KoCheckGPT: 한국어 초거대언어모델 작성 글 판별기

강명훈[1°], 이정섭[1], 이승윤[1], 홍성태[1], 박정배[2*] 임희석[1,2*]
[1]고려대학교 컴퓨터학과, [2]Human-inspired AI 연구소
{chaos8527, omanma1928, dltmddbs100, ghdchlwls123, insmile, limhseok}@korea.ac.kr

KoCheckGPT: Korean LLM written document detector

Myunghoon Kang[1°], Jungseob Lee[1], Seungyoon Lee[1], Seongtae Hong[1], Jeongbae Park[2*] Heuiseok Lim[1,2*]
[1]Department of Computer Science and Engineering, Korea University, [2]Human-inspired AI Research

요 약

초거대언어모델(LLM)의 도래에 따라 다양한 과업들이 도메인 관계 없이 제로샷으로 추론이 가능해짐에 따라서 LLM이 다양한 산업분야에 적용되고 있다. 대표적으로 ChatGPT와 GPT-4는 상용 API로 서비스를 제공하여 용이한 서비스 접근으로 다양한 이용층을 끌어들이고 있다. 그러나 현재 상용 API로 제공되고 있는 ChatGPT 및 GPT-4는 사용자의 대화 내역 데이터를 수집해 기업의 보안 문제를 야기할 수 있고 또한 생성된 결과물의 환각 문제로 인한 기업 문서의 신뢰성 저하를 초래할 수 있다. 특히 LLM 생성 글은 인간의 글과 유사한 수준으로 유창성을 확보한만큼 산업현장에서 LLM 작성 글이 판별되지 못할 경우 기업 활동에 큰 제약을 줄 수 있다. 그러나 현재 한국어 LLM 작성 글 탐지 서비스가 전무한 실정이다. 본 논문에서는 한국어 초거대언어모델 작성 글 판별기: KoCheckGPT 를 제안한다.KoCheckGPT는 산업현장에서 자주 사용되는 문어체, 개조식 글쓰기로 작성된 문서 도메인을 목표로 하여 글 전체와 문장 단위의 판별 정보를 결합하여 주어진 문서의 LLM 작성 여부를 효과적으로 판별한다. 다국어 LLM 작성 글 판별기 ZeroGPT와의 비교 실험 결과 KoCheckGPT는 우수한 한국어 LLM 작성 글 탐지 성능을 보였다.

주제어: 초거대언어모델, 인공지능작성글판별, 모델 설계

1. 서론

지시문을 따르도록 학습하는 Instruction Tuning과 파라미터 크기의 증대로 초거대언어모델(LLM)시대가 도래했다. Instruction Tuning은 단순 x, y의 입력과 출력 형태로 제공되었던 기존 과업을 지시문 x_I에 따른 출력 y를 생성하는 학습 방법을 말한다 [1]. 사용자의 의도에 맞는 출력물을 생성하도록 학습하는 Instruction Tuning은 과업별 추가 학습 없이도 과업 특화 미세조정 모델의 성능과 비견된다는 점에서 LLM의 일반화 성능을 증대시키는 주요 요인으로 알려져 있다 [2, 3]. 이에 다양한 분야에서 과업 특화 미세조정 모델을 학습하지 않고 LLM으로 문제를 해결하고자 하는 노력들이 등장했다. ChatGPT [4]와 GPT-4 [5]는 대표적인 Instruction Tuning으로 학습된 LLM으로 상용 API 제공을 통해 학계 및 일반 이용자들의 꾀어내고 있다[1].

한편, 사용자의 의도에 부합하는 생성 결과물 출력과 도메인 관계 없이 다양한 과업에 적용할 수 있다는 LLM의 특성에 기인해 산업 현장에서의 LLM 도입이 증가하고 있다. 특히 보고서 및 에세이를 작성하는 사무영역[2]에서부터 마케팅 및 광고 카피라팅 같은 창의 영역까지[3]까지 LLM의 사용범위는 산업 전반에 확대되고 있다. 또한 LLM이 업무 영역에 결부될 때 적절한 보조 소프트웨어가 지원될 경우 최대 56%의 영역에서 업무 효율성이 증대될 수 있다는 연구결과도 존재한다 [6].

그러나 현재의 상용 API 형태로 제공되는 LLM을 산업현장에서 사용하는 데에는 다양한 제약 사항이 존재한다. 먼저 기업의 정보 유출문제, 보안 문제다. 현재 상용 API로 제공되는 ChatGPT 및 GPT-4는 모델의 추가 학습을 목적으로 사용자가 LLM과 주고 받은 대화 내역 데이터를 수집한다[4]. 따라서 ChatGPT, GPT-4를 일상 업무에 사용할 경우 기업의 정책이나 대외비, 혹은 개인정보가 기입된 입력은 OpenAI에 제공되어 기업의 보안 문제를 야기할 수 있다. 다음으로 생성된 문서의 환각(Hallucination) 문제다. 사람의 선호를 보상체계로 삼는 강화학습을 이용하는 Instruction Tuning의 방법론의 근원적 문제 [7]와 임의적인 학습 데이터 선정과정으로 인한 입력 출력 데이터의 불일치 문제 [8] 등으로 인하여 LLM은 생성한 글의 환각문제로부터 자유롭지 못하다. 이러한 상황에서 적절힌 조치 없이 LLM을 이용하여 업무 문서 작성 및 수정을 지시할 경우 사실과 다른 내용으로 업무 수행과정과 그 결과물에 큰 오류가 발생할 수 있다. 마지막으로 문서의 신뢰성 문제이다.

*교신저자(Corresponding author)

[1]https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[2]https://www.cbsnews.com/news/chatgpt-chatbot-tiktok-ai-artificial-intelligence/

[3]https://www.fastcompany.com/90833253/ryan-reynolds-used-chatgpt-to-make-a-mint-mobile-ad-and-the-results-were-mildly-terrifying

[4]https://openai.com/policies/terms-of-use

인간의 글과 유사한 수준으로 유창하게 글을 작성하는 LLM의 능력을 고려할 때 작성물에 존재하는 환각 문제가 유창성으로 판별되기 어려운 점이 존재한다 [9, 10]. 따라서 LLM의 보급 이후로 기업 작성 문서의 진위 여부 판단 및 신뢰성 보장 문제가 대두된다.

이러한 LLM의 한계에도 불구하고 광범위하게 사용되는 사회환경을 고려할 때 산업현장에 LLM을 효과적으로 도입하기 위해서는 LLM 생성 결과물의 진위 여부를 판별 및 탐지하는 시스템이 요구된다. 영어의 경우 OpenAI에서 자체 제공하는 AI classifier[5] 및 ZeroGPT[6]등의 판별 서비스가 존재하나 한국어의 경우 LLM 탐지 서비스가 전무한 실정이다. 이러한 문제를 해결하기 위해 본 논문은 한국어 초거대언어모델 작성 글 판별기: KoCheckGPT 를 제안한다. KoCheckGPT는 LLM이 글을 생성할 때 사용하는 어휘, 어구의 사용 패턴을 파악하는 방식으로 학습된 LLM 작성 글 판별기이다. KoCheckGPT는 산업현장에서 자주 사용되는 문어체, 개조식 글쓰기로 작성된 문서 도메인을 목표로 하며 글 전체와 문장 단위의 판별 정보를 결합하여 주어진 문서의 LLM 작성 여부를 판별하게 된다. 본 논문은 인간이 작성한 글과 LLM이 작성한 글을 정확하게 판별하기 위해 공개된 한국어 데이터셋을 활용해 인간의 작성글을 패러프레이징 하도록 LLM에게 프롬프팅하여 학습 및 평가 데이터셋을 구축했다. 이를 바탕으로 KoCheckGPT를 학습하여 ZeroGPT와 비교실험을 진행한 결과 한국어 LLM 작성 글 판별 정확도, F1-score 및 모든 분류 측정 항목에서 우수한 성능을 달성했다.

2. 관련 연구

2.1 AI 생성 글 판별기

생성형 언어모델의 발달로 인간이 작성한 글과 AI가 작성한 글의 구분이 어려워지자 AI가 작성한 글을 판별할 수 있는 판별기 연구가 최근 확장되고 있다. 기존 연구로는 BERT [11]와 GPT-2 [12]의 토큰 생성 확률 및 top-k 생성확률을 이용하여 주어진 문서의 단어별 이상치를 시각화 하는 GLTR [13]이 존재한다. 또 다른 연구로는 LLM의 작성글을 판별하기 위한 연구로 원문이 p모델로 작성되었는지 확인하기 위해 T5 [14]로 원문을 훼손하고 p모델로 훼손된 글과 원문의 로그 확률을 비교하여서 LLM작성 글을 판별하는 DetectGPT [15]가 있다.

2.2 생성 글 품질 평가

한편 AI가 생성한 글 자체의 품질을 평가하는 연구들이 존재한다. LLM이 생성한 글과 원본의 글과 비교하여 생성된 글에서 원본의 글을 잘못 표상한 오류 타입을 13가지로 분류하여 제시하는 연구 [16]와 생성 글의 일관성을 확인하기 위한 측도로 질의 생성(QG)와 질의 응답(QA)를 활용한 DecompEval [17] 연구가 존재한다. 또한 GPT-3를 이용하여 n개의 샘플링 글의 일치도를 바탕으로 환각 정도를 파악하는 SelfCheckGPT [18] 연구가 있다. 하지만 현재 LLM이 한국어로 작성한 글에 대한 품질 평가 및 인간 작성 글과의 구분을 위한 판별기 개발은 미진한 실정이다.

[5]https://platform.openai.com/ai-text-classifier
[6]https://www.zerogpt.com/

Prompt Objective	Text
System 1	아래의 ‘Document‘를 당신의 스타일로 페러프레이즈 하여 동일한 의미의 document로 변환해주세요. 단, 당신이 작성한 document만 출력해야합니다. {document}
System 2	아래의 ‘Document‘를 당신의 스타일의 새로운 document를 작성하세요. 단, 당신이 작성한 ‘Document‘는 이전의 ‘Document‘와 상관 없어야 하며, 당신이 작성한 document만 출력해야합니다. {document}
System 3	아래의 ‘Document‘를 당신의 스타일의 보고서 형식의 document로 페러프레이즈 해주세요. 단, 당신이 작성한 document만 출력해야합니다. {document}
System 4	아래의 ‘Document‘를 바탕으로 새로운 보고서를 만들어주세요. 단, 당신이 작성한 새로운 보고서만 출력해야합니다. {document}

표 1. 데이터셋 구축용 프롬프트 예시

3. 제안 방법

제안하는 KoCheckGPT는 한국어로 작성된 LLM 생성 글을 판별하는 이진 분류기이다. KoCheckGPT는 LLM이 사용하는 언어 패턴을 문장 단위, 글 전체 단위 등 2가지 단위로 파악한 뒤 해당 정보를 결합하여 문어체, 개조식으로 작성된 글의 LLM 작성 여부를 판별한다. 본 논문에서는 KoCheckGPT를 학습하기 위한 데이터셋 구축과 이를 이용한 KoCheckGPT 학습방법을 제안한다.

3.1 데이터셋 구축

LLM이 작성한 글은 유창성이 인간의 글과 유사한 수준으로 발전된 점을 고려할 때 LLM 작성 글 판별기의 학습 및 검증 데이터셋은 1)인간의 글과 매우 유사하면서 2)일부의 구조적, 언어적 차이가 존재해야한다. 1)은 실제 판별기가 맞닿을 검증 환경과 연관이 있으며 2)는 LLM의 언어 패턴과 인간의 언어 패턴의 구분 신호를 학습하기 위함이다. 이러한 LLM 작성 글 판별기의 요구 명세를 바탕으로 KoCheckGPT 학습 및 검증 데이터셋을 구축했다. KoCheckGPT 학습 및 검증 데이터셋은 대표적인 상용 API로 제공되는 ChatGPT를 이용하여 제작되

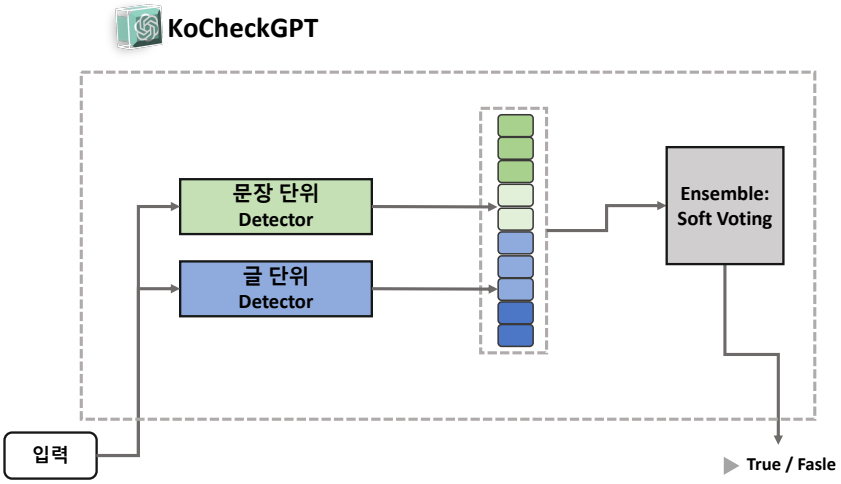

그림 1. KoCheckGPT 구조도

었다. 공개된 한국어 데이터셋을 인간이 작성한 글로 선정하고 해당 글의 일부 내용을 변형 및 패러프레이징 하도록 ChatGPT에게 프롬프트로 지시를 하여 데이터셋을 구축했다. 데이터셋 구축 과정에서 사용된 프롬프트는 표 1에 나타나 있다. 제시된 프롬프트는 LLM의 언어 패턴이 잘 드러나도록 '당신의 스타일로' 키워드와 기존의 문서의 의미를 유지하도록 '동일한 의미의' 키워드가 포함되도록 구성했다. System 1,3 프롬프트는 문장 단위의 데이터셋 구축을 위함이며 System 2,4, 글 전체 단위 데이터셋 구축을 위함이다.

3.2 KoCheckGPT

제안하는 KoCheckGPT의 구조는 그림 1과 같다. KoCheckGPT는 문서 D_i 입력으로 받아 문장 단위, 글 단위별 판별기(Detector)에서 추출된 두 로짓값을 소프트 보팅(Soft Voting) 방법으로 앙상블하여 주어진 문서의 LLM 작성 여부를 판별하게 된다. 먼저 문장 단위 판별기 Sent_{detect}, 글 단위 판별기 Doc_{detect}는 3.1에서 구축된 문장 단위, 글 단위 판별 데이터셋을 활용해 사전학습언어모델을 미세조정한다. 이후 주어진 문서를 두 판별기에 투과하여 각 분류기에서 로짓 값을 추출한다. 마지막으로 두 로짓 값에 반영비 파라미터 λ를 고려한 가중합을 취한 뒤 소프트 보팅 앙상블 방법을 이용하여 최종 분류를 하게 된다. 최종 판별식은 아래와 같다.

$$\hat{y}_i = \text{argmax}(\lambda \text{Sent}_{detect}(D_i) + \lambda \text{Doc}_{detect}(D_i)) \in \{0, 1\} \quad (1)$$

4. 실험

4.1 실험환경

KoCheckGPT를 학습 및 검증하기 위해 구축한 데이터셋과 이를 실험한 환경은 아래와 같다. 먼저 KoCheckGPT를 학습

	학습 데이터	검증 데이터
행정문서대상기계독해	17,042	63
에세이글평가데이터/글짓기	5,780	6
에세이글평가데이터/대안제시	4046	0
에세이글평가데이터/설명글	2442	13
에세이글평가데이터/주장	3022	1
논문자료 요약	7,914	10
요약문및레포트생성	15,074	7
문서 수	55,320	100

표 2. KoCheckGPT 학습 및 검증 데이터셋 통계정보

및 검증하기 위한 데이터셋의 요약 통계가 표 2에 나타나 있다. 데이터셋은 AI hub에 공개된 행정 문서 대상 기계독해 데이터[7], 에세이 글 평가 데이터[8], 논문자료 요약[9], 요약문 및 레포트 생성 데이터[10] 등 개조식과 문어체로 작성된 글을 원천 데이터셋으로 삼았다. 학습 데이터의 경우 원천 데이터셋 대비 전체 데이터셋 비율에 따라서 계층적 샘플링을 이용해 샘플을 추출했으며 검증 데이터는 랜덤 샘플링을 통해 데이터셋을 구성했다. 레이블링은 사람이 작성한 경우 레이블은 0, ChatGPT가 패러프레이징 하거나 새로 작성한 문서의 경우는 1로 진행했으며 레이블별 비율은 1:1이다.

KoCheckGPT는 klue/roberta-large 사전학습 모델 [19]을

[7] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=569
[8] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=545
[9] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=90
[10] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=realm&dataSetSn=582

Model	Accuracy	F1	Recall	Precision
ZeroGPT	56.00	50.00	56.00	61.00
KoCheckGPT	**91.00**	**89.80**	**88.00**	**91.67**

표 3. ZeroGPT, KoCheckGPT 비교실험 결과

학습 데이터셋을 활용하여 미세조정 했다. 미세조정에 사용한 하이퍼 파라미터는 batch size = 32, epochs = 3, learning rate = 5e-5, num warmup steps = 500, max length = 256으로 설정했으며 RAM 378GB, NVIDIA Quadro RTX 8000(48GB) GPU 컴퓨팅 자원을 활용하여서 학습을 진행했다. 비교 실험 대상은 영어 및 다국어 LLM 작성 글 판별기인 ZeroGPT[11]이며 이진 분류 성능을 정확도(Accuracy), F1-score(F1), 재현율(Recall), 정밀도(Precision)기준으로 평가했다.

4.2 실험 결과

표 3에 비교 실험 결과가 정리되어 있으며 각 모델의 이진분류 성능이 Accuracy, F1, Recall, Preicsion으로 표현되어 있다. 베이스라인 모델인 ZeroGPT의 Accuracy와 F1-score는 각각 56.00, 50.00로 random sampling의 결과와 근접한 정도로 낮은 성능을 보였으나 KoCheckGPT는 91.00, 89.80으로 우수한 LLM 작성 글 판별 성능을 보였다. 특히 ZeroGPT의 경우 주어진 한국어 문서를 대부분 LLM이 작성한 글, 즉 1로 예측하는 경향이 두드러져 한국어 작성 문서의 판별능력이 매우 떨어지는 것을 확인했다. 반면 KoCheckGPT는 페러프레이징된 한국어 데이터셋을 문장, 글 단위의 분류결과를 효과적으로 앙상블하여서 ZeroGPT대비 우수한 판별 성능을 보여준다. 이는 향후 도메인을 확장하여 추가 학습을 진행할 경우 모델의 일반화 능력을 보다 확보할 수 있을 것으로 기대된다.

5. 결론

본 논문은 문장, 글 전체 단위의 판별 정보를 앙상블하여 효과적으로 한국어 LLM 작성글을 판별하는 KoCheckGPT를 제안한다. 또한 KoCheckGPT를 학습 및 검증하기 위한 데이터셋도 구축하여서 향후 한국어 LLM 작성글을 판별 연구에도 기여했다. KoCheckGPT는 다국어 LLM 작성 글 판별 모델 ZeroGPT대비 우수한 한국어 LLM 작성글 판별 성능을 보였으며 인간 작성 글과 LLM 작성 글 두 경우 모두에서 강건한 분류 성능을 보인다. 다만 본 논문의 초점은 사무직 산업현장에서 쓰이는 문어체, 개조식 글쓰기 형태의 LLM 작성 글 판별을 목표로 하기 때문에 다른 도메인 문서에 대해서는 판별 성능을 단정하기 어렵다. 향후 연구에서는 다중 도메인 문서에 대한 데이터셋 구축과 추가 학습을 진행하여 문서 도메인에 강건한 한국어 LLM 작성글 판별기를 제안할 계획이다.

[11] https://www.zerogpt.com/

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2023-2018-0-01405). 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기술기획평가원의 지원을 받아 수행된 연구임 (No. 2020-0-00368, 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발). 이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425).

참고문헌

[1] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *International Conference on Learning Representations*, 2021.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27 730–27 744, 2022.

[3] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," *ACM Computing Surveys*, 2022.

[4] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[5] R. OpenAI, "Gpt-4 technical report," *arXiv*, pp. 2303–08 774, 2023.

[6] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "Gpts are gpts: An early look at the labor market impact potential of large language models," *arXiv preprint arXiv:2303.10130*, 2023.

[7] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *arXiv preprint arXiv:2307.15217*, 2023.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1–38, 2023.

[9] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, May 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.229

[10] C. Zhou, G. Neubig, J. Gu, M. Diab, F. Guzmán, L. Zettlemoyer, and M. Ghazvininejad, "Detecting hallucinated content in conditional neural sequence generation," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1393–1404, Aug. 2021. [Online]. Available: https://aclanthology.org/2021.findings-acl.120

[11] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of naacL-HLT*, Vol. 1, p. 2, 2019.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, Vol. 1, No. 8, p. 9, 2019.

[13] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical detection and visualization of generated text," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 111–116, Jul. 2019. [Online]. Available: https://aclanthology.org/P19-3019

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[15] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "Detectgpt: Zero-shot machine-generated text detection using probability curvature," *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., Vol. 202, pp. 24 950–24 962, 2023. [Online]. Available: https://proceedings.mlr.press/v202/mitchell23a.html

[16] K. Murugesan, S. Swaminathan, S. Dan, S. Chaudhury, C. Gunasekara, M. Crouse, D. Mahajan, I. Abdelaziz, A. Fokoue, P. Kapanipathi, S. Roukos, and A. Gray, "MISMATCH: Fine-grained evaluation of machine-generated text with mismatch error types," *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4485–4503, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.findings-acl.274

[17] P. Ke, F. Huang, F. Mi, Y. Wang, Q. Liu, X. Zhu, and M. Huang, "DecompEval: Evaluating generated texts as unsupervised decomposed question answering," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9676–9691, Jul. 2023. [Online]. Available: https://aclanthology.org/2023.acl-long.539

[18] P. Manakul, A. Liusie, and M. J. Gales, "Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models," *arXiv preprint arXiv:2303.08896*, 2023.

[19] S. Park, J. Moon, S. Kim, W. Cho, J. Han, J. Park, C. Song, J. Kim, Y. Song, T. H. Oh, J. Lee, J. Oh, S. Lyu, Y. Jeong, I. Lee, S. Seo, D. Lee, H. Kim, M. Lee, S. Jang, S. Do, S. Kim, K. Lim, J. Lee, K. Park, J. Shin, S. Kim, E. L. Park, A. Oh, J. Ha, and K. Cho, "KLUE: korean language understanding evaluation," *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, J. Vanschoren and S. Yeung, Eds., 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/98dce83da57b0395e163467c9dae521b-Abstract-round2.html