

MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing

- <https://aclanthology.org/2024.findings-acl.300.pdf>
- ACL 2024
- 밈 연구로 방향을 틀게 되어서 읽어본 논문
 - 밈이라는게 이미지, 텍스트 모두 이해해야하는 채널이라 멀티모달 LLM과의 조합이 좋지 않을까 싶음
 - Multimodal COT라는 논문을 읽어야 할듯
 - 모듈식 구성이라 조금 복잡한 느낌이지만, 잘 작성된듯함.
 - 엄청 다양한 실험을 한 것이 인상깊음 (여러 요소를 고려한듯)

 [연구 배경 및 문제 정의](#)

 [본 논문의 제안](#)

 [주요 기여 \(Contributions\)](#)

[Abstract](#)

[1. Abstract](#)

[2. Related Work](#)

[Studies on Memes](#)

[Visual Question Answering \(VQA\)](#)

[Multimodal Large Language Models](#)

[3. The MemeMQACorpus Dataset](#)

[Prompting for Question Diversification](#)

[4. The ARSENAL Model](#)

[System Architecture](#)

[5. Experiments](#)

[6. Benchmarking MemeMQA](#)

[7. Robustness Analysis](#)

[8. Conclusion and Future Work](#)

연구 배경 및 문제 정의

- 밈(meme)은 이미지와 텍스트를 결합하여 사회적, 정치적 메시지를 강력하게 전달하는 수단으로 진화함.
- 하지만 이들은 **풍자, 비유, 문화적 맥락**, 나아가 **허위 정보와 혐오 표현**의 매개체가 되기도 함.
- 기존의 밈 분석 연구는 주로 분류나 캡션 생성에 국한되어 있으며, ****시각적·언어적 의미의 교차 추론(cross-modal reasoning)****이나 **정확한 설명 생성**에는 한계가 있음.

본 논문의 제안

- **MemeMQA Task**: 밈 이미지에 대해 역할 기반 다중 선택 질문을 제시하고, 정답과 함께 그 이유(설명)를 생성하는 **복합적 멀티모달 QA 작업** 제안.
- **MemeMQACorpus**: 1,122개 밈 이미지와 1,880개 구조화 질문으로 구성된 **신규 QA 데이터셋** 구축.
- **ARSENAL 모델**:
 - 2단계 모듈식 멀티모달 추론 프레임워크.
 - ① 답변 예측 (Answer Prediction)
 - ② 정답에 대한 설명 생성 (Explanation Generation)
 - LLaVA-7B로 생성한 중간 근거(**R_generic** , **R_specific**)와 T5-large 기반의 텍스트 생성 모델을 연동해 높은 추론 품질 달성.

주요 기여 (Contributions)

1. **MemeMQA Task**: 기존 VQA나 멀티모달 추론과 차별화된 **역할기반 다중선택 + 설명 생성** 과제 정식화.
2. **MemeMQACorpus**: 사회적 의미를 내포한 밈에 대해 **질문 다양화, 역할 프레이밍(영웅/악당/피해자)** 기반 QA 데이터셋 공개.
3. **ARSENAL 프레임워크**:
 - LLaVA와 T5를 결합한 **2단계 추론 구조** 제안.
 - **CoT 기반 멀티모달 prompting 전략**과 중간 근거 삽입 구조.
4. **강건성 실험 설계**:
 - 다양한 질문 재구성 (Question Diversification)
 - 혼란 변수(confounders) 주입 실험을 통해 **모델의 일반화 능력** 평가.

Abstract

- 밈
 - 다양한 커뮤니케이션의 매체로 진화
 - 잠재적 위험성 탐구 필요성 증가
 - 기존 연구: 밈의 폐쇄된 환경에서 해악을 감지, 의미 레이블을 적용, 자연어 설명을 제공하는 분석
- 본 연구:
 - MemeMQA 멀티모달 질문-응답 프레임워크 소개
 - 구조화된 질문에 대한 정확한 응답을 유도하고 일관된 설명을 제공
 - MemeMQACorpus: 1,880개 질문과 1,122개 밈에 대한 적절한 답변-설명 쌍
 - ARSENAL: MemeMQA 문제를 다루기 위해 LLM의 추론 능력을 활용하는 새로운 이단계 멀티모달 프레임워크
 - 응답 예측 정확도를 약 18% 향상
 - 최고 기준선보다 독특한 텍스트 생성을 보여주는 우수성을 입증
 - ARSENAL's robustness, diversification of question-set, confounder-based evaluation regarding MemeMQA's generalizability, and modality-specific assessment

1. Abstract

- **밈:** 적절한 형식이나 공식 언어의 전통적인 의존성이 없이도 강력한 정보 전파를 위한 접근 가능한 형식을 제공
 - 초보 콘텐츠 제작자와 경험이 풍부한 전문가 모두에게 정보를 전파할 수 있는 기회를 제공
 - 일반 대중에게 해롭기도 한 정보를 퍼뜨릴 위험
 - 기존 연구: 혐오 발언, 사이버 괴롭힘 등 다양한 형태의 해악성 탐구, 보통 블랙박스 설정, 모델의 해석 가능성을 높이고 콘텐츠 조절을 위한 효과적인 도구 역할
- **본 연구**

- 다양한 엔티티에 할당된 의미적 역할에 대한 구조화된 질문과 함께 밈을 주어진 경우 올바른 답변 엔티티를 추론
- 답변에 대한 간결한 설명을 생성
- 밈의 맥락화된 의미 분석을 탐구

• 분야

- 잘 알려진 개인과 정치적 인물과 같은 개체의 내러티브 프레이밍을 탐구
- 선거 또는 팬데믹과 같은 중요한 사건 동안에 특히 중요
 - 증오 발언 및 허위 정보와 같은 해로운 콘텐츠가 퍼질 위험이 높아져 효과적인 조정의 필요성이 강조
- 밈의 피해화, 찬양 및 악마화의 의도를 분석 → '영웅', '악당', '희생자'와 같은 용어를 채택
- 밈에 대한 이해를 심화하고 소셜 미디어를 보다 안전하게 만드는 데 기여

• MemeMQA 프레임워크

- 소셜 미디어 사용자와 사실 확인자가 밈의 해로움을 평가할 수 있도록 돕도록 설계
- 밈을 분석하는 것은 고급 추론을 요구하는 미묘한 의미 때문에 복잡: 상식과 문화적 이해가 포함
- 그림1) 밈의 핵심 주제와 이민자가 다양성을 풍요롭게 하는 것에 대한 암시된 메시지를 이해하는 것이 필수적

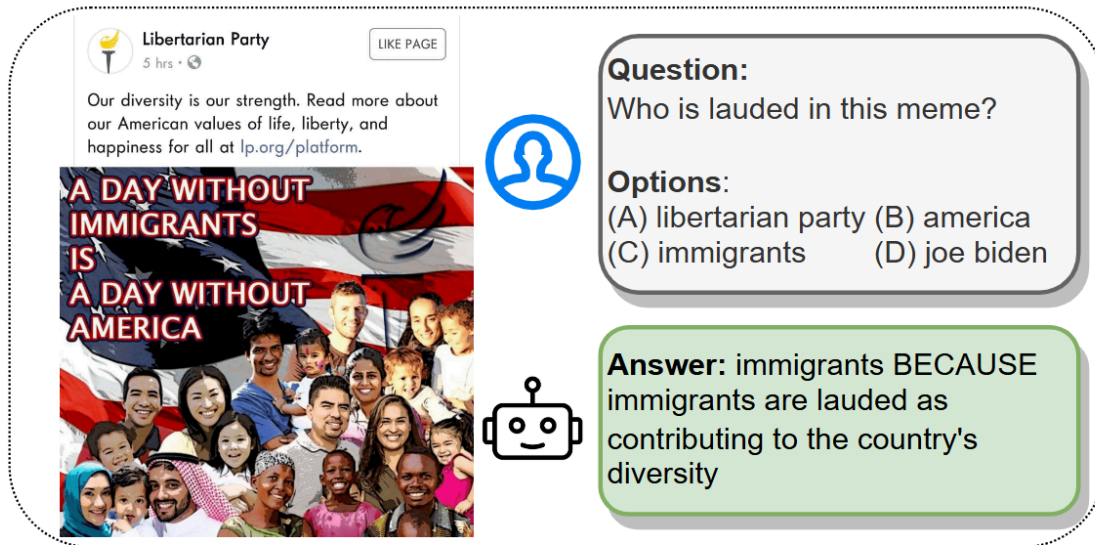


Figure 1: The MemeMQA task: Given an input meme and multiple choices, identify the correct answer and justify.

• 정리/요약

- 정치적 밈에 대한 다선택 질문을 답변하고 설명하는 새로운 작업을 소개
- ExHVV 데이터셋을 사용하여 1,122개의 밈에 대해 1,880개의 질문이 포함된 데이터셋(MemeMQACorpus)을 생성
- MemeMQACorpus를 벤치마킹
- 다중 모달 LLM의 추론 능력을 활용하는 새로운 **모듈식** 접근 방식인 ARSENAL을 제안
 - 이론적 근거, 답변 예측 및 설명 생성을 위한 모듈을 포함
 - 베이스라인과 비교 분석하여 그 강점과 한계를 강조
- 기여
 - MemeMQA: 밈의 맥락에서 다중 모달 질문-답변 설정을 도입하는 새로운 작업 형식화
 - MemeMQACorpus: MemeMQA를 위한 다양한 질문 및 다중 선택 설정을 도입하기 위해 이전에 사용할 수 있었던 데이터셋을 확장

- ARSENAL: MemeMQA를 위한 다중 모달 LLM 생성 이론적 근거를 활용하는 모듈식 프레임워크 시스템 아키텍처
- 개방형 연구: 벤치마킹, 프롬프트 평가, 다양한 질문의 상세 분석, 혼란 기반 교차 검증, 다중 모달리티의 함의 및 제안된 해결책의 한계에 대한 포괄적인 연구를 포함

2. Related Work

Studies on Memes

- 다양한 밈 분석 측면 포함: 개체 식별, 감정 예측, 혐오 밈 탐지 등: Visual BERT, UNITER 및 이중 스트림 인코더
- 다중 모달 증거 예측, 역할 레이블 설명 및 혐오 밈의 의미 분석
- 주석의 스키마와 품질에 의해 제약을 받으며, 밈 현상에 대한 개방형 탐색을 제한

Visual Question Answering (VQA)

- VQA 연구의 발전을 탐구
 - Antol et al. (2015): 개방형 질문과 후보 답변을 강조 → 답변을 분류하기 위해 이미지와 질문 표현을 결합
 - 교차 모달 상호작용을 탐색
 - 공동 주의(co-attention), 소프트 주의(soft-attention), 하드 주의(hard-attention)
 - 일반적인 상식 추론을 통합하기 위한 노력: UpDn, LXMERT와 같은 모델은 비선형 변환 및 변환기(Transfomers)를 활용하여 VQA(Visual Question Answering)에 적용
 - 표준 비주얼-질문-답변 프레임워크에서는 이미지가 관련 질문과 함께 제공, 설정에 따라 여러 선택지 옵션 추가 가능
 - 밈(meme)은 자주 일치하지 않는 텍스트 콘텐츠와 이미지를 결합 → 작업 어려움, 직관적 X

Multimodal Large Language Models

- **대규모 언어 모델(LLMs)의 출현**은 자연어 이해 및 추론에 상당한 발전
 - 비주얼-언어 기반 작업에 대한 멀티모달 증강에 대한 친화성도 반영
 - 융합 기반 어댑터 레이어를 통해 LLMs를 증강하여 VQA에서 멀티모달 대화에 이르기까지 다양한 작업에서 뛰어난 성능을 발휘
- 그러나 기존의 멀티모달 LLM들인 LLaVA, miniGPT4 등: **میم**에서 나타나는 시각-언어적 불일치에서 **풍자와 아이러니 같은 뉘앙스**를 이해하는 데 한계
 - 유사한 작업 중 일부가 **میم** 관련 작업을 다루고 있지만
 - 주로 캡션 생성과 VQA의 비주얼-언어적 기초 설정에 한정
 - LLM의 고유한 한계, 즉 사전 학습 편향 및 환각(hallucinations) (Zhao et al., 2023) 등으로 인해 한계
- **MemeMQA의 두 가지 목표**
 - 답변 예측 및 설명 생성을 포함
 - Multimodal CoT(MM-CoT) 모델을 포함
 - 기존 방법들 미흡
 - DETR 기반 시각 인코딩(Carion et al., 2020)과 통합 QA T5 모델에서의 텍스트 인코딩/디코딩을 결합한 2단계 프레임워크
 - 답변 예측에서는 우수하지만 설명에서는 부족
 - 멀티모달 LLM(LLaVA, InstructBLIP, miniGPT4): **میم** 의미를 이해하는 데 가능성, 질문별 정확성에서는 고군분투, 보다 넓은 **میم** 맥락을 우선시하여 세부 정확한 답변보다 중점
 - 본 연구: MemeMQA 작업이 제기하는 복잡한 시각-의미 추론과 관련된 문제를 다루는 데 집중, 멀티모달 LLM과 질문-답변을 위한 신경 추론 설정(neural reasoning setups)의 한계를 고려

3. The MemeMQACorpus Dataset

- 현재의 밈 데이터셋: 일반적으로 범주형 레이블, 그에 대한 설명을 포함
- 전통적인 비주얼 질문 답변(VQA) 프레임워크 존재, 밈의 미묘한 복잡성 부족
 - 탐지, 분할, 조건부 멀티모달 모델링(예: 캡션 생성, 비주얼 질문 답변, 다중 선택 VQA) 및 강력한 시각-언어적 통합 등
 - 멀티모달 추론, 추상 아이디어 표현, 언어의 뉘앙스 있는 메커니즘 사용(예: 말장난, 유머, 비유 등) 등 복잡성을 다루는 데 부족
- MemeMQA 코퍼스를 소개
 - 자유 형식 질문 및 다중 선택 응답을 모방하도록 설계된 새로운 데이터셋
 - 기존 멀티모달 데이터셋인 ExHVV(Sharma et al., 2023)를 보완
 - 3K 밈에 대해 자동으로 구성된 구조적 질문을 사용, 영웅, 악당, 피해자 등 세 가지 개체 유형의 함축적 역할에 대한 언어적 설명을 제공
 - 목표: QnA 설정을 통해 밈 해석 및 의사소통의 복잡성을 모사하는 것
 - 다중 선택 설정에서 정확히 하나의 개체만을 올바른 답변으로 이끌어내는 역할 기반 쿼리를 만드는 것
 - 그림2)
 - 기존 데이터셋: 엔티티: 민주당, 역할: 악당
 - 역할 기반 질문 설계: "이 밈에서 비방당하는 것은 무엇인가?"
 - 오답(방해 선택지, Distractors) 선택: 정답과 같은 역할로 분류된 엔티티는 오답으로 선택하지 않음
 - 질문 다양성 확보 (Free-form Questioning): 역할명(영웅, 악당, 피해자)의 동의어(synonyms)를 활용, "악당"을 대신해 "비난받는 대상", "비방의 주체" 등의 표현
 - MemeMQACorpus의 추가 변형
 - (a) 질문 다각화, (b) 역할 분석을 수행 → 섹션7

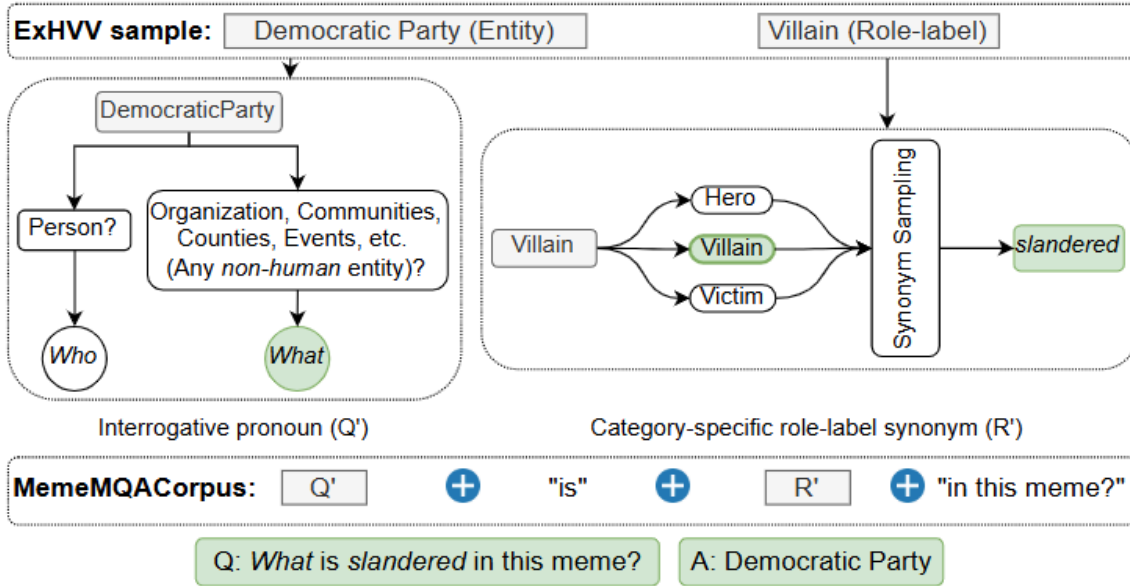


Figure 2: A schematic diagram showing question-answer construction process in MemeMQACorpus, using entity and role-label information from ExHVV.

Prompting for Question Diversification

- framing of the original questions 의 다양성 달성을 위해
 - 사전 훈련된 LLM인 Llama-2-7b-chat가 제로 샷 프롬프트를 통해 추론하는 데 활용
 - LLM은 질문 설정에 대한 맥락을 제공 받음
 - 그 뒤에 의미를 변경하지 않고 질문을 여러 방식으로 다시 작성하도록 모델에 요청
 - = 원래의 의미와 따라서 원래 옵션 세트의 유효성이 유지되도록 보장
 - LLM이 제공하는 재구성된 질문 중 하나가 무작위로 선택되고, 이 선택된 질문이 원래 질문을 대체

Prompt :

I will give you a question which will be about one specific entity in one specific meme and its categorisation. The question should not ask about any purpose or concept, it should always be about the entity. Paraphrase and rewrite this question in five ways keeping the meaning and answer intact and such that the new questions are longer than the original: "{question}"
Just give a numbered list of new questions and nothing else.
New questions:

Original Question :



Who is denounced in this meme?

Generated Questions:

1. Which entity is being denounced in this meme?
2. In this meme, which character is denounced?
3. What specific entity is being criticized in this meme?
4. Who is the subject of denunciation in this meme?
5. In this particular meme, what entity is being condemned?

Figure 3: Description of the prompting setup for free-form synthetic question generation using the LLM, Llama-2-7b-chat. The randomly chosen question option is highlighted in yellow.

4. The ARSENAL Model

- meme 기반 입력을 사용하여 다양한 프롬프트 구성을 분석하고 최적의 전략을 결정
- **Prompting Configurations**
 - CoT 추론을 사용한 다중 모드 질문-응답에서는 질문, 맥락(이미지와 관련된 텍스트), 옵션, 강의(상세 일반 맥락), 설명(간결한 맥락적 진술), 답변, 중간 생성 텍스트가 포함 (QCMLEAG)

- 프롬프트 구성: QCMLEAG의 요소를 결합하여 input → output으로 표현
 - 이전의 일단계 접근법(QCM→LA 또는 QCM→AL)은 한계
 - 성능이 향상된 2단계 설정으로 유도
 - MemeMQA는 ScienceQA보다 더 복잡한 추론을 포함하므로, 우리는 강의(L)를 상세한 역할 정의로 사용
 - 11개의 프롬프트 구성을 먼저 검토
 - 하나/두 개의 단계 방법과 base 모델 unifiedqa-t5-base/large (AT5B/L) 및 t5-large (T5L)를 사용하여 진행
 - MemeMQA에 대한 2단계 프레임워크의 적용 가능성을 확인

System Architecture

- 입력과 출력
 - 입력
 - (i) meme 이미지
 - (ii) OCR 텍스트
 - (iii) 해당하는 여러 옵션이 있는 질문
 - 출력
 - (i) 답변
 - (ii) 설명
- ARSENAL을 위해 다단계 설정을 제안하여 MM-CoT 및 다중 모드 LLM의 개별 강점을 MemeMQA의 전체 목표에 활용
 - 프레임워크는 답변 예측 및 설명 생성을 포함하는 두 단계의 프로세스
 - 모듈식 설계
 - 두 단계 모두 LLM 유추 근거를 통합
 - 초기 단계: 중간 근거rationale 생성을 포함한 두 단계로 구성
 - 두 번째 단계: 설명 생성을 중심으로 진행

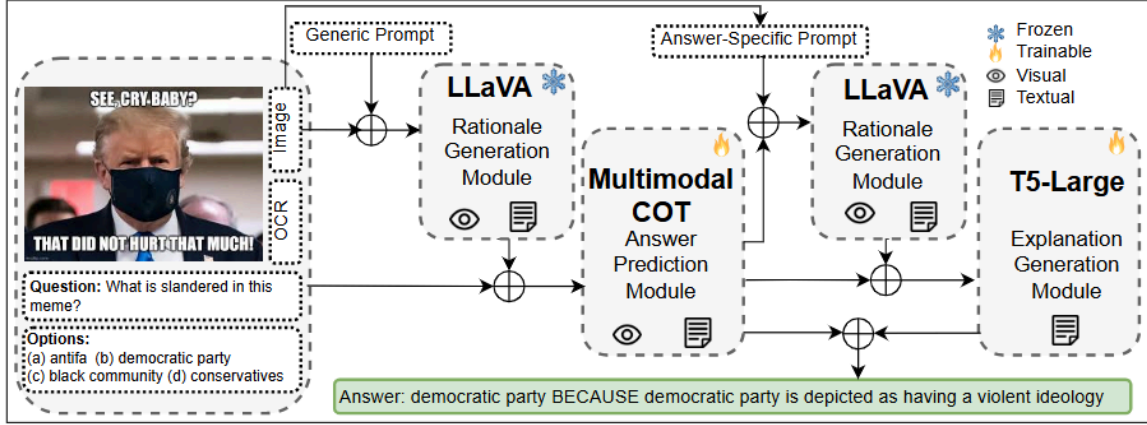


Figure 5: A schematic diagram of ARSENAL for the MemeQA task (\oplus : fusing the information via concatenation).

- Rationale Generation

- 첫 번째 단계에서 "일반적인 근거": $R_{generic}$ LLaVA-7B으로 생성, meme에 대한 의미적 정보를 텍스트 형태로 제공

- 일반적인 OCR 정보만으로는 충분 X
- $P_{generic}$: "이 meme을 상세히 설명하십시오."

$$R_{generic} = Model_{LLaVA}(M_{memeI}, P_{generic})$$

- 두 번째 단계: LLaVA-7B 모델을 다시 사용하여 첫 번째 단계에서 생성된 답변과 답변 특정 프롬프트 $P_{specific}$ 의 조합으로 "답변 특정 근거"인 $R_{specific}$ 를 생성

- $P_{specific}$ 는 "어떻게 [답변] [재구성된 질문]" 형태
- 질문의 첫 두 단어를 제거하여 구성
- 예) 질문 Q: "이 meme에서 피해를 입은 사람은 누구인가?", 답변 'Joe Biden'
- → "Joe Biden은 이 meme에서 어떻게 피해를 입었는가?"

- Stage 1 - Answer Prediction

- 두 단계의 훈련을 가진 멀티모달 CoT 모델을 구현
- T5-large 모델을 사용하여 QCM→LE 후 QCMG→A의 프롬프트 전략

- 모델은 DETR 모델에서 얻은 임베딩 형태의 시각적 데이터를 제공
- T5 모델의 인코더 스택에 게이트가 있는 교차 주의 레이어를 추가하여 사용

$$H_{\text{use}} = (1 - \lambda) \cdot H_{\text{language}} + \lambda \cdot H_{\text{attn vision}}$$

- λ 는 융합된 이미지 + 텍스트 임베딩의 시그모이드 활성화 출력, H_{language} 텍스트 입력 임베딩, $H_{\text{attn vision}}$: 텍스트 + 비전 교차 주의의 출력

- Rgeneric에서 추가적인 맥락 정보를 제공

1. 텍스트 G를 생성하는 것을 목표로 하는 텍스트 생성 작업으로 처음 5 에포크 동안 MemeMQACorpus에서 미세 조정
2. Yanswer 생성을 위해 미세 조정

- Stage 2 - Explanation Generation

- 이전 단계에서 얻은 답변에 대한 설명 생성을 중심으로 진행
- LLaVA-7B 모델을 다시 사용하여 특정 답변에 대한 합리적인 근거 R_{specific} 를 생성
- 하지만, 예상된 설명의 구조와 간결성이 결여
- R_{specific} 와 질문, 정답과 함께 단일 모달 T5-large 모델에 제공, 텍스트-투-텍스트 생성에 사용
- 2 에포크 동안 미세 조정
- 최종 결과(설명): "답변: [답변] 때문에 [설명]"

5. Experiments

- ARSENAL
 - 다양한 모델 비교
 - 결과는 5회 실행을 통해 평균화 (~~와우 부럽 시간과 비용~~)
- MemeMQA 작업은 두 가지 구성 요소: 답변 예측 및 설명 생성 → 다른 메트릭
 - 답변 예측: 정확도
 - 설명 품질: BLEU-1, BLEU-4, ROUGE-L, METEOR, CHRF, BERTScore

- 베이스라인: 단일 모달(텍스트, 이미지) 및 다중 모달 설정을 포함
- ARSENAL의 강건성 평가

6. Benchmarking MemeMQA

| Type | Models | Accuracy | BLEU-1 | BLEU-4 | ROUGE-L | METEOR | CHRF | BERTScore |
|------|---|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| UM | UM.TEXT.T5 | 0.53 | <u>0.59</u> | <u>0.15</u> | 0.44 | 0.41 | 0.35 | 0.901 |
| | UM.TEXT.GPT3.5 | 0.28 | - | - | - | - | - | - |
| | UM.IMAGE.ViT.BERT.BERT | 0.46 | 0.51 | 0.10 | 0.45 | 0.44 | 0.38 | <u>0.911</u> |
| | UM.IMAGE.BEiT.BERT.BERT | 0.40 | 0.50 | 0.11 | 0.44 | 0.44 | 0.38 | <u>0.909</u> |
| MM | MM.ViT.BERT.BERT | 0.45 | 0.51 | 0.11 | 0.46 | 0.44 | 0.38 | <u>0.911</u> |
| | MM.BEiT.BERT.BERT | 0.44 | 0.48 | 0.09 | 0.45 | 0.45 | 0.39 | 0.910 |
| | MM-CoT (w/o OCR) | 0.59 | 0.58 | 0.13 | 0.53 | 0.50 | 0.47 | 0.891 |
| | MM-CoT | 0.67 | <u>0.59</u> | 0.12 | <u>0.54</u> | <u>0.51</u> | 0.49 | 0.894 |
| | ViLT | 0.43 | - | - | - | - | - | - |
| | •MM-CoT (w/ Lecture) | 0.69 | <u>0.59</u> | 0.13 | 0.54 | <u>0.51</u> | 0.49 | 0.895 |
| | miniGPT4 (ZS) | 0.32 | <u>0.09</u> | 0.00 | 0.14 | <u>0.21</u> | 0.23 | 0.753 |
| | miniGPT4 (FT) | 0.28 | 0.12 | 0.00 | 0.16 | 0.23 | 0.26 | 0.771 |
| | LLaVA (ZS) | - | 0.05 | 0.00 | 0.09 | 0.17 | 0.18 | 0.837 |
| | MM-CoT (QCM _L →A, w/ LLaVA rationales) | 0.66 | 0.59 | 0.12 | <u>0.54</u> | <u>0.51</u> | 0.49 | 0.896 |
| | ARSENAL (w Entity-Specific Rationale) | 0.87 | 0.58 | 0.17 | 0.53 | 0.56 | 0.48 | 0.932 |
| | ★ARSENAL (w Generic Rationale) | 0.87 | 0.63 | 0.19 | 0.55 | 0.56 | 0.46 | 0.934 |
| | $\Delta_{\star \rightarrow \bullet}(\%)$ | 18↑ | 4↑ | 4↑ | 1↑ | 5↑ | 1↓ | 2↑ |

Table 2: Benchmarking results for MemeVQA, comparing the proposed approach vs unimodal and multimodal baselines. Table Footnotes: **highest**, second-highest, •: MM-CoT (w Lecture) – Best Baseline, and ★: ARSENAL (proposed approach). ARSENAL variants – (a). *w Entity-Specific*: Utilizes rationale conditioned upon the answer predicted by the first module; and (b). *w Generic*: Utilizes generic rationale.

- 단일 모달
 - T5 기반의 텍스트 전용 모델
 - 0.53의 정확도로 답변 예측에서 우수한 성능
 - 이미지 및 다중 모달 모델보다 뛰어남
 - 설명은 불완전하고 반복적이며 일관성이 없어 ROUGE-L (0.44), CHRF (0.35), METEOR (0.41) 점수 낮음
 - ViT 모델 단일 모달 이미지
 - 낮은 답변 예측 정확도와 T5 기준과 유사한 유창하지만 반복적인 설명을 제공
- 멀티모달
 - ViT + BERT
 - 답변 예측 정확도: 0.45 (ViT 단독과 유사)

- 설명 품질: ViT 단독 모델보다 약간 더 나음
- BEiT + BERT
 - 답변 예측 정확도: 0.44
 - 설명 품질: BEiT 단독보다 개선됨
- 핵심 인사이트:
 - ViT가 BEiT보다 강력한 성능을 보이며, 단일 모달 및 멀티모달 환경 모두에서 더 견고함
 - 이는 Vicuna 기반 miniGPT-4 및 LLaVA 기반 ARSENAL에도 긍정적인 영향을 줌
- **LLM 기반**
 - miniGPT-4 설명 품질: 구체성 부족, 설명이 너무 길고 비구체적
 - GPT-3.5 답변 예측 정확도: 0.28 / 설명 품질: miniGPT-4와 유사, 구체성 부족 및 평가 점수 저조
- **LLaVA 기반 모델 (ARSENAL)**
 - BERTScore: 0.837 (가장 높은 점수)
 - 설명 품질: 높은 일관성과 구체성 유지, 밈 콘텐츠와의 높은 적합성
 - OCR 텍스트의 중요성: OCR 텍스트 제거 시 정확도 8% 감소
 - 일반적 강의(Generic Lectures, L) 추가 시 정확도 2% 증가
 - MM-CoT 모델과의 비교:
 - MM-CoT 모델: 밈 이해에서 성능 저조
 - ARSENAL (신규 모델):
 - Rationale Generation Module로 인해 더 높은 정확도와 설명 품질 달성
 - 밈 콘텐츠의 맥락적 이해를 강화
- Discussion
 - 60개의 무작위 테스트 샘플에 대한 우리 분석은 답변 품질, 설명 일관성 및 모달리티별 뉘앙스 측면에서 ARSENAL을 다른 방법과 비교

- ARSENAL: LLaVA 접근 방식을 통해 다양한 multimodal리티의 세부사항을 효과적으로 통합하여 추론하고 설명하는 데 뛰어남
- MM-CoT 모델: 구문적 및 문법적 정확성에서 어려움,
 - T5 기반의 텍스트 전용 모델: 종종 일관성이 없고 불완전한 출력을 생성
 - UM.IMG.ViT.BERT.BERT 모델: 문맥화와 정렬에서 문제, 설명은 의미적으로 관련이 있지만 무관
- 이미지 전용 접근 방식과 다중 modal 기준선은 어휘적 편향
- MM.ViT.BERT.BERT 다중 modal 설정은 유창성을 추구하더라도 복잡한 추론에서 실패, 일반적인 설명을 초래

7. Robustness Analysis

- MemeMQA와 같은 작업에 대한 모델의 효능을 나타낼 것으로 기대되는 핵심 요소는 질문/답변 형식 내 변동성에 대한 **강건성**
- **Question Diversification**
 - ARSENAL과 현재의 기준선들이 MemeMQACorpus보다 더 자연스럽게 구성된 질문을 사용하여 성과를 평가
 - 질문 다양성은 Llama-2-7b-chat 모델을 사용하여 각 원본 질문에 대한 다섯 개의 독창적인 변형을 생성함으로써 달성
 - 다양한 질문 형식을 보장하는 것에 대해, 새롭게 다양한 질문에 대해 훈련하고 테스트한 결과, 0.82의 답변 예측 정확도 얻음
 - → ARSENAL이 질문 구조 설정에서 중요한 변형과 다양성을 수용할 수 있음 입증, 다른 모델들이 이러한 변화에 대해 그만큼 강력하지 않다는 것

| Experiment | Q_{div} | Yes/No | None (All) | None (Train) |
|------------------|-----------|--------|------------|--------------|
| UM.TXT.T5 | 0.351 | 0.805 | 0.461 | 0.457 |
| UM.ViT.BERT.BERT | 0.273 | 0.373 | 0.328 | 0.253 |
| MM.ViT.BERT.BERT | 0.341 | 0.295 | 0.474 | 0.438 |
| ARSENAL | 0.818 | 0.769 | 0.692 | 0.721 |

Table 4: Robustness Analysis: (a) Question Diversification (Q_{div}); (b) Confounder Setting (three scenarios).

- **Confounding Analysis**

- **Confounder A – Yes/No 변환**

- 데이터셋을 이진 질문(예/아니오) 형태로 변환(50% 확률).
- "예" 답변: "[정답]이 [재구성된 질문]인가?" 형태로 변경.
- "아니오" 답변: 역할 레이블(role labels) 변경하여 혼란을 유발.

- **Confounder B – 모든 데이터셋에서 'None' 샘플링**

- 전체 데이터셋에서 20%의 답변을 'None'으로 대체.
- 역할 레이블을 바꿔 일관성을 유지하면서도 혼란을 추가.
- 새로운 데이터셋: $M_{new} = \{M, \text{None}\}$
 - 질문: "이 밈에서 비방받는 대상은 조 바이든인가" / 정답: "None"


- **Confounder C – 학습 데이터에서만 'None' 샘플링**

- 훈련 데이터셋에서만 20%의 답변을 무작위로 'None'으로 변경.
- 테스트 데이터는 기존 그대로 유지하여 모델이 'None'에 적응하는지 평가.
- 새로운 데이터셋: $M_{new} = \{M, \text{None}\}$

8. Conclusion and Future Work

- MemeMQA: 밈의 시각적·언어적 의미를 파악하기 위한 질문-응답 시스템 개발

- 새로운 데이터셋인 **MemeMQACorpus**를 구축하여 모델의 **추론 능력(reasoning ability)** 평가
- ARSENAL 프레임워크: 2단계 학습 프레임워크 제안, 기존 언어 모델과 다중 모달 LLM의 성능을 극대화
- 향후 연구 방향
 - 다각적(multi-perspective) 질문 세트 개발을 통해 단순한 QnA를 넘어서는 최적의 다중 모달 솔루션 구축을 목표
 - 더 정교한 질문 유형과 복잡한 데이터셋을 통해 추론 능력 및 일반화 성능 강화 예정
- Limitations
 - LLaVA의 생성 능력: 몇 가지 경우에서 LLaVA의 이론은 서로 다른 키워드의 동시 발생으로 인한 유도 편향 존재
 - LLaVA의 편향된 추론이 부정확한 맥락화로 인해 ARSENAL의 출력을 희석
 - MM-CoT는 아마도 LLM 기반 이론 대신 표준화된 정의를 사용하여 정확하게 답변을 도출



Rationale: In the meme, the **Democratic Party** is portrayed as a joke by using a humorous symbol associated with the party. The image features a green triangle with the words "The Ban," which likely refers to the **Democratic Ban on certain fireworks**, and a party symbol. This visual humor is intended to make the viewer laugh and possibly make a mockery of the political party and its policies. The meme's intention is to create a negative and amusing perception of the **Democratic Party**, which can be considered as an example of political humor or satire. However, it is important to note that humor and satire can be subjective, and different people might find different aspects funny or offensive.

Question : What is vilified in this meme?

Options : ['democratic party', 'green party', 'scott morrison', 'bernie sanders']

MM-COT : The answer is (B) BECAUSE the green party is portrayed as una **unacceptable much restrictive** thes lives.

ARSENAL: **Democratic party** BECAUSE democratic party is depicted as having a **ban on fireworks**.

Figure 8: An example of the error-type committed by ARSENAL (proposed approach) vs. the correct inferencing by the MM-COT based approach.