다중 LLM 응답과 이중 정제를 통한 환각 탐지 데이터 생성 프레임워크*

*유하영[1], 김경현[1], 김영화[1], 권준형[1], 김영빈[1,2]
[1]중앙대학교 AI대학원, [2]중앙대학교 첨단영상대학원
e-mail : {bluebarry37, khyun8072, movie112, dirchdmltnv, ybkim85}@cau.ac.kr

A Dataset Construction Framework for Hallucination Detection via Multi-LLM Responses and Dual-Stage Filtering

*Hayeong Ryu[1], Kyeonghyun Kim[1], Yeonghwa Kim[1], JuneHyoung Kwon[1] and YoungBin Kim[1,2]
[1]Department of Artificial Intelligence, Chung-Ang University
[2]Graduate School of Advanced Imaging Science, Multimedia & Film Chung-Ang University

Abstract

To effectively detect hallucinations in Large Language Model(LLM), it is essential to utilize high-performance detection models in conjunction with reliable training data that includes both faithful and hallucinated outputs. This study introduces a novel framework for synthetic dataset construction, which leverages multiple LLM for response generation and applies a dual-stage filtering process to automatically generate high-quality faithful and hallucinated samples. Experimental results show that a T5-base model fine-tuned on the proposed dataset achieves a 24.5% improvement in F1 score compared to a baseline constructed using GPT-4 responses. Furthermore, the proposed approach enables cost-efficient data acquisition without relying on expensive API calls.

* 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(RS-2021-II211341, 인공지능대학원지원(중앙대학교))을 받아 수행된 연구임.

I. 서론

대규모 언어 모델(Large Language Model, LLM)의 자연어 생성(Natural Language Generation) 능력은 비약적으로 발전하였으나, 여전히 입력과 무관하거나 사실에 반하는 정보를 생성하는 환각(Hallucination) 현상이 지속되고 있다. 이러한 현상은 사용자에게 잘못된 정보를 제공하거나, 신뢰성이 중요한 응용 분야에서 심각한 오류로 이어질 수 있다. 따라서, 생성 응답의 사실성과 정합성을 정밀하게 탐지할 수 있는 환각 탐지 모델(Hallucination Detection Model)의 개발이 요구되고 있다[1].

신뢰할 수 있는 환각 탐지 모델 구축을 위해서는 환각에 대한 명확한 기준이 포함된 고품질 학습 데이터의 확보가 선행되어야 한다. 이에 따라 최근에는 인간의 개입 없이, 생성된 응답에 자동으로 라벨을 부여하는 합성 주석(Synthetic Annotation) 방식이 도입되었다[2]. 대표적인 예로, HaluEval[3]은 응답의 주요 엔티티(Entity)를 다른 값으로 치환하여 환각 응답을 생성하고, 이를 활용해 합성 주석 데이터로 구성한다.

하지만 이러한 방식은 환각의 표현 범위를 엔티티 수준에 국한하므로, 실제 LLM에서 나타나는 미묘하고

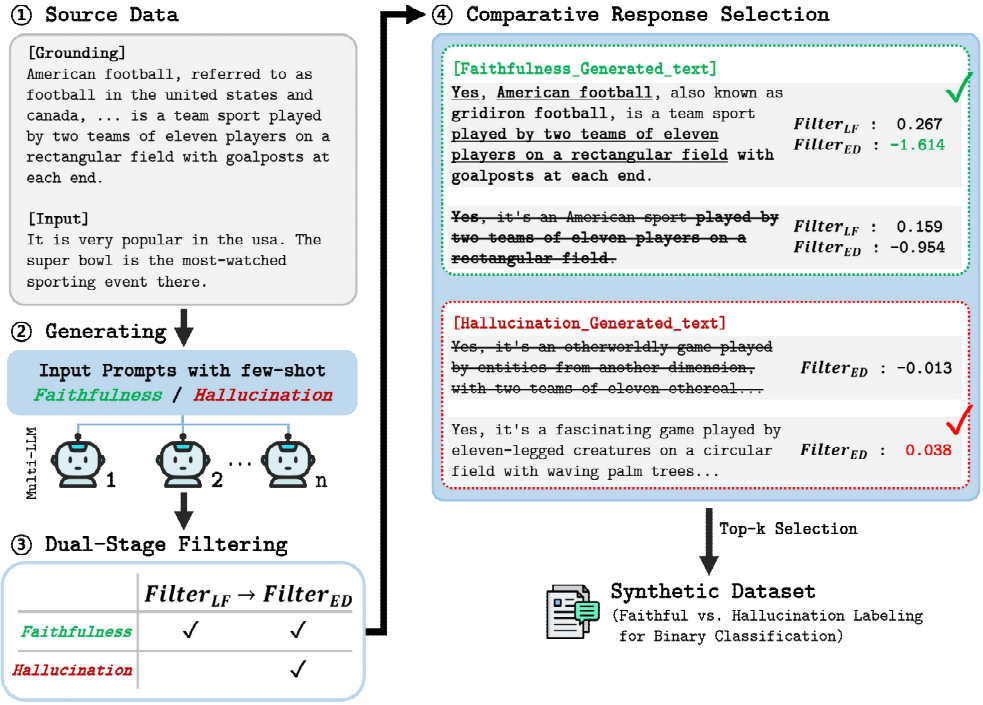

그림 1. 다중 LLM 및 이중 필터링 기반의 합성 데이터셋 구축 프레임워크

복합적인 환각 양상을 반영하기 어렵다는 제약이 있다. 이에 대한 대안으로, GPT-4와 같은 고성능 LLM에 프롬프트(Prompt)를 적용하여 명시적인 규칙 없이도 자연스러운 응답 생성이 가능한 합성 데이터(Synthetic Data) 방식이 제안되었다[4].

그럼에도 불구하고, 단일 고성능 LLM 기반의 생성 방식은 여전히 한계를 지닌다. 단일 모델에 의존할 경우, 모델 고유의 편향이 생성 결과에 그대로 반영되어 표현의 다양성을 충분히 확보하기 어렵다[5]. 또한, 대부분의 고성능 모델이 상용 API 형태로 제공되기 때문에 반복적인 호출에 따른 비용 부담도 발생한다. 이에 본 연구에서는 다중 개방형 LLM을 활용하여 동일 입력에 대해 다양한 응답을 생성하는 합성 데이터 구축 전략을 제안한다. 하지만 이러한 접근은 표현의 다양성을 확보하는 데 효과적이나, 모델 간 표현 방식의 차이와 소형 모델의 높은 환각률로 인해 응답 간 일관성이 저하될 수 있다[6]. 따라서, 생성 데이터의 품질과 일관성을 확보하기 위해, 본 연구에서는 정교한 이중 필터링 절차를 함께 설계하였다.

본 연구의 주요 기여는 다음과 같다:

- 다중 LLM과 이중 필터링 기법을 결합한 새로운 합성 데이터셋 구축 프레임워크를 제안한다.
- 환각 감지 벤치마크(Benchmark) 데이터셋인 BEGIN[7]을 기반으로 합성 데이터를 생성한다. 이후 생성된 응답에 대해 두 단계의 필터링을 도입한다. (1) 입력 및 참조 문장과 응답의 정합성 점수를 계산하여 품질이 낮은 응답을 제거하고, (2) 일관성과 다양성이 균형을 이루는 응답을 선별한다.
- T5-base[8] 모델을 기존 연구인 GPT-4 기반 데이터와 본 연구가 생성한 합성 데이터로 각각 미세 조정한 결과, 후자가 전자 대비 F1 점수에서 24.5% 높은 성능을 기록하였다.

II. 본론

본 장에서는 환각 탐지 모델의 학습을 위한 합성 데이터 구축 프레임워크를 제안한다. 그림 1은 제안한 데이터 생성 절차의 전체 흐름을 시각화한 것이다.

2.1 다중 LLM을 활용한 합성 데이터 생성

환각 탐지 모델의 탐지 성능을 높이기 위해서는 고품질의 학습 데이터가 필요하다. 기존의 합성 데이터 생성 방식은 단일 고성능 LLM에 의존하고 있어, 편향, 비용, 표현의 다양성 측면에서 한계를 지닌다. 따라서 본 연구에서는 다중 개방형 LLM을 활용하여 API 호출 없이 낮은 비용으로 다양한 관점의 데이터를 생성한다. 이 과정에서 사실 기반(Faithfulness) 응답과 환각 응답을 각각 효과적으로 생성할 수 있도록 목적에 특화된 프롬프트를 설계하였다. 사실 기반 생성 프롬프트는

주어진 지식과 대화 이력에만 근거하여 응답을 생성하도록 유도하였고, 환각 생성 프롬프트는 입력 정보와 불일치하는 정보를 일부 수정하거나 삽입하여 응답의 사실성을 저해하도록 설계되었다.

2.2. 이중 필터링을 통한 데이터 정제

다중 개방형 LLM을 활용한 데이터 생성은 모델 간 표현 방식의 차이로 인해 응답 품질의 편차가 발생할 수 있다. 이를 완화하고자, 본 절에서는 이중 필터링 기법을 제안한다.

2.2.1 내용 정합성 및 사실 기반 필터링

$Filter_{LF}$는 생성된 응답 문장 R이 입력 문장 I와 참조 문장 G에 대해 내용 정합성 및 사실 일관성을 충족하는지를 정량적으로 평가한다.

1) 경량 정합성 점수

$L-Score$는 생성된 응답 문장 R과 참조 정보 G 간의 내용 정합성을 엔티티 수준에서 평가한다. $G \cup I$에서 추출한 엔티티 e들의 집합 E는 $E = ENT(G \cup I)$로 정의된다[9]. $ENT(\cdot)$는 개체명 인식(Named Entity Recognition, NER)을 통해 G와 I로부터 의미 있는 정보 요소를 추출한다. 이 중 E에 포함된 엔티티 중 R에 실제로 등장하는 엔티티들의 부분집합을 $R_e = \{e \in E | e \subset R\}$로 정의한다.

최종적으로 $L-Score$는 참조 문장으로부터 추출된 엔티티 수 $|E|$ 대비, 응답에 포함된 엔티티 수 $|R_e|$의 비율로 계산된다[10]. 해당 값이 클수록, 생성된 응답이 참조 문장의 핵심 엔티티를 충실히 반영한 것으로 간주한다.

$$L-Score = \frac{|R_e|}{|E|} \tag{1}$$

2) 사실 충실도 점수

$F-Score$는 응답 문장 R이 입력 문장 I 및 참조 문장 G와 사실적으로 일관되는지를, NLI 기반의 Encoder 모델로 정량 평가한 값이다[11]. 먼저, G와 I의 합집합 $G \cup I$를 문장 단위로 분할하여 집합 $S_G = \{g_1, g_2, \dots, g_n\}$를 구성하고, R 또한 문장 단위로 분할하여 집합 $R = \{r_1, r_2, \dots, r_n\}$로 변환한다. 이후, 각 응답 문장 r에 대하여 문장 g와의 코사인 유사도(Cosine Similarity) $s(g,r) = \cos(v_g, v_r)$를 계산하고, 유사도가 가장 높은 상위 k개의 문장을 $Top-k(r) = \underset{g \in S_G}{argmax}^{(k)} s(g_i, r_i)$로 선별한다. $Top-k(r)$ 문장 쌍에 대해 사전학습된 NLI 모델을 적용하여 함의(Entailment) 확률 p_{ent}와 모순(Contradiction) 확률 p_{con}를 $p(g \Rightarrow r) = f_{NLI}(g,r) = (p_{ent}, p_{con})$로 산출한다. 이를 토대로 사실적 정합성 점수 $\Phi(g,r)$를 다음과 같이 계산한다.

$$\Phi(g,r) = s(g,r)[p_{ent} - p_{con}] \in [-1,1] \tag{2}$$

$\Phi(g,r)$ 값이 클수록 응답 문장 r이 참조 문장 G에 의해 사실적으로 정합함을 의미한다. 전체 응답 R에 대한 $F-Score$는 선택된 상위 k개 문장 쌍들 점수 $\Phi(g,r)$의 평균으로 계산되며, 이 값이 클수록 생성된 응답이 I와 G 간의 높은 일관성을 보인다고 할 수 있다.

$$F-Score = \frac{1}{k} \sum_{(G,R) \in TOP-k} \Phi(g,r) \tag{3}$$

3) 통합 품질 점수

산출된 $L-Score$와 $F-Score$를 가중합하여, 생성된 응답의 전반적인 품질을 평가하는 통합 품질 지표는 다음과 같이 정의한다.

$$Score_{LF} = w \cdot F-Score + (1-w) \cdot L-Score \tag{4}$$

이를 통해 생성된 데이터에서 내용이 부정확하거나 사실성과 정합성이 부족한 응답을 필터링할 수 있다.

2.2.2 일관성 - 다양성 응답 최적화

앞서 제안한 $Filter_{LF}$ 기법은 개별 응답의 품질 선별에는 효과적이나, 응답 집합 내 의미적 중복 및 불균형이 존재할 수 있다. 따라서, 본 논문에서는 GMoA (Gated Mixture of Agents)에서 제안된 Eigen Divergence 기반 필터링을 적용한다[12]. 먼저, 동일한 입력에 대해 다중 LLM이 생성한 총 M개의 응답을 수집하고 각 응답을 문장 임베딩 모델을 통해 고차원 벡터 $z_i \in \mathbb{R}^d$로 변환한다. 변환된 임베딩 벡터 집합 $\{z_1, z_2, \cdots, z_M\}$을 임베딩 행렬 $Z \in \mathbb{R}^{M \times d}$로 변환한 뒤, 공분산 행렬 Σ를 계산한다. 이후 상위 k개의 고윳값 $\{\lambda_1, \lambda_2, \cdots, \lambda_M\}$으로 전체 응답의 일관성을 정량화한 Eigen Score $E(\Sigma)$를 구한다. 여기서 α는 수치적 안정성을 위한 작은 상수이다. 각 응답 r_i이 전체 일관성에 미치는 영향을 평가하기 위해, 해당 응답을 제거한 임베딩 집합 Z_{-i}의 공분산 행렬 Σ_{-i}와 대응하는 $E(\Sigma_{-i})$를 계산한다.

$$E(\Sigma) = \frac{1}{K} \sum_{i=1}^{K} \log(\lambda_i + \alpha) \tag{5}$$

$$E_{-i}(\Sigma_{-i}) = \frac{1}{K-1} \sum_{j=1}^{K-1} \log(\lambda_{-ij} + \alpha) \tag{6}$$

이를 기반으로 r_i에 대한 Eigen Divergence D_i는 다음과 같이 정의한다.

$$D_i = E(\Sigma) - E_{-i}(\Sigma_{-i}) \tag{7}$$

D_i가 클수록 의미적 일관성을 해치는 응답이며, 작을수록 다양성과 일관성 간의 균형을 이루는 일관성이 높은 응답이다.

2.3. 학습 데이터셋 구축 프레임워크

본 절에서는 2.1절의 합성 데이터 생성 전략과 2.2절의 이중 필터링 기법을 바탕으로 최종 학습 데이터셋을 구축하는 단계별 프레임워크를 설명한다.

입력 샘플 하나당 사실 기반 응답 9개, 환각 응답 9개를 생성하여 각각 응답 후보군으로 구성한다. 이후 첫 번째 필터링 단계인 $Filter_{LF}$에서는 사실 기반 응답 후보군에만 $Score_{LF}$를 수행하고, 정합성 점수가 가장 낮은 3개를 제거한 후 상위 6개를 선별한다. 이때, 환각 응답은 의도된 정보 변형을 포함하고 있어, 사실성 및 정합성 기준이 오히려 표현 다양성을 제한할 수 있으므로 이 단계에서 제외된다. 이는 이후 3.3절에서 실험적으로 검증한다.

다음으로 $Filter_{LF}$를 통해 선별된 사실 기반 응답 6개와 생성된 환각 응답 후보군 9개에 대하여, 각각 $Filter_{ED}$를 적용하여 D_i를 산출한다. 각 입력에 대해 사실 기반 응답은 오름차순으로 상위 3개를 환각 응답은 내림차순으로 상위 3개를 선별하고, 최종적으로 입력 샘플 하나당 사실 기반 응답 3개와 환각 응답 3개로 구성된 데이터셋을 구축한다. 위 절차로 구성된 데이터셋은 응답의 사실성 여부를 판단하는 이진 분류(Binary Classification) 기반의 환각 탐지 모델을 학습하는 데 사용된다.

Ⅲ. 실험 구성 및 결과

3.1. 데이터셋

학습을 위한 실험 데이터셋은 다음과 같이 구성한다. 먼저, 원천 데이터(Source Data)는 Wikipedia 등 다양한 문서를 기반으로 구성된 BEGIN 데이터셋을 활용한다. 본 논문에서는 이 데이터에서 추출한 참조 문장 G와 입력 문장 I 쌍을 하나의 샘플로 사용한다.

각 샘플에 대해서는 세 가지 개방형 LLM (Llama-3-8B-Instruct[13], Qwen2.5-7B-Instruct[14], Mistral-7B-Instruct-v0.2[15])을 활용하여 초기 응답 데이터셋을 생성했다. 생성된 초기 데이터에 대하여 본 논문 2.2절에서 제안하는 이중 필터링 전략을 적용하여 최종 학습 데이터셋을 구성한다.

3.2. 실험 환경

환각 탐지 모델은 223M 파라미터 규모의 T5-base 인코더-디코더 구조를 기반으로 하며, 본 연구에서 구축한 합성 데이터셋으로 미세조정되었다. 학습에는 Learning Rate 1e-4, Batch Size 4, Epochs 1을 적용하였다. 모델의 효율성을 높이기 위해 Low-Rank Adaptation[16]을 활용하였으며, 이때 Rank $r=16$, Scaling Factor $\alpha=32$, 최적화 대상은 Query q와 Value v 모듈로 설정했다.

$Filter_{LF}$의 $L-Score$ 계산에는 NER 도구인 spaCy[17]를 활용하여 주요 엔티티를 추출하였고, $F-Score$ 계산에는 DeBERTa-v3-base[18] 모델을 통해 의미적 정합성을 평가한다. 두 점수는 $Filter_{LF}$의 수식에 따라 통합되며 이때 가중치는 $w=0.7$로, $F-Score$에서 선택하는 문장 쌍 상위 k는 3으로 설정하였다.

모델 성능 평가는 F1 점수를 기준으로 수행했다. Faithfulness-F1은 모델이 사실에 기반한 응답을 정확히 식별하는 능력을, Hallucination-F1은 환각 응답을 정밀하게 탐지하는 능력을 평가한다. Faithfulness-F1, Hallucination-F1 두 지표 모두 각 라벨에 대한 F1 Score를 산출한 값이다.

3.3 실험 결과

LLM	F1 Score	Accuracy
GPT-4	0.493	0.555
Multi-LLM (ours)	**0.614**	**0.657**

표 1. 단일 LLM(GPT-4)과 다중 LLM 기반 환각 탐지 성능 비교

표 1은 기존 연구인 GPT-4 기반 데이터셋[5]과 본 연구의 데이터셋을 활용해 학습한 환각 학습탐지 모델 간의 성능을 비교한 결과이다. 실험 결과, GPT-4 기반 데이터로 학습한 모델은 F1 점수 0.493, 정확도 0.555를 기록한 반면, 제안된 데이터셋으로 학습한 모델은 F1 점수 0.614, 정확도 0.657을 달성하였다. 이는 기존 성능 대비 24.54% 개선된 결과로, 제안한 전략이 의미적 정합성과 표현적 다양성을 효과적으로 반영한 응답을 선별함으로써 모델 성능을 실질적으로 개선했음을 보여준다. 또한, GPT-4 API를 활용해 합성 데이터셋의 응답을 생성할 경우 응답 당 약 0.006 USD의 비용이 발생하는 반면, 본 연구는 다중 개방형 LLM을 활용함으로써 별도의 API 비용 없이 합성 데이터를 구축할 수 있다.

다음으로 표 2는 다중 LLM으로 생성된 합성

LLM + Method	F1 Score			
	Macro	Micro	Faithfulness	Hallucination
GPT-4	0.493	0.555	0.315	0.670
Multi-LLM	0.462	0.513	0.296	0.628
+ Filter_{LF}	0.478	0.511	0.348	0.608
+ Filter_{ED}	0.467	0.483	0.375	0.558
+ $\text{Filter}_{ED} \rightarrow \text{Filter}_{LF}$	0.563	0.610	0.421	0.706
+ $\text{Filter}_{LF} \rightarrow \text{Filter}_{ED}$	**0.614**	**0.657**	0.487	**0.742**
+ $\text{Filter}^{*}_{LF} \rightarrow \text{Filter}_{ED}$	0.604	0.630	**0.502**	0.705

표 2. 필터링 전략에 따른 환각 탐지 성능 변화

데이터에 다양한 필터링 전략을 적용했을 때, 환각 탐지 모델의 성능 변화를 분석한 지표이다. 실험 결과, 본 연구에서 제안한 $Filter_{LF} \rightarrow Filter_{ED}$의 경우 모든 지표에서 가장 우수한 성능을 나타냈다. 또한, 본 논문에서 제안한 모든 필터링 기법은 기존 연구인 GPT-4보다 Macro F1 점수에서 더 높은 수치를 기록하였다. 생성된 데이터를 무작위로 선별해 미세조정한 Multi-LLM은 가장 낮은 성능을 보였으나, 여기에 $Filter_{LF}$ 또는 $Filter_{ED}$를 단독으로 적용한 경우 각각 +0.016, +0.005를 보였다. 한편, 필터링 적용 순서를 반대로 한 $Filter_{ED} \rightarrow Filter_{LF}$는 -0.051 성능을 기록했다. 또한 $Filter^{*}_{LF} \rightarrow Filter_{ED}$ 조합은 환각 응답에도 $Filter_{LF}$를 확장한 조합으로, 실험 결과 유의미한 성능 향상은 보이지 않았다. 이는 환각 응답이 본래 의도된 정보 변형을 포함하고 있기 때문에, 해당 응답에 사실성 및 정합성 기준을 적용할 경우 표현 다양성이 제한되어 오히려 모델 학습에 부정적인 영향을 줄 수 있음을 시사한다.

결론적으로, 다중 LLM 기반 합성 데이터는 적절한 필터링 과정을 통해 고품질의 데이터를 구성하여, 환각 탐지 모델의 성능을 유의미하게 향상할 수 있다. 특히, $Filter_{LF}$와 $Filter_{ED}$를 순차적으로 적용하는 이중 필터링 전략이 가장 효과적인 접근임을 확인하였다.

Ⅳ. 결론 및 향후 연구 방향

본 논문은 다중 개방형 LLM과 이중 필터링을 결합한 합성 데이터 구축 프레임워크를 제안하였다. 실험 결과, 제안된 데이터셋으로 학습한 환각 탐지 모델은 기존 GPT-4 기반 데이터셋 대비 F1 점수를 향상시키며 API 비용 없이도 우수한 성능을 달성할 수 있음을 입증하였다.

그러나 이중 필터링 메커니즘은 참조 문장에 기반하여 작동하고 NLI와 같은 외부 평가 모델에 의존하도록 설계되어있다. 이에 특정 모델이나 도메인에 따라 성능이 좌우될 수 있어 범용적인 적용에는 한계가 있다. 향후 연구로는 도메인 적응성과 모델 다양성에 따른 데이터셋 구축 또는 입력 특성에 따라 동적으로 조정되는 적응형 필터링 메커니즘의 개발이 주요 과제로 남아 있다.

참고문헌

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., and Zhang, Y., “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, pp. 1 – 38, Dec. 2023.

[2] Padhi, I., McCandlish, S., and Amodei, D., “Granite guardian,” arXiv preprint arXiv:2412.07724, Dec. 2024.

[3] Li, J., Cheng, X., Zhao, X., Nie, J. Y., and Wen, J. R., “HaluEval: A large-scale hallucination evaluation benchmark for large language models,” in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6449 – 6464, Singapore, January 2023.

[4] Zhang, D., Gangal, V., Lattimer, B., and Yang, Y., “Enhancing hallucination detection through perturbation-based synthetic data generation in system responses,” in *Proc. Conf. Association for Computational Linguistics (ACL)*, pp. 13321 – 13332, Bangkok, Thailand, July 2024.

[5] Lauscher, A., and Glavaš, G., “How much do LLMs hallucinate across languages? On multilingual estimation of LLM hallucination in the wild,” arXiv preprint arXiv:2502.12769, Feb. 2025.

[6] Wang, C., Tang, F., Zhang, Y., Wu, T., and Dong, W., “Towards harmonized regional style transfer and manipulation for facial images,” Computational Visual Media, 9, 2, 351 – 366, April 2023.

[7] Dziri, N., Rashkin, H., Linzen, T., and Reitter, D., “Evaluating attribution in dialogue systems: The BEGIN benchmark,” Transactions of the Association for Computational Linguistics, 10, – , 1066 – 1083, December 2022.

[8] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J., “Exploring the limits of transfer learning with a unified text-to-text transformer,” Journal of Machine Learning Research, 21, 140, 1 – 67, January 2020.

[9] Kryscinski, W., McCann, B., Xiong, C., and Socher, R., “Evaluating the factual consistency of abstractive text summarization,” in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332 – 9346, Online, November 2020.

[10] Nan, F., Nallapati, R., Wang, Z., Nogueira dos Santos, C., Zhu, H., Zhang, D., McKeown, K., and Xiang, B., “Entity-level factual consistency of abstractive text summarization,” in *Proc. Conf. European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2727 – 2733, Online, April 2021.

[11] Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A., “SummaC: Re-visiting NLI-based models for inconsistency detection in summarization,” Transactions of the Association for Computational Linguistics, 10, – , 163 – 177, March 2022.

[12] Abdulaal, A., Jin, C., Montaña-Brown, N., Gema, A. P., Castro, D. C., Alexander, D. C., Teare, P. A., Diethe, T., Oglic, D., and Saseendran, A., “Balancing act: Diversity and consistency in large language model ensembles,” in *Proc. Conf. International Conference on Learning Representations (ICLR)*, –, Vienna, Austria, May 2025.

[13] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[14] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[15] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[16] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W., “LoRA: Low-rank adaptation of large language models,” in *Proc. Conf. International Conference on Learning Representations (ICLR)*, –, Virtual, April 2022.

[17] Park, E. L., Hagiwara, M., Milajevs, D., and Tan, L. (eds.), “Proceedings of the Workshop for NLP Open Source Software (NLP-OSS),” in Proc. Workshop for NLP Open Source Software (NLP-OSS), –, Melbourne, Australia, July 2018.

[18] https://huggingface.co/microsoft/deberta-v3-base