Evaluating Generalization Capability of Language Models across Abductive, Deductive and Inductive Logical Reasoning

Yu Sheng[1,2], **Wanting Wen**[1], **Linjing Li**[1,2,3], **Daniel Zeng**[1,2]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Wenge Technology Co., Ltd, Beijing, China
`{shengyu2021, wanting.wen, linjing.li, dajun.zeng}@ia.ac.cn`

Abstract

Transformer-based language models (LMs) have demonstrated remarkable performance on many natural language tasks, yet to what extent LMs possess the capability of generalizing to unseen logical rules remains not explored sufficiently. In classical logic category, abductive, deductive and inductive (ADI) reasoning are defined as the fundamental reasoning types, sharing the identical reasoning primitives and properties, and some research have proposed that there exists mutual generalization across them. However, in the field of natural language processing, previous research generally study LMs' ADI reasoning capabilities separately, overlooking the generalization across them. To bridge this gap, we propose UniADILR, a novel logical reasoning dataset crafted for assessing the generalization capabilities of LMs across different logical rules. Based on UniADILR, we conduct extensive investigations from various perspectives of LMs' performance on ADI reasoning. The experimental results reveal the weakness of current LMs in terms of extrapolating to unseen rules and inspire a new insight for future research in logical reasoning.[1]

1 Introduction

Logical Reasoning (LR) refers to the cognitive process of applying deterministic logical rules and known facts to derive valid conclusions, can be observed in diverse domains, such as scientific inquiry, decision-making, and everyday discourse (Nunes, 2012). Recent research in language models (LMs) have achieved significant progress in many natural language (NL) tasks, such as text summarization (Jin et al., 2024), question answering (Louis et al., 2024) and commonsense reasoning (Krause and Stolzenburg, 2023). However, whether LMs are proficient in grasping the underlying logical rules implied in complicated contexts have not been explored adequately.

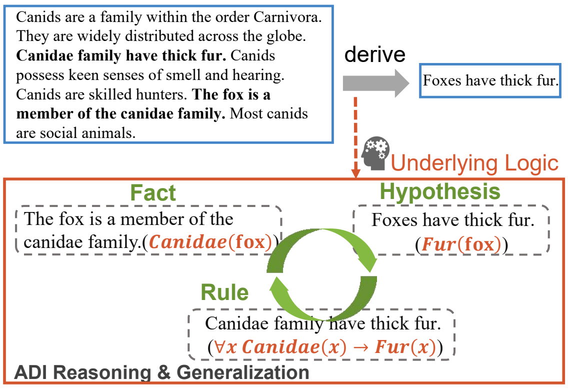

Canids are a family within the order Carnivora. They are widely distributed across the globe. **Canidae family have thick fur.** Canids possess keen senses of smell and hearing. Canids are skilled hunters. **The fox is a member of the canidae family.** Most canids are social animals.
derive
Foxes have thick fur.
Underlying Logic
Fact
The fox is a member of the canidae family.(Canidae(fox))
Hypothesis
Foxes have thick fur. (Fur(fox))
Rule
Canidae family have thick fur. (∀x Canidae(x) → Fur(x))
ADI Reasoning & Generalization

Figure 1: ADI reasoning within specific contexts. The sharing primitives motivates studying generalization across them.

In the field of classic logic, LR is categorized into three primary types: abductive reasoning, deductive reasoning, and inductive reasoning (ADI) (Peirce, 1974). This categorization has long been widely recognized, providing a foundation for investigating different modes of logical reasoning. Many studies on LMs have sought to evaluate their proficiency in learning these reasoning rules (Young et al., 2022; Saparov and He, 2023; Saparov et al., 2023). However, there remains some limitations of existing benchmarks. For real-world scenarios requiring LR, reasoners generally discern underlying logic from complex contexts and select sound facts as premises for reasoning. Yet, various previous benchmarks are conditioned by providing all logical premises explicitly (Clark et al., 2020; Bostrom et al., 2021), where LMs are prone to draw conclusion through shortcuts like directly splicing premises, potentially resulting in an over-estimated of the reasoning capability.

Moreover, as shown in Figure 1, we can observe that ADI share fundamental reasoning primitives (*Rule*, *Fact*, *Hypothesis*) from the point of a uni-

---

[1] The codes and datasets are available at `https://github.com/YuSheng-00/UniADILR`

Proceedings of the 31st International Conference on Computational Linguistics, pages 4945–4957
January 19–24, 2025. ©2025 Association for Computational Linguistics

fied framework grounded in classical logic. The primitives constitute various logical rules via different direction of inference (Pauwels et al., 2012). The distinctive characteristic motivates a potential for investigating the generalization across them, which has been discussed in the field of logic (EGOIRE and Lakhdar, 1996; Rivera and Becker, 2007; Grégoire, 2001) but overlooked in LMs' reasoning. Specifically, some valuable research questions, such as *"For supervised learning, how performance varying when the distribution of logical rule in test datasets shifts from that in the training phase"* and *"How logical rules presented in demonstrations impacting the large LMs' few-shot performance"* remains unexplored. In prior datasets, logical rules either appear in isolation or are coupled together in a way that makes them inseparable, hindering the comprehensive evaluation on comparison of different reasoning capabilities and generalization.

To tackle these issues, we propose a novel dataset serving as a unified benchmark to evaluate ADI logical reasoning (**UniADILR**) capabilities of LMs. We specify multiple reasoning primitive rules and then expand them to obtain examples in NL to imply different logical types. To alleviate the influence caused by various shortcuts, we define a unified domain involving fictional concepts and predicates to construct reasoning examples. The reasoning rule appearing in each example is restrained strictly to avoid logical confusion. UniADILR requires the models to select premises which derive the given hypothesis within a complicated context including multiple distracting facts, reflecting the capability of LMs to identify underlying logic and perform reasoning.

To furnish a comprehensive understanding of LMs' proficiency on ADI reasoning and generalization across them, we conduct exhaustive experiments from a wide range of perspectives. (1) Different learning paradigm: We compare two main-stream learning paradigms, in-context learning (ICL) and supervised fine-tuning (SFT), evaluating performance of various LMs with different size when the logical type in testing datasets outside the distribution of (OOD) the learning stage. Furthermore, we compare the respective strengths and weakness of the two methods. (2) Number of logical rules: We change the number of logical rules appearing in the learning stage and analyse the test accuracy achieved by LMs. (3) OOD Vocabulary: To explore whether the LMs take hold of the rationale of the reasoning rules or merely imitate the recurring pattern in the training data, we manually annotate 1500 additional examples containing more abundant vocabularies and context closer to real-world for testing only. The results reveal a significant decline in the performance of LMs when the complexity of the vocab lifts in the testing sets even if the logical rules remains identical. (4) In addition, we conduct an in-depth analysis of the reasoning bias and performance gap in terms of the different logical types. Overall, the investigations highlight the challenge faced by LMs in emulating human-like generalization between similar inference rules, inspiring future work to tackle this issue. Our contributions are outlined as follows:

- A natural language logical reasoning dataset encompassing abduction, deduction and induction within the unified framework and domain.
- To the best of our knowledge, it is the first research to propose the generalization across the fundamental ADI reasoning types and assess them on a unified benchmark.
- We conduct comprehensive experiments to evaluate various LMs from a wide range of axes and reveal their capabilities of generalization across ADI reasoning rules.

2 Preliminaries

2.1 Definition of ADI Reasoning

In this paper, we define a ternary reasoning framework grounded in first-order logic (FOL), designed to operate ADI reasoning rules based on the Rule-Fact-Hypothesis primitives. Variables are represented by a lowercase letter x (To simplify, our examples generally involve unary relation). And we denote sets of defined predicates using uppercase letters $\mathbb{P} = \{F, G, H, \dots\}$ and constants using lowercase letters $\mathbb{K} = \{a, b, c, \dots\}$. A basic rule takes the form of "$\forall \mathrm{x}\ F(\mathrm{x}) \rightarrow G(\mathrm{x})$" which means *if something satisfies* F*, then it can be inferred that it satisfy* G. Facts are known statements, and hypotheses are conjectures to be proven, both of which arise from instantiating variables within predicates. Based on this structure, we define ADI reasoning as follows:

Deductive Reasoning is a logical process that substantiates specific hypothesis from general rules and known facts:

$$\forall \mathrm{x}\ F(\mathrm{x}) \rightarrow G(\mathrm{x}) \wedge F(a) \implies G(a). \tag{1}$$

Abductive Reasoning is the process of making plausible explanation or hypothesis to account for observed facts:

$$\forall \mathrm{x}\ F(\mathrm{x}) \rightarrow G(\mathrm{x}) \wedge sup.\ G(a) \implies F(a), \tag{2}$$

where $F(a)$ is considered as the cause of $G(a)$, and $sup.$ means supposing $G(a)$ is observed.

Inductive Reasoning is the process of drawing general rule or pattern based on specific observations or known facts:

$$\exists a\ (f(a) \wedge g(a)) \implies \forall \mathrm{x}\ F(\mathrm{x}) \rightarrow G(\mathrm{x}). \tag{3}$$

Notably, deductive reasoning aims to provide logically valid results but inductive reasoning and abductive reasoning does not guarantee absolute certainty, but providing probable conclusions based on available evidence.

2.2 ADI Generalization

It can be intuitively observed that under the unified reasoning framework, ADI reasoning share the fundamental primitives and characteristics, with the only difference being the direction of inference. The conversion across ADI logical type can be represented by a triangle with the vertices labeled as rule, hypothesis and fact, which has been proposed by Pierce as a logical cycle (Pauwels et al., 2012). It is natural to discuss whether, after being taught merely one or two rules, it is possible to automatically generalize to the others unseen. Several studies in the field of philosophy and logic have investigated this property. For instance, Grégoire (2001) argue that certain characteristics between inductive and abductive reasoning allow them to be mutually convertible and EGOIRE and Lakhdar (1996) observed that, in some cases, a deductive result could also be correctly interpreted by induction. However, in the field of NL reasoning, whether LMs can automatically generalize to other inference rules has not yet been researched. Therefore, we propose ADI generalization and evaluate this capability of LMs by creating a novel dataset using the defined logical framework.

3 UniADILR Dataset and Evaluation

We construct a novel dataset, UniADILR, to evaluate the performance of LMs on ADI reasoning and investigate the generalization between them. Using the framework defined in Section 2.1, we derive formal rules for each reasoning type and programmatically extend the underlying logic into natural language examples. Additionally, we manually annotate extra examples that are closer to real-world language distribution, reserved specifically for testing. Subsequently we conduct systematic evaluation of LMs' logical reasoning capabilities from multiple dimensions.

3.1 Task Definition

UniADILR defines the *proof generation* task to measure the reasoning capabilities of diverse LMs. As shown below, the model is offered a statement $\mathcal{S}$ and a NL context $\mathcal{C} = \{f_s, f_d\}$ consisting of correct logical premises f_s and distracting sentences f_d. UniADILR requests the reasoning model to trace logical clues f_s from $\mathcal{C}$ to support the statement $\mathcal{S}$ via performing ADI reasoning. The proof p is organized in the format of "`sent1 & sent2 & ... & sentk` $\rightarrow \mathcal{S}$".

An Example in UniADILR

▶ **Context** $\mathcal{C}$:
sent1: Bright vumpuses are zumpuses. sent2: Sam is bright. sent3: Sam is a sterpus. $\cdots$ sent6: Sam is a vumpus. $\cdots$ sentk: Every kind vumpus is a brimpus.
▶ **Statement** $\mathcal{S}$: Sam is a bright vumpus.
▶ **Proof** p: sent2 & sent6 $\rightarrow$ Sam is a bright vumpus.
▶ **Reasoning Type**: Deduction

3.2 Synthetic Data based on FOL Formulas

Formula Extension. We create FOL formulas implying logical rules according to the defined framework. We define a *domain* consisting of predicates $\mathbb{P}$ and constants $\mathbb{K}$, which is uniformly organized based on virtualized concepts (e.g. "wumpus" instead of "cat") to prevent the evaluating bias caused by shortcuts of commonsense knowledge. We apply $\mathbb{P}$ and $\mathbb{K}$ to instance each logical formula to derive specific examples implying underlying logic (e.g. $\forall \mathrm{x}\ Canidae(\mathrm{x}) \rightarrow Fur(\mathrm{x}) \wedge Canidae(wumpus) \implies Fur(wumpus)$). Afterward, the instantiated logical expressions are programmatically expanded into NL sentences (e.g. "*Caidae family have thick fur; Wumpus is a member of the canidae family; Wumpus have thick fur.*").

Adding Distractors. Distracting contexts are developed by manipulating the predicates and constants in the logical expressions. For example, for the fact $Canidae(wumpus)$, we transform the predicate to obtain $Bright(wumpus)$ or the constant to obtain $Canidae(rifpist)$. Generally, we do not modify all predicates and constants in a logical statement simultaneously, ensuring a certain degree of the similarity with the original statement. This process generates misleading facts for LMs, preventing the model from identifying reasoning premises directly through heuristic shortcuts, such as word overlap. For each example in UniADILR, we create 10–20 distracting sentences.

3.3 Human Annotation

Furthermore, to explore whether the LMs grasp the rationale of logical rules or just imitate the recurring pattern in the training datasets, we expand UniADILR to include an additional manual subset containing 1500 examples with more complex and closer to real-world contexts. These examples are reserved exclusively for test. We use an iterative, human-in-the-loop process to annotate effectively. Initially, human experts manually annotate some examples demonstrating the corresponding reasoning rule as D_{ex}. Then, we instruct a GPT-3.5 (OpenAI, 2023a) to generate more examples involving the consistent logical rule within the prompt, which incorporates a brief task description and five data points randomly selected from D_{ex}. Following this, human experts meticulously review and correct the model's output to ensure the accurate logical relationship and coherence. The refined examples are then added into D_{ex} and the process continues until a total of 500 examples for each reasoning type are obtained. Detailed examples for prompts and data can refer to Appendix A.

Retrieval and Paraphrasing. Subsequently, we augment annotated examples by a pre-trained paraphrasing model PEGASUS (Zhang et al., 2020) to improve the diversity and variability of our dataset. Finally, we retrieve the distracting facts f_d from a corpus of English *Wikipedia*[2] article text regarding the S as the query. We employ *term frequency-inverse document frequency* (TF-IDF) (Sparck Jones, 1972) to search 10-20 sentences presumed to most potentially confuse the LMs.

3.4 Statistics

Overall, **UniADILR** consists of two subsets: program-synthesized **UniADILR-PSy** and human-GPT annotated **UniADILR-HGc**. For each type of ADI reasoning, we generate 3200 samples, split into train/validation/test sets in a ratio of 7:1:2 in UniADILR-PSy, and 500 examples with complex contexts in UniADILR-HGc, solely for testing. In subsequent sections, we denote the data splits in the format of `[Reasoning Type Abbreviation]-[Split]`. For example, `Ab-In-train` represents the training split in UniADILR-PSy, which includes a mixture of abductive and inductive reasoning. As shown in Table 1, we comprehensively analyze the statistics of the two subsets. Detailed illustrations can be found in Appendix B. Intuitively, UniADILR-HGc exhibits a more abundant lexicon compared to UniADILR-PSy, enabling it more suitable for evaluating whether an LM has merely memorized patterns in the learning phase or instead mastered the underlying logic.

UniADILR	-PSy	-HGc
# Samples	9600	1500
Average Context Length	70.5	274.0
Vocab Size	417	9652
Gunning Fog Index	11.3	41.2
Retrieval Baseline	16.7%	8.7%

Table 1: The statistics of UniADILR.

3.5 Evaluation Settings

SFT. According to the reasoning type involved in the examples, we split UniADILR into three test subsets, which contains only one specific type of ADI reasoning. Taking the `Ab-test` as an example, we set four data splits to fine-tune the LMs: (a) training set involving the consistent logical type as the testing set, `Ab-train`; (b) training set involving single one reasoning type inconsistent with testing set, `De-train` and `In-train` respectively; (c) training set involving mixed reasoning types of ADI excluding which in testing set, `De-In-train`. Such an extensive setting enables a clear investigation of how the performance of SFT LMs will evolve when the distribution of logical rules in the training data shifts from that in the testing data.

ICL. For ICL, we delve into the LMs' performance by varying the rule distribution of the sampled demonstration examples. The models are provided with a query and a demonstration set C which includes a task instruction I and k demonstration examples. Therefore, $C = \{I, s(x_1, y_1), \cdots, s(x_k, y_k\}$, where $s(x_k, y_k)$ is an example randomly selected from the prompt set P.

---

[2] `https://en.wikipedia.org/wiki/Wikipedia`

We vary k and P to test how the demonstration influences the models' performance:

- $\boldsymbol{k = 0}$: $C = \{I\}$ only includes the task description.
- $\boldsymbol{k = 1}$: $C = \{I, s(x_1, y_1)\}$; $s(x_1, y_1) \in$ `Ab-train`, `De-train`, `In-train` respectively for three different settings.
- $\boldsymbol{k = 3}$: $C = \{I, s(x_1, y_1), s(x_2, y_2), s(x_3, y_3)\}$; $s(x_1, y_1) \in$ `Ab-train`, $s(x_2, y_2) \in$ `De-train` and $s(x_3, y_3) \in$ `In-train`.

Number of rules presented during the training stage. As for ADI reasoning can be viewed as permutations of similar primitives in a triangular direction, it is reasonable to expect an improvement of the LMs' generalization abilities to derive the remaining reasoning rule when they have learned along another two reasoning directions. To verify this assumption, we control the number of reasoning types appearing in the SFT training stage and demonstration examples for ICL. This allows us to investigate whether increasing exposure to a greater number of logical rules can enhance the generalisation across ADI reasoning.

OOD Vocabulary. Additionally, we test the LMs fine-tuned with the synthesized data on the UniADILR-HGc, which encompasses a wider range of vocabulary distribution and more intricate, longer contents. The results reveal the robustness of the LMs' reasoning capabilities acquired by learning from synthesized datasets.

Comparison between different learning strategies. Furthermore, we compare the overall performance of the LMs learning through the SFT and ICL methods. Through the detailed comparison and further analysis, we summarize their respective strengths and weakness in terms of the reasoning accuracy and generalization capability.

4 Experiments and Analysis

4.1 Implementation Details

For SFT settings, we fine-tune the pretrained **Flan-T5-large** (780M) and **Flan-T5-XL** (3B) (Raffel et al., 2020) implementations from the `HuggingFace Transformers`[3] library. Experiments of fine-tuning the pre-trained T5 models are conducted on 8 NVIDIA A100 GPUs. We utilize Adam optimizer and pick the model giving the best proof accuracy on the dev set. As for ICL, we compare the reasoning capabilities of **LLaMA2-7B**, **LLaMA2-13B**, **LLaMA2-70B** (Touvron et al., 2023), **GPT-3.5** and **GPT-4** (OpenAI, 2023b) under different demonstration settings. The detailed hyper-parameters are illustrated in Appendix C. We take the *proof accuracy* (PA) as the evaluation metric in the following experiments. If premises P_{pred} selected by the model match exactly with the label P_{gold}, the prediction is signed as True.

4.2 Performance IND and OOD

SFT LMs internalize rules seen during training very well but struggle to generalize to novel reasoning types. As demonstrated in Table 2, under the SFT setting, the models that fine-tuned on the training data involving the completely consistent reasoning rule with the testing data exhibit exceptional performance with the PA nearly all reaching 100%. However, when the distribution of reasoning type shifts from the training data, the performance experiences a significant decline, with the worst case even dropping below 10%. The results suggest that models fine-tuned with supervision may merely learn specific data patterns but struggle to capture the intrinsic connections among the basic logical reasoning primitives, making it challenging to generalize the reasoning rules to other types, akin to the nuanced reasoning processes observed in human cognition.

ICL large LMs performs stably when faced with the distribution variation of the reasoning type presented in the demonstration examples. As shown in Figure 2, the PAs within each bar cluster reflect the performance difference of each LM when the demonstration examples sampled from various distributions. We found that despite the value of PA changed along with the reasoning rules involved in the demonstration examples, the performance fluctuation is not as dramatic as which in SFT settings, generally not exceeding 5%. The results demonstrate that larger LMs through ICL can be more robust than smaller-scale LMs fine-tuned when attempting to generalize directly to OOD reasoning rules. Throughout a more meticulous observation, we found that LLaMA2 performs better with an increasing number of demonstration examples, while GPTs seems to excel when the reasoning type in the demonstration aligns consistently with the testing set. Among all tested LMs, GPT-4 significantly surpasses others and achieves a competitive PA even under the few-shot settings without any tuning compared to the models fine-

---

[3] `https://huggingface.co/`

Model	Test/Train	Ab-train	De-train	In-train	Complement-train
T5-780M	Ab-test	100.0%	9.8%(↓90.2%)	11.2%(↓88.8%)	55.7%(↓44.3%)
	De-test	32.3%(↓67.7%)	100.0%	48.8%(↓51.2%)	60.8%(↓39.2%)
	In-test	10.7%(↓89.3%)	40.8%(↓59.2%)	100.0%	70.0%(↓30.0%)
T5-3B	Ab-test	100.0%	20.0%(↓80.0%)	29.7%(↓70.3%)	59.2%(↓40.8%)
	De-test	66.9%(↓32.8%)	99.7%	55%(↓44.7%)	53.3%(↓46.4%)
	In-test	28.6%(↓71.4%)	66.6%(↓33.4%)	100.0%	69.3%(↓30.7%)

Table 2: Proof accuracy on SFT T5-780M and T5-3B. A significant performance drop can be observed when the reasoning type differs in the training and testing dataset.

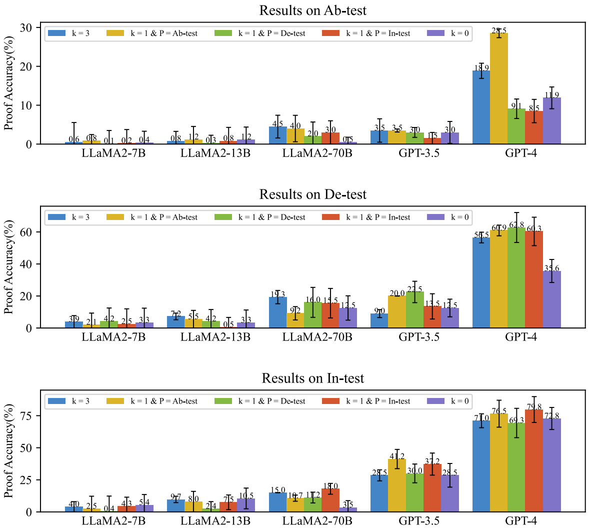


Figure 2: Proof accuracy on ICL LMs under various demonstration settings. GPT-4 performs best and the influence brought by the reasoning type visible in learning stage are not as pronounced as observed in SFT.

tuned on the OOD rules. However, for our presented UniADILR, none of the tested LMs attain a satisfactory level at extrapolating to unseen reasoning rules possessing the shared properties as we expect.

4.3 Number of Logical Rules

Increasing exposure to a greater number of logical rules during training can significantly enhance generalization among ADI reasoning for fine-tuned LMs, but appears to have little impact under the ICL settings. Comparing the PA of LMs fine-tuned on the testing set involving complementary reasoning types (in column 6 of Table 2) with that only involving the single OOD reasoning type (in column 4 and column 5 of Table 2), we observe an increase ranging from about 10% to 40%. Although the performance remains notably lower than that training in-distribution, these results suggest that incorporating a broader range of reasoning types might enhance the LMs' capability to generalize to other reasoning rules. We consider the phenomenon may be attributed to the combinationality nature of ADI reasoning. Additionally, we found better generalization between deductive and inductive reasoning, aligning with the findings in EGOIRE and Lakhdar (1996). However, as mentioned before, larger-scaled LMs under the few-shot settings exhibit relatively much more robustness to the number of learnable reasoning rules, as the fluctuation of PA does not exceed 5% along with the number of the logical rules presented in the demonstration examples.

4.4 OOD Vocabulary

As demonstrated in Figure 3, the PA of the fine-tuned T5 experiences a substantial drop on UniADILR-HGc even if the logical rules involved in the testing data remains consistent with that in the training data. A similar conclusion also applies to LLaMA2. These findings highlight that despite the fine-tuned LMs demonstrate a strong fit to characteristics within the training set, the acquired reasoning capabilities are exceptionally fragile and struggle with generalization to more intricate contexts. Surprisingly, we observed that GPT-3.5 achieved an even higher PA on the UniADILR-HGc compared with that on the UniADILR-PSy. We consider this might be attributed to UniADILR-PSy's use of fictional names for all concepts, preventing LMs from accessing commonsense knowledge. This observation implicitly suggests that the superior performance achieved by LMs on some existing logical reasoning datasets may rely on the commonsense context, instead of mastering the logical skills actually.

4.5 Overall Comparison of SFT and ICL

As shown in Figure 4, we demonstrate the overall performance of SFT and ICL for LMs and summarize their respective advantages and disadvan-

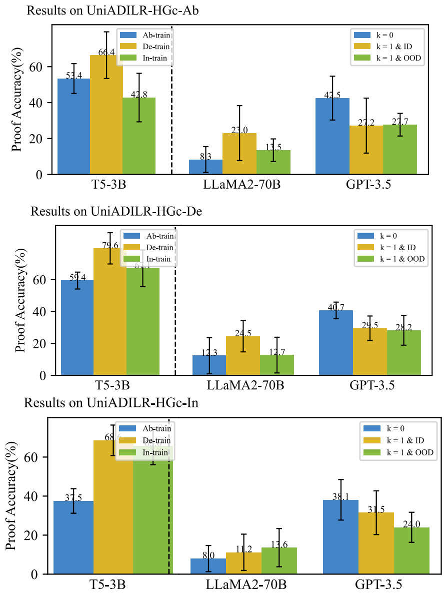


Figure 3: Performance of SFT T5-3B and ICL LLaMA2-70B, GPT-3.5 on UniADILR-HGc.

	Model/Type	Ab	De	In	Avg
SFT	T5-780M	47.6	49.9	$53.3^{\dagger}$	50.3
	T5-3B	$\mathbf{65.1}^{\dagger}$	**62.1**	61.1	**62.9**
ICL	LLaMA2-7B	0.4	3.2	$3.3^{\dagger}$	2.3
	LLaMA2-13B	0.9	4.1	$7.6^{\dagger}$	4.2
	LLaMA2-70B	2.8	$14.5^{\dagger}$	11.3	9.5
	GPT-3.5	2.9	15.5	$33.1^{\dagger}$	17.2
	GPT-4	11.9	55.1	$\mathbf{73.9}^{\dagger}$	46.7

Table 3: Average proof accuracy (%) of diverse models on ADI reasoning respectively. The **bolded** results represent which LLM achieves the highest PA, and the results with † indicate each LLM's highest PA achieved in which reasoning type.

tages. Firstly, ICL demonstrates more robustness, as its overall performance don't change dramatically along with the various reasoning rules involved in the demonstration examples. In contrast, although SFT LMs effectively internalize seen logical rules, its performance drops dramatically once the data distribution shifts OOD. Another notable advantage of ICL is its few-shot learning ability, not requiring a large amount of supervised labels costly in experts' time and labor. However, its overall effectiveness still falls behind the supervised settings. As which can be seen intuitively, only the best-performing GPT-4 is comparable to the supervised smaller-scaled models fine-tuned on the complementary training set. The results suggest a potential direction of integrating the two methods for future work.

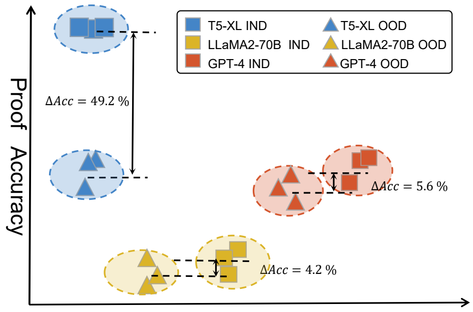


Figure 4: An intuitive representation demonstrates the generalization ability and overall performance of the SFT medium LM, open-source and close-source large LMs.

4.6 Proficiency Gap in ADI Reasoning

The results in Table 3 reveals that the proficiency of LMs in handling ADI rules varies indeed. Abductive reasoning emerges as the most challenging, with the highest average accuracy on GPT-4 reaching only 11.9%. In contrast, LMs tend to perform better on deductive and inductive reasoning. We attribute this phenomenon to the lack of available abductive reasoning data extracted from the web.

4.7 Reasoning Type Bias

Futhermore, we discuss whether LMs can distinguish what kind of logical reasoning type employed during their proofs. LMs are queried about the reasoning type used to substantiate the hypothesis in parallel with the proof generation process, and the response options include `abduction`, `induction`, and `deduction`. Results in Figure 5 demonstrate a noticeable bias in the distribution of predicted reasoning types. Deductive reasoning appears most frequently, particularly for GPTs. Additionally, we observed that when the model is presented with only one example, it is more inclined to predict the reasoning type as that appeared in the demonstration example, underscoring the significant impact of prompts on ICL performance, as noted in previous research (Zhao et al., 2021; Lu et al., 2022). The obvious bias reveals that even if the large LMs predict the reasoning clues successfully, they might not be aware of the reasoning rationale they are applying.

5 Related Work

5.1 Datasets of Textual Logical Reasoning

Logical reasoning is a fundamental reasoning ability and has attracted an increasing number of studies. LogicNLI (Tian et al., 2021) and FOLIO (Han

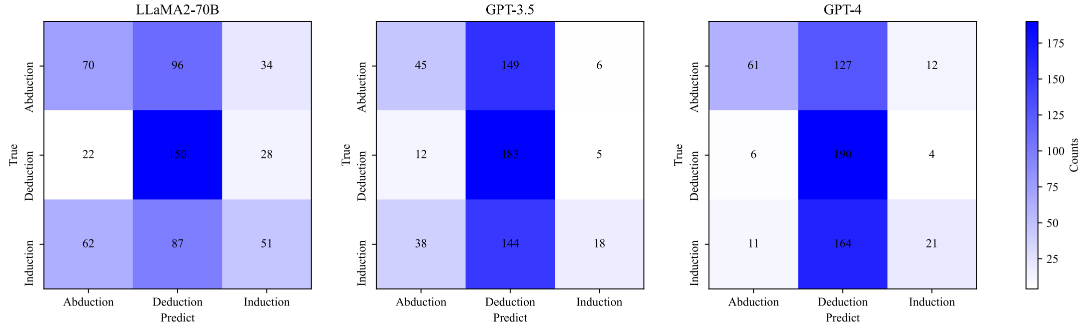


Figure 5: Confusion matrices reflecting the prediction of reasoning types under ICL settings. LLMs are observed to prefer generating deduction and the bias is particularly pronounced in GPTs.

et al., 2022) are two NLI-style datasets focusing on the first-order logic. EntailmentBank (Dalvi et al., 2021) proposes the proof generation task to represent the reasoning process explicitly, effectively improving the interpretability of the textual reasoning. However, these datasets cover mixed logical reasoning rules and are not intended to test ADI reasoning in isolation, therefore not proper for our evaluation.

Meanwhile, the significance of disentangling the reasoning types has also attracted much attention in recent years. RuleTaker (Clark et al., 2020) reason deductively over first-order logic knowledge bases expressed in natural language, determining the truth or falsity of given statements. PARAPATTERN (Bostrom et al., 2021) develop deductive sentence pairs from *Wikipedia* corpus using dependency template. PRONTOQA (Saparov and He, 2023) and PRONTOQA-OOD (Saparov et al., 2023) generate deductive examples according to propositional logical rules. AbductionRules (Young et al., 2022) compiles a synthesized dataset for training and testing natural-language abduction and CLUTRR (Sinha et al., 2019) is a dataset for inductive reasoning from text, but only induces family relations as rules, something that transformers are already proven to do well. By contrast, UniADILR involves broader inductive rules and therefore are more challenging.

However, these previous research have consistently focused on an isolated aspect of the three fundamental reasoning types defined in philosophy, ignoring their inter-connections and contrasts. In comparison, to the best of our knowledge, UniADILR stands out as the first benchmark designed to assess ADI logical reasoning within a unified domain and set of concepts. A detailed comparison can refer to Appendix D.

5.2 Generalization Ability of LMs

A number of recent works have measured the generalization ability of LMs from varying perspectives, involving length generalization (Anil et al., 2022; Kazemi et al., 2023; Saparov and He, 2023), width generalization (Saparov et al., 2023), compositional generalization (Wu et al., 2021; Kudo et al., 2023) and systematic generalization (Sinha et al., 2019) etc.. Diverging from previous works, our experiments explore generalization from a novel perspective — the generalization between different logical reasoning types. Inspired by philosophy research where ADI reasoning can be regarded as permutations in the varying order of unified reasoning elements, it is essential to assess extrapolation of learned reasoning skills by LMs to another one. Moreover, Our exploration can also be viewed to make contributions for investigating whether LMs understand the logical relationship between reasoning primitives or merely mimic common patterns in the training data (Xu et al., 2020).

6 Conclusion

We propose UniADILR, a natural language reasoning dataset containing examples demonstrating the fundamental logical rules in ADI. Motivated by evaluating the generalization across ADI reasoning, we design the subsets in UniADILR featuring the same formalization and concepts except for the specific reasoning rule. Experiments conducted on UniADILR indicate that the LMs struggle to extrapolate learned logic rules to a novel one in ADI reasoning even if the unique ranging factor between them is the reasoning direction. In addition, the results of SFT and ICL learning methods on LMs suggest their respective strengths and weakness faced with the reasoning rules out-of-distribution. Our work provides a novel perspective and benchmark for assessing the LMs' capability of generalizing to unseen logical rules.

Limitations

In this paper, we pay emphasis on evaluating the comparison and generalization of different ADI reasoning types in this paper, so how to enhance the prompt strategy to improve the proof accuracy is not our focus. Therefore, we solely employ the prompt in the plain form without delving into the potential change in proof accuracy resulting from an increase in demonstration numbers or meticulous design of example selection and ordering.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 72293575.

References

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 38546–38556. Curran Associates, Inc.

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*, pages 6266–6278, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

Eric GR EGOIRE and SA Lakhdar. 1996. Inductive reasoning is sometimes deductive.

Éric Grégoire. 2001. Generalization v.(induction v. abduction). In *Proc. of the Int. Conf. on Artificial Intelligence (IC-CAI'2001)*, volume 2, pages 868–873.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. LAMBADA: Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada. Association for Computational Linguistics.

Stefanie Krause and Frieder Stolzenburg. 2023. Commonsense reasoning and explainable artificial intelligence using large language models. In *European Conference on Artificial Intelligence*, pages 302–319. Springer.

Keito Kudo, Yoichi Aoki, Tatsuki Kuribayashi, Ana Brassard, Masashi Yoshikawa, Keisuke Sakaguchi, and Kentaro Inui. 2023. Do deep neural networks capture compositionality in arithmetic reasoning? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1351–1362, Dubrovnik, Croatia. Association for Computational Linguistics.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22266–22275.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

OpenAI. 2023a. Chatgpt: Optimizing language models for dialogue.

OpenAI. 2023b. GPT-4 Technical Report. *arXiv e-prints*, arXiv:2303.08774.

Pieter Pauwels, Ronald De Meyer, and Jan Van Campenhout. 2012. Towards a simulation of peirce's cycle of abductive, deductive and inductive reasoning. In *Model-Based Reasoning in Science and Technology*.

Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*, volume 1. Harvard University Press.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

FD Rivera and Joanne Rossi Becker. 2007. Abduction–induction (generalization) processes of elementary majors on figural patterns in algebra. *The Journal of Mathematical Behavior*, 26(2):140–155.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using OOD examples. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Vikas Thada and Vivek Jaglan. 2013. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2(4):202–205.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yuhuai Wu, Albert Jiang, Jimmy Ba, and Roger Baker Grosse. 2021. {INT}: An inequality benchmark for evaluating generalization in theorem proving. In *International Conference on Learning Representations*.

Keyulu Xu, Jingling Li, Mozhi Zhang, Simon S. Du, Ken ichi Kawarabayashi, and Stefanie Jegelka. 2020. What can neural networks reason about? In *International Conference on Learning Representations*.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A The Implementations for HMC annotation

The pipeline of the HMC annotation is illustrated in Figure 6. And additionally, we display an example of prompts used to instruct chatGPT to annotate more examples containing specific reasoning rules in Figure 7. These responses will be checked and corrected manually to retain logically sound examples.

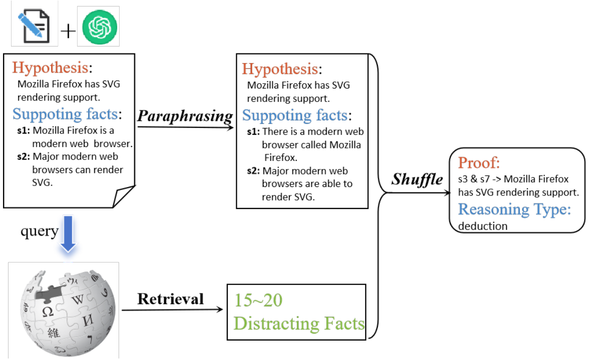

+
Hypothesis:
Mozilla Firefox has SVG rendering support.
Suppoting facts:
s1: Mozilla Firefox is a modern web browser.
s2: Major modern web browsers can render SVG.
Paraphrasing
Hypothesis:
Mozilla Firefox has SVG rendering support.
Suppoting facts:
s1: There is a modern web browser called Mozilla Firefox.
s2: Major modern web browsers are able to render SVG.
Shuffle
Proof:
s3 & s7 -> Mozilla Firefox has SVG rendering support.
Reasoning Type:
deduction
query
Retrieval
15~20 Distracting Facts

Figure 6: Pipeline of HMC annotation.

B Statistics

As shown in Table 1, we comprehensively analyze the statistics of the two subsets. The average sentence length in UniADILR-HGc is four times that in UniADILR-PSy. Unlike PSy, which uses a monotonous vocabulary from a limited set of predefined concepts, the vocabulary in HGC is sampled from the distribution of GPT's lexicon, resulting in more abundant and realistic word choices. We include additional visualizations of the dataset complexity using the Gunning Fog Index (GFI), defined as the proportion of sentence length to complex words (words containing three or more syllables). Intuitively, the proportion of complex vocabulary in HGC is significantly higher than in PSy. A retrieval-based naive baseline is provided for comparison with LLMs. We test three approaches relying on text overlap (Jaccard similarity (Thada and Jaglan, 2013) and TF-IDF (Sparck Jones, 1972)) or semantic embeddings (Word2Vec (Mikolov et al., 2013)), and calculate their average. These characteristics demonstrate that UniADILR-HGc is suitable for evaluating whether an LM has merely memorized patterns in the training set or has truly mastered the rules of logical reasoning.

C Parameters and Prompts

C.1 Parameters

Experiments of fine-tuning the pre-trained T5 models are conducted on NVIDIA A100 GPUs. We utilize Adam optimizer and pick the model that gives best proof accuracy on the dev set. During training, the hyperparameters are set as follows:

- T5-780M: initial learning rate = 1×10^{-4}, epochs = 300, batch size = 8, max number of input and output tokens = 1024/600 respectively, warm-up steps = 500
- T5-3B: initial learning rate = 5×10^{-4}, epochs = 30, batch size = 1, max number of input and output tokens = 1024/600 respectively, warm-up steps = 500

As for LLaMA2, we use the official open-source code provided by Meta[4]. The parameters for generation are set as follows: temperature = 0.9, top_p = 0.9, max_seq_len = 2048, max_gen_len = 200. For GPTs, we directly utilize the API provided by OpenAI [5] and used the default parameters.

C.2 Prompt Examples

A specific demonstration example for context learning (k=1) used in our experiments is illustrated in Figure 8. The demonstration consists of a task description, an example and a query question.

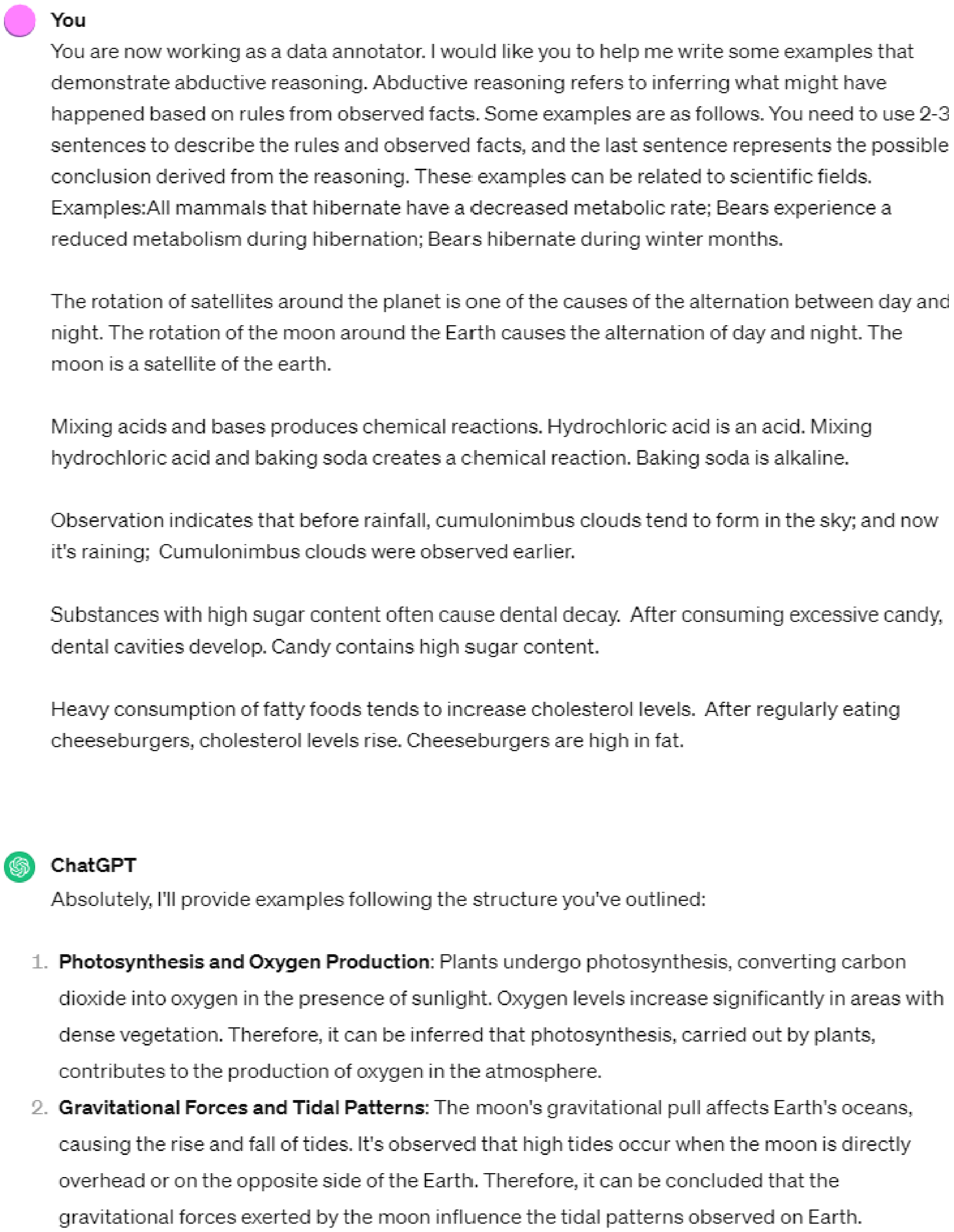
D Related Work: Datasets Comparison

In the field of NLR, several datasets have been established to investigate the capabilities of LMs in ADI logical reasoning. Deduction, acknowledged as the most prevalent human reasoning ability, has garnered significant attention. RuleTaker (Clark et al., 2020) reason deductively over first-order logic knowledge bases expressed in natural language, determining the truth or falsity of given statements. PARAPATTERN (Bostrom et al., 2021) develop deductive sentence pairs from *Wikipedia* corpus using dependency template. PRONTOQA (Saparov and He, 2023) and PRONTOQA-OOD (Saparov et al., 2023) generate deductive examples according to propositional logical rules. Relatively speaking, datasets dedicated to abductive and inductive reasoning are relatively scarce. AbductionRules (Young et al., 2022) compiles a synthesized dataset for training and testing natural-language abduction, demonstrating that transformers can perform abductive reasoning via fine-tuning. CLUTRR (Sinha et al., 2019) is a dataset for inductive reasoning from text, but only induces family relations as rules, something that transformers are already proven to do well. Previous research have consistently focused on an isolated aspect of the three fundamental reasoning types defined in philosophy, ignoring their inter-connections and contrasts. In comparison, to the best of our knowledge, UniADILR stands out as the first benchmark designed to assess ADI logical reasoning within a unified domain and set of concepts. A detailed comparison can refer to Table 4.

---

[4] `https://huggingface.co/meta-llama`

[5] `https://platform.openai.com/docs/overview`



You

You are now working as a data annotator. I would like you to help me write some examples that demonstrate abductive reasoning. Abductive reasoning refers to inferring what might have happened based on rules from observed facts. Some examples are as follows. You need to use 2-3 sentences to describe the rules and observed facts, and the last sentence represents the possible conclusion derived from the reasoning. These examples can be related to scientific fields. Examples:All mammals that hibernate have a decreased metabolic rate; Bears experience a reduced metabolism during hibernation; Bears hibernate during winter months.

The rotation of satellites around the planet is one of the causes of the alternation between day and night. The rotation of the moon around the Earth causes the alternation of day and night. The moon is a satellite of the earth.

Mixing acids and bases produces chemical reactions. Hydrochloric acid is an acid. Mixing hydrochloric acid and baking soda creates a chemical reaction. Baking soda is alkaline.

Observation indicates that before rainfall, cumulonimbus clouds tend to form in the sky; and now it's raining; Cumulonimbus clouds were observed earlier.

Substances with high sugar content often cause dental decay. After consuming excessive candy, dental cavities develop. Candy contains high sugar content.

Heavy consumption of fatty foods tends to increase cholesterol levels. After regularly eating cheeseburgers, cholesterol levels rise. Cheeseburgers are high in fat.

ChatGPT

Absolutely, I'll provide examples following the structure you've outlined:

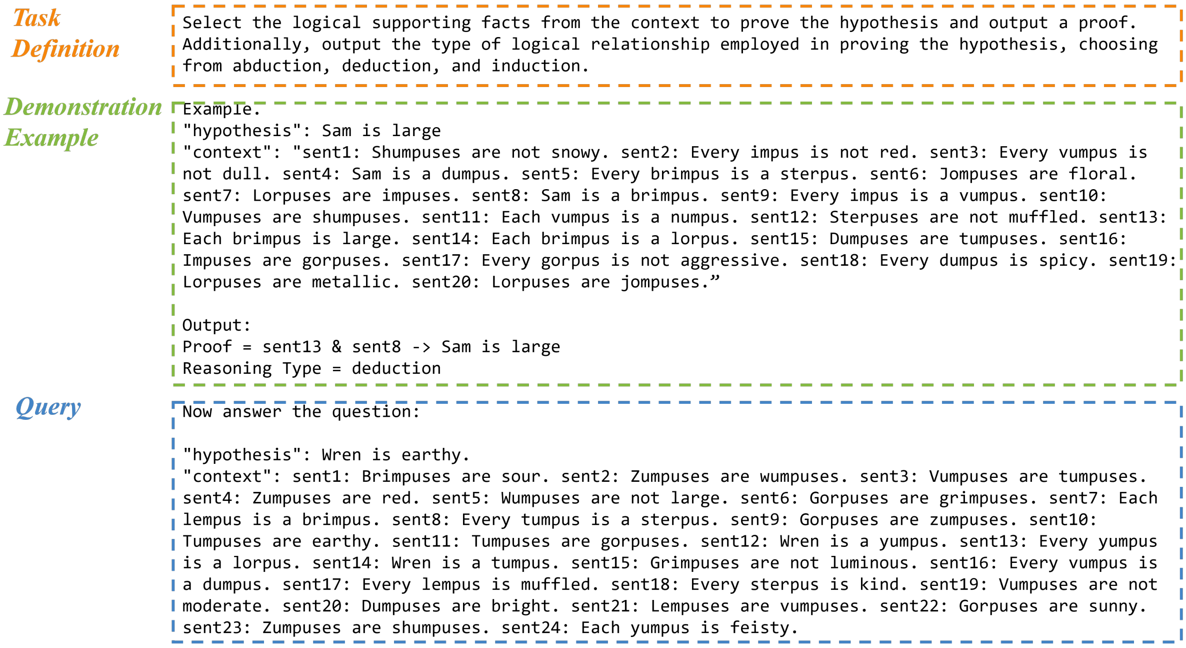
1. **Photosynthesis and Oxygen Production**: Plants undergo photosynthesis, converting carbon dioxide into oxygen in the presence of sunlight. Oxygen levels increase significantly in areas with dense vegetation. Therefore, it can be inferred that photosynthesis, carried out by plants, contributes to the production of oxygen in the atmosphere.
2. **Gravitational Forces and Tidal Patterns**: The moon's gravitational pull affects Earth's oceans, causing the rise and fall of tides. It's observed that high tides occur when the moon is directly overhead or on the opposite side of the Earth. Therefore, it can be concluded that the gravitational forces exerted by the moon influence the tidal patterns observed on Earth.

Figure 7: An example of the prompt we used to instruct GPT to generate reasoning examples involving the target reasoning type.

Dataset → **Property ↓**	**CLUTRR**[1]	**RuleTaker**[2]	**ParaPattern**[3]	**AbductionRules**[4]	**ProntoQA-OOD**[5]	**UniADILR (Ours)**
Reasoning Types	Induction	Deduction	Deduction	Abduction	Deduction	Abduction Deduction Induction
Task Formalization	Question Answering	Textual Entailment	Conclusion Derivation	Conclusion Derivation	Proof Generation	Proof Generation
Data Source	Program	Program	Wikipedia	Program	Program	Program & Human-annotated
Multiple Types	✗	✗	✗	✗	✗	✓
Generalization Between Types	✗	✗	✗	✗	✗	✓

[1] Sinha et al. (2019); [2] Clark et al. (2020); [3] Bostrom et al. (2021); [4] Young et al. (2022); [5] Saparov et al. (2023).

Table 4: A comparison of **UniADILR** with other related datasets involving ADI.



Task Definition

```
Select the logical supporting facts from the context to prove the hypothesis and output a proof.
Additionally, output the type of logical relationship employed in proving the hypothesis, choosing
from abduction, deduction, and induction.
```

Demonstration Example

```
Example.
"hypothesis": Sam is large
"context": "sent1: Shumpuses are not snowy. sent2: Every impus is not red. sent3: Every vumpus is
not dull. sent4: Sam is a dumpus. sent5: Every brimpus is a sterpus. sent6: Jompuses are floral.
sent7: Lorpuses are impuses. sent8: Sam is a brimpus. sent9: Every impus is a vumpus. sent10:
Vumpuses are shumpuses. sent11: Each vumpus is a numpus. sent12: Sterpuses are not muffled. sent13:
Each brimpus is large. sent14: Each brimpus is a lorpus. sent15: Dumpuses are tumpuses. sent16:
Impuses are gorpuses. sent17: Every gorpus is not aggressive. sent18: Every dumpus is spicy. sent19:
Lorpuses are metallic. sent20: Lorpuses are jompuses.”

Output:
Proof = sent13 & sent8 -> Sam is large
Reasoning Type = deduction
```

Query

```
Now answer the question:

"hypothesis": Wren is earthy.
"context": sent1: Brimpuses are sour. sent2: Zumpuses are wumpuses. sent3: Vumpuses are tumpuses.
sent4: Zumpuses are red. sent5: Wumpuses are not large. sent6: Gorpuses are grimpuses. sent7: Each
lempus is a brimpus. sent8: Every tumpus is a sterpus. sent9: Gorpuses are zumpuses. sent10:
Tumpuses are earthy. sent11: Tumpuses are gorpuses. sent12: Wren is a yumpus. sent13: Every yumpus
is a lorpus. sent14: Wren is a tumpus. sent15: Grimpuses are not luminous. sent16: Every vumpus is
a dumpus. sent17: Every lempus is muffled. sent18: Every sterpus is kind. sent19: Vumpuses are not
moderate. sent20: Dumpuses are bright. sent21: Lempuses are vumpuses. sent22: Gorpuses are sunny.
sent23: Zumpuses are shumpuses. sent24: Each yumpus is feisty.
```

Figure 8: Prompt example for ICL in our experiments.