

Linguistic Rule Induction Improves Adversarial and OOD Robustness in LLMs

COLING 2024

Shuoran Jiang, Youcheng Pan, Qingcai Chen*, Yang Xiang, Yukang Lin

Harbin Institute of Technology, Shenzhen & Peng Cheng Laboratory

May 4th, 2024



Content

1 Abstract & Introduction

2 Robustness Metrics

3 Method

4 Experiment Analysis

5 Conclusion

Content

1 Abstract & Introduction

2 Robustness Metrics

3 Method

4 Experiment Analysis

5 Conclusion

Abstract & Introduction

- Ensuring **robustness** is especially important when AI is deployed in responsible or **safety-critical** environments.
- **Different models perform variously for robustness.** ChatGPT can perform brilliantly in both adversarial and out-of-distribution (OOD) robustness, while other open-source LLMs, like LLaMA, 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 study 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 robustness.
- Furthermore, this paper explore a low-cost way of aligning LLMs with linguistic rules.

Background

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.

- Current state of LLMs in handling adversarial noises and OOD shifts.
- Gap in performance between ChatGPT and other LLMs like LLaMA-2, ERNIE.

Content

- 1 Abstract & Introduction
- 2 Robustness Metrics

- 3 Method
- 4 Experiment Analysis
- 5 Conclusion

Adversarial and OOD Robustness Metrics

Robustness refers to the capacity of a system to endure disruptions (OOD shifts) or external factors (adversarial disturbances) that could lead to its malfunction.

Adversarial Robustness

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y) \in \mathcal{D}} \max_{|\delta| \leq \epsilon} [f(x + \delta), y] \quad (1)$$

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

Out-of-Distribution (OOD) Robustness

$$\min_{f \in \mathcal{H}} \mathbb{E}_{e \sim \mathcal{Q}} \mathbb{E}_{(x,y) \in \mathcal{D}^{e\ell}} [f(x), y] \quad (2)$$

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

Content

- 1 Abstract & Introduction
- 2 Robustness Metrics

- 3 Method
- 4 Experiment Analysis
- 5 Conclusion

Linguistic Rule Induction (LingR)

- Definition: Systematic identification of grammatical patterns and structures.
- Application: Use these patterns to improve LLMs' text comprehension and reasoning capabilities.
- Implementation: Creation of LingR dataset derived from Universal Dependencies English EWT.

Linguistic Rule Induction (LingR)

The LingR instructions are constructed by distilling the ChatGPT with 71 linguistic questions.

51 for the syntactic structure parsing (SSP) tree

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?

Linguistic Rule Induction (LingR)

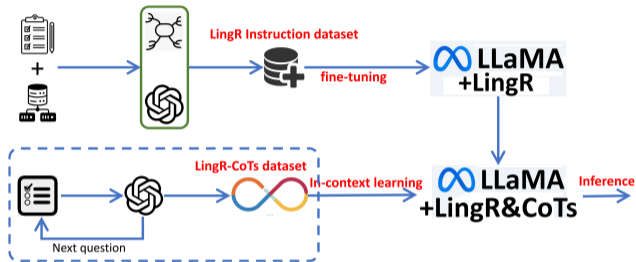
The LingR instructions are constructed by distilling the ChatGPT with 71 linguistic questions.

51 for the syntactic structure parsing (SSP) tree

20 for the semantic dependency parsing (SDP) tree

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?

Methodology



LingR Construction Templates around SSP & SDP

Pure sentences from Universal Dependencies

Semantic structure & dependency parsers

Task-relevant questions to distill CoTs from ChatGPT

ChatGPT API

Lingr-CoTs Instruction Algorithm

图: The overall framework includes LingR and LingR-CoTs construction and two-phase fine-tuning

Methodology

Algorithm 1 Instruction construction

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

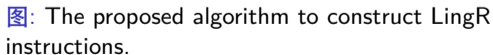
- 1: **Initialize** LingR = [] # Instruction Set
- 2: **for** x, y in \mathcal{D} **do**
- 3: item = {} # instruct set
- 4: **for** q in Q **do**
- 5: response \leftarrow request (ChatGPT; x, y, q)
- 6: item['cot'] \pm response['answer']
- 7: **if** response['decision'] == True **then**
- 8: LingR \pm [item]
- 9: **end if**
- 10: **end for**
- 11: **end for**
- 12: # the request() function
- 13: request (ChatGPT; x, y, q) = {
- 14: model=gpt-4;
- 15: messages=give answer for q on x ;
- 16: can y be made from answer of q }

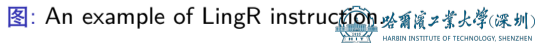

Instruction:

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.

 **图:** The proposed algorithm to construct LingR instructions.

 **图:** An example of LingR instruction  哈尔滨工业大学(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

Linguistic Rule Chain-of-Thoughts (LingR-CoTs)

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, this paper designed a task-specific linguistic rule-based chain-of-thoughts (LingR-CoTs) for few-shot fine-tuning and in-context learning.

The association (\mathcal{A}) level questions distilled

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?

The intervention (\mathcal{I}) level questions distilled

The counterfactual (\mathcal{C}) level questions distilled

Linguistic Rule Chain-of-Thoughts (LingR-CoTs)

The association (\mathcal{A}) level questions distilled

The intervention (\mathcal{I}) level questions distilled

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.

The counterfactual (\mathcal{C}) level questions distilled

Linguistic Rule Chain-of-Thoughts (LingR-CoTs)

The association (\mathcal{A}) level questions distilled

The intervention (\mathcal{I}) level questions distilled

The counterfactual (\mathcal{C}) level questions distilled

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, ...

Content

- 1 Abstract & Introduction
- 2 Robustness Metrics

- 3 Method
- 4 Experiment Analysis
- 5 Conclusion

Experiments-Main Results

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)	<u>40.5</u>	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	<u>20.2</u>
LingR-OPT (13B)	40.5	<u>26.9</u>	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	<u>53.6</u>	76.8	20.6
LingR-LLaMA-2 (13B)	<u>40.5</u>	30.3	<u>42.9</u>	30.6	29.6	55.5	<u>95.0</u>	0.9

Experiments-Ablation Study

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	<u>91.2</u>	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	<u>69.2</u>
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)	<u>86.1</u>	<u>91.5</u>	91.5	71.2
LingR-LLaMA2 (1000-shot)	<u>82.7</u>	<u>81.5</u>	83.0	86.5

Experiments-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.

Contributions

- Analysis of how linguistic rules contribute to LLMs' robustness.
- The novel approach of using linguistic rules to enhance robustness.
- Impact on the development of more resilient AI systems, and potential areas for further research and application in real-world scenarios..

Content

- 1 Abstract & Introduction
- 2 Robustness Metrics

- 3 Method
- 4 Experiment Analysis
- 5 Conclusion

Conclusion

Summary of key findings and their significance.

Reflection on the journey from hypothesis to proven results.

Encouragement for continued exploration and application of LingR methods.

Questions & Answers

Thank you for your attention!
Any questions or comments?