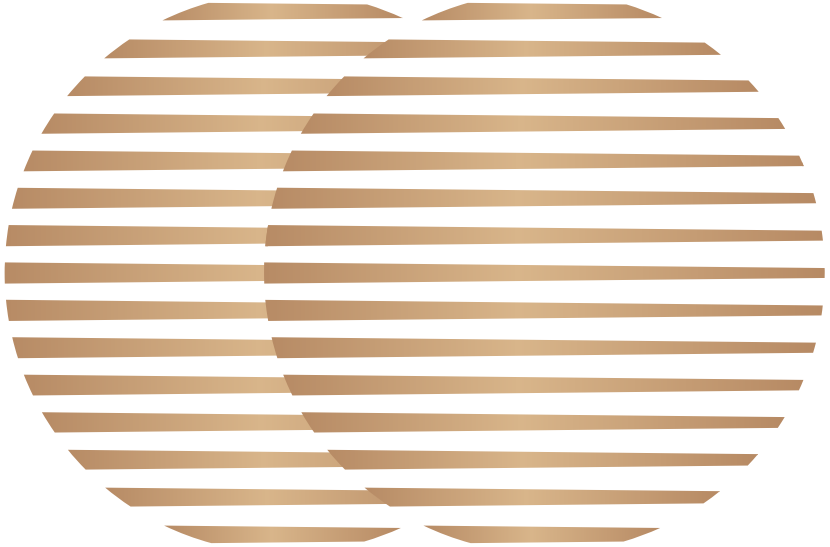
2024년 2월 2일 Study Meeting전 중간 결과 보고

ChatGPT 실험

Can GhatGPT distinguish Logical Fallacy?

정지원
성균관대학교 인공지능학과
석사과정
jwjw9603@g.skku.edu



성균관대학교
SUNG KYUN KWAN UNIVERSITY(SKKU)

Contents

01 Progress

- 진행 내용 Overview
- Experiment Setting and Method
- Prompt
- Data Distribution
- Result
- Result(Table)

진행 내용 Overview

- 실험 진행
  - Zero-shot, 2-shot
  - Multi class classification with no fallacy
- LM with KG
  - 논문 리뷰
  - How to??

실험 세팅 및 방법

1. 지식 그래프가 활용이 되면 도움이 될 것 같다고 판단한 클래스에 대해서 분류를 진행 함.
2. 총 4개의 벤치마크 데이터셋에 대해서 실험을 진행 함.
    a. Argotario : Hasty Generalization, Irrelevant Authority
    b. LOGIC : Faulty Generalization, False Causality, Fallacy of Credibility
    c. COVID-19 : Cherry picking, False Causality, Hasty Generalization, False Authority, Post Hoc
    d. CLIMATE : Cherry picking, False Causality, Hasty Generalization, False Authority, Post Hoc
3. Gpt-3.5-turbo(Chatgpt) 모델을 사용하고, 모델의 성능을 평가하기 위해 zero-shot, 2-shot learning으로 진행 함
4. Argotario, COVID-19, CLIMATE 데이터셋은 No Fallacy 클래스가 있지만 LOGIC 데이터는 없음
    a. LOGIC데이터는 총 13개의 클래스가 있으며 나머지 10개의 클래스를 No Fallacy 클래스로 지칭 함.
    b. No Fallacy 클래스를 추가했을 때의 prompt는 크게 바뀐게 없으며(class label만 추가), LOGIC데이터셋에서의 prompt만 바뀜(다음 페이지)

Prompting(LOGIC)

SYSTEM

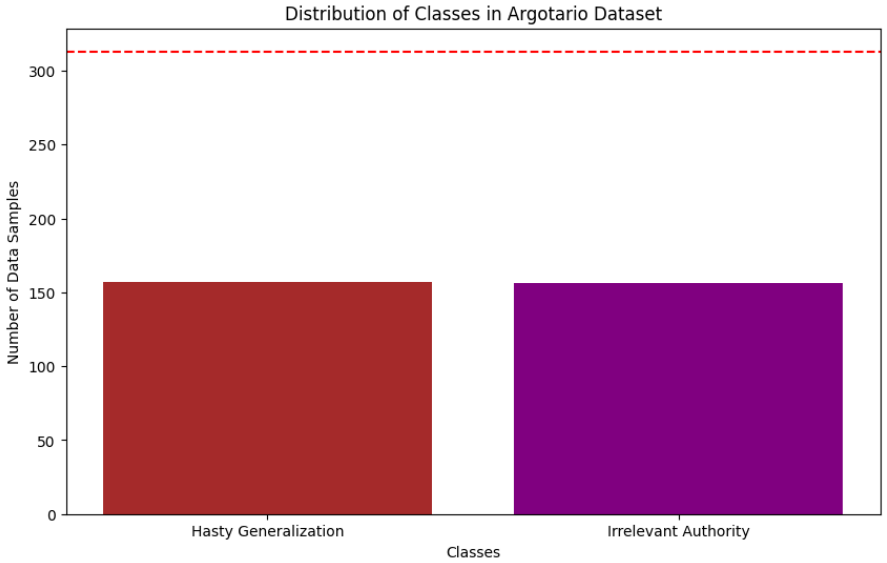
Your task is to detect a fallacy in the Text. The label can be 'Faulty Generalization', 'Fallacy of Credibility', 'False Causality' and 'No Fallacy'.
We refer to those with logical errors but not falling into the previous classes('Faulty Generalization', 'Fallacy of Credibility', 'False Causality') as 'No Fallacy'.

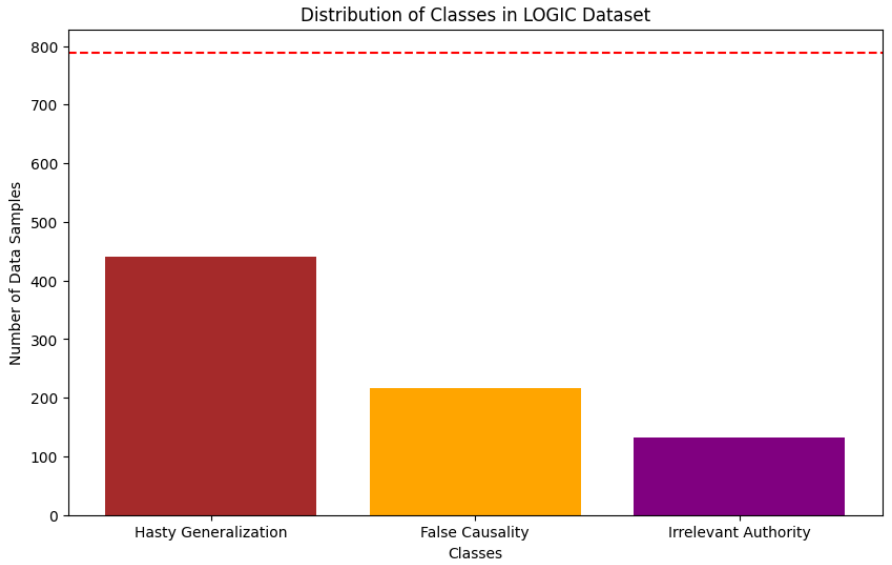
SYSTEM

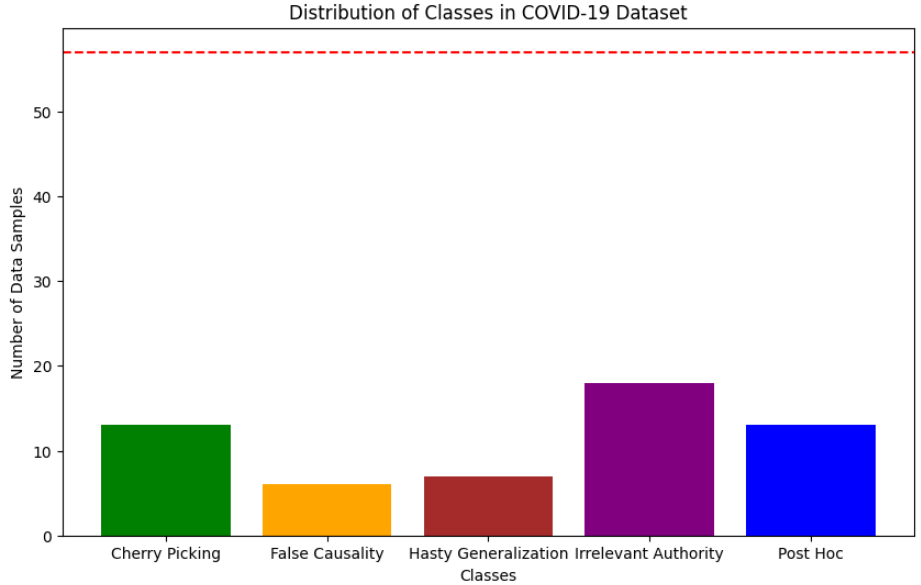
Your task is to detect a fallacy in the Text. The label can be 'Faulty Generalization', 'Fallacy of Credibility', 'False Causality' and 'No Fallacy'.
We refer to those with logical errors but not falling into the previous classes('Faulty Generalization', 'Fallacy of Credibility', 'False Causality') as 'No Fallacy'.
'No Fallacy' includes 'circular reasoning', 'appeal to public', 'ad hominem', 'deductive reasoning', 'appeal to emotion', 'false dilemma', 'equivocation', 'fallacy of extension', 'fallacy of relevance', 'fallacy of credibility', 'intentional'.
Given the two texts(examples) for each fallacy type, determine which of the following fallacies occur in the text.
1) Faulty generalization
1. My friend said her Math ...h classes must be hard!
2. If I don't take this A.P... up for the A.P. class.

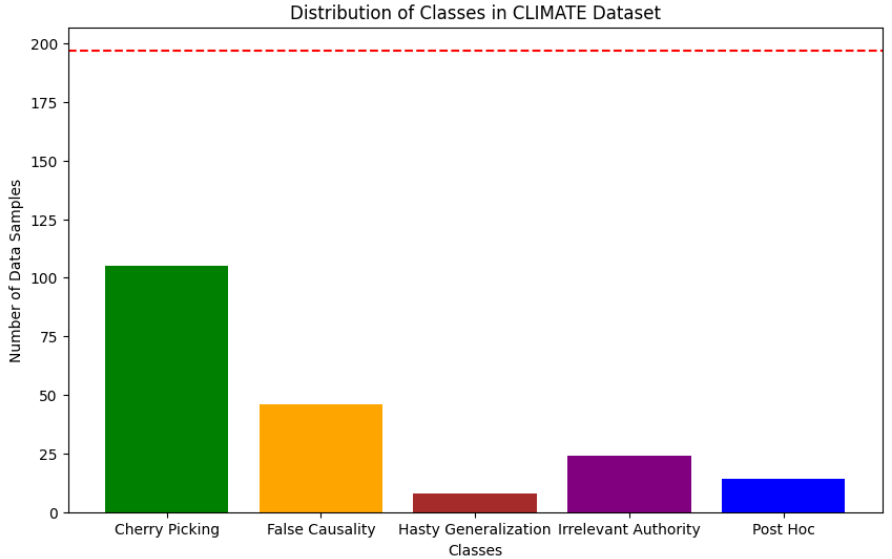
2) Fallacy of credibility
1. We're going to protest a...glesias said it's okay.
2. If I made a reference to... likely be utilizing...
3) False causality
1. I started exercising twi...sing leads to new jobs!
2. It's cold on a summer da...obal warming is a hoax.

데이터 분포(only Fallacy)

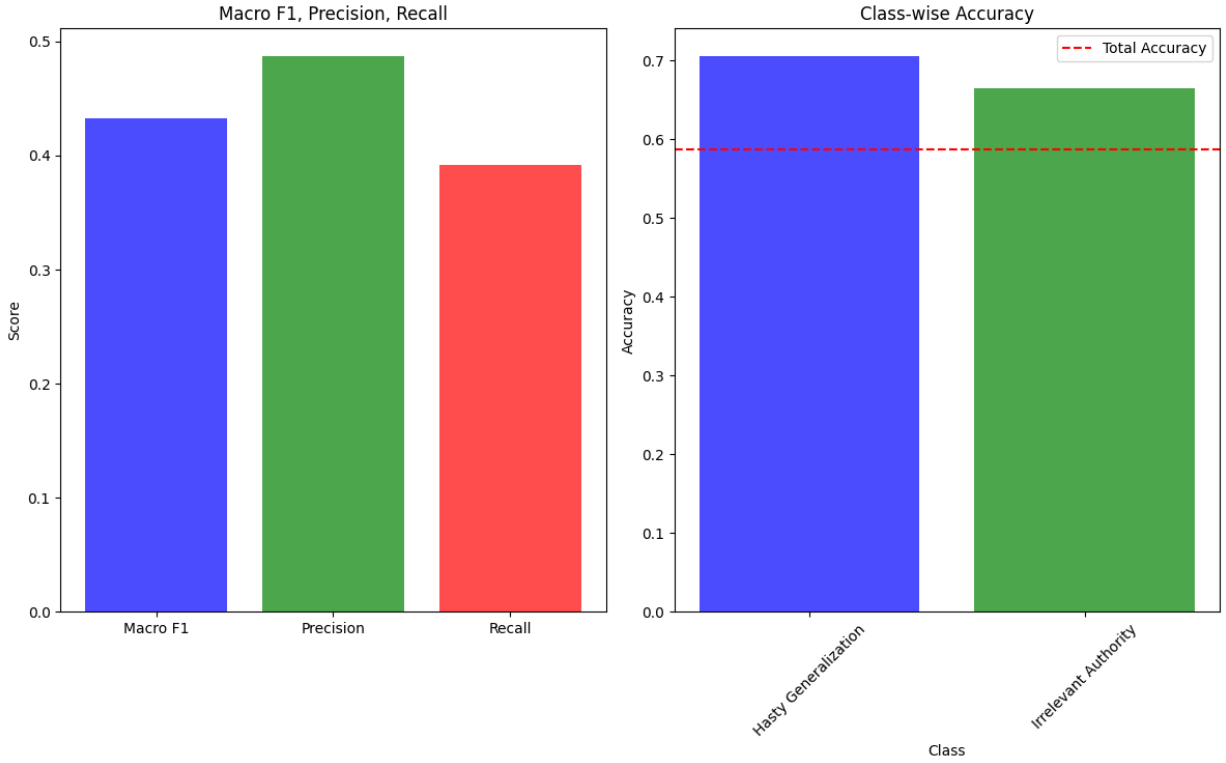




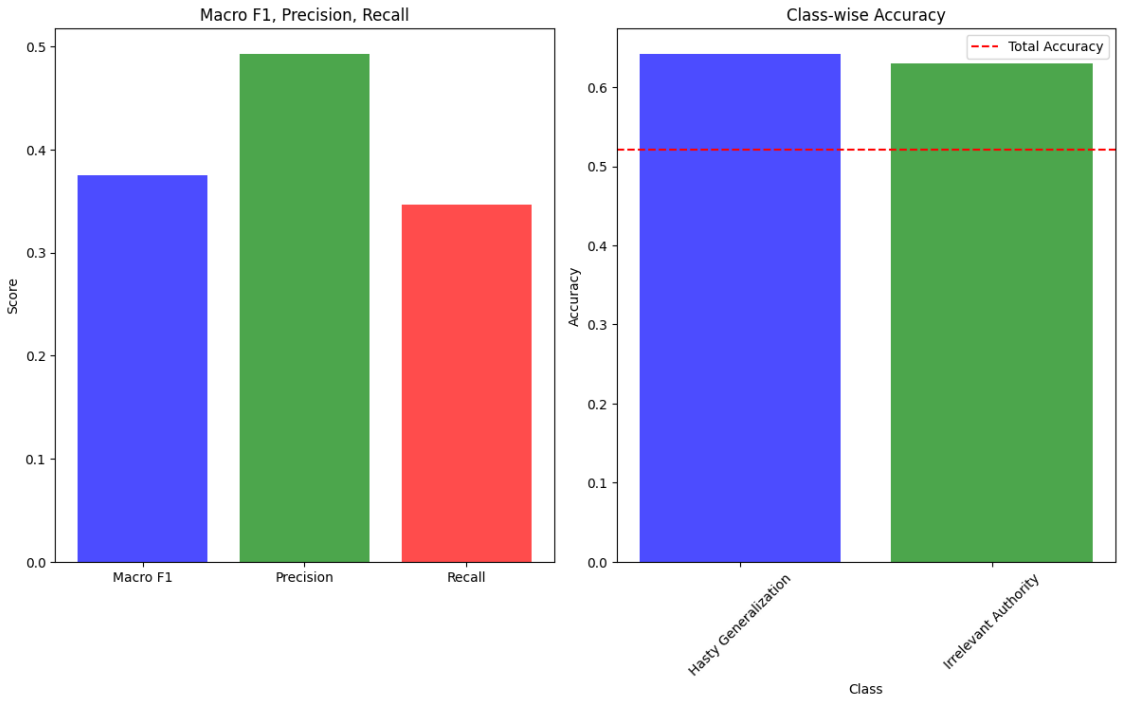




결과(Argotario)

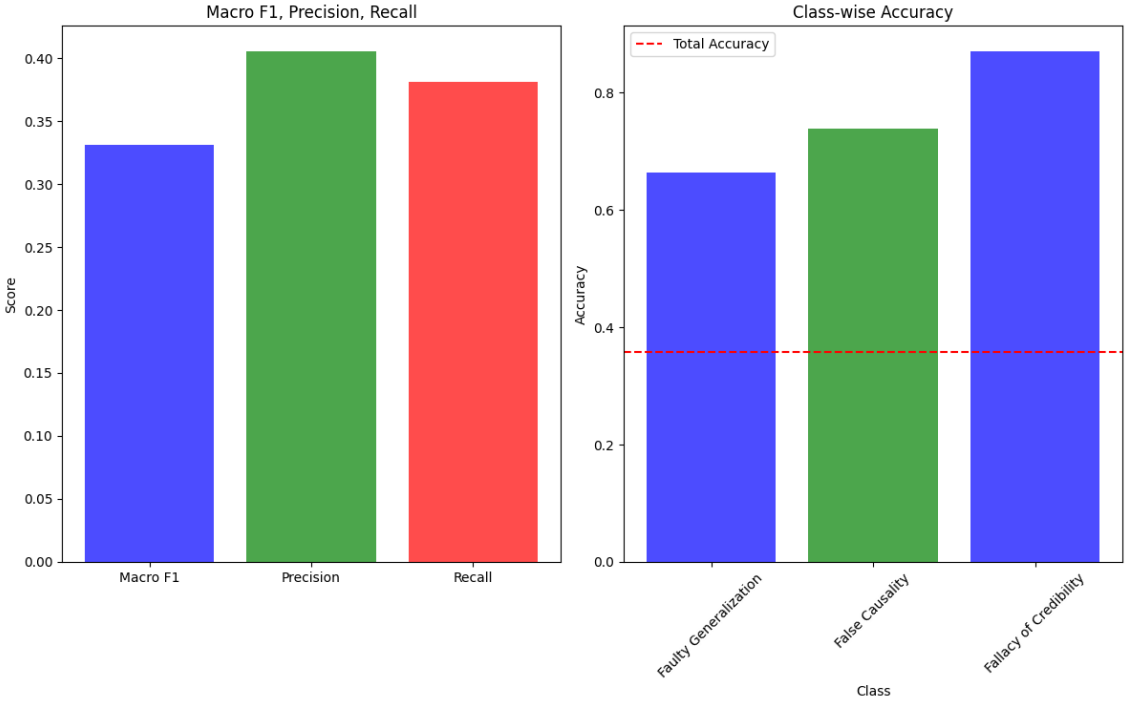


Zero-Shot

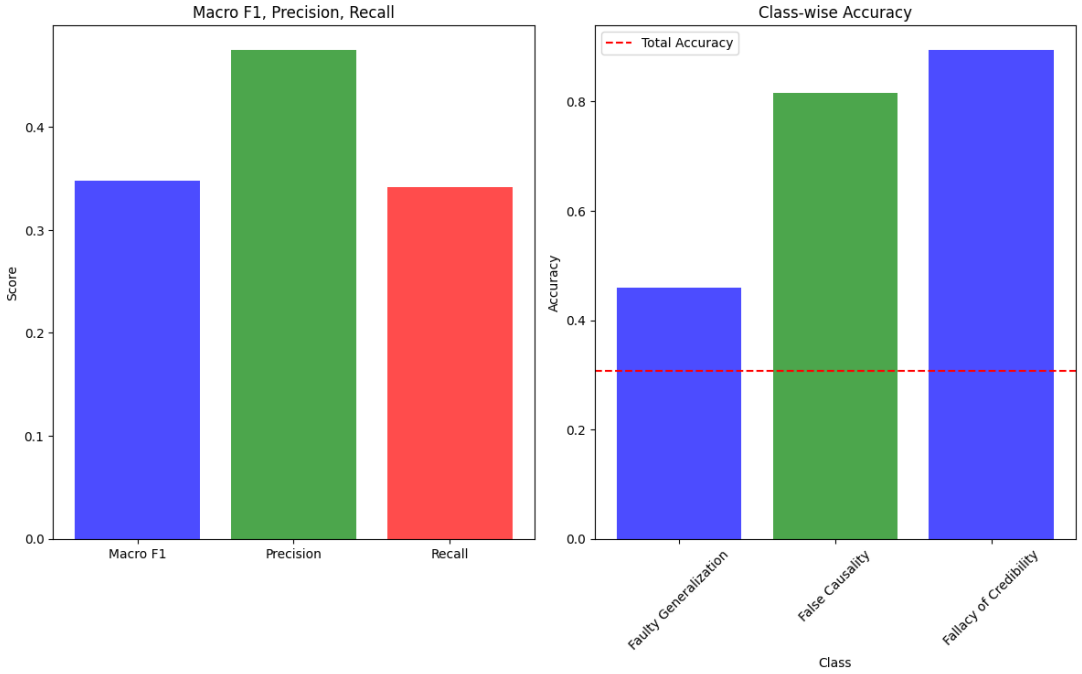


Two-Shot

결과(LOGIC)

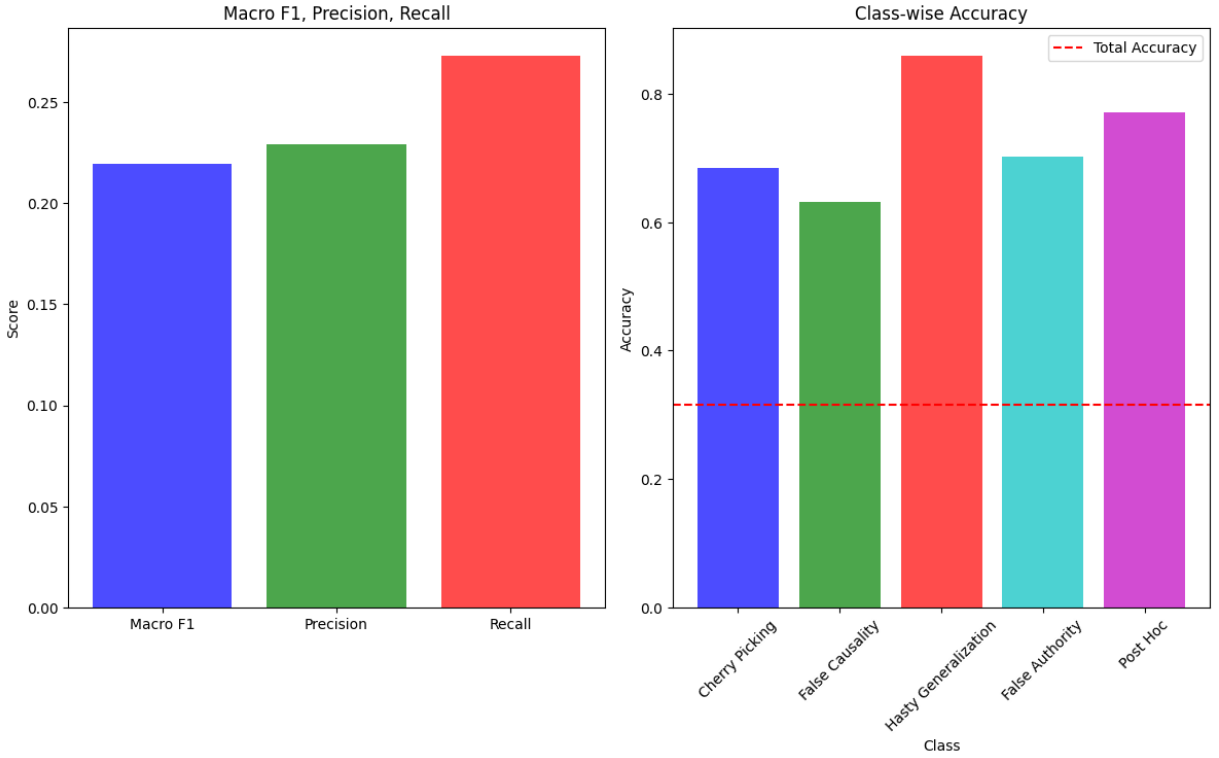


Zero-Shot

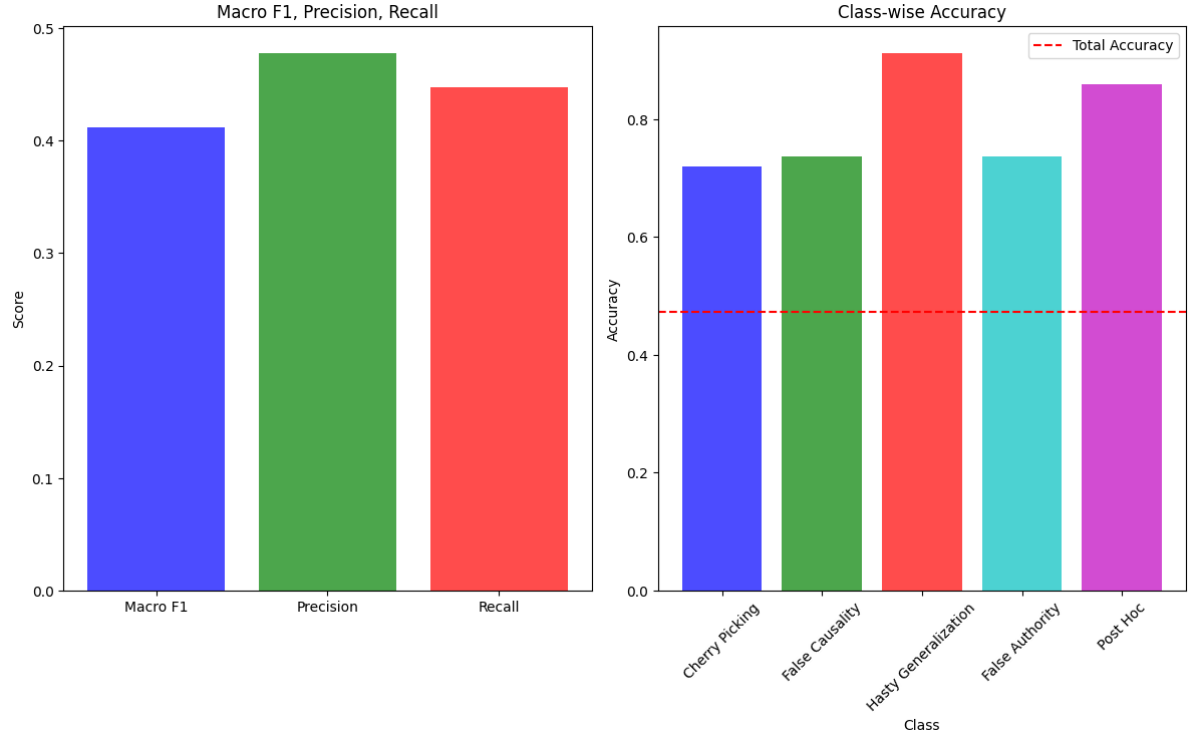


Two-Shot

결과(COVID-19)

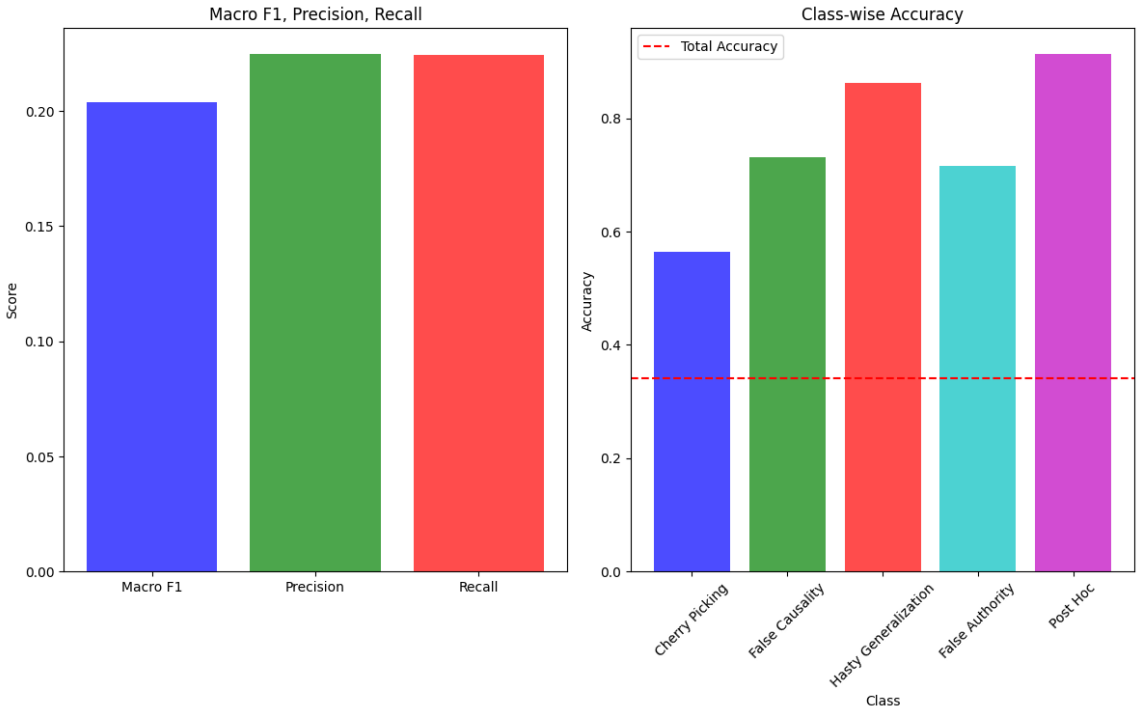


Zero-Shot

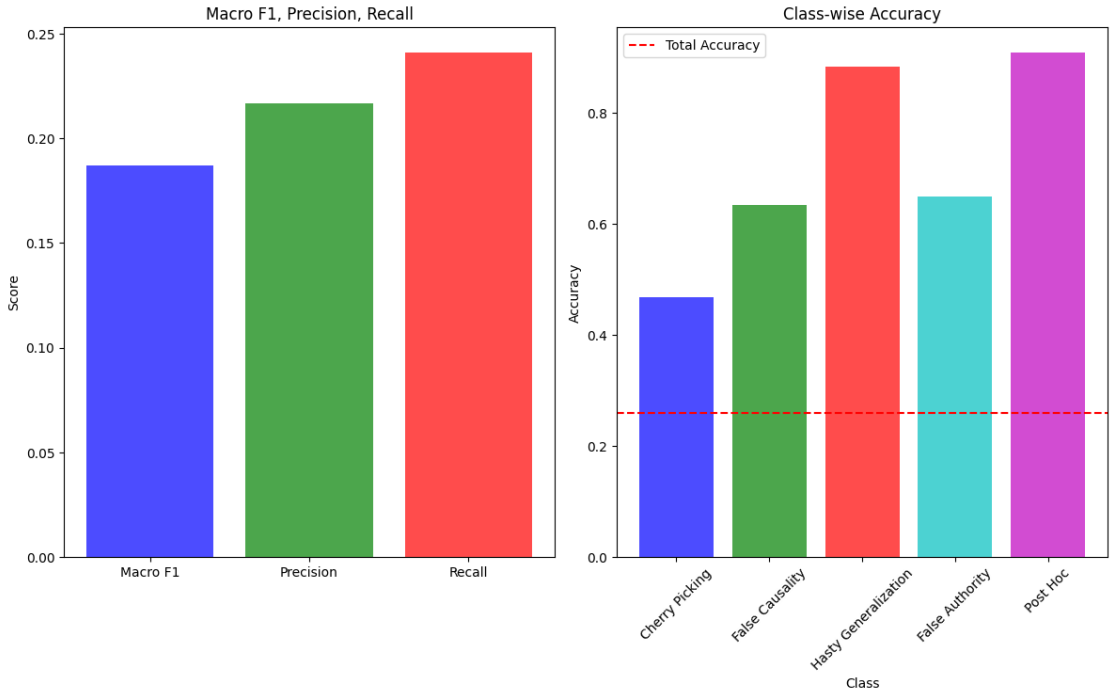


Two-Shot

결과(CLIMATE)

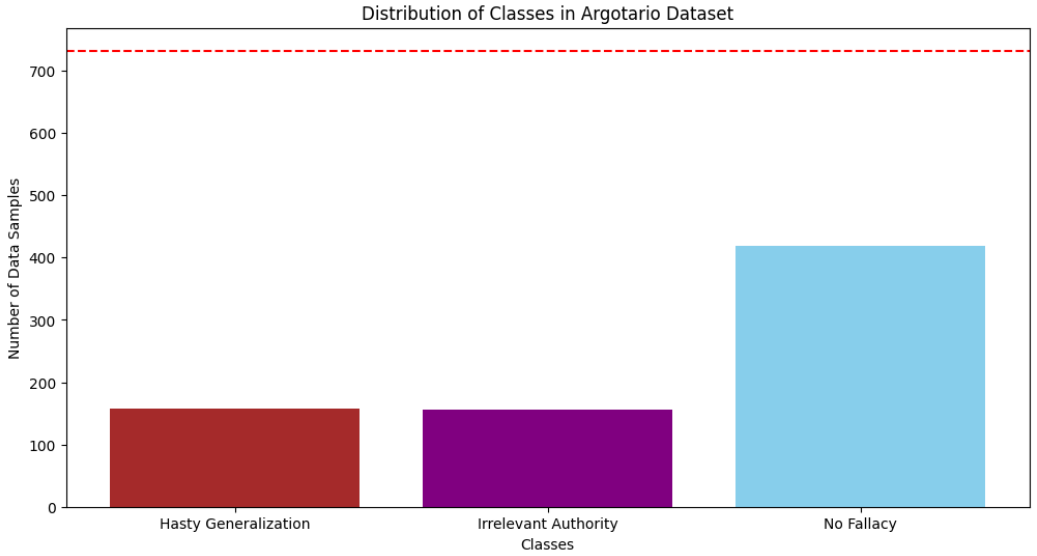


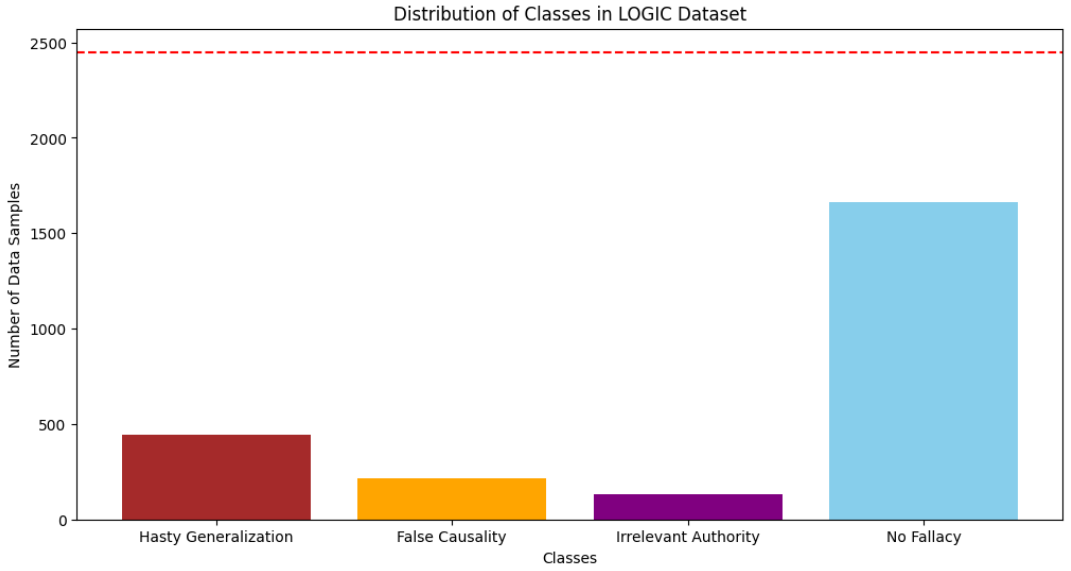
Zero-Shot

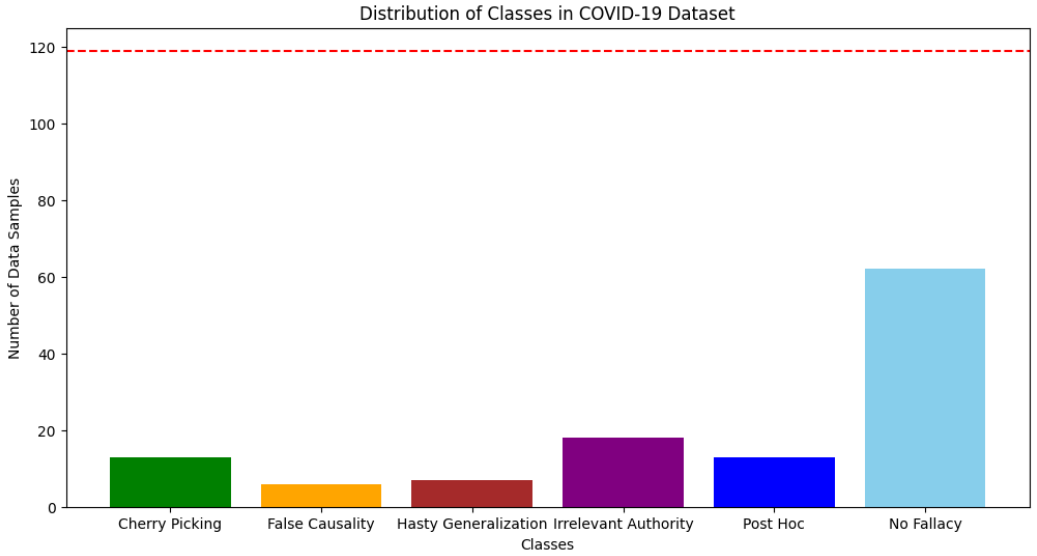


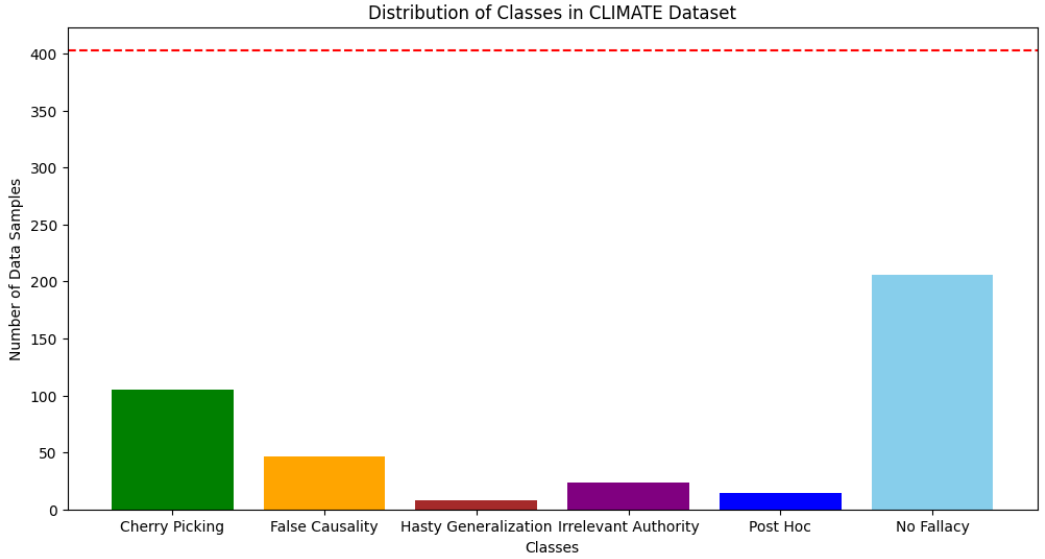
Two-Shot

데이터 분포(with No Fallacy)

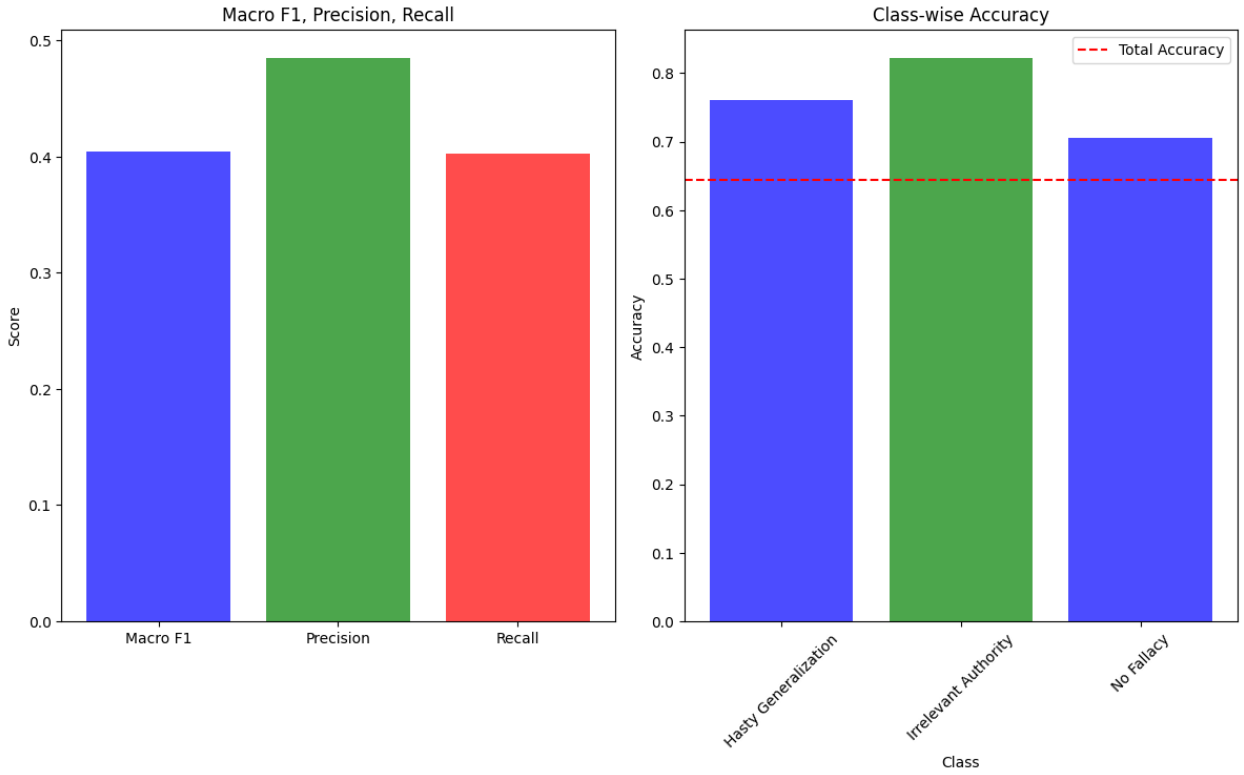




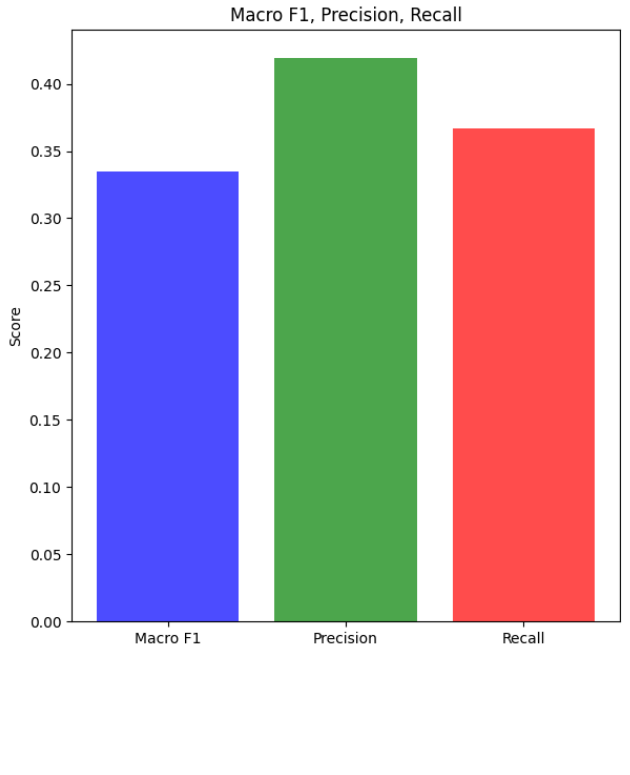


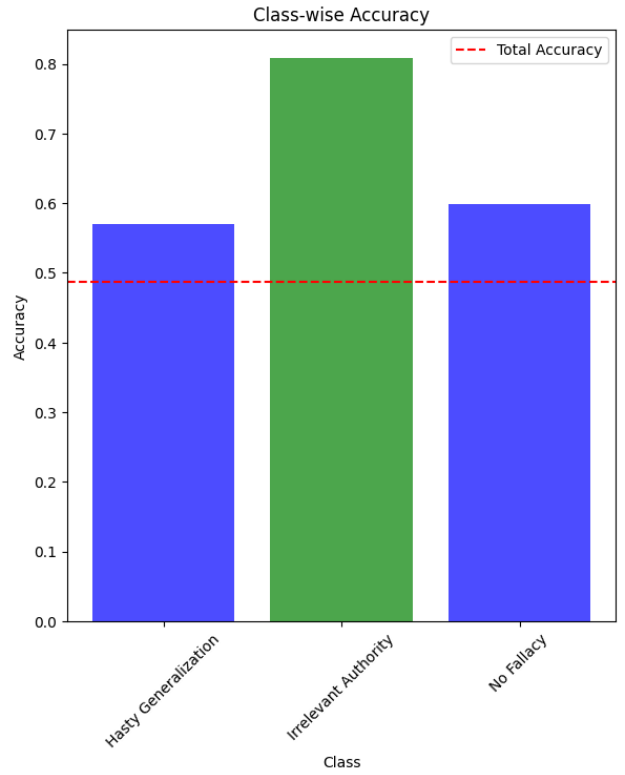


결과(Argotario)



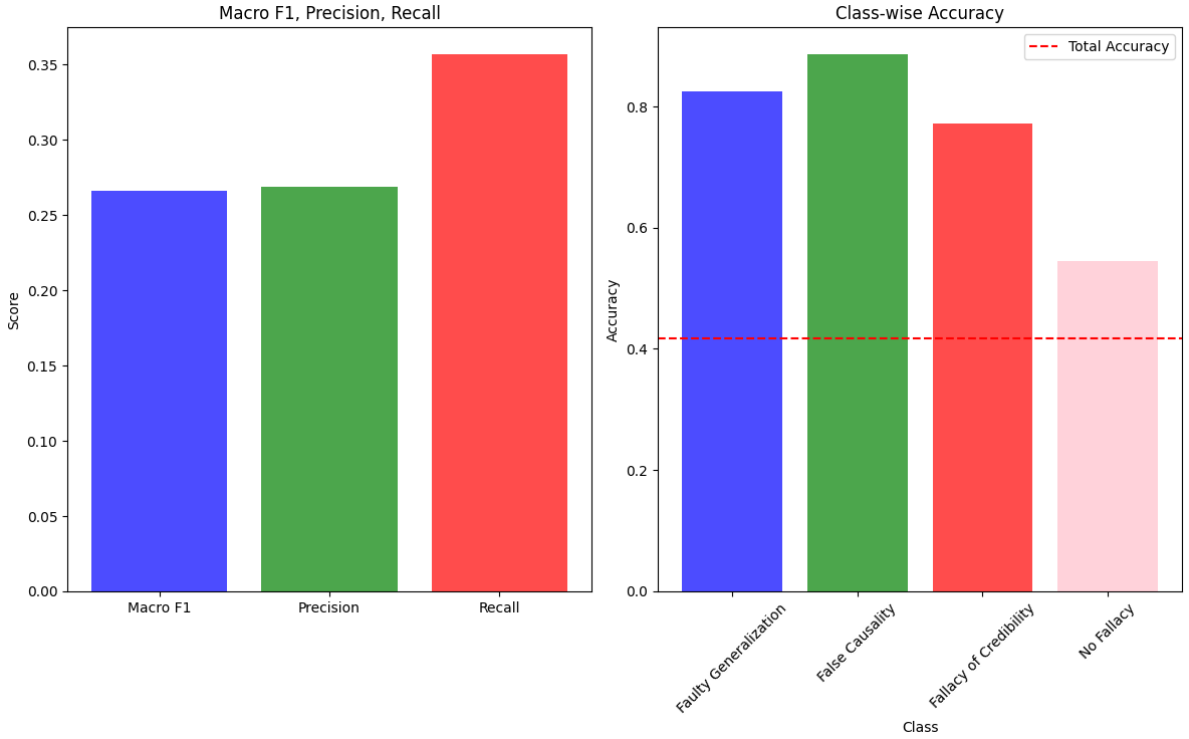
Zero-Shot



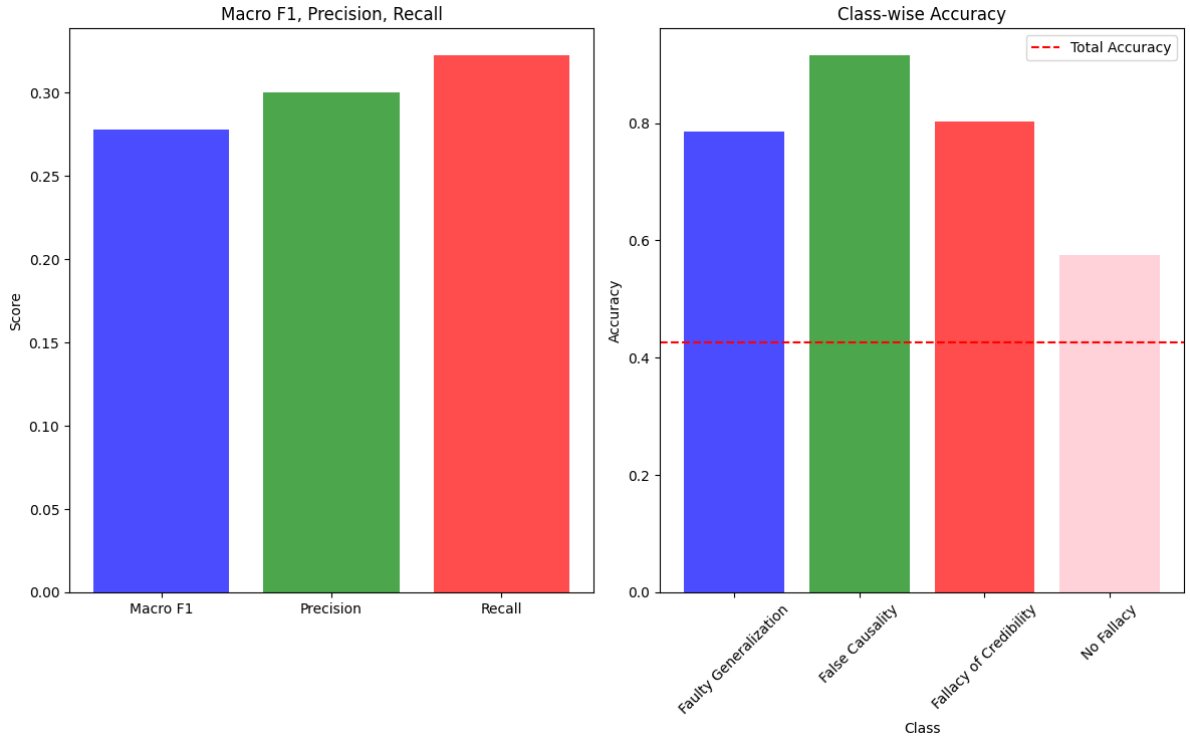


Two-Shot

결과(LOGIC)

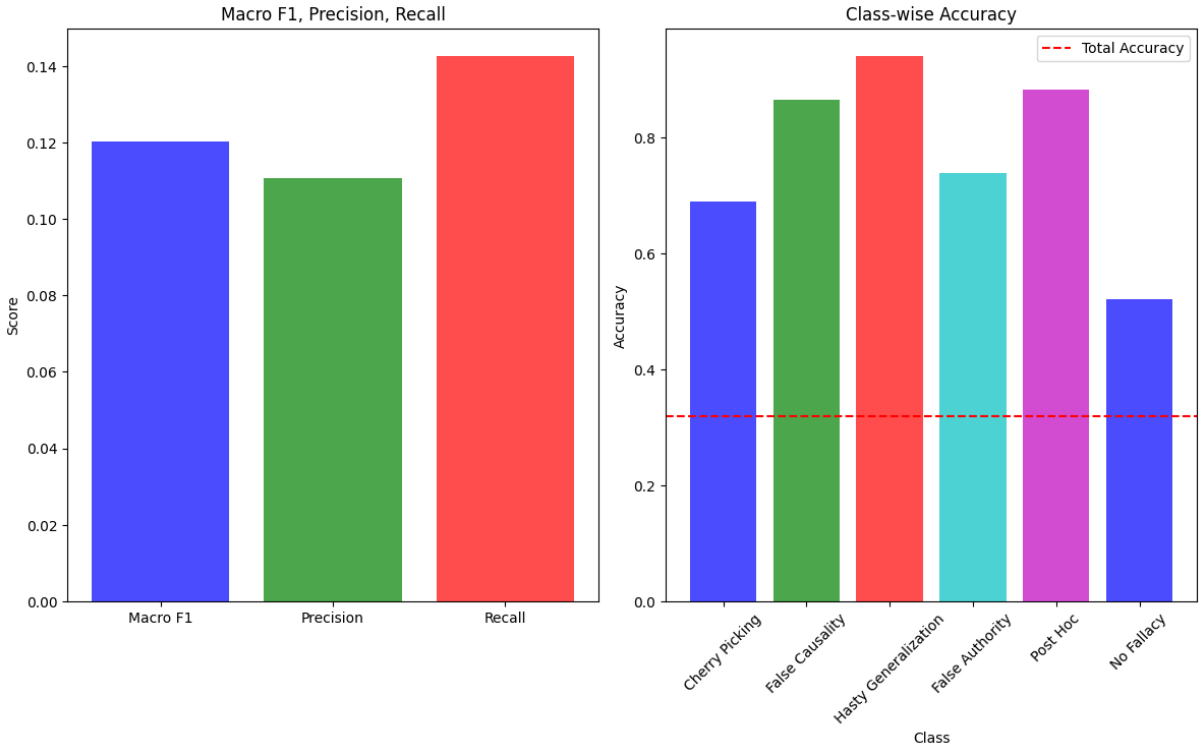


Zero-Shot

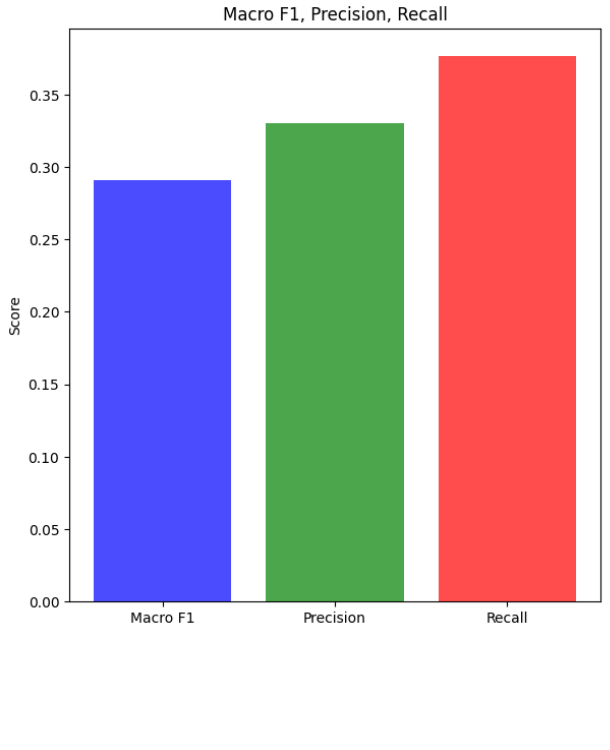


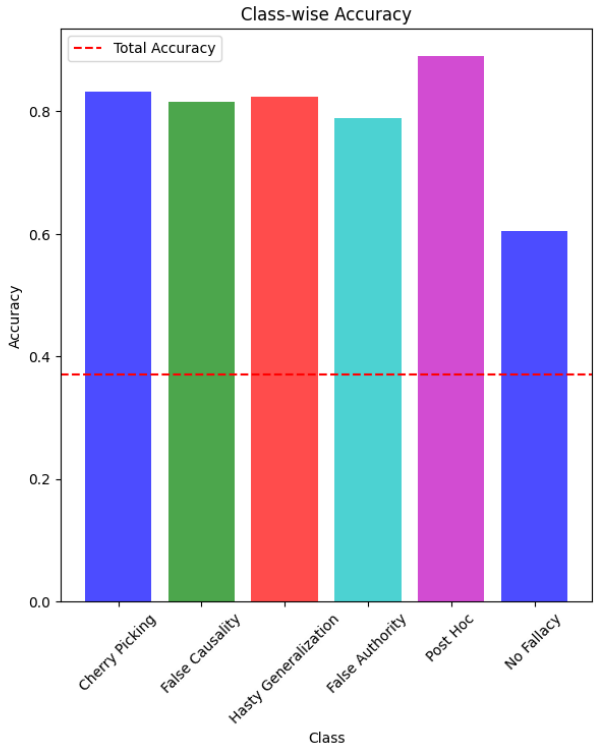
Two-Shot

결과(COVID-19)



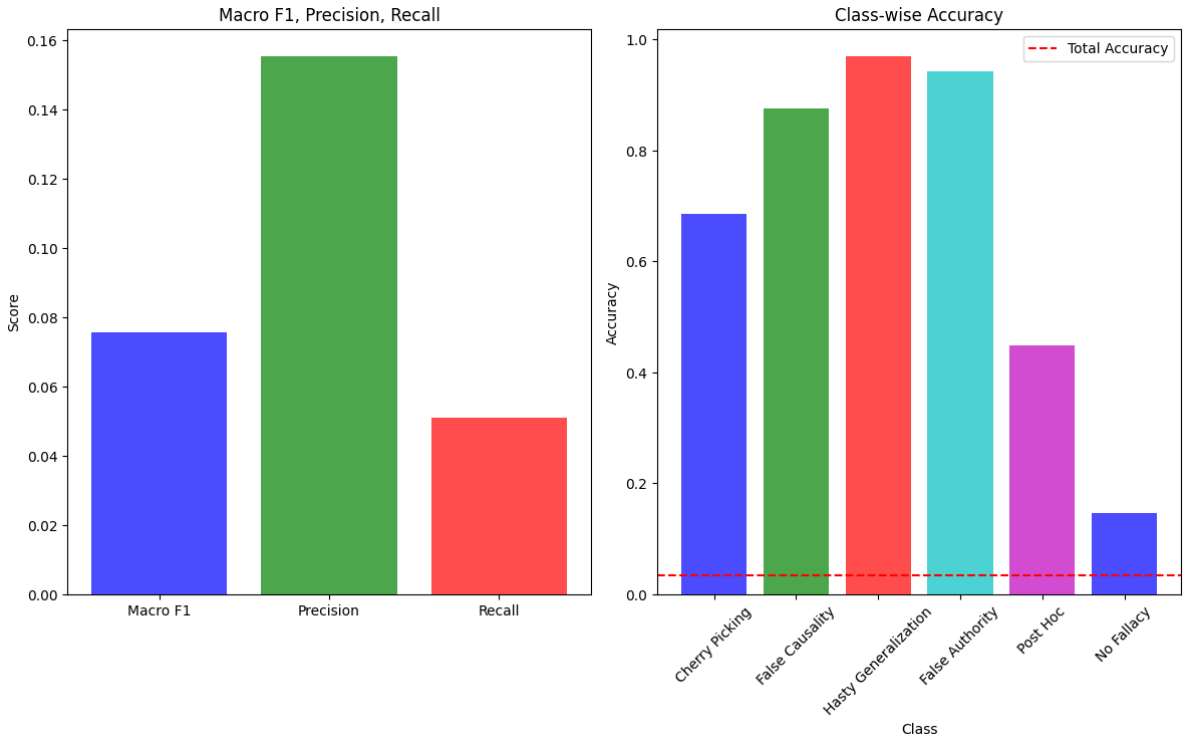
Zero-Shot



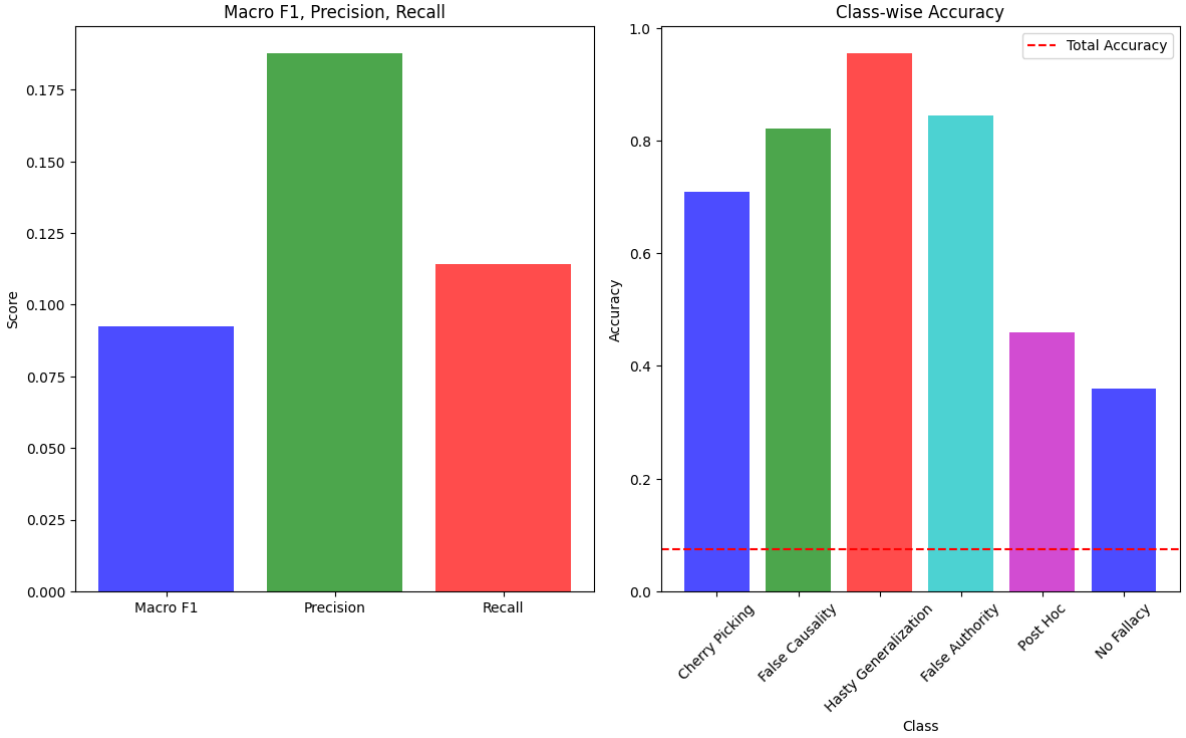


Two-Shot

결과(CLIMATE)



Zero-Shot



Two-Shot

결과(Table)-without No Fallacy

Dataset	Total Accuracy	Precision	Recall	Macro-F1
Argotario (Zero-Shot)	0.5879	0.4871	0.3917	0.4324
Argotario (Two-Shot)	0.5208	0.4929	0.3467	0.3751
LOGIC (Zero-Shot)	0.3574	0.4053	0.3807	0.3311
LOGIC (Two-Shot)	0.3067	0.4749	0.3414	0.3481
COVID-19 (Zero-Shot)	0.3158	0.2290	0.2731	0.2196
COVID-19 (Two-Shot)	0.4737	0.4779	0.4474	0.4119
CLIMATE (Zero-Shot)	0.3401	0.2249	0.2245	0.2037
CLIMATE (Two-Shot)	0.2589	02167	0.2412	0.1869
AVG(Zero-Shot)	0.3902	0.3519	0.3099	0.2969
AVG(Two-Shot)	0.3964	0.4174	0.3415	0.3313

2-shot 평균이 높은 이유는 COVID데이터 때문

결과(Table)-without No Fallacy

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Faulty Generalization (441)	0.6641	0.4601
False Causality (216)	0.7389	0.8150
Irrelevant Authority (132)	0.8707	0.8948
Total (789)	0.3574	0.3067

LOGIC

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Cherry Picking (13)	0.6842	0.7193
False Causality (6)	0.6316	0.7368
Hasty Generalization (7)	0.8596	0.9123
False Authority (18)	0.7018	0.7368
Post Hoc (13)	0.7719	0.8596
Total (57)	0.3158	0.4737

COVID-19

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Cherry Picking (105)	0.5635	0.4670
False Cause (46)	0.7310	0.6345
Hasty Generalization (8)	0.8629	0.8832
False Authority (24)	0.7157	0.6497
Post Hoc (14)	0.9137	0.9086
Total (197)	0.3401	0.2589

CLIMATE

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Hasty Generalization (157)	0.7061	0.6422
Irrelevant Authority (156)	0.6645	0.6294
Total (313)	0.5879	0.5208

Argotario

결과(Table)-with No Fallacy

Dataset	Total Accuracy	Precision	Recall	Macro-F1
Argotario (Zero-Shot)	0.6434	0.4851	0.4029	0.4039
Argotario (Two-Shot)	0.4877	0.4194	0.3667	0.3348
LOGIC (Zero-Shot)	0.4173	0.2688	0.3570	0.2661
LOGIC (Two-Shot)	0.4263	0.2999	0.3225	0.2776
COVID-19 (Zero-Shot)	0.3193	0.1108	0.1427	0.1203
COVID-19 (Two-Shot)	0.3697	0.3302	0.3769	0.2908
CLIMATE (Zero-Shot)	0.0347	0.1555	0.0510	0.0756
CLIMATE (Two-Shot)	0.0744	0.1878	0.1142	0.0923
AVG(Zero-Shot)	0.3537	0.2551	0.2384	0.2165
AVG(Two-Shot)	0.3395	0.3093	0.2951	0.2489

2-shot 평균이 높은 이유는 COVID데이터 때문

결과(Table)-with No Fallacy

LOGIC

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Faulty Generalization (441)	0.8248	0.7852
False Causality (216)	0.8865	0.9163
Irrelevant Authority (132)	0.7722	0.8024
No Fallacy (1660)	0.5439	0.5741
Total (789)	0.4173	0.4263

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Cherry Picking (13)	0.6891	0.8319
False Causality (6)	0.8655	0.8151
Hasty Generalization (7)	0.9412	0.8235
False Authority (18)	0.7395	0.7899
Post Hoc (13)	0.8824	0.8908
No Fallacy (62)	0.5210	0.6050
Total (57)	0.3193	0.3697

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Cherry Picking (105)	0.6849	0.7097
False Cause (46)	0.8759	0.8213
Hasty Generalization (8)	0.9702	0.9553
False Authority (24)	0.9429	0.8437
Post Hoc (14)	0.4491	0.4591
No Fallacy (206)	0.1464	0.3598
Total (197)	0.0347	0.0744

CLIMATE

Class	Zero-Shot Accuracy	Two-Shot Accuracy
Hasty Generalization (157)	0.7609	0.5697
Irrelevant Authority (156)	0.8224	0.8087
No Fallacy (419)	0.7049	0.5984
Total (732)	0.6434	0.4877

COVID-19

Argotario