.org/2022.emnlp-main.598

[126] 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'22)*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 465–475. https://doi.org/10.18653/v1/2022.findings-naacl.35

[127] Kyle Richardson and Ashish Sabharwal. 2022. Pushing the limits of rule reasoning in transformers through natural language satisfiability. In *36th AAAI Conference on Artificial Intelligence (AAAI'22), 34th Conference on Innovative Applications of Artificial Intelligence (IAAI'22), The 12th Symposium on Educational Advances in Artificial Intelligence (EAAI'22*. AAAI Press, 11209–11219. https://doi.org/10.1609/AAAI.V36I10.21371

[128] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06*. AAAI. http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2418

[129] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics (EMNLP'20) (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4661–4675. https://doi.org/10.18653/v1/2020.findings-emnlp.418

[130] Dagobert D. Runes. 2001. *The Dictionary of Philosophy*. Citadel Press.

[131] Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 7399–7409. https://doi.org/10.18653/v1/2022.acl-long.511

[132] Marzieh Saeidi, Max Bartolo, Patrick S. H. Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 2087–2097. https://doi.org/10.18653/v1/d18-1233

[133] Swarnadeep Saha, Sayan Ghosh, Shashank Srivastava, and Mohit Bansal. 2020. PRover: Proof generation for interpretable reasoning over rules. In *EMNLP*. Association for Computational Linguistics, 122–136.

[134] Swarnadeep Saha, Yixin Nie, and Mohit Bansal. 2020. ConjNLI: Natural language inference over conjunctive sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 8240–8252. https://doi.org/10.18653/v1/2020.emnlp-main.661

[135] Swarnadeep Saha, Prateek Yadav, and Mohit Bansal. 2021. multiPRover: Generating multiple proofs for improved interpretability in rule reasoning. In *NAACL-HLT*. Association for Computational Linguistics, 3662–3677.

[136] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An adversarial winograd schema challenge at scale. In *The 34th AAAI Conference on Artificial Intelligence (AAAI'20), The 32nd Innovative Applications of Artificial Intelligence Conference (IAAI'20), The 10th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI'20)*. AAAI Press, 8732–8740. https://ojs.aaai.org/index.php/AAAI/article/view/6399

[137] Soumya Sanyal, Zeyi Liao, and Xiang Ren. 2022. RobustLR: Evaluating robustness to logical perturbation in deductive reasoning. *CoRR* abs/2205.12598 (2022). https://doi.org/10.48550/arXiv.2205.12598 arXiv:2205.12598

[138] Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. FaiRR: Faithful and robust deductive reasoning over natural language. In *ACL*. Association for Computational Linguistics, 1075–1093.

[139] Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu, and Xiang Ren. 2022. APOLLO: A simple approach for adaptive pretraining of language models for logical reasoning. *CoRR* abs/2212.09282 (2022). https://doi.org/10.48550/arXiv.2212.09282 arXiv:2212.09282

[140] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for If-Then reasoning. In *AAAI*. AAAI Press, 3027–3035.

[141] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*. Association for Computational Linguistics, 5477–5490.

[142] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *EMNLP/IJCNLP*. Association for Computational Linguistics, 4462–4472.

[143] Abulhair Saparov and He He. 2022. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *CoRR* abs/2210.01240 (2022).

[144] Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling multi-step reasoning capabilities of large language models into smaller models via semantic decompositions. *CoRR* abs/2212.00193 (2022). https://doi.org/10.48550/arXiv.2212.00193 arXiv:2212.00193

[145] Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *EMNLP/IJCNLP*. Association for Computational Linguistics, 4505–4514.

[146] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, Satinder Singh and Shaul Markovitch (Eds.). AAAI Press, 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[147] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR* abs/2206.04615 (2022).

[148] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL'19), Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 6418–6428. https://doi.org/10.18653/v1/p19-1644

[149] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-bench tasks and whether chain-of-thought can solve them. *CoRR* abs/2210.09261 (2022).

[150] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *ACL/IJCNLP (Findings) (Findings of ACL, Vol. ACL/IJCNLP 2021)*. Association for Computational Linguistics, 3621–3634.

[151] Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. *CoRR* abs/2210.12217 (2022). https://doi.org/10.48550/arXiv.2210.12217 arXiv:2210.12217

[152] Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18), Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 641–651. https://doi.org/10.18653/V1/N18-1059

[153] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT'19), Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4149–4158. https://doi.org/10.18653/v1/n19-1421

[154] 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. In *NeurIPS*.

[155] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks'21)*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/3ef815416f775098fe977004015c6193-Abstract-round1.html

[156] Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *ACL*. Association for Computational Linguistics, 3878–3887.

[157] Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 57–66. https://doi.org/10.18653/v1/d18-1006

[158] Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "What if..." reasoning over procedural text. In *EMNLP/IJCNLP*. Association for Computational Linguistics, 6075–6084.

[159] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2020. Is Multihop QA in DiRe condition? Measuring and reducing disconnected reasoning. In *EMNLP*. Association for Computational Linguistics, 8846–8863.

[160] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Trans. Assoc. Comput. Linguistics* 10 (2022), 539–554. https://doi.org/10.1162/tacl_a_00475

[161] Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/786.html

[162] Gladys Tyen, Hassan Mansoor, Peter Chen, Tony Mak, and Victor Carbune. 2023. LLMs cannot find reasoning errors, but can correct them! *CoRR* abs/2311.08516 (2023). https://doi.org/10.48550/ARXIV.2311.08516 arXiv:2311.08516

[163] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[164] David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *ACL*. Association for Computational Linguistics, 960–966.

[165] Douglas N. Walton. 1990. What is reasoning? What is an argument? *J. Philos.* 87, 8 (1990), 399–419.

[166] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS'19)*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 3261–3275. https://proceedings.neurips.cc/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html

[167] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations (ICLR'19)*. OpenReview.net. https://openreview.net/forum?id=rJ4km2R5t7

[168] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *CoRR* abs/2203.11171 (2022).

[169] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *CoRR* abs/2206.07682 (2022). https://doi.org/10.48550/arXiv.2206.07682 arXiv:2206.07682

[170] 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 (2022).

[171] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguistics* 6 (2018), 287–302.

[172] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *ICLR (Poster)*.

[173] Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'18), Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 1112–1122. https://doi.org/10.18653/v1/n18-1101

[174] Tomer Wolfson, Mor Geva, Ankit Gupta, Yoav Goldberg, Matt Gardner, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Trans. Assoc. Comput. Linguistics* 8 (2020), 183–198. https://doi.org/10.1162/TACL_A_00309

[175] Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 2660–2676. https://doi.org/10.18653/v1/2022.acl-long.190

[176] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *The 10th International Conference on Learning Representations (ICLR'22)*. OpenReview.net. https://openreview.net/forum?id=RdJVFCHjUMI

[177] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING'20)*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 4762–4772. https://doi.org/10.18653/V1/2020.COLING-MAIN.419

[178] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP@ACL'19)*, Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes (Eds.). Association for Computational Linguistics, 31–40. https://doi.org/10.18653/v1/W19-4804

[179] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the 8th Joint Conference on Lexical and Computational Semantics (SEM@NAACL-HLT'19)*, Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria (Eds.). Association for Computational Linguistics, 250–255. https://doi.org/10.18653/v1/s19-1027

[180] Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. *CoRR* abs/2205.12443 (2022). https://doi.org/10.48550/arXiv.2205.12443 arXiv:2205.12443

[181] Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *CoRR* abs/2212.10923 (2022). https://doi.org/10.48550/arXiv.2212.10923 arXiv:2212.10923

[182] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*. Association for Computational Linguistics, 2369–2380.

[183] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *NAACL-HLT*. Association for Computational Linguistics, 535–546.

[184] Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *CoRR* abs/2211.13892 (2022). https://doi.org/10.48550/arXiv.2211.13892 arXiv:2211.13892

[185] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 2115–2129. https://doi.org/10.18653/v1/2021.emnlp-main.162

[186] Wenpeng Yin, Dragomir R. Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP'21) (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 4913–4922. https://doi.org/10.18653/v1/2021.findings-acl.435

[187] 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'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 218–227. https://doi.org/10.18653/v1/2022.findings-acl.19

[188] Jianxing Yu, Wei Liu, Shuang Qiu, Qinliang Su, Kai Wang, Xiaojun Quan, and Jian Yin. 2020. Low-resource generation of multi-hop reasoning questions. In *ACL*. Association for Computational Linguistics, 6729–6739.

[189] Ping Yu, Tianlu Wang, Olga Golovneva, Badr AlKhamissy, Gargi Ghosh, Mona T. Diab, and Asli Celikyilmaz. 2022. ALERT: Adapting language models to reasoning tasks. *CoRR* abs/2212.08286 (2022). https://doi.org/10.48550/arXiv.2212.08286 arXiv:2212.08286

[190] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *ICLR*. OpenReview.net.

[191] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. *CoRR* abs/2203.14465 (2022).

[192] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, 93–104. https://doi.org/10.18653/v1/d18-1009

[193] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence?. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL'19), Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 4791–4800. https://doi.org/10.18653/v1/p19-1472

[194] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. *CoRR* abs/2205.11502 (2022).

[195] Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *EMNLP*. Association for Computational Linguistics, 4630–4639.

[196] Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. GreaseLM: Graph REASoning enhanced language models. In *ICLR*. OpenReview.net.

[197] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *CoRR* abs/2210.03493 (2022).

[198] Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Transformer-XH: Multi-evidence reasoning with eXtra hop attention. In *ICLR*. OpenReview.net.

[199] Chen Zheng and Parisa Kordjamshidi. 2020. SRLGRN: Semantic role labeling graph reasoning network. In *EMNLP*. Association for Computational Linguistics, 8881–8891.

[200] Victor Zhong and Luke Zettlemoyer. 2019. E3: Entailment-driven extracting and editing for conversational machine reading. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL'19), Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2310–2320. https://doi.org/10.18653/v1/p19-1223

[201] Wanjun Zhong, Tingting Ma, Jiahai Wang, Jian Yin, Tiejun Zhao, Chin-Yew Lin, and Nan Duan. 2022. Disentangling reasoning capabilities from language models with compositional reasoning transformers. *CoRR* abs/2210.11265 (2022).

[202] Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Yining Chen, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2022. Analytical reasoning of text. In *Findings of the Association for Computational Linguistics (NAACL'22)*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 2306–2319. https://doi.org/10.18653/v1/2022.findings-naacl.177

[203] Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. Learning to decompose: Hypothetical question decomposition based on comparable texts. *CoRR* abs/2210.16865 (2022). https://doi.org/10.48550/arXiv.2210.16865 arXiv:2210.16865

[204] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *CoRR* abs/2205.10625 (2022).

[205] Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, and Xiang Ren. 2021. RICA: Evaluating robust inference capabilities based on commonsense axioms. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 7560–7579. https://doi.org/10.18653/v1/2021.emnlp-main.598

Received 6 May 2023; revised 9 March 2024; accepted 26 April 2024