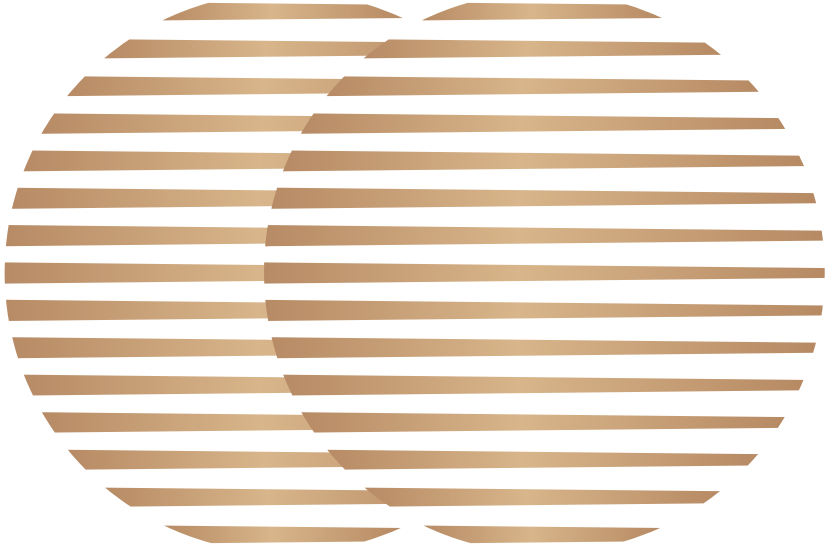
2024년 3월 11일 Study Meeting

논리 오류를 다시 물어보자

Let's ask about the logical error again

정지원
성균관대학교 인공지능학과
석사과정
jwjw9603@g.skku.edu



성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

Contents

01 Progress

- 진행 내용 Overview
- Review - Faulty Generalization
- False Causality
- Irrelevant Authority
- General Question
- Specific Question
- Experiment
- Analysis
- ToDo

진행 내용 Overview

- 지식 그래프가 사용되기에 적합한 fallacy
    1) Faulty Generalization
    2) False Causality
    3) Irrelevant Authority
- General Question vs Specific Question
- 실험

지식 그래프가 사용되면 좋을 것 같은 fallacy

1. 일전에 지식 그래프가 사용되기에 적합한 fallacy들을 언급했었다.
    1) Faulty Generalization, False Causality, Irrelevant Authority, Post Hoc, Cherry Picking
    2) Post Hoc는 False Causality의 부분집합, **Post Hoc ⊂ False Causality**
    3) Cherry Picking은 오직 두 개의 데이터셋에만 존재(COVID-19, CLIMATE)
    4) 따라서, 세 개의 클래스(Faulty Generalization, False Causality, Irrelevant Authority)에 집중함
    5) **Faulty Generalization, False Causality, Irrelevant Authority 은 4개의 데이터셋(argotario, LOGIC, COVID-19, CLIMATE)에서 총 1/3을 차지함(2059개)**
2. Faulty Generalization에 대해서 알아보았으니, False Causality와 Irrelevant Authority에 대해 알아보고 규칙을 찾아보자.
3. 규칙을 기반으로 세 개의 클래스에 적용될 수 있는 General한 질문을 만들자.
4. Classwise Question도 만들자.
5. 3번과 4번의 결과를 비교해보자.

Review-Faulty Generalization

1. "Annie must like Starbucks because all white girls like Starbucks."
- 한국어 : "모든 백인 소녀들이 스타벅스를 좋아하기 때문에 애니도 스타벅스를 좋아할 것입니다."
- A : Annie must like starbucks
- Q : All white girls like starbucks
- A(애니) < Q(백인소녀)

2. It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
- 한국어 : 올해 라스베거스는 작년에 비해 더 따뜻합니다. 그러므로 지구 온난화는 급속히 가속화되고 있습니다.
- A : global warming is rapidly accelerating.
- Q : It is warmer this year in Las Vegas as compared to last year.
- A > Q
- Question : Does the fact that Las Vegas is warmer this year compared to last year necessarily imply that global warming is rapidly accelerating?
- Question2 : Is global warming rapidly decreasing?

3. "The two courses I took at UF were not very interesting. I don't think its a good university."
- 한국어 : "내가 University of Florida의에서 수강한 두 과목은 그다지 흥미롭지 않았습니다. 좋은 대학이 아니라고 생각합니다."
- A : I don't think its a good university.
- Q : The two courses I took at UF were not very interesting.
- A > Q
- **Question**: Did the lack of interest in the two courses you took at UF lead you to believe it's not a good university?
- **Question2**: What was your overall experience like at UF besides these two courses? or Isn't the University of Florida a good university?

- Q : 샘플 데이터에서 특정 특성이나 속성을 가진 개체 또는 사례, 예시를 의미
- A : 주어진 문장, 문맥에서 관심있는 특징이나 주제를 의미
- **A > Q문장은** 경험한, 주위 사람에게 들은, 본 내용(Q)을 기반으로 일반화적인 주장(A)을 하는 경우
- **A < Q문장은** 예외를 무시하고 대중적인 주장, 관념(Q)을 기반으로 소수의 경우, 주장(A)을 일반화 시키는 경우

Review-Faulty Generalization

1. **A < Q 문장은** 단순히 Q가 사실인지를 되묻는 형태의 질문이면 된다. 왜냐하면 이 문장은 소수(한 명, 친구, 가족) 에서 주장하는 내용(A)을 대중적인 관념(Q)을 기반으로 사용하기 때문에, 근본적으로 Q가 사실인지를 직접적으로 물으면 논리 오류를 해결할 수 있음.

2. **A > Q 문장은** 경험한, 주위 사람에게 들은, 본 내용(Q)을 기반으로 일반화 적인 주장(A)를 하는 경우로, 이런 문장 같은 경우에는 A와 Q의 관계를 묻거나(Question), A를 되묻는 형태의 질문(Question2)을 하면 된다.

    1) Question : A와 Q의 관계를 묻는 형태로 일반화된 주장에 대한 논리적 결함을 직접적으로 다룸

        a. 주장의 근거와 일반화된 결론 간의 관계를 더 명확하게 이해할 수 있지만, 지식 그래프가 사용되기에 쉬운 질문 형태가 아님

        b. 질문을 만들 때 고유명사(사람 이름, 회사, 앱 등)는 지식 그래프에 없을 경우가 있음

    2) Question2 : A를 되묻는 경우로 주장된 결과를 의심하거나 부정할 수 있도록 유도하는 형태

        a. 일반화적인 주장을 직접적으로 묻는 형태로서, A에 대한 답을 추론하는 과정에 적절함

        b. 주장(A)으로만 구성되어 있는 문장들에게 적합함.(e.g. All four year olds talk too much.)

3. **A > Q 문장은** A를 되묻는 형태만으로 충분할 경우에는 Question2, 아닐 경우에는 Question으로 다루면 어떨까?

False Causality

1. 데이터셋마다 False Causality를 어떻게 명칭하고 정의하는지 알아보자.

2. LOGIC
   - 명칭 : False Causality
   - Def : A statement that jumps to a conclusion implying a causal relationship **without supporting evidence**
   - This fallacy occurs when an argument assumes that **since two events are correlated, they must also have a cause and effect relationship.** -〉 상관관계가 있다고 원인과 결론(인과관계)라고 생각하는 오류, 상관관계 ⊃ 인과관계

3. COVID-19, CLIMATE
   - 명칭 : False Cause(Causal Oversimplification)
   - Def : X is identified as the cause of Y when another factor Z causes both X and Y OR X is considered the cause of Y when actually it is the opposite -〉 X가 Y의 원인으로 잘못 간주되는 상황
   - 원인과 결론의 관계에 대한 질문을 하면 되지 않을까?
   - **Post Hoc 도 Causal Oversimplification이라 함**
     - Def : 사건 A가 일어난 후 B가 발생하기 때문에 B가 A때문에 발생한다고 가정된다. 다시 말해, 단순한 상관관계가 있는데 원인-결론 관계로 해석되는 경우.
   - 상관관계인지, 인과관계인지를 확인하는 질문이면 되지 않을까?

4. Wikipedia에서 Causal Oversimplification def: Assuming **a single cause** or reason when there are actually multiple causes for an issue
   - Causal Oversimplification 같은 경우도 결국 원인과 결론의 관계를 확인하는 질문이면 된다.

False Causality

False Causality

- **Definition:** This fallacy occurs when an argument assumes that since two events are correlated, they must also have a cause and effect relationship.
- **Example:** We observed an increase in ice cream sales at the same time as air conditioner sales increased. Therefore, we can conclude that selling more ice cream causes more air conditioners to be sold.
- **Synonyms or Subtypes:** Post hoc ergo propter hoc, Cum hoc ergo propter hoc, Regression Fallacy, Consecutive Relation, Magical Thinking, Gambler's Fallacy (rarely called temporal flaw/temporal fallacy), Ludic Fallacy.

False Causality

5. 즉, 인과관계의 오류는 원인과 결과 사이의 관계가 잘못되었거나, 불충분한 경우에 발생하는 오류이다.
6. 잘못 되었거나 -> 가짜 인과관계,Post Hoc
7. 불충분 -> Oversimplification
8. 위 내용을 기반으로 False Causality의 format을 정리한다면 다음과 같다 : (https://en.wikipedia.org/wiki/Post_hoc_ergo_propter_hoc)
    - A occurred
    - B occurred
    - Therefore, A caused B
9. LOGIC 데이터셋에서 예시를 살펴보자

False Causality

1. text : "Every time I go to sleep the sun goes down. Therefore, my sleeping causes the sun to set."
- 한국어 : "나는 잠에 들 때마다 해가 진다. 그러므로 나의 잠이 해를 지게 한다."
- A : I go to sleep the sun goes down.
- B : Therefore, my sleeping causes the sun to set.
- 질문 : "나는 잠에 들 때마다 해가 진다는데, 이것이 진짜로 내 잠이 해를 지게 만드는 것인가요?"
- Question: Does the fact that the sun goes down every time you go to sleep necessarily mean that your sleeping causes the sun to set?

2. text : Since Anna Camacho became vice president of the parent-teacher association, student performance has declined and teacher morale is down. We on the school board believe that Camacho bears sole responsibility for the downtrend.
- 한국어 : 안나 카마초(Anna Camacho)가 학부모-교사 연합의 부회장이 된 이후로 학생의 성적은 떨어지고 교사의 사기도 떨어졌습니다. 우리 교육청에서는 Camacho가 하락세에 대한 전적인 책임을 지고 있다고 믿습니다.
- A : Anna Camacho became vice president of the parent-teacher association, student performance has declined and teacher morale is down.
- B : We on the school board believe that Camacho bears sole responsibility for the downtrend.
- 질문 : "안나 카마초가 학부모-교사 연합의 부회장이 된 이후로 학생 성적이 하락하고 교사의 사기가 낮아졌다는데, 이것이 정말로 카마초의 전적인 책임인가요?"
- Question: Just because Anna Camacho became vice president and student performance declined, does it mean that Camacho's appointment caused the decline in student performance and teacher morale?

3. text : Children who play violent video games act more violently than those who don't.
- 한국어 : 폭력적인 비디오 게임을 하는 어린이는 그렇지 않은 어린이보다 더 폭력적으로 행동합니다.
- A: 폭력적인 비디오 게임
- B: 폭력적인 행동
- 질문 : 폭력적인 비디오 게임을 하는 어린이들이 모두 폭력적으로 행동하나요?
- Question: Is it accurate to conclude that all children who play violent video games act violently?

4. text : Eighty-five percent of those surveyed said that it is important to have a cell phone.
- 한국어 : 설문조사에 참여한 사람들 중 85%는 휴대전화를 갖는 것이 중요하다고 답했습니다.
- A : 85%는 휴대전화를 갖는 것이 중요
- B : 휴대전화를 갖는 것이 중요
- 질문 : 설문조사에 참여한 사람들 중 85%가 휴대전화 소유가 중요하다고 답했습니다. 이 결과는 항상 모든 사람들의 의견을 반영할까요?
- Question: Does the fact that 85% of those surveyed believe that having a cell phone is important mean that it is always important to have a cell phone?

Irrelevant Authority

1. 데이터셋마다 Irrelevant Authority를 어떻게 명칭하고 정의하는지 알아보자.

2. LOGIC
   - 명칭 : Fallacy of Credibility
   - Def : An appeal is made to some form of ethics, authority, or credibility.
   - some form of ethics, authority, or credibility 을 기반으로 하는 주장이 맞는지를 확인하는 질문

3. COVID-19, CLIMATE
   - 명칭 : False Authority
   - Def : An appeal to authority is made where the it lacks credibility or knowledge in the discussed matter or the authority is attributed a tweaked statement.

4. Argotario
   - Def : 논쟁적 담화에서 권위를 사용하는 것은 본질적으로 오류가 아니지만, 토론 중인 주제와 관련이 없는 경우.

Irrelevant Authority

Fallacy of Credibility

- **Definition:** This fallacy is when an appeal is made to some form of ethics, authority, or credibility.
- **Example:** If mailing a hand-written letter was good enough in the past, then you don't need those pesky computers (appeal to tradition).
- **Synonyms or Subtypes:** Appeal to authority, Appeal to to nature, Naturalistic fallacy, Appeal to tradition, Chronological snobbery (reverse of tradition), Appeal to novelty, Ipse dixit, Etymological fallacy, Appeal to poverty, Appeal to accomplishment.

Irrelevant Authority

5. 즉, 무관한 권위로 인한 오류는 주어진 분야의 지식에 대한 정당한 권위가 없는 경우에 발생하는 질문이다.
6. 증거로 사용하는 권위자의 의견이 주제와 같은 분야인지를 확인하는 질문이면 된다.
7. 위 내용을 기반으로 format을 정리하면 다음과 같다 : (https://en.wikipedia.org/wiki/Argument_from_authority)
    - Person A claims that X is true.
    - Person A is an expert in the field concering X.
    - Therefore, X should be believed.
8. LOGIC 데이터셋 예시를 살펴보자

Irrelevant Authority

1. You ask your mother if you can go to the mall with your friends. She says "no". You ask why? She says, "because I'm the mom and I say so".
   - Question : Is being a mother directly relevant to the decision of whether going to the mall with friends is allowed?
   - 친구들과 함께 쇼핑몰에 가는 것을 허락하는 결정과 엄마가 되는 것이 직접적으로 관련이 있나요?
2. Famous actors like Alex Gonzaga and Philip Salvador support Oplan Tokhang, so it must be an effective operation.
   - Question : Do the endorsements of famous actors directly correlate with the effectiveness of Oplan Tokhang?
   - 유명 배우들의 지지가 오플란 토칭의 효과성과 직접적으로 관련되나요?
3. We should move to the Midwest because Mujtaba from the Wall Street Journal says the cost of living is cheaper there.
   - Question : Does Mujtaba's expertise in journalism from the Wall Street Journal necessarily qualify him to give accurate information about the cost of living in the Midwest?
   - 월스트리트 저널의 무즈타바가 미들웨스트의 생활비에 대한 정확한 정보를 제공할 자격이 있는가요?
4. “We should abolish the death penalty. Many respected people, such as Imran Kader, have publicly stated their opposition to it.”
   - Question : Are the opinions of respected individuals, like Imran Kader, directly related to the efficacy or morality of the death penalty?
   - 임란 카더와 같은 존경받는 인물들의 의견이 사형제도의 효능이나 도덕성과 직접적으로 관련이 있나요?
5. Since Dwayne Johnson (aka The Rock) endorsed this protein powder, it must work wonders.
   - Question : Does Dwayne Johnson's endorsement of a protein powder guarantee its effectiveness?
   - 드웨인 존슨이 추천한 단백질 파우더의 효과를 보장하나요?

General Question

1. Faulty Generalization, False Causality, Irrelevant Authority 세 개의 클래스의 공통점이 무엇인가?
   - 원인과 결과, 두 사건, 전제와 결론에서의 관계에 관한 문제들이다.
   - 즉, 관계를 묻는 질문을 만들자.
2. Create one question for each text that ask about the relationship between key events within the text rather that directly asking what a logical fallacy is.

Classwise Question

1. Faulty Generalization, False Causality, Irrelevant Authority 세 개의 클래스마다 적합한 prompt를 사용해서 질문을 만들자.
2. 이 방법은 추후에 further analysis, ablation study에 사용될 수 있음.

Experiment

1. LOGIC데이터셋에서 Faulty Generalization(441), False Causality(216), Irrelevant Authority (132)세 개의 클래스를 분류하는 multi-class classification task를 수행함
2. General question을 만들 때는 zero-shot으로 만들었으며, Classwise question은 zero,1, 2, 5-shot으로 만듦
3. 질문을 만들 때는 gpt-3.5-turbo-instruct 모델을 사용
4. 분류 작업은 zero-shot으로 진행함, gpt-3.5-turbo 모델을 사용

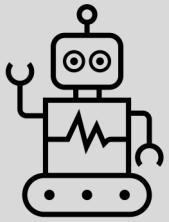
Prompt(General-Question)

User

I'll give you some texts. This text contains one of the following logical fallacy : {fallacy_type}.
Create one question for each text that ask about the relationship between key events rather that directly asking what a logical fallacy is.

Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
Question :

Question : `How does the fact that it is warmer in Las Vegas this year than last year support the claim that global warming is rapidly accelerating?`



Assistant

Prompt(Classwise Question– Faulty Generalization)

User

A faulty generalization often follows the following format:

The proportion Q of the sample has attribute A. Therefore, the proportion Q of the population has attribute A.

Extracting Q and A from each text allows us to determine the scope of the claims made by Q and A.
When A > Q, it refers to cases where a minority of people make generalized claims based on their experiences, hearsay, or observations. In such texts, it's appropriate to formulate questions that inquire about or challenge the claim A, or about the relationship between A and Q.
When A < Q, it refers to cases where questions are formulated to inquire about the content of Q.
I'll provide you with some texts, along with their Q and A, comparisons of Q and A, and the Q and A extracted from the texts.
These texts contain the logical fallacy of Faulty Generalization.
Considering the above, Please create one question for each text.

Here are some examples:
- text : "Annie must like Starbucks because all white girls like Starbucks." - Q : All white girls like starbucks. - A : Annie must like starbucks. - A < Q - question : Do all white girls like Starbucks?
- Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.

Question : `Does the temperature change in Las Vegas necessarily indicate a global trend?(zero-shot)`

Assistant

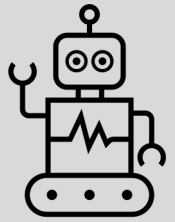
Prompt(Classwise Question– False Causality)

User

A false causality format often follows the following format:
A occurred
B occurred
Therefore, A caused B
I'll give you some Texts. These texts contain the logical fallacy of False Causality.
Create one question for each text that asks whether the relationship between A and B is correct.
It would be good to refer to the text format.

Here are some examples:
- text : "Every time I go to sleep the sun goes down. Therefore, my sleeping causes the sun to set."
- - question : Does the fact that the sun goes down every time you go to sleep necessarily mean that your sleeping causes the sun to set?
- `Text : Children who play violent video games act more violently than those who don't.`
- `Question :`

Question : `Does playing violent video games directly cause children to act more violently?`
`(zero-shot)`



Assistant

Prompt(Classwise Question– Irrelevant Authority)

User

A irrelevant authority format often follows the following format:

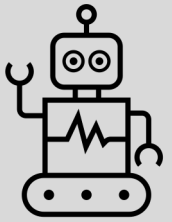
Person A claims that X is true.
Person A is an expert in the field concerning X.
Therefore, X should be believed

I'll give you some texts. This text contains the logical fallacy of irrelevant authority.
Create one question for each text to verify the credibility of Person A claiming X.
It would be good to refer to the text format.

Here are some examples :

- text : You ask your mother if you can go to the mall with your friends. She says "no". You ask why? She says, "because I'm the mom and I say so".
- Question : How does the fact that someone is a mother justify their decision to not allow their child to go to the mall with friends?
- Text : `Lebron James, one of the most decorated basketball players of all time, says you need to eat breakfast so you need to eat breakfast`

Question : `Can we trust Lebron James' claim that breakfast is necessary based solely on his expertise in basketball?(zero-shot)`



Assistant

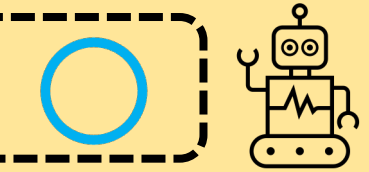
Prompt(Method-Question)

User

Your task is to detect a fallacy in the Text.
The label can be 'Faulty Generalization' and 'False Causality' and 'Irrelevant Authority'.
Please detect a fallacy in the text based on the Question.

Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
Question : Does the fact that Las Vegas is warmer this year compared to last year necessarily imply that global warming is rapidly accelerating?
Label :

Faulty Generalization



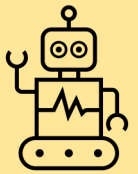
Assistant

Prompt(Method-No Question)

User

Your task is to detect a fallacy in the Text. The label can be “Faulty Generalization” and 'False Causality' and 'Irrelevant Authority'.
Text : It is warmer this year in Las Vegas as compared to last year; therefore, global warming is rapidly accelerating.
Label :

The Other Fallacy

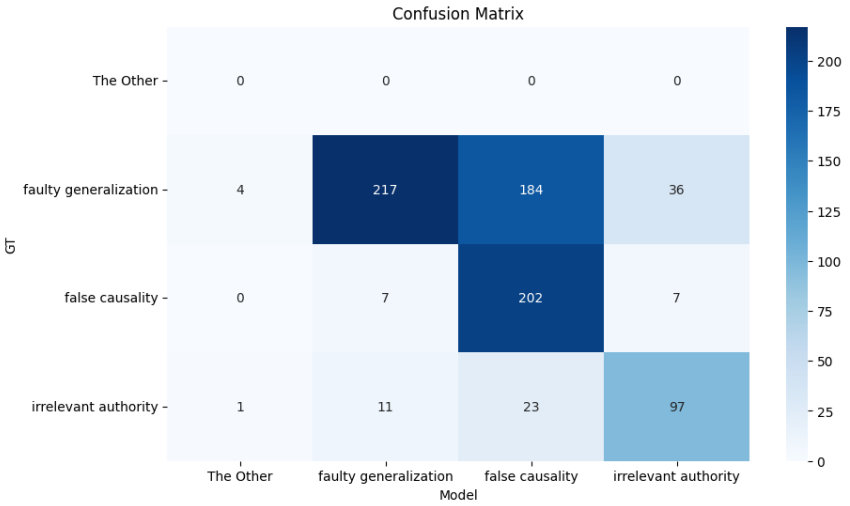


Assistant

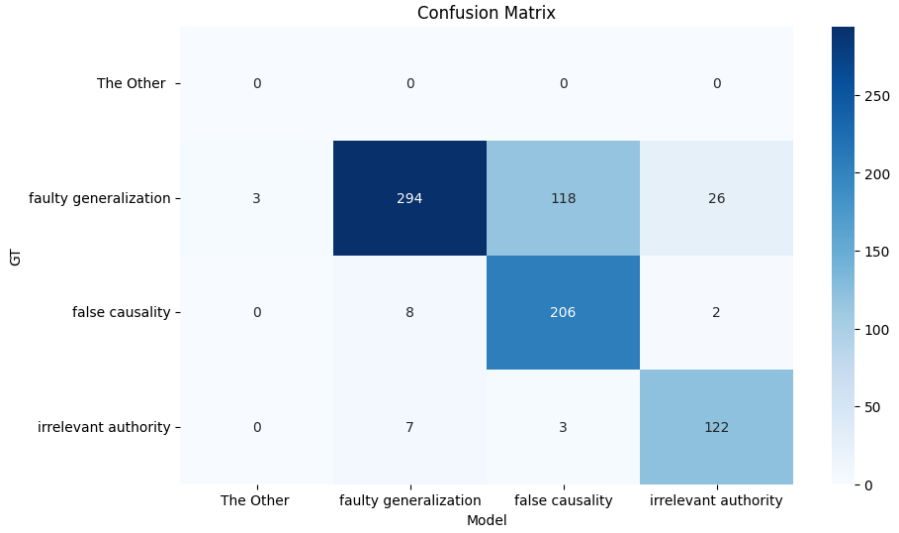
Result

Type/Metric	Total Accuracy	Precision	Recall	F1	FG acc	FC acc	IA acc
No Question	0.64	0.52	0.54	0.50	0.68(±0.01)	0.71(±0.01)	0.91
General Question	0.73(±0.01)	0.56(±0.01)	0.57(±0.01)	0.55(±0.01)	0.77(±0.01)	0.78(±0.01)	0.92(±0.01)
Classwise Question - zero	0.78(±0.01)	0.60	0.63	0.60	0.79(±0.01)	0.83(±0.01)	0.95
Classwise Question - one	0.78	0.59	0.62	0.59	0.79	0.83	0.94
Classwise Question - two	0.76(±0.01)	0.58	0.60	0.57(±0.01)	0.78(±0.01)	0.81(±0.01)	0.93(±0.01)
Classwise Question - five	0.72(±0.01)	0.56	0.58(±0.01)	0.55	0.74(±0.01)	0.78(±0.01)	0.93

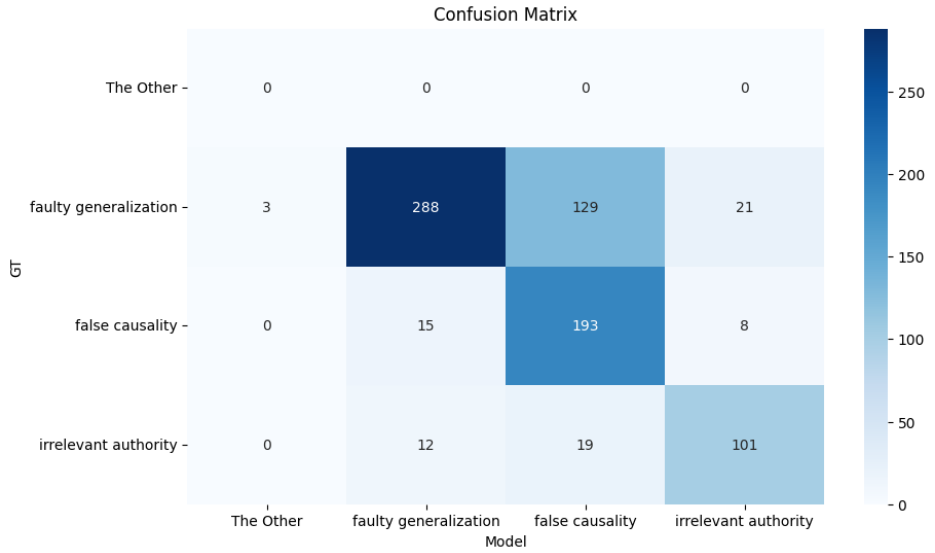
Result(Seed1)



Base



Classwise Question



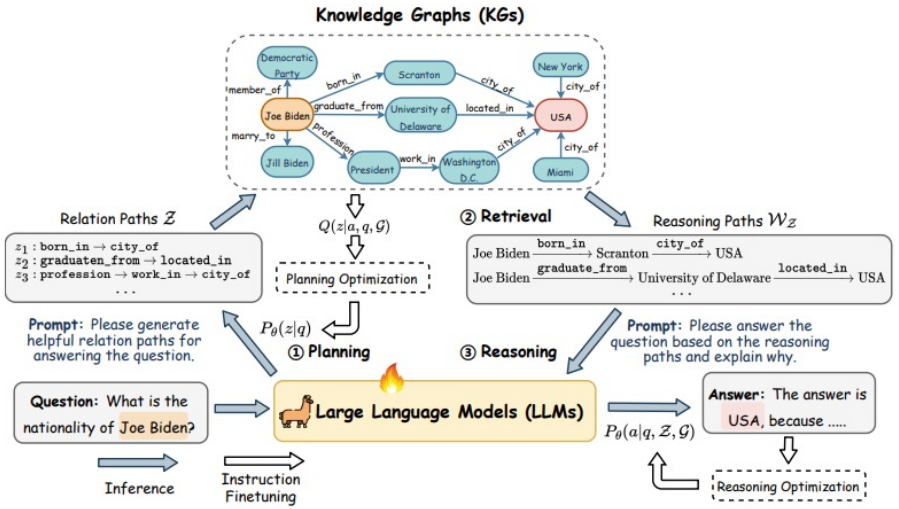
General Question

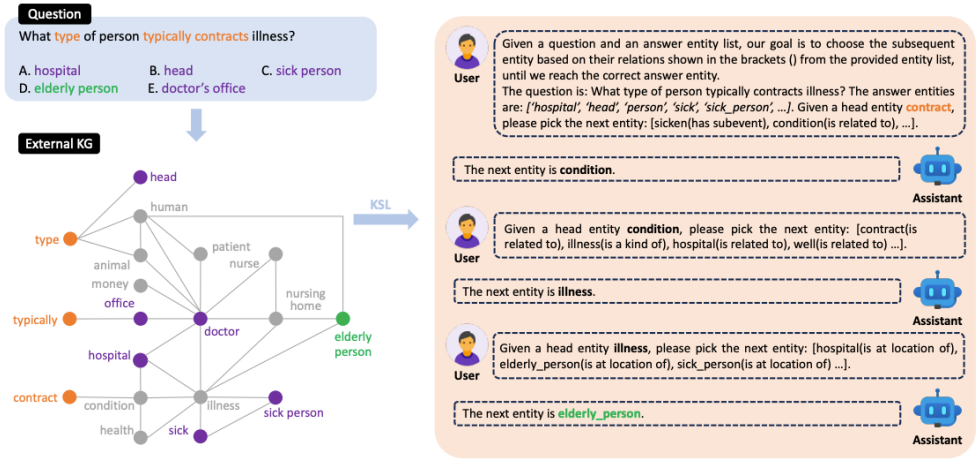
Analysis

1. LOGIC 데이터셋에 대해 진행했으며, 총 789개의 데이터셋으로 이루어진 3-class classification을 수행함
2. 질문 없이 testing 했을 때의 결과 대비, General question 결과가 9%정도의 향상이 있었으며, Class wise question은 최대 14% 성능 향상이 있었음
    1) Faulty Generalization 개수가 상대적으로 나머지 두 클래스 대비 개수가 많음
    2) No question도 False Causality를 비교적 잘 맞춤(recall=0.9)
    3) Irrelevant authority는 개수가 적어서 acc가 높음
    4) General Question은 False Causality에 큰 영향을 많이 못 줌
    - ➢ 나머지 데이터셋에서도 실험이 필요함
3. LLM(Chatgpt)는 상대적으로 False Causality는 잘 구분함, Faulty Generalization을 제일 구분 못함
4. Classwise question을 생성할 때, 예시를 안 주었을 때(zero-shot) 성능이 가장 좋았으며, 예시를 줄수록 성능이 떨어짐
5. 5-shot classwise question은 General question과 성능이 거의 같아짐

ToDo

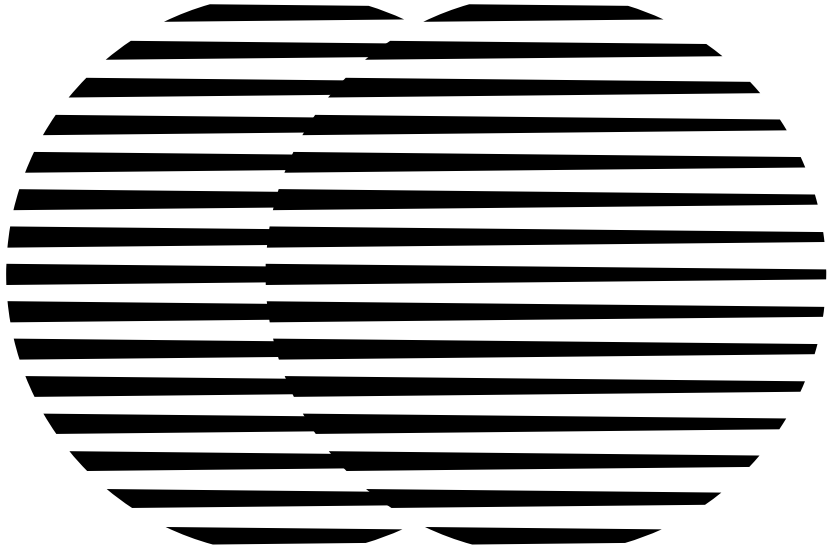
1. 다른 데이터셋에서도 실험, 4개의 데이터셋을 한 번에 실험 진행하기
    1) 결과를 다음 미팅 전에 공유
2. 지식 그래프 방법론 구체화 및 적용
    1) RoG 방법론 : 나오는 Reasoning Path로부터 selection&inference 사용하기
    2) Knowledge Solver : Subgraph를 생성하는 방법





감사합니다

발표 경청해 주셔서 감사합니다



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

성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)