Linguistic Rule Induction Improves Adversarial and OOD Robustness in Large Language Models

Shuoran Jiang[1], Youcheng Pan[2], Qingcai Chen[1,2,⋆], Yang Xiang[2,⋆], Yukang Lin[1]
[1]Harbin Institute of Technology, Shenzhen, [2]Peng Cheng Laboratory
shuoran.chiang@gmail.com, panyoucheng4@gmail.com, qingcai.chen@hit.edu.cn
xiangy@pcl.ac.cn, yukanglinlearner@gmail.com

Abstract

Ensuring robustness is especially important when AI is deployed in responsible or safety-critical environments. ChatGPT can perform brilliantly in both adversarial and out-of-distribution (OOD) robustness. Still, other popular large language models (LLMs), like LLaMA-2, ERNIE, and ChatGLM, do not perform satisfactorily in this regard. Therefore, it is valuable to study what efforts play essential roles in ChatGPT, and how to transfer these efforts to other LLMs. This paper experimentally finds that linguistic rule induction is the foundation for identifying the cause-effect relationships in LLMs. Accurately processing the cause-effect relationships in LLMs can improve their adversarial and OOD robustness. Furthermore, we explore a low-cost way of aligning LLMs with linguistic rules. Specifically, we constructed a linguistic rule instruction dataset to fine-tune LLMs. To further energize LLMs for reasoning step-by-step with the linguistic rules, we propose the task-relevant LingR-based chain-of-thoughts. Experiments showed that LingR-induced LLaMA-13B achieves comparable or better results with GPT-3.5 and GPT-4 on various adversarial and OOD robustness evaluations.

Keywords: Adversarial, Out-of-distribution, Robustness, Linguistic Rule, Cause-effect, Chain-of-thoughts

1. Introduction

Large language models (LLMs) have performed remarkably well in various natural language processing (NLP) tasks (Li et al., 2022; Wang et al., 2023). These foundation models show powerful in-context learning capabilities, and their prompting learning pattern has become a widely adopted paradigm for various NLP tasks (Zhou et al.). Among all LLMs, the generative pre-trained Transformers (GPT) family language models, such as GPT-3.5 and GPT-4, have made significant strides and even shown great potential in replacing traditional search engines (Mukherjee et al., 2023).

However, there still needs to be more research on whether LLMs are plagued by the inherent vulnerability of deep neural networks to data disturbances (Wang et al., 2021a). These disturbances include adversarial noises and out-of-distribution (OOD) shifts (Milbich et al., 2021). The robustness of these disturbances is an important metric for trustworthy and secure AI research (Yang et al., 2023). Wang et al. (2023) studied the state-of-the-art adversarial and OOD robustness in ChatGPT. However, due to the secrecy of ChatGPT training (Liu et al., 2023), researchers still cannot determine what efforts enhance the robustness of ChatGPT (Liu et al., 2023). Is it the larger pre-training corpus, the larger number of model parameters, or a corpus designed explicitly for the language model robustness?

Up to now, other newest LLMs, such as LLaMA 2 (Touvron et al., 2023b), ChatGLM (Du et al., 2022),

⋆ Corresponding Author

GPT-3.5: No, the sentence "The president advised the doctor." does not necessarily imply that "The doctor advised the president." *Linguistically, the two sentences have different subjects and objects, and their meanings depend on the specific context in which they are used.* One does not necessarily imply the other without additional information or context.

GPT-4: No, it does not imply that "The doctor advised the president." *From a linguistic perspective, sentences express actions and relationships between entities (subjects, objects, etc.), and part of this expression is the direction of the actions – who does what to whom. In English, this is usually done with word order: the subject (the one doing the action) comes before the verb, and the object (the one receiving the action) comes after the verb.*

LLaMA-2: The sentence "The president advised the doctor" does not necessarily imply that the doctor advised the president in return. *From a linguistic perspective, we would need more context or specific phrasing to determine if there was a reciprocal advising relationship between the president and the doctor.*

ChatGLM: The president advised the doctor, so the doctor would have to have advised the president.
Therefore, the answer is yes.

ERNIE: No, this does not imply that "The doctor advised the president." *From the linguistic perspective, the sentence "The president advised the doctor" implies that the president was the advisor and the doctor was the advice.} It means that the president gave advice to the doctor, not the other way around.*

Figure 1: LLMs respond to an NLI question "The president advised the doctor. Does this imply that "The doctor advised the president."? Yes, no, or maybe? Give your thoughts on this decision from a linguistic perspective."

ERNIE (Wang et al., 2021b) et. al., do not catch up with ChatGPT regarding adversarial and OOD robustness (Wang et al., 2023). An effective way to improve adversarial and OOD robustness in machine learning models is by identifying cause-effect relationships (Zhang et al., 2021). For NLP, the cause-effect pairs can be linked through linguistic patterns (Nazaruka, 2020; Yang et al., 2022), such

LREC-COLING 2024, pages 10565–10577
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

as the semantic dependency parsing tree. Our preliminary experiments also found that LLaMA-2, ChatGLM, and ERNIE have a significant gap in their ability to answer linguistic questions compared to GPT-3.5 and GPT-4. Herein, it inspires us to explore whether linguistic rule induction plays an essential role in the robustness of language models (Ellis et al., 2022; Goldstein et al., 2022). Language rule induction can be used to identify the patterns of word usage, sentence structure, verb conjugation, or any other grammatical component of a language (Ellis et al., 2022). Whether LLMs can reason with linguistic rules step-by-step is a sign of interpretable AI services in NLP (Belinkov et al., 2020).

Our test in Figure 1 used a natural language inference (NLI) example from HANS (McCoy et al., 2020) to evaluate whether popular LLMs have the linguistic dependence parsing ability to answer the NLI question accurately. This NLI example has high word overlaps in its premise and hypothesis, and an LLM will respond to incorrect answers if it cannot correctly identify the difference between subject and object in two sentences. In this example, even though GPT-3.5, GPT-4, LLaMA-2, and ERNIE (Wang et al., 2021b) give the correct answers, only GPT-3.5 and GPT-4 reason the question with linguistic rules, e.g., the semantic dependencies. As GPT-3.5 and GPT-4 can analyze linguistic rules, they outperform other LLMs in adversarial and OOD robustness.

To explore the significance of linguistic rule induction for enhancing the adversarial and OOD robustness in LLMs, we construct the Linguistic Rule (LingR) instruction dataset to fine-tune LLMs. Specifically, the LingR dataset is constructed on pure text from the Universal Dependencies (UD) English EWT (Liu et al., 2018) dataset, in which each sentence can be used to generate instructions corresponding to different linguistic questions. By fine-tuning with LingR instructions, the LLMs acquire the basic linguistic rules. Further, to leverage linguistic knowledge to enhance the robustness of LLMs in downstream NLP tasks, we also construct the LingR chain-of-thoughts (LingR-CoTs) dataset for LLMs fine-tuning and in-context learning.

Experiments thoroughly evaluated adversarial robustness across the AdvGLUE Wang et al. (2021) and ANLI Nie et al. (2020) benchmarks, and OOD robustness across the Flipkart reviews Adane et al. (2023), DDXPlus medical diagnosis datasets Fansi Tchango et al. (2022), Heuristic Analysis (HANS) for NLI systems McCoy et al. (2019), and Paraphrase Adversaries from Word Scrambling (PAWS) Zhang et al. (2019) benchmarks. Our results demonstrate that LingR fine-tuned LLaMA-13B (LingR-LLaMA) possesses basic linguistic rules. LingR-CoTs-based few-shot and in-context learning can achieve comparable robustness with GPT-3.5 and GPT-4 in downstream NLP tasks. These results confirm that linguistic rule induction is crucial in enhancing LLMs' robustness, and our study proposed a low-cost way to improve it.

Our contributions are as follows:

- We empirically found that whether LLMs can analyze input text from the linguistic perspective plays an important role in adversarial and OOD robustness.
- To align the linguistic rule induction among LLMs, we propose a linguistic instruction constructing method based on knowledge distillation from ChatGPT.
- Linguistic instructions enable LLMs to acquire the ability to analyze text linguistically. On this basis, this study also proposes a linguistic chain-of-thoughts (LingR-CoTs) construction method. It further improves the adversarial and OOD robustness on downstream NLP tasks.

2. Related Works

2.1. Large Language Models

Large language models (LLMs), like the popular GPT-3.5, GPT-4, LLaMA, and ERINE, unify NLP tasks into the instruction learning paradigm (Chang et al., 2023). As LLMs are pre-trained on a large-scale corpus, they show significant performance on a wide of downstream NLP tasks such as sentiment analysis, question answering, logical reasoning, and automatic diagnosis (Wang et al., 2023). Most popular LLMs use an auto-regressive language model in their decoder-only architectures (Min et al., 2023), such as GPT-3.5, GPT-4, LLaMA-2, and ERNIE. Given a context sequence X, the auto-regressive LM objective is to maximize the log-likelihood of the next word while given previous words.

$$\mathcal{L}(X;\boldsymbol{\theta}) = \prod_{t=1}^{T} p(x_t | x_{t-1}, \cdots x_1) \tag{1}$$

where T is the sequence length.

LLMs use prompt engineering (Clavié et al., 2023; White et al., 2023; Zhou et al.) to interact with users, where users provide specific prompt texts to guide LLMs in generating desired responses.

2.2. Chain-of-Thoughts Instruction

Chain-of-thoughts (CoTs) instruction is a two-tiered querying strategy to elicit a sequence of intermediate reasoning steps for each query from LLMs (Chen et al., 2023; Wei et al., 2022). The CoTs

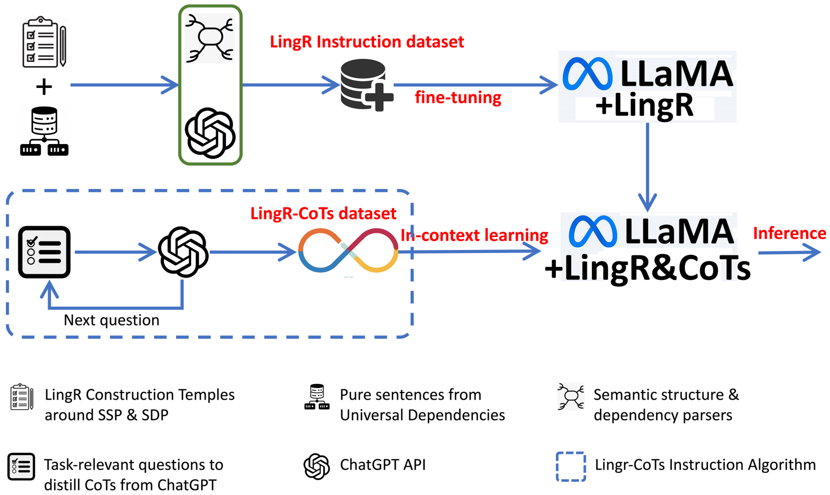

Figure 2: The overall framework includes LingR and LingR-CoTs construction and two-phase fine-tuning.

instructions have significantly bolstered the performance of prompting in tackling intricate tasks (Cai et al., 2023). By simply appending the instruction "Let's think step-by-step, ..." to each query, the CoTs instruction significantly improves GPT-3's math reasoning accuracy from 17.7% to 78.7% (Wei et al.; Chen et al., 2023).

2.3. Robustness of Language Models

Robustness refers to the capacity of a system to endure disruptions or external factors that could lead to its malfunction (Wang et al., 2023). These disruptions include adversarial disturbances and OOD shifts.

The goal of adversarial robustness in language models is summarized as follows,

$$\min_{f\in\mathcal{H}}\mathbb{E}_{(x,y)\in\mathcal{D}}\max_{|\delta|\leqslant\epsilon}\ell\left[f\left(x+\delta\right),y\right] \tag{2}$$

where ϵ represents the imperceptible changes σ of an input text sequence x, and y is the learning objectives.

The objective of OOD robustness is the average risk on all possible OOD shifts,

$$\min_{f\in\mathcal{H}}\mathbb{E}_{e\sim\mathcal{Q}}\mathbb{E}_{(x,y)\in\mathcal{D}^e}\ell\left[f\left(x\right),y\right] \tag{3}$$

where e represents the OOD shift from the distribution $\mathcal{Q}$ of training data.

3. Methods

This section constructs two instruction datasets. The first is Linguistic Rule (LingR) instructions for pure text in the universal dependencies English EWT (UD-English-EWT) (Silveira et al., 2014) dataset, and the second is LingR chain-of-thoughts (LingR-CoTs) instructions for all downstream NLP tasks evaluated in our experiment. As shown in Figure 2, our proposed linguistic rule induction for LLMs includes three phases: i. task-irrelevant LingR instruction learning phrase, ii. task-specific LingR-CoTs for few-shot fine-tuning phrases, and iii. the in-context learning (ICL) to inference robustness evaluations.

3.1. Linguistic Rule Instruction

This study designs the LingR instruction dataset with pure text upon the UD-English-EWT dataset. UD-English-EWT comprises 16,621 sentences from weblogs, newsgroups, emails, reviews, and Yahoo answers. The dependency trees in UD-English-EWT are automatically converted into Stanford Dependencies (Liu et al., 2018) and then hand-corrected to Universal Dependencies. Each sentence in UD-English-EWT can be used to construct instructions by distilling the ChatGPT for the 72 linguistic questions (51 for the syntactic structure parsing (SSP) tree and 20 for the semantic dependency parsing (SDP) tree).

Distilling SSP knowledge. SSP focuses on the formal structure of sentences, which describes how words and phrases are combined into more complex structures. An SSP tree usually represents the syntactic structure, where nodes are words or phrases, and edges represent syntactic relationships such as subjects, objects, modifiers, etc. 51 questions upon the SSP tree are listed from easy to difficult in Table 1. These questions are categorized into 4 groups: (i) "What/Which is/are the ... ?" (ii) "How many ... ?" (iii) "Is/Are there ... ?" (iv) "If there are ... ?"

Distilling SDP knowledge. Identifying cause-effect relationships can help language models understand the connection between events, actions, or situations, and thus accurately interpret, predict,

What/Which is/are the … ?	1. root verb? 2. function of this noun? 3. subject? 4. object? 5. prepositional phrase? 6. direct object of this verb? 7. indirect object of this verb? 8. complement of this verb? 9. participial phrase? 10. gerund phrase? 11. infinitive phrase? 12. adverbial phrase? 13. prepositional phrase? 14. noun clause? 15. root of the dependency tree? 16. direct object of this sentence? 17. predicate of this sentence? 18. indirect object of this sentence? 19. subject complement of this sentence? 20. object complement of this sentence? 21. subordinate clause? 22. modifier of the subject? 23. modifier of the direct object? 24. modifier of the indirect object? 25. modifier of the subject complement? 26. modifier of the object complement? 27. modifier of the adverbial phrase? 28. modifier of the prepositional phrase? 29. modifier of the subordinate clause? 30. head of the subject phrase? 31. head of the object phrase? 32. head of the predicate? 33. main subject of the sentence? 34. verb is being used in the sentence? 35. direct object in the sentence? 36. indirect object in the sentence? 37. adjective modifying in the sentence? 38. nature of the pronoun? 39. being modified by prepositions? 40. are affected by the passive voice, if any? 41. being negated by "not" or its equivalent? 42. being compared by "like" or its equivalent? 43. being emphasized by italics or boldface?
How many … ?	44. conjunctions are used in the sentence, and what is their function? 45. noun phrases are in the sentence, and what are their relationships to each other? 46. nouns are in the sentence? 47. parts of the sentence interact to convey meaning?
Is/Are there … ?	48. a subordinate clause? If so, what is its relationship to the principal clause? 49. a participial phrase? If so, what is its relationship to the rest of the sentence? 50. any ellipses or omissions, and if so, what is their effect on the syntax?
If there are … ?	51. multiple clauses in the sentence, what is the relationship between them?

Table 1: Distilling SSP knowledge with 51 questions.

What/Which is/are the … ?	1. main predicate? 2. subjects of main predicate? 3. relationship between the arguments and main predicate? 4. argument structure of main predicate? 5. dependency relationship between subject and main predicate? 6. arguments are modified as a core argument? 7. discourse function of the different arguments? 8. sentence participate in the same event description? 9. semantic roles of the arguments? 10. scope of negations? 11. types of modality expressions? 12. aspectual profile of the verb?
Are there any … ?	13. clauses in the sentence? If so, what is their relationship to the main predicate? 14. negations? 15. comparative or superlative expressions?
How do/does/is /are the … ?	16. sentence express causation? 17. sentence express temporality? 18. arguments relate to one another? 19. information in the sentence presented? 20. Is it new information or old information? 21. argument structures of the verb vary?

Table 2: Distillating SDP knowledge with 21 questions.

and generate text. Cause-effect relationships can be represented in text in a variety of ways: (i) connective words and phrases, such as "because", "due to", "as a result of", "therefore", "hence", "consequently", "so", "thus" and so on; (ii) verbs and verb phrases like in the sentence "The loud noise alarmed the birds."; (iii) nouns and noun phrases like in the sentence "The storm was the cause of the power outage."; (iv) passive voice, such as the "by" phrase in "The town was destroyed by the hurricane." (v) context or common sense; (vi) modifiers and adverbial phrases.

Understanding and identifying cause-and-effect relationships in text is critical for general NLP tasks, such as (i) answering the "Why ... ?" in question answering, (ii) understanding the cause-effect relationships in premise and hypothesis in natural language inference, (iii) understanding the reason for emotions and viewpoints in sentiment analysis, (iv) the cause and effect of an event in event extraction. For LLMs, accurately identifying and processing these relationships is critical to achieving higher adversarial robustness and OOD generalization.

SDP is a powerful tool to capture cause-effect relationships in the text as it can provide transparent semantic relationships and complex structure analysis. Table 2 shows 21 questions to distill SDP knowledge.

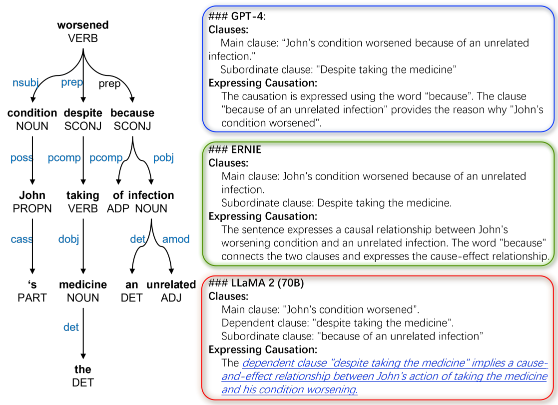

Figure 3: LLMs responses for request: "For the following sentence, are there any clauses in the sentence? If so, what is their relationship to the main predicate? Are there any negations? Are there any comparative or superlative expressions? How does sentence express causation? `Despite taking the medicine, John's condition worsened because of an unrelated infection.`" GPT and ERNIE correctly recognize the cause-effect relationship in it, but LLaMA-2 reasons the spurious causal dependency.

Leveraging SDP knowledge in language models can improve robustness against spurious correlations in the text. (i) SDP reveals the true relationships between words in a sentence, beyond just their position in the sentence or their syntactic relations. (ii) SDP trees can aid the language models in recognizing the core information of a sentence (like subjects and predicates) and peripheral information (like modifiers or additional details). (iii) SDP offers a more precise and structured context, mitigating dependence on spurious correlations or noise in the data. These advantages ensure that the language models don't over-rely on peripheral information for decision-making. Without proper SDP knowledge, as shown in Figure 3, LLaMA-2 might mistakenly associate "taking the medicine" with "John's condition worsened" and respond that the medicine led to the deterioration in John's health. However, the sentence explicitly states that the deterioration was due to an "unrelated infection." By correctly parsing semantic dependencies, the language model can recognize that 'Descite' represents a relationship despite this, while 'reason' represents the true reason. In this way, the model will not be misled by the spurious correlation between the drug and the condition but will understand that the true cause is an infection unrelated to the drug.

$\mathcal{A}$	Whether the premise and hypothesis have the same 1. root verbs? 2. subjects? 3. objects? 4. prepositional phrases? 5. roots of their the dependency trees. 6. Which words are compared by 'like' or equivalent in premise and hypothesis? 7. Are there any ellipses or omissions in both the premise and hypothesis, and if so, what is their effect on the syntax? 8. Which words are emphasized by italics or boldface in the premise and hypothesis?
$\mathcal{I}$	9. Are there subordinate clauses in both premise and hypothesis? If so, are its relationships to the principal clauses the same? 10. Which parts of both the premise and hypothesis are affected by the passive voice? If any, are these parts in both sentences the same? 11. Which words are negated by 'not' or its equivalent in both premise and hypothesis? If any, are these words in both sentences the same? 12. Are there any comparative or superlative expressions in the premise and hypothesis? And whether these expressions in the premise and hypothesis are the same.
$\mathcal{C}$	13. Whether both premise and hypothesis have indirect objects of the root verbs? If so, are two indirect objects the same? 14. Whether the predicates of the premise and hypothesis are the same. 15. What is their relationship (e.g., co-ordinate, subordinate) if multiple clauses exist in the premise and hypothesis? 16. How do the various parts of both premise and hypothesis interact to convey their overall meaning? 17. How do the premise and hypothesis express causation respectively? 18. Whether the causations are the same in premise and hypothesis. 19. What are the relationships between the arguments and the main predicate in premise and hypothesis? And whether these relationships in premise and hypothesis are the same. 20. Are there any negations in the premise and hypothesis? If so, are these negations in the premise and hypothesis the same, and are the scopes of negations in the premise and hypothesis the same? 21. What types of modality expressions are present in the premise and hypothesis? And whether these types of premises and hypotheses are the same. 22. What are the aspectual profiles of the verbs in the premise and hypothesis respectively? And whether these aspectual profiles are the same in premise and hypothesis.

Table 3: Knowledge distillation question on NLI tasks in association ($\mathcal{A}$), intervention ($\mathcal{I}$) and counterfactual ($\mathcal{C}$) on cause-effect semantic relationships.

Our preliminary experiments verified that GPT-4 masters the ability of SSP and SDP analysis on text. We use SSP and SDP questions in Tables 1 and 2 to distill the **ling**uistic **r**ule (LingR) instructions from GPT-4 by Algorithm 1. Because not all sentences contain the grammatical structures in the above 72 questions, distilled instructions with no answers are filtered out automatically and manually.

The final LingR dataset contains 301,286 instructions. An instruction example for an SSP question upon an English sentence is listed in Table 4.

Algorithm 1 Instruction construction

Input: Dataset $\mathcal{D}=\{x_i, y_i\}_{i=1}^{N}$, questions $\mathcal{Q}=\{q_m\}_{m=1}^{M}$

Initialize `LingR` $=[]$ # Instruction Set
for x, y in $\mathcal{D}$ **do**
  `item` $=\{\}$ # instruct set
  for q in $\mathcal{Q}$ **do**
    `response` $\leftarrow$ `request` (`ChatGPT`; x, y, q)
    `item` [‘`cot`’] $\pm$ `response` [‘`answer`’]
    if `response` [‘`decision`’] $==$ `True` **then**
      `LingR` $\pm$ [`item`]
    end if
  end for
end for
# the `request`() function
`request` (`ChatGPT`; x, y, q) $=\{$
    model=gpt-4;
    messages=give answer for q on x;
    can y be made from answer of q $\}$

Instruction:
For the given sentence: “You wonder if he was manipulating the market with his bombing targets.”.
What is the prepositional phrase in this sentence?",
Input:
Response:
“with his bombing targets” is the prepositional phrase in this sentence.

Table 4: An example of instruction on SSP questions.

3.2. LingR-CoTs Instruction

After the LingR instruction fine-tuning, how to improve the linguistic reasoning step-by-step in LLMs on downstream NLP tasks is another key in our study. To this end, we designed a task-specific linguistic rule-based chain-of-thoughts (LingR-CoTs) for few-shot fine-tuning and ICL. As the tasks used for adversarial and OOD robustness evaluations are all included in the GLUE benchmark, we used the proposed LingR-CoTs construction summarized in the Algorithm 1 to generate few-shot examples. Take the NLI task as an example, there are 22 questions to distill linguistically step-by-step thoughts from ChatGPT as shown in Table 3 with three levels of causal reasoning.

The association ($\mathcal{A}$) level questions are related to whether there is a certain correlation between premise and hypothesis sentences in NLI. At this level, the distilled chain-of-thoughts mainly focuses on the surface structure and content of two sentences. The intervention ($\mathcal{I}$) level questions consider the ability to change the meaning of a sentence through some intervention, such as adding or removing certain components. The counterfactual ($\mathcal{C}$) level questions focus on the deep meanings of sentences, considering the authenticity of sentences in other contexts or conditions.

For QQP, MNLI, QNLI, RET, and SST-2 tasks in

Model & #Param	Adversarial robustness (ASR) ↓						OOD robustness (F1) ↑	
	SST-2	QQP	MNLI	QNLI	RTE	ANLI	Flipkart	DDXPlus
Random baseline	50.0	50.0	66.7	50.0	50.0	66.7	20.0	4.0
BERT-B (110M)	67.0	62.1	71.3	60.2	59.5	N/A	N/A	N/A
RoBERTa-B (125M)	41.5	38.6	48.2	47.5	54.6	N/A	N/A	N/A
DeBERTa-L (435M)	66.9	39.7	64.5	46.6	60.5	69.3	60.6	4.5
BART-L (407M)	56.1	62.8	58.7	52.0	56.8	57.7	57.8	5.3
GPT-J (6B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	2.4
Flan-T5-L (11B)	40.5	59.0	48.8	50.0	56.8	68.6	58.3	8.4
OPT (13B)	47.6	53.9	60.3	52.7	58.0	58.3	44.5	0.3
OPT-ICL (13B)	50.0	41.0	67.8	50.0	50.4	65.4	75.4	1.2
LLaMA (13B)	67.3	71.0	56.8	61.7	45.3	68.0	67.8	6.3
LLaMA-ICL (13B)	63.9	52.3	52.6	50.0	36.7	64.6	76.1	11.2
LLaMA-2 (13B)	55.1	47.1	54.8	55.3	61.4	56.5	77.1	0.2
LLaMA-2-ICL (13B)	52.7	44.3	48.8	41.5	38.9	60.0	78.0	6.8
GPT-NEOX (20B)	52.7	56.4	59.5	54.0	48.1	70.0	39.4	12.3
BLOOM (176B)	48.7	59.0	73.6	50.0	56.8	66.5	28.0	0.1
text-davinci-002 (175B)	46.0	28.2	54.6	45.3	35.8	68.8	57.5	18.9
text-davinci-003 (175B)	44.6	55.1	44.6	38.5	34.6	62.9	57.3	19.6
GPT-3.5-turbo (175B)	**39.9**	**18.0**	**32.2**	34.5	24.7	55.3	60.6	20.2
LingR-OPT (13B)	40.5	26.9	50.4	50.0	61.1	**49.0**	**95.5**	1.2
LingR-LLaMA (13B)	46.6	62.8	52.0	48.7	**22.2**	53.6	76.8	**20.6**
LingR-LLaMA-2 (13B)	40.5	30.3	42.9	**30.6**	29.6	55.5	95.0	0.9

Table 5: All results about adversarial robustness on AdvGLUE, ANLI, Flipkart, and DDXPlus. The best and top-2 results are highlighted with bold types and underlines respectively.

GLUE, the LingR-CoTs are constructed with step-by-step answers incrementally distilled from ChatGPT with related questions. Table 6 demonstrates two instruction examples to fine-tune LLMs with step-by-step CoTs distilled from ChatGPT.

Instruction of NLI task: Are the following premise and hypothesis sentences entailment, neutral or contradiction with each other? Premise: Conceptually cream skimming has two basic dimensions - product and geography. Hypothesis: Product and geography are what make cream skimming work. Answer me with 'entailment', 'neutral' or 'contradiction'., **### Let's think step by step.** 1. From a linguistic perspective, the premise and hypothesis sentences do not have the same root verbs. ... ... 16. There are no ellipses or omissions in either premise or hypothesis sentences. 17. Therefore: the answer is neutral. **### Response:** neutral
Instruction of sentiment analysis task: The movie review below is positive or negative? Give an answer with 'positive' or 'negative'. tapestry woven of romance, dancing, singing, and unforgettable characters. **### Let's think step by step.** 1. Here's the linguistic analysis step by step. The review uses words like "tapestry woven," which connotes something that is intricately and beautifully made. This choice of words suggests a positive connotation. ... ... 6. The review does not contain words that typically indicate a negative sentiment, such as "boring," "dull," "poor," etc. 7. Therefore: the answer is positive. **### Response:** Positive

Table 6: Examples of LingR-CoTs on NLI and sentiment analysis tasks.

4. Experiments

4.1. Implementation details

To evaluate the effectiveness of our proposed LingR&CoTs (LingR + LingR-CoTs), we fine-tuned OPT-13B (Zhang et al., 2022), LLaMA-13B (Touvron et al., 2023a) and LLaMA-2-13B (Touvron et al., 2023b) with the efficient low-rank adaptation (LoRA) (Hu et al., 2021) on two NVIDIA A100 GPUs with 160GB memory. The LoRA hyper-params are set as $r = 8$, $\alpha = 16$ and dropout $p = 0.05$, and the optimizer is AdamW (Loshchilov and Hutter) with a learning rate $r = 1e - 5$ and a batch size of 8, and the maximum length of prompt input is $2,048$. The number of training epochs is set to 5.

4.1.1. Benchmark and Metric

In order to verify the effectiveness of LingR&CoTs in improving LLMs' adversarial robustness, we chose AdvGLUE Wang et al. (2021) and ANLI Nie et al. (2020) as benchmarks. The metric for the adversarial robustness is adopted as the attack success rate (ASR) (a small ASR means well adversarial robustness). For OOD robustness, we chose Flipkart reviews Vaghani and Thummar, DDXPlus medical diagnosis datasets Fansi Tchango et al. (2022), Heuristic Analysis (HANS) for NLI systems McCoy et al. (2019), and Paraphrase Adversaries from Word Scrambling (PAWS) Zhang et al. (2019) as benchmarks. On Flipkart and DDXPlus, the F1 score is chosen as the metric, and the accuracy is chosen as the metric for HANS and PAWS.

As GPT-3.5 may include the entire GLUE datasets in its training data, we performed few-shot fine-tuning (FFT) on LLaMA-13B for evaluation fairness. In the FFT phrase, the numbers of few-shot LingR-CoTs examples are 500 for SST-2, and 1000 for QQP, MNLI, QNLI, RTE, and ANLI respectively.

4.1.2. Baselines

We evaluated several state-of-the-art LLMs as baselines, including i. supervised fine-tuning (FT) LLMs, like the BERT (Kenton and Toutanova, 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), BART (Lewis et al., 2020), GPT-J (Wang, 2021), Flan-T5 (Chung et al., 2022) and ii. the in-context learning (ICL) based LLMs, like the LLaMA, GPT-NEOX (Black et al., 2022), OPT (Zhang et al., 2022), BLOOM (Scao et al., 2022) and ChatGPT (text-davinci-002, text-davinci-003, gpt-3.5-turbo, gpt-4).

4.2. Result Analysis

4.2.1. Main Results

Table 5 lists all experimental results about the adversarial robustness of AdvGLUE and ANLI, and part of the OOD robustness on Flipkart and DDXPlus.

(1) These results show that the latest GPT models (including text-davinci-002/003, GPT-3.5-turbo, and GPT-4) and our proposed LingR-OPT, LingR-LLaMA, LingR-LLaMA-2 show consistent improvements for adversarial and OOD evaluations. (2) The results of zero-shot learning on OPT-13B, LLaMA-13B, and LLaMA2-13B are very close to the random baseline regarding adversarial robustness. (3) The ICL on three LLMs shows improvement on most of these benchmarks. However, there is still a significant gap compared to GPT-3.5. (4) Our LingR-CoTs fine-tuned OPT-13B, LLaMA-13B, and LLaMA2-13B perform well against adversarial robustness, improved by $\geqslant 9.5$ ASR compared with their vanilla ones. Even though LingR-LLaMA did not achieve the top-1 adversarial robustness on SST-2 and MNLI, it also performed comparably with the earlier versions (text-davinci-002/003) of GPT-3.5. (5) On OOD benchmarks, LingR-OPT-13B, LingR-LLaMA-13B, and LingR-LLaMA2-13B improve at least 17.9 F1-score on Flipkart. Specifically, even though DDXPlus is an unseen domain dataset from all fine-tuning data, LingR-LLaMA achieves a better result than GPT-4. The source code will be released soon.

Table 7 lists OOD robustness evaluations on HANS and PAWS. LingR-LLaMA-13B beats all compared models on HANS and PAWS with test accuracies of 93.8% and 71.2%. The OOD accuracy on HANS exceeds its in-distribution counterpart (dev-matched in MNLI).

Model	OOD Robustness (Acc) ↑			
	ID dev-m	OOD HANS	ID QQP	OOD PAWS
Random baseline	66.7	50.0	50.0	50.0
OPT (full-shot)	85.5	70.8	91.2	47.5
OPT (1000-shot)	46.5	50.3	64.0	58.9
LLaMA (full-shot)	85.3	75.3	90.5	46.9
LLaMA (1000-shot)	59.3	49.6	65.8	57.2
LLaMA2 (full-shot)	**87.3**	70.7	90.7	69.2
LLaMA2 (1000-shot)	85.2	56.3	83.4	58.5
LingR-LLaMA (1000-shot)	84.7	**93.8**	82.7	68.7
LingR-LLaMA (full-shot)	86.1	91.5	**91.5**	**71.2**
LingR-LLaMA2 (1000-shot)	82.7	81.5	83.0	86.5

Table 7: OOD robustness on HANS and PAWS, where all compared foundation models are 13B.

(i) Under the few-shot (1000 examples) learning setting, LingR-LLaMA-13B achieves a comparable level with LLaMA-2-13B on the in-distribution benchmarks and significantly outperforms LLaMA-13B, OPT-13B. Most importantly, the OOD generalization level of LingR-LLaMA-13B far exceeds vanilla OPT-13B, LLaMA-13B, and LLaMA-2-13B respectively. (ii) Under the fine-tuning of the total training data, the in-distribution test accuracy of LingR-LLaMA-13B is close to the other three vanilla LLMs. At the same time, it maintains its absolute advantage on the OOD benchmarks.

The above experimental results prove that LingR instructions significantly improve the adversarial and OOD robustness without the incremental pre-training on this related corpus. In addition, compared with GPT-3.5 and GPT-4, the LingR-LLaMA performs equivalently and saves 162 billion parameters.

4.2.2. Ablation study

LLaMA-13B	SST2	QQP	MNLI	QNLI	RTE	ANLI
FFT (1000-shots)	67.3	71.0	56.8	61.7	45.3	68.0
w/ LingR	66.4	71.8	55.4	64.3	40.3	68.5
w/ LingR-CoTs	53.4	62.8	52.0	58.7	22.2	53.6
w/ LingR&CoTs	50.2	60.0	48.5	55.5	22.4	50.7
w/ LingR&CoTs&ICL	50.2	59.4	48.3	55.7	23.6	51.4
FFT (2000-shots)	66.8	72.4	54.5	62.2	40.2	66.7
w/ LingR	67.2	68.9	52.1	60.1	40.2	68.0
w/ LingR-CoTs	50.3	61.1	53.6	59.2	**21.8**	54.1
w/ LingR&CoTs	**46.5**	50.7	**48.2**	**55.0**	24.6	49.1
w/ LingR&CoTs&ICL	47.9	**50.4**	**48.2**	54.8	22.3	**48.3**

Table 8: Ablation study evaluates the effectiveness of LingR, LingR-CoTs, LingR&CoTs, in which all models are fine-tuned with 1000-shots and 2000-shots respectively.

In order to verify the effectiveness of our proposed LingR&CoTs dataset for linguistic rule induction and robustness in LLMs, we designed this ablation study. This experiment performed adversarial robustness (ASR) evaluations in (1) FFT LLaMA-13B with 1000-shots and 2000-shots, (2) FFT LLaMA-13B with LingR-only instructions (w/ LingR),

(3) FFT LLaMA-13B with LingR-CoTs-only instructions (w/ LingR-CoTs), (4) FFT LLaMA-13B with LingR&CoTs instructions (w/ LingR&CoTs) and (5) in-context learning LLaMA-13B with LingR&CoTs instructions (w/ LingR&CoTs&ICL).

Table 8 shows that 2000-shots of LingR&CoTs have significantly improved the robustness and generalization compared with LingR-only and LingR-CoTs-only, especially for the ASRs of SST-2, QQP, MNLI, QNLI, and ANLI with declines of 3.8, 10.4, and 3.9, 4.2, 5.0 respectively. LingR&CoTs and LingR&CoTs&ICL achieve closed results, with a maximum ASR difference of 2.3 across all datasets. This result is reasonable and interpretable according to Dai et al. (2023) that ICL behaves similarly to explicit fine-tuning at the prediction, representation, and attention behavior levels. In addition, the more examples of FFT, the more adversarial robustness our proposed LingR&CoTs achieve.

4.3. Case Study

To explain how our proposed LingR&CoTs improve adversarial and OOD robustness in an interpretable way, we analyze the chain-of-thoughts output from the inference stage in LLaMA2-13B and LingR-LLaMA2-13B, respectively. The details are shown in the following Tables 9, 10 and 11.

LingR-LLaMA2 responds: The phrase "one to jump" from the linguistic perspective green is the adjective clause and implies that this movie is something to avoid or miss out on, even for those who are fans of horror films. Answer: Negative
LLaMA2 responds: The movie review below is positive or negative? Give an answer with 'positive' or 'negative'. This one is definitely one to jump, even for horror movie fanatics.

Table 9: Responses on AdvSST2

Table 9 illustrates a semantic analysis request: "Below is an instruction that describes a task. Write a response that appropriately completes the request. The movie review below is positive or negative? Answer with 'positive' or 'negative', and the gold rationale for making this decision. This one is definitely one to jump, even for horror movie fanatics." The word "fanatics" in this request usually strongly correlates with positive emotions. In common expressions, like weblogs, news, and online movie reviews, sentences with this strongly positive word usually express positive emotions. In this example, LLaMA2 does not understand the instruction intent and answer the question. While, Ling&CoTs fine-tuned LLaMA2 answers the question correctly and gives the linguistic rationale.

Table 10 exhibits an NLI request: Below is an instruction that describes a task. Write a response that appropriately completes the request. In Nigeria, by far the most populous country in sub-Saharan Africa, over 2.7 million people are infected with HIV. Based on that information, is the sentence "2.7 percent of the people infected with HIV exist in Africa." entailments or not entailments, and give the Gold Rationale for making this decision? The premise and hypothesis have high word overlaps, and two sentences often have an implicit relationship in common expressions. LLMs pre-trained on a corpus collected from the Internet will often predict that two sentences are mutually entailed only based on lexical overlaps. For example shown in Tabel 10, LingR-LLaMA2-13B responds with a reasonable rationale and correct answer, but LLaMA2-13B cannot give any information in response.

LingR-LLaMA2 responds: Not entailments. The original statement only indicates the number of people infected with HIV in Nigeria but does not provide the overall percentage for the entire African region.
LLaMA2 responds: Not entailment. While the statement is true for Nigeria alone, it cannot be generalized to the entire sub-Saharan African region without additional contextual information.

Table 10: Responses on AdvRTE.

LingR-LLaMA2 responds: No. The first sentence uses "lead to," which implies a cause-and-effect relationship. The second sentence uses "help with," which suggests a supportive or beneficial relationship.
LLaMA2 responds: Yes, the original sentence you provided is a paraphrase of the same idea as the second sentence you mentioned. From a linguistic perspective, the choice to use "precise advancement for dinner" instead of "eating dinner earlier in the evening" reveals a more formal and technical tone. This phraseology suggests a more scientific or medical approach to weight loss, emphasizing the precision and accuracy involved in determining the optimal time for dinner.

Table 11: An example of LingR&CoTs-LLaMA2-13B and LLaMA2-13B response in Flipkart.

Table 11 shows a review classification request: Write a response that appropriately completes the request. Can eating precisely advance for dinner lead to weight loss? Is that a paraphrase of the following sentence? Does eating dinner earlier in the evening help with weight loss? Yes or No? Give the gold rationale in linguistic perspective why make this decision. Even though LLaMA2-13B can analyze the request from a linguistic perspective, it does not capture the key grammatical components in the text to reason and answer questions. In contrast, LingR-LLaMA2 accurately captures the difference between "precisely" and "earlier" as time adverbs.

5. Conclusion

This paper finds that the linguistic rule induction plays an important role in improving the LLMs' robustness, and further proposes a low-cost knowledge distillation for aligning ChatGPT and other LLMs with linguistic rules. We proposed LingR and LingR-CoTs instruction datasets to enhance the linguistic rule induction in LLMs. By fine-tuning LLMs via these two datasets, LLMs' adversarial and OOD robustness shows consistent improvements. In addition, we provided a low-cost way to design linguistic rule instruction and task-specific linguistic CoTs. This study is of great significance for further research to improve the LLMs' robustness in multilingual domains.

6. Acknowledgements

This work is jointly supported by grants from the National Key R&D Program of China (No. 2022ZD0116002), Shenzhen Science and Technology Plan (No. ShenKeJiChuangXinZhi[2023]87, No. KJZD20230923115113026), the Science and Technology Department of Guizhou Province (No. Qiankehe Support[2022]General019), the National Social Science Foundation - Major Project (No. 20&ZD226), the Project funded by China Postdoctoral Science Foundation (No. 2023M741843), the Shenzhen Development and Reform Commission (No. XMHT20190108009), the National Natural Science Foundation of China (No. 62276075, 62106115, 62006062 and 62176076), the Guangdong Provincial Key Laboratory (No. 2022B1212010005), the Major Key Project of PCL (No. PCL2022D01, PCL2023A09), the Key Laboratory of Intelligent Computing in Network Environment.

7. Bibliographical References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 1–5.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *Challenges & Perspectives in Creating Large Language Models*, page 95.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2023. Large language models as tool makers. *arXiv preprint arXiv:2305.17126*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Jiuhai Chen, Lichang Chen, Heng Huang, and Tianyi Zhou. 2023. When do you need chain-of-thought prompting for chatgpt? *arXiv preprint arXiv:2304.03262*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Benjamin Clavié, Alexandru Ciceu, Frederick Naylor, Guillaume Soulié, and Thomas Brightwell. 2023. Large language models in the workplace: A case study on prompt engineering for job type classification. In *International Conference on Applications of Natural Language to Information Systems*, pages 3–17.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Kevin Ellis, Adam Albright, Armando Solar-Lezama, Joshua B Tenenbaum, and Timothy J O'Donnell. 2022. Synthesizing theories of human language with bayesian program induction. *Nature communications*, 13(1):5024.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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 Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Bjorn Ommer. 2021. Characterizing generalization under out-of-distribution shifts in deep metric learning. *Advances in Neural Information Processing Systems*, 34:25006–25018.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Erika Nazaruka. 2020. An overview of ways of discovering cause-effect relations in text by using natural language processing. In *Evaluation of Novel Approaches to Software Engineering: 14th International Conference, ENASE 2019, Heraklion, Crete, Greece, May 4–5, 2019, Revised Selected Papers 14*, pages 22–38. Springer.

Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-argument based bi-encoder for paraphrase identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014.

A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021a. Infobert: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*.

Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Shuohuan Wang, Yu Sun, Yang Xiang, Zhihua Wu, Siyu Ding, Weibao Gong, Shikun Feng, Junyuan Shang, Yanbin Zhao, Chao Pang, et al. 2021b. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2112.12731*.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, Timothy J Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64(5):1161–1186.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. 2021. Adversarial robustness through the lens of causality. In *International Conference on Learning Representations*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

8. Language Resource References

Padma Adane, Avanti Dhiran, Shruti Kallurwar, and Sushmita Mahapatra. 2023. Sentiment analysis of product reviews from amazon, flipkart, and twitter. In *International Conference on Information and Communication Technology for Competitive Strategies*, pages 373–381. Springer.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in Neural Information Processing Systems*, 35:31306–31318.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A Smith. 2018. Parsing tweets into universal dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448. Association for Computational Linguistics (ACL).

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 Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

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*. Association for Computational Linguistics.

Nirali Vaghani and Mansi Thummar. Flipkart product reviews with sentiment dataset, 2023. *URL https://www. kaggle. com/dsv/4940809*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

9. Appendix

As it can not be sure that the entire GLUE benchmark is not in the training data for OpenAI's models, we conducted another experiment, as shown in Table 12, to evaluate fine-tuning LLMs with the GPT-4-generated examples of the downstream task. This comparison would be stronger to demonstrate the inductive ability of our proposed LingR&CoTs instruction dataset.

As Table 12 shows the GPT-4-generated samples fine-tuned LLMs do not exhibit significant improvements compared to the zero-shot learning LLMs. Even though combining the GPT-4-generated dataset with LingR&CoTs together to fine-tune LLMs, there are no further improvements in the results.

This experiment demonstrates that our proposed linguistic rule instructions can distill the reasoning logic of an LLM into other LLMs from a linguistic perspective, instead of the simple data augmentation.

Fine-tuning setting	Model (#Param)	Adversarial robustness (ASR) ↓						OOD robustness (F1) ↑	
		SST-2	QQP	MNLI	QNLI	RTE	ANLI	Flipkart	DDXPlus
Zero-shot	OPT (13B)	47.6	53.9	60.3	52.7	58.0	58.3	44.5	0.3
	LLaMA (13B)	67.3	71.0	56.8	61.7	45.3	68.0	67.8	6.3
	LLaMA-2 (13B)	55.1	47.1	54.8	55.3	61.4	56.5	77.1	0.2
GPT-4-generated 1000-shots	OPT (13B)	44.5	30.8	68.6	51.8	30.9	59.2	57.1	2.4
	LLaMA (13B)	45.3	60.3	75.2	50.0	58.0	69.4	62.7	6.5
	LLaMA-2 (13B)	45.3	59.0	51.2	48.6	51.9	59.1	66.3	3.7
LingR&CoTs	OPT (13B)	**40.5**	**26.9**	50.4	50.0	61.1	**49.0**	**95.5**	1.2
	LLaMA (13B)	46.6	62.8	52.0	48.7	**22.2**	53.6	76.8	**20.6**
	LLaMA-2 (13B)	**40.5**	30.3	42.9	30.6	29.6	55.5	95.0	0.9
Combined	OPT (13B)	43.9	29.5	51.2	50.0	61.7	51.3	93.4	2.6
	LLaMA (13B)	48.6	57.7	53.7	52.7	32.1	55.2	79.7	17.9
	LLaMA-2 (13B)	**40.5**	43.6	**41.3**	**30.4**	35.8	53.4	95.2	4.5

Table 12: Comparison about the zero-shot, GPT-4-generated samples, LingR&CoTs and GPT-4-generated+LingR&CoTs fine-tuned LLMs for adversarial robustness on AdvGLUE, ANLI, Flipkart, and DDXPlus.