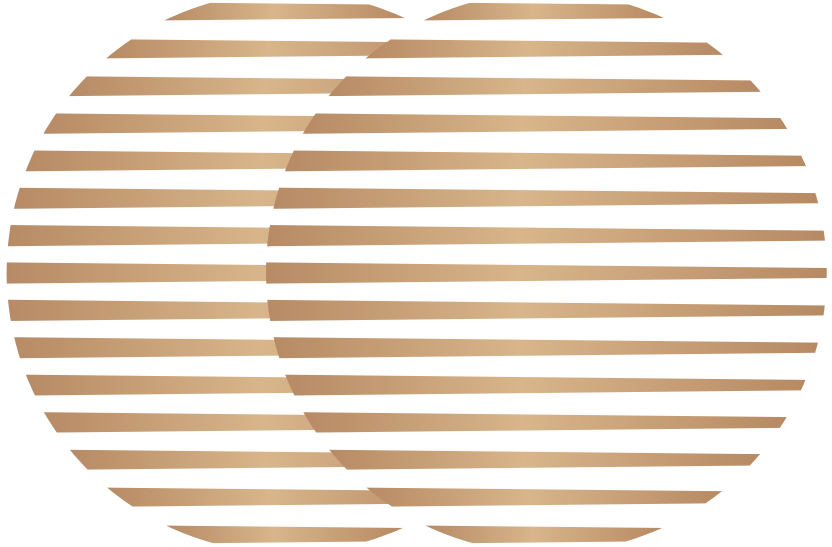
2024년 2월 20일 Study Meeting

LLM과 지식 그래프를 사용한 논리 오류 감지

Logical Fallacy Detection with LLMs and KGs

정지원
성균관대학교 인공지능학과
석사과정
jwjw9603@g.skku.edu



성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

Contents

01 Progress

- 진행 내용 Overview
- 연구 motivation
- Logical fallacy type 정의
- Related work
- Paper Follow-Up
- Goal
- My Method
- References

진행 내용 Overview

- 본 연구(주제)의 동기, 목적, 목표가 뭔가?
  - 연구의 배경, 목적
  - 동기를 충족시키기 위해 어떻게 접근해야 할까?
- 논문 요약 / 정리
  - 논문들의 소개보단 흐름 파악
- My Model Methodology
  - 어떻게 모델을 구성하며, 모델의 part별 정당성, 타당성 확보 필요

Motivation

1. 논리 오류를 찾아내고 파악하는 것은 논쟁, 대화에서의 퀄리티와 정당성을 부여하는데 중요하다.
2. 최근 LLM의 발달로 자연어 처리에서의 엄청난 발전을 이뤄내고 있다. 특히, 복잡한 추론 과정에서도 좋은 성능을 보여주고 있다.
    a. 또한 LLM의 부족한 구조적 정보를 보충하고자 지식 그래프를 함께 사용하는 경우도 많다.
3. 하지만, LM은 논리 오류를 인식하는데 다양한 genre, domain, fallacy type, dataset에 따라 명확한 한계점을 보이고 있다.
    a. 기존 연구들은, 논리 오류 데이터셋을 직접 만들고, 주로LLM이 아닌 LM을 대상으로 평가를 진행하지 않고 있다.
        1) 데이터셋의 레이블링에 대한 편향이 있다(e.g. multi-label, wrong label).
    b. 우리는 LLM(ChatGPT-3.5)로 논리 오류 감지 작업을 다양한 데이터셋에 진행했는데 성능이 매우 떨어지는 것을 보이고 있으며 클래스 불균형을 확인할 수 있다. -> LLM의 한계
    c. 특히, 다양한 논리 오류 중 일상 대화에서 쉽게 발생하는 오류들의 개수가 다른 클래스 대비 많으며 잘 구분하지 못하고있다(e.g. hasty generalization, false causality, cherry picking).
    d. 이러한 논리 오류들의 공통점으로는 문장, 대화 내 전제와 결론 간의 잘못된 정보, 연결 관계를 보여주고 있다. -> 지식 그래프를 사용하자
4. 이러한 한계점은 LLM의 실용성, 적용 가능성에 큰 타격을 줄 수 있다.
5. 본 연구는 현존하는 논리 오류 데이터셋 중 hasty generalization, false causality, cherry picking과 같은 논리 오류를 대상으로 다양한 LLM과 지식 그래프를 활용하여 논리 오류 추론 능력을 발전시키고자 한다.

Motivation

6. LLM과 지식 그래프를 함께 사용해서 간단하면서도 강력한 추론 능력을 보이고(zero-shot), LLM을 fine-tuning시켜서 논리 오류 감지/추론의 generalizable을 보이고자 한다.
7. 새로운 데이터셋(hasty generalization that easily occur in real conversations)으로 fine-tuning 시킨 LLM을 평가하면서 실용성, 적용가능성 까지 검증하고자 한다.

Logical Fallacy type 정의

1. Hasty generalization
   - ✓ 지식 그래프는 넓은 범위의 상식과 일반 지식을 포괄하며, 특정 주장이나 개념에 대해 다양한 사례와 상황을 제공할 수 있음. 이를 통해 사용자는 제한된 데이터나 사례에 근거한 일반화의 타당성을 평가하고, 더 넓은 맥락에서의 일반화가 적절한지 여부를 판단할 수 있음
2. False Causality
   - ✓ 지식 그래프는 개념 사이의 다양한 유형의 관계를 명시적으로 모델링 함. 이를 통해 사용자는 두 사건이나 현상 사이의 직접적인 인과관계가 있는지, 그저 상관관계에 불과한지를 분석할 수 있음.
3. Cherry Picking
   - ✓ 지식 그래프는 주제에 대한 다양한 관점과 정보를 제공할 수 있음. 이를 통해 사용자는 특정 데이터나 사실을 선택적으로 사용하는 대신, 주장이나 결론에 대해 더 폭넓고 균형 잡힌 시각을 갖출 수 있음.
4. Irrelevant Authority
   - ✓ 지식 그래프는 개념과 그 관계를 모델링하지만, 특정 권위자의 의견이 특정 문제에 대해 관련성이 있는지 판단하는 데는 직접적인 정보를 제공하지 않을 수 있음. (하지만, 저번 미팅에서 봤던 예시는 도움이 될 수 있음)
5. Post Hoc
   - ✓ 지식 그래프는 사건, 현상, 개념 간의 관계를 모델링 할 수 있지만, 시간적 순서와 인과관계를 직접적으로 구분하는 것은 더 복잡한 추론을 요구함. 특히, 사건 간의 인과관계를 정립하기 위해서는 단순한 시간적 순서를 넘어서는 깊은 분석과 맥락적 이해가 필요함.

정리하자면,,,

1. Considering the Motivation :

   ① 데이터셋의 relabeling(뺄 수도 있음) or 실용성 평가용 데이터셋 만들기

      a. 데이터셋을 만든다는 것은 실제 대화, 실제로 흔히 발생하는 대화 형태의 데이터를 넣어서 fine-tuned 모델 평가하기 -> 사실상 최종 결론에 해당할 듯

   ② 논리 오류 중 텍스트 내 관계에서 발생하는 문제에 대한 해결 필요성

      a. LLM(ChatGPT)에서의 낮은 성능, 4개의 데이터셋에서 가장 많은 비중을 차지, 사람들이 흔히 실수하는 논리

   ③ 2번 내용을 해결하기 위해 지식 그래프를 어떻게 사용하는지(LLMs with KGs) ★

   ④ 3번 내용을 통해 zero-shot model(gpt-3.5)은 inference, fine-tuned model(LLaMA)은 1번의 데이터로 실용성 평가

2. 3번이 해결되면 전체적으로 내용을 만들 수 있다.

Related work

1. 본 연구와 관련이 있을 수 있는 연구들의 주제는 크게 4가지 있다.
    1) Logical Fallacy
    2) LLM Evaluation
    3) LLM with KGs
    4) Prompting engineering
2. 기존 미팅에서는 Logical fallacy, LLM Evaluation, LLM with Kgs에 관련된 논문 소개가 많았다.
3. 지금까지 소개한 논문들의 특징과 흐름, prompting engineering 논문의 특징까지 짚고 넘어가자.
    1) Prompting engineering 에 관한 논문들은 요약본으로 정리

Paper Follow – UP(Logical Fallacy)

ACL 2021

2021.08

Breaking Down the invisible wall of informal fallacies in online discussions

- Reddit 소셜미디어 데이터셋

2022.02

EMNLP 2021

Logical Fallacy Detection

- 13가지의 타입을 가지고 있는 논리 오류 데이터셋(LOGIC)
- 논리 오류의 구조적 형태를 학습하고자 함

Multitask Instruction-based Prompting for Fallacy Recognition

EMNLP 2022

Data	Ex	F	Genre	Domain
ARGOTARIO	880	5	Dialogue	General
PROPAGANDA	5.1k	15†	News	Politics
LOGIC	4.5k	13	Diverse	Education
COVID-19	621	9‡	SocMed/News	Covid-19
CLIMATE	477	9‡	News	Climate

2022.12

IJCAL 2023

2023.01

Case-based reasoning with Language Models for classification of Logical fallacies

- Case base reasoning기반으로 LOGIC 데이터셋 학습 및 평가
- 논리 오류를 발생시키는 문장에 counterarguments, goals, explanations, structure와 같은 추가적인 정보를 주고 CBR알고리즘 + attention 매커니즘 적용

2023.08

EMNLP 2023

Argument-based Detection and Classification of Fallacies in Political Debates

- 정치 연설문 데이터셋 생성
- Detection을 진행하기 위해 fallacy가 발생하는 시작, 끝 부분을 찾기 위해 Pos-tagging 방법 사용

2023.11

Preprint 2023

A closer Look at the Self-Verification abilities of Large Language Models in Logical Reasoning Paper

- LLM의 logical reasoning을 제대로 이해하는지 파악하기 위해 최신 LLM에 대해 성능 평가를 진행하였고, FALLACIES라는 데이터셋 생성
- Formal : 24개 타입, Informal : 208개 타입

2023.11

Preprint 2023

Large Language Models are Few-Shot Training Example Generators : A case Study in Fallacy Recognition

- Multitask instruction-based prompting for fallacy recognition 논문의 저자가 낸 논문
- 클래스별 불균형을 줄이기 위해 prompting을 통해 data augmentation진행

2023.12

Detecting argumentative fallacies in the Wild : Problems and Limitations of Large Language Models

- 기존의 논리 오류 관련 논문들은 각자 데이터셋을 만들고 LM을 평가하는 목적으로 나옴.
- 이러한 점은 일반적으로 실제 상황에서 적용 가능성과 유용성을 이해하는데 많은 한계점을 보임.
- 그래서, natural language argumentation schemes로 이루어진 검증 코퍼스를 만듦

데이터셋 출처와 LLM을 대상으로 fallacy dataset 평가

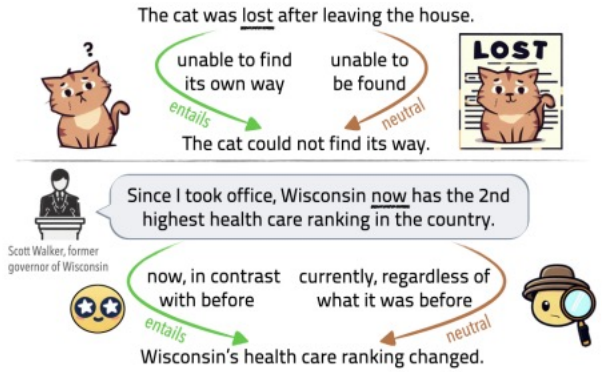
Paper Follow – UP(LLM Evaluation)

EMNLP 2023

2023.04

We' re Afraid Langauge Models aren't Modeling Ambiguity

- 중의성(애매모호함)을 가지는 텍스트들을 모아 데이터셋을 만듦
- 전제와 가설, 두 애매모호한 문장을 보여주고 이 문장 간의 관계를 확인한다.



2023.08

EMNLP 2023

This is not a Dataset : A Large Negation Benchmark to Challenge Large Language Models

- 부정어(negation)을 가지는 텍스트들을 모아 데이터셋을 만듦.
- 데이터는 wornet에서 11가지의 relation을 정하고, 그 relation이 들어가는 triple을 추출함.
- 추출한 triple을 기반으로 template(prompt)를 만들고 template에 맞춰 데이터를 생성함.
  - Triple : ⟨part, bill, bird⟩
  - Template : ⟨noun1+(e)s⟩[are commonly | may be] part of ⟨noun2 +(e)s⟩.
- 만들어진 데이터셋은 두 명의 native speakers들이 데이터셋으로부터 220개의 문장을 랜덤 샘플링을 진행해서 평가함.

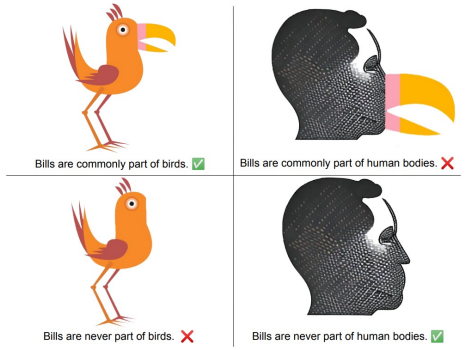


Figure 1: Affirmative and negative sentences in the dataset.

Paper Follow – UP(LLM with KGs)

Unifying Large Language Models and Knowledge Graphs : A Roadmap

- LLM과 KGs의 통합을 위한 전망적인 로드맵 제시

MindMap : Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models

- KG를 사용해서LLM을 최신 지식과 연결하고 LLMs의 추론 경로를 유도하기 위한 방법을 탐구함
- Evidence graph mining -> Evidence graph aggregation -> LLM reasoning on the mindmap
- Entity linking 방법은 얻어갈만한 정보

Knowledge-Driven CoT : Exploring faithful reasoning in LLMs for knowledge-intensive Question Answering

- CoT Collection을 미리 만들고 이를 기반으로 Retrieve-reader-verifier 모듈을 거침

ICLR 2024 — 2023.06

IEEE 2023 — 2023.06

EMNLP 2023 — 2023.08

Preprint 2023 — 2023.08

Preprint 2023 — 2023.08

Preprint 2023 — 2023.08

Boosting Language Models Reasoning with Chain-of Knowledge Prompting

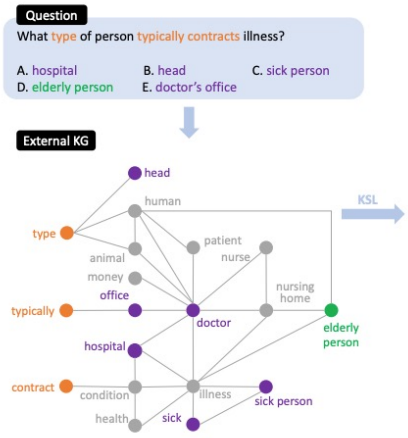
- Evidence triple, explanation hints 사용
- F2-verificiation

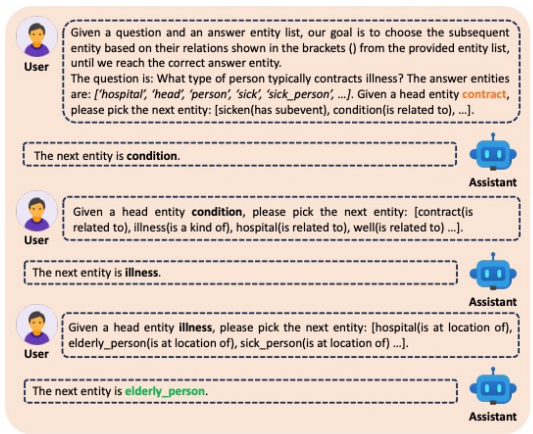
Reasoning on Graphs : Faithful and interpretable large language model reasoning

- Plan-and-solve를 차용한 planning-retrieval-reasoning framework 제시
- ELBO를 사용하여 수식적으로 설명

Knowledge Solver : Teaching LLMs to search for domain Knowledge from Knowledge graphs

-> LLM이 entity를 Step by Step방법으로 생각하도록 구성함





아이디어 및 방법론

Paper Follow – UP(Prompting Engineering)

Post Hoc Explanations of Language Models Can Improve Language Models

- Post-hoc explanation을 활용하여 각 입력 특성이 모델 예측에 미치는 영향을 잡아내는 점수를 출력함. 이를 통해 수정 신호를 보냄

Re-Reading Improves Reasoning in Language Models

$$y \sim \sum_{z \sim P(z|\,c^{(\mathrm{cot})}(t,\mathrm{re2}(x));\theta^{(\mathrm{llm})})} P(y|\,c^{(\mathrm{cot})}(t,\mathrm{re2}(x),z);\theta^{(\mathrm{llm})}) \cdot P(z|\,c^{(\mathrm{cot})}(t,\mathrm{re2}(x));\theta^{(\mathrm{llm})}).$$

ICLR 2024 — 2022.05

2023.05 — NeurIPS 2023

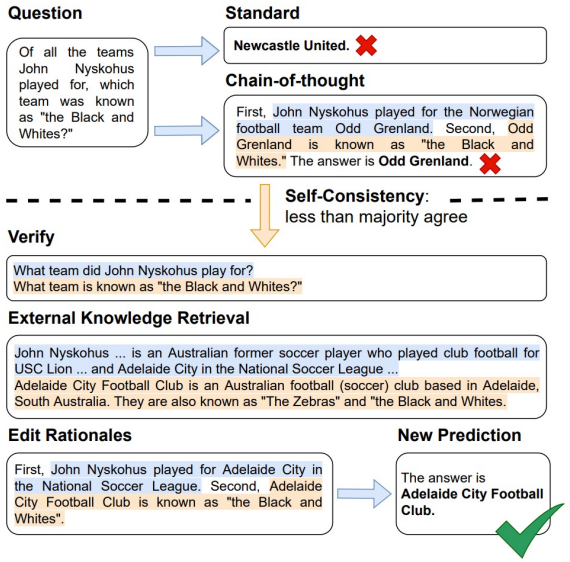
ACL 2023 — 2023.05

2023.09 — ICLR 2024

Selection-Inference : Exploting Large Language Models for Interpretable Logical Reasoning

- Few-shot example로부터 선택(selection)을하고, 선택된 정보를 기반으로 추론(inference)를 하는 prompt

Verify-and-Edit : A knowledge-Enhanced Chain-of Thought Framework




```
Question: {Input Query}
Read the question again: {Input Query}
#Thought-eliciting prompt (e.g.,“Let’s think step by step”)#
```

-> 지식 그래프가 아닌 Wikipedia, google 사용

아이디어 및 방법론

Goal

- 논리 오류(Logical Fallacy)가 발생하는 텍스트에서 지식 그래프를 활용해서 텍스트의 논리 오류를 더 잘 인식하고 분류하는 것을 목적으로 두고 있다.
- 여러가지 논리 오류가 있지만 **지식 그래프**가 활용되기 위해서는 **텍스트 내에 연결관계**에서의 정보 오류로 인한 문제가 있는 논리 오류에 집중을 하자.
- 지식 그래프를 활용하면서 LLM이 **step-by-step**으로 reasoning하도록 하고싶다.

My Method

1. Original Text로부터 premise, conclusion으로 분할한다. -> 논리 오류를 일으키는 문장의 구성 파악
    1) Prompting 사용
2. Premise, conclusion 각각 keyword를 추출한다. -> For using KG and 문장 내 연결 관계 파악 사전 단계
    1. Prompting LLMs to extract the key entities from the question query Q via in-context learning
    2. 1단계에서 생성되는 엔티티들을 Mset
    3. Mset이 실제 그래프에 존재하지 않을수도 있으니 entity linking을 수행함
    4. Entity linking은 KG의 모든 엔티티들과 Mset의 모든 엔티티들을 Bert Encoder를 사용해 임베딩 H_G, H_M을 만듦
    5. Cosine Similarity를 비교해서 Mset에 있는 각 엔티티들을 KG의 가장 가까운 이웃 엔티티에 링크해서 최종적인 keyword entity 생성
3. 각 텍스트 별 키워드들을 head entity로 지정하고 지식 그래프(conceptnet)로부터 linking을 진행해서 relation path를 생성한다.
    1) Relation path는 relation만 포함한다.(e.g. head entity -> relation -> tail entity 중 relation만 가져온다는 뜻임)
    2) Relation path를 선택한 이유 : relation은 특정한 지식 영역에서의 근본적인 관계를 나타내기 때문에 상대적으로 더 안정적일 수 있으며, 관계 기반의 문제에 더 적합하다 판단했다.
4. 생성된 relation path를 지식 그래프와 매칭해서 실제 reasoning path를 생성한다. -> 문장 내에 연결관계 파악 및 Step-by-Step reasoning
    1) 3번에서 4번과정이 일종의 **검증 과정**이라 볼 수 있다. 왜냐하면 3번에서의 relation path가 4번에서 지식 그래프에 없으면 제외되기 때문
5. Premise, Conclusion로부터 생성된 reasoning path를 aggregate한다. -> Selection & Inference & Re-question
6. Considering the reasoning path of both premise and conclusion, please answer me what kind of logical fallacy in the original text

My Method

7. Zero-shot & Fine-tuning
    1) Zero-shot : 방법론은 정답을 틀릴 경우에 진행함
    2) Fine-tuning : 다양한 데이터셋을 통합해서 학습하고 new dataset으로 실용성, 적용가능성 평가
8. New dataset(applicability & practicality)

Reference

1. SAHAI, Saumya; BALALAU, Oana; HORINCAR, Roxana. Breaking down the invisible wall of informal fallacies in online discussions. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021. p. 644-657.
2. ALHINDI, Tariq, et al. Multitask Instruction-based Prompting for Fallacy Recognition. *arXiv preprint arXiv:2301.09992*, 2023.
3. JIN, Zhijing, et al. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*, 2022.
4. GOFFREDO, Pierpaolo, et al. Argument-based Detection and Classification of Fallacies in Political Debates. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. p. 11101–11112.
5. SOURATI, Zhivar, et al. Case-based reasoning with language models for classification of logical fallacies. *arXiv preprint arXiv:2301.11879*, 2023.
6. HONG, Ruixin, et al. A Closer Look at the Self-Verification Abilities of Large Language Models in Logical Reasoning. *arXiv preprint arXiv:2311.07954*, 2023.
7. PAN, Shirui, et al. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
8. WANG, Jianing, et al. Boosting Language Models Reasoning with Chain-of-Knowledge Prompting. *arXiv preprint arXiv:2306.06427*, 2023.

Reference

9. ALHINDI, Tariq; MURESAN, Smaranda; NAKOV, Preslav. Large Language Models are Few-Shot Training Example Generators: A Case Study in Fallacy Recognition. *arXiv preprint arXiv:2311.09552*, 2023.

10. LIU, Alisa, et al. We're Afraid Language Models Aren't Modeling Ambiguity. *arXiv preprint arXiv:2304.14399*, 2023.

11. GARCÍA-FERRERO, Iker, et al. This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models. *arXiv preprint arXiv:2310.15941*, 2023.

12. LUO, Linhao, et al. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*, 2023.

13. WEN, Yilin; WANG, Zifeng; SUN, Jimeng. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

14. FENG, Chao; ZHANG, Xinyu; FEI, Zichu. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*, 2023.

15. WANG, Keheng, et al. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023.

Reference

16. CRESWELL, Antonia; SHANAHAN, Murray; HIGGINS, Irina. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

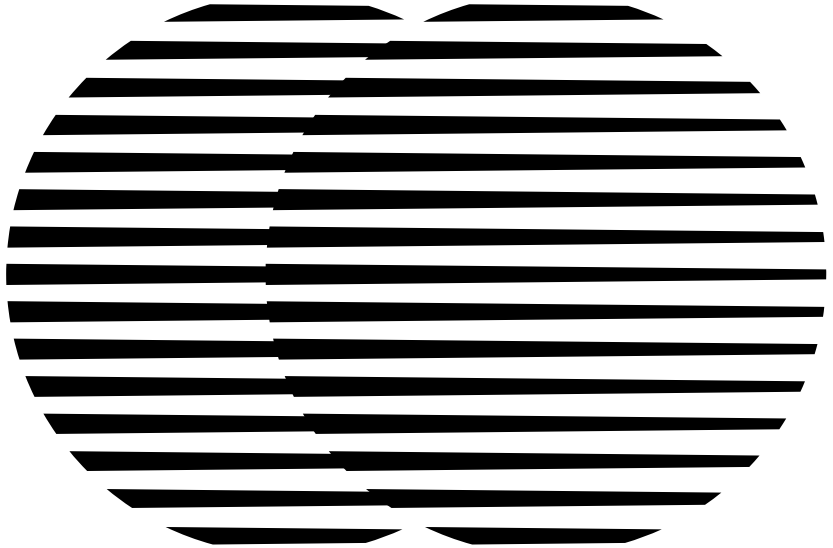
17. KRISHNA, Satyapriya, et al. Post hoc explanations of language models can improve language models. *Advances in Neural Information Processing Systems*, 2024, 36.

18. ZHAO, Ruochen, et al. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*, 2023.

19. XU, Xiaohan, et al. Re-reading improves reasoning in language models. *arXiv preprint arXiv:2309.06275*, 2023.

감사합니다

발표 경청해 주셔서 감사합니다



정지원 성균관대학교 인공지능학과 석사 과정
jwjw9603@g.skku.edu

성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)