어체에 따른 초거대언어모델의 한국어 환각 현상 분석

문현석†, 어수경†, 서재형†, 박찬준§, 허윤아‡*, 임희석†‡*
고려대학교 컴퓨터학과†, Upstage§, Human-inspired AI 연구소‡
{glee889, djtnrud, seojae777, limhseok}@korea.ac.kr†
chanjun.park@upstage.ai§, yj72722@korea.ac.kr‡

Empirical Study on the Hallucination of Large Language Models Derived by the Sentence-Closing Ending

Hyeonseok Moon†, Sugyeong Eo†, Jaehyung Seo†, Chanjun Park§, Yuna Hur‡*, Heuiseok Lim†‡*
Department of Computer Science and Engineering, Korea University†, Upstage§, Human-inspired AI Research‡

요 약

초거대 언어모델은 모델의 학습 없이 학습 예시만을 입력에 추가함으로써 목표하는 작업을 수행한다. 이런 방식은 상황 내 학습 (In-Context Learning, ICL)이라 불리며, 초거대 언어모델 활용의 사실상의 표준으로 사용되고 있다. 하지만 이러한 모델은, 환각현상 등 사용상의 한계가 발생하는 상황이 다수 발생한다는 연구 결과가 나오고 있다. 본 연구에서는 초거대언어모델을 한국어 작업에서 사용하는 경우, 매우 간단한 수준의 종결어미 변환만으로도 성능 편차가 매우 크게 발생함을 확인하였다. 우리는 이에 대한 분석을 통해, 학습 예시의 어체와 추론 대상의 어체의 변환에 따라 초거대언어모델의 효용성이 크게 변함을 발견하고 이에 대해 분석한다. 나아가 우리는 본 실험 결과를 바탕으로, 어체에 대한 일관성이 유지된 형태의 한국어 데이터 구축이 이루어져야 함을 제안한다.

주제어: 한글 및 한국어 정보처리 학술대회, tex 양식

1. 서론

언어모델의 크기는 시간이 지남에 따라 점차 커지고 있다 [1, 2]. 60M개의 파라미터 수를 가지고 있는 BERT에서, 현재는 175B의 사이즈를 가진 모델까지, 모델의 사이즈는 커져가며 그에 따라 성능도 함께 올라가고 있다 [3]. 특히, 이러한 초거대언어모델은 task에 대한 설명 및 예시를 입력단에 추가하는 것 만으로도, 추가적인 학습 없이 목표하는 task를 수행할 수 있어 그 활용가치가 매우 크다[4]. 이러한 방법은 상황 내 학습 (In-Context Learning, ICL)이라 불리며 [3, 5], 입력예시와 출력예시로 구성되는 학습예시 데이터를 초거대언어모델의 입력에 추가함으로써, 모델 전체가 해당 데이터로 학습한 것과 유사한 역할을 수행하도록 한다. 이러한 방법은 GPT3에서 제안된 이후 현재 ChatGPT까지, LLM을 활용하는 데 있어서 사실상의 표준으로 사용되고 있다 [6]. 이러한 방법론을 적용할 때는 prompt선택에 따라 결과가 달라지며, 학습 예시(shot)를 선정하는 방법에 따라서도 매우 큰 성능 편차가 발생하기에, 적절한 활용 방법 및 가장 효과적인 학습 예시 선택 방법 등이 집중적으로 연구되고 있다 [7, 8].

이러한 방법은 다양한 분야에서 매우 높은 적합도를 보이며, 특정 작업의 경우, 사람보다 더 높은 성능을 낼 정도로 효과적임이 밝혀졌으나[3, 9, 10] 연구가 지속되며 여러 한계점이 지적되고 있다[11]. 이중 가장 대표적인 문제는 환각(Hallucination)과 관련한 문제이다 [12]. 이는 LLM의 생성 능력이 과하게 뛰어남에 따라 발생하는 문제로, 그럴듯한 문장으로 보이나 실제로는

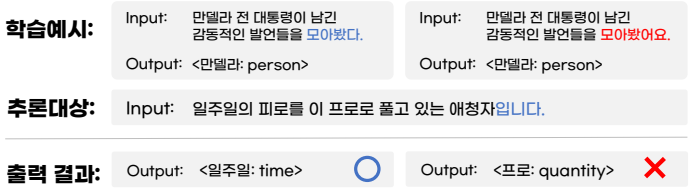

그림 1. 초거대언어모델의 활용 예시 (In-context learning). 우리는 학습예시의 어체에 따라 출력 결과가 달라지는것을 실험적으로 확인하였다.

사실과 거리가 있거나 정답이 아닌 답변을 생성해내는 현상을 의미한다 [13]. 이러한 환각 현상이 발생하는 상황은 매우 다양한 것으로 밝혀지고 있다. 이는 생략어구에 의한 것 [14, ?], 정보 과다에 의한 것[15], 단순 랜덤변수에 의한 영향[16] 등, 자칫 사소하고 단순하다고 보일 수 있는 요소들에 의해 생성 능력이 매우 크게 영향을 받는 모습을 보여준다.

본 연구에서는 한국어 작업을 수행함에 있어 발생할 수 있는 환각 현상 한가지를 지적하고자 한다. 이는 어체와 관련한 연구로, 우리는 ICL에서 학습 예시로 구어체 문장을 넣었을 때 초거대언어모델의 한국어 작업 수행 능력이 매우 큰 폭으로 저하되는 것을 확인하였다. 특히, 그림 1에서와 같이, 경어체 형태의 문장을 이해하는 작업을 수행할 때, 경어체의 학습 예시를 사용하는 경우에는 비교적 잘 수행할 수 있으나, 구어체 학습 예시를 사용하는 경우에는 큰 폭의 성능 저하가 발생함을 확인하였다. 이러한 현상은 초거대 언어모델을 한국어 작업에 활용함에 있어서는, 영어에서 고려되지 않는 종결어미나 어체

*교신저자(Corresponding author)

와 관련하여서도 특별한 고려가 필요함을 보여준다.

우리는 이와 같은 논의를 발전시켜, 입력 예시 문장의 어체와, 수행 대상 문장의 어체의 변화에 따른 초거대언어모델의 수행 능력을 분석한다. 이는 기존 어체에 대해 일관성 없이 구축되어 있는 한국어 데이터에 대하여, 이를 동일한 의미를 가진 구어체, 경어체 문장으로 변환사켜 수행했을 경우 발생하는 성능 변화 및 환각 현상에 대한 분석을 의미한다. 특히, 사용하는 단어의 변화(ex. 밥-식사 등)는 온전히 제외하고 종결어미에 의한 변화만을 분석 대상으로 하여, 문장 대부분의 요소를 보전한 매우 단순한 수준의 변화가 초거대언어모델의 성능에 미치는 영향을 분석한다. 이를 위해 우리는 현재 공개되어있는 KLUE 벤치마크[17]의 개체명인식 학습용 데이터를 활용하고, 이후 구어체와 경어체로 각각 변환하여 학습 예시 데이터와 검증용 데이터를 구축한다. 우리는 본 연구를 통해, 종결어미의 변화만으로도 초거대언어모델의 성능이 크게 변화할 수 있음을 밝히며, 이러한 혼란을 방지하기 위하여, 데이터 구축시 어체를 고려한 구축이 이루어져야 함을 제안한다.

2. 관련연구

환각 현상은 생성 결과물이 준수하여 사실처럼 보이나, 실제로는 사실과 거리가 먼 텍스트를 출력하게 되는 경우를 의미한다[18]. 대표적으로는 주어진 정보를 제대로 반영하지 못한 채 생성하거나 [19], 사실 정보에 완전히 반대되는 정보를 생성하는 경우가 포함된다[20, ?]. 이러한 환각 현상이 발생하는 요인은 매우다양하다고 알려져 있다 [2]. 이는 단순하게 모델 자체의 이해 성능 부족으로 인해 주어진 정보를 제대로 이해하지 못하거나 생성 단계에서 주어진 정보를 반영하지 못해 발생하는 경우도 다수 존재하며[21, 22], 특정하게는 생략어구를 제대로 이해하지 못하거나, 일반상식과 관련한 문장을 이해하지 못하는 등, 다방면의 요인들이 분석되어왔다 [13].

이러한 연구들은 주로 영어 문장들을 대상으로 이루어지고 있다. 하지만 한국어는 영어 문장과 다르게, 완전히 동일한 명사와 동사를 동일한 어순으로 조합하더라도 종결어미에 따라 문장의 뉘앙스가 바뀌며 사용 대상이 바뀐다 [23]. 이러한 한국어의 특징은 영어와는 다른 분석을 가능하게 한다. 특히 우리는 이 종결어미의 변화가 초거대언어모델의 이해 성능에도 영향을 끼치는지 확인하며, 특히 이로 인해 발생할 수 있는 언어모델 환각 현상에 대해 초점을 맞춰 분석한다.

3. Prompt 설계

3.1 실험 목적 및 환경 설정

우리는 어체 변화에 따른 초거대언어모델의 언어 이해 성능 분석에 목적을 둔다. 특히 학습 예시 데이터의 어체와 추론 대상이 되는 데이터의 어체가 다르거나 같을 시 발생하는 성능 차이를 중점으로 분석한다. 본 연구에서의 실험 파이프라인은 그림 2와 같다. 우리는 첫번째로 작업에 대한 설명과 함께 5개의 학습 예시를 제공해주는 세팅을 실험함으로써, n-shot 예시로써 어체의 작용을 확인한다. 두번째로는, 학습 예시로 제공한 문장과 동일한 문장을, 어체만 바꾸어 추론하게 지시함으로써 어체에 따라 발생하는 초거대언어모델의 환각 현상에 대해 분석한다.

이를 위하여 우리는 KLUE benchmark의 NER데이터셋을 활용한다 [17]. 개체명분류군은 총 6가지로 시간(time), 인명(person), 지명(locateion, 수량(quantity), 날짜(date), 기관명(organization)으로 구성되어있다. KLUE NER데이터는 훈련/검증 데이터 내에 어체에 대한 고려가 이루어지지 않았으며, 구어체, 문장형태, 인터넷어체 등 다양한 문장구조로 구축되어 있다. 우리는 실험을 위해 훈련 데이터로부터 5개의 데이터를 임의 추출하여 ICL을 위한 입력 샘플을 구성하였고, 검증 데이터로부터 100개의 데이터를 임의 추출하여 테스트 데이터로 삼았다. 개체명 분류군(entity type)의 분포에 따른 편향을 완화하기 위해, ICL을 위한 데이터와 테스트 데이터에는 모든 개체명 분류군이 한종류 이상 포함되도록 구성하였다.

3.2 종결어미 변환

우리는 종결어미 변환을 위해 GPT3.5를 활용한다. 이는 어체 변환에 있어서 일관적인 기준을 적용하기 위함이며, ChatGPT생성 이후 결과물은 모두 사람이 직접 검수하여 종결어미의 변환 이외의 문장 변환이 이루어졌는지 확인한다. 이 단계를 위한 프롬프트(prompt)의 설계를 위해, 우리는 기존 연구들에서 적용하는 방법과 같이 학습 예시 뿐 아니라 작업에 대한 지시를 함께 포함했으며, 구체적으로 활용한 프롬프트는 그림 3과 같다.

3.3 개체명 인식

우리는 현재 대표적으로 활용되고있는 GPT3.5를 활용하여 환각현상을 검증한다. 영어로 프롬프트를 설계한 이유는 학습 언어 자원 중 가장 많은 비중을 차지하는 영어로 프롬프트를 지시하는 것이 더 좋은 성능을 낸다는 연구 결과에 기인한다 [24]. 우리는 다양한 어체로 변환한 학습예시 문장들을 활용하여 프롬프트를 구축하고, 이를 통해 초거대언어모델이 개체명인식을 수행하도록 지도한다. 프롬프트 설계에 있어서 우리는 개체명 인식 작업에 대한 지시를 함께 포함했으며, 최종적으로 구축한 프롬프트는 그림 4과 같다. 우리는 보다 공정한 실험을 위하여 동일한 세팅에 대하여 3번씩 검증을 수행했고, 3번의 검증의 평균 성능을 해당 세팅의 최종 성능으로 간주한다.

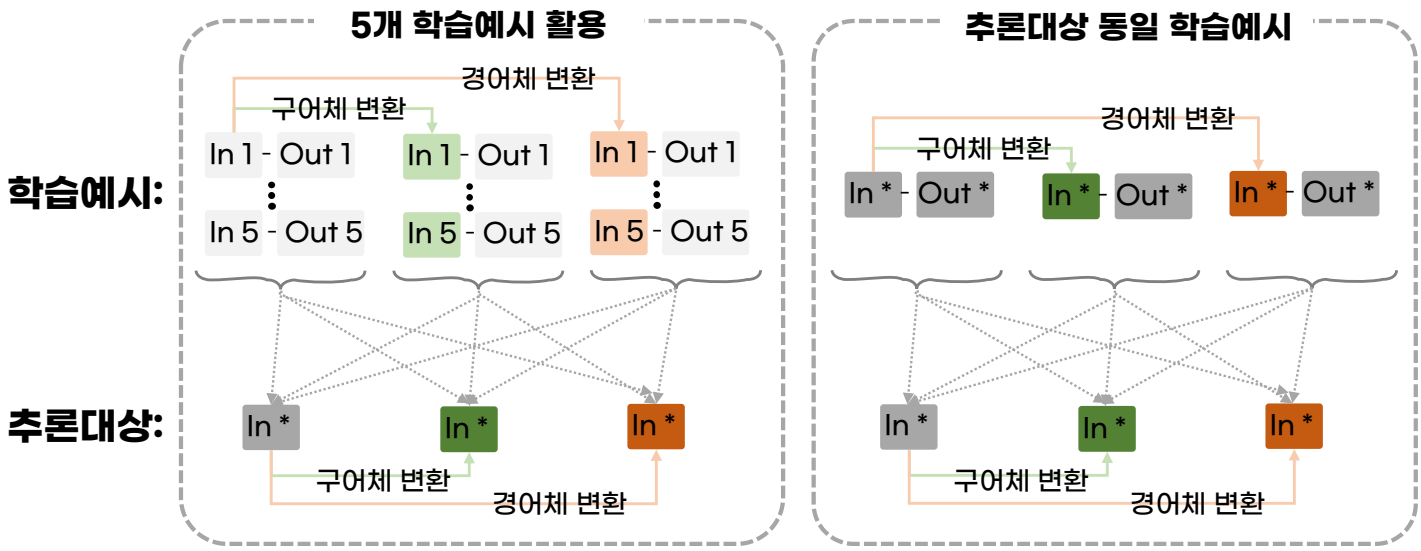

그림 2. 실험 파이프라인. 우리는 종결어미에 따른 언어모델 이해 성능 분석을 위해 어체만을 변환시킨 5개의 학습예시를 활용한 5-shot성능을 다양한 어체의 추론 대상을 통해 분석하였다. 또한, 추론 대상과 동일한 문장을 어체를 바꿔가며 학습예시로 활용했을 때 초거대언어모델의 작업 수행 능력을 평가함으로써, 한국어 작업에 있어 초거대언어모델의 환각 현상에 대해 분석한다.

5-shot		추론 대상		
		기존 데이터	구어체	경어체
학습 예시	기존 데이터	45.79 (43.73 / 52.97)	46.46 (44.38 / 53.96)	**48.04** (45.60 / 56.28)
	구어체	39.52 (36.86 / 47.73)	**49.08** (45.97 / 58.59)	47.51 (45.56 / 54.73)
	경어체	47.33 (45.07 / 54.40)	46.66 (44.43 / 54.75)	**49.06** (46.51 / 57.50)

표 1. 5개 학습 예시를 활용한 검증 결과. 각 수치는 각각 F1-score (Precision / Recall)를 의미한다.

You are an expert in Korean.
Your answer should be in Korean.
Your task is to change the format of the given statement.
Statement is given to you with any form of "경어체" or "구어체", and you should transform it into "**경어체**".
Statement can be in a form of a noun phrase, and in this case, you also should transform it into a complete form of a "**경어체**" sentence. If a given statement is already in a "**경어체**", you may not revise it.
You must return the output only, without any explanations.

For example,
… …

그림 3. 어체 변환을 위해 사용한 프롬프트. 위는 경어체로의 변환을 위한 프롬프트를 보여주며, 구어체로의 변환을 위하여는 붉게 표시된 부분을 "구어체"로 변환하여 수행하였다. 우리는 직접 생성한 어체 변환 예시를 추가하여 최종 프롬프트를 설계한다.

You are an expert in Korean.
Your answer should be in Korean.
Your task is to distinguish named-entities (also can be perceived as the proper nouns) among the given statement.
You should find every named entity in a given sentence, and categories each named entity to the correct entity type.
Entity type should be one of the following set: "person", "location", "organization", "date", "time" and "quantity"
You should return the output only, without any explanations.

For example,
… …

그림 4. 개체명 인식에 사용한 프롬프트. 우리는 이에 훈련 데이터에서 추출한 학습 예시를 추가하여 실제 개체명인식 검증을 위한 최종 프롬프트를 설계한다.

4. 실험 및 분석

4.1 5개 학습예시 활용

여기서 추론 대상으로 사용한 5개 문장과 학습 예시의 문장은 중복되지 않는 문장으로, 100개 추론 대상 문장을 추론하는 데 있어서 모두 동일한 학습 예시가 활용되었다. 단, 실험 세팅에

1-shot 어체변환		추론 대상		
		기존 데이터	구어체	경어체
학습 예시	기존 데이터	**76.11** (73.21 / 83.52)	72.50 (68.60 / 81.28)	75.58 (71.35 / 85.86)
	구어체	72.20 (68.50 / 81.93)	72.18 (68.46 / 81.97)	**74.79** (70.62 / 85.19)
	경어체	74.77 (71.23 / 82.86)	74.80 (71.23 / 82.84)	**76.41** (72.79 / 85.92)

표 2. 5개 학습 예시를 활용한 검증 결과. 각 수치는 각각 F1-score (Precision / Recall)를 의미한다.

따라 다양한 어체로 변환된 학습 예시가 사용된다. 실험시 기존 데이터는 KLUE 벤치마크에서 공개된 원본 데이터를 의미하며, 이는 어체에 대한 통일 없이 다양한 어체로 구성되어있다. 이에 대한 실험 결과는 표 1와 같다.

실험 결과, 학습 예시로서 구어체 데이터를 활용하는 경우 매우 큰 폭의 성능 하락이 있음을 확인할 수 있었다. 이는 구어체 추론 대상에 대해서는 뛰어난 정확도로 추론하나, 기존 데이터를 제대로 다루지는 못하는 것을 알 수 있다. 또한, 기존 데이터를 활용하는 경우에 있어서, 학습 예시를 모두 경어체로 통일시켜주는 것이 일반적인 성능 향상을 이끄는 요인이 될 수 있음을 보여준다. 이는 어체에 관계없이 모든 추론 대상들에 대해 일반적으로 높은 성능을 보여주었다. 이는 언어모델의 학습 예시로서 사용할 때 어체를 경어체로 바꾸는 것이 효과적인 전략이 될 수 있음과, 모델의 입력으로 주어지는 어체에 따라 언어모델의 성능이 일관적이지 못하게 측정될 수 있음을 보여준다.

4.2 학습예시 어체변환

본 실험에서는 추론 대상과 동일한 문장을 학습 예시로 활용하는 경우의 성능을 검증한다. 이는 추론 대상에 대한 정답 출력 결과물이 이미 학습 예시로서 포함되어있는 경우를 의미하며, 어체에 따른 언어모델의 이해 및 생성 능력의 변화를 확인한다. 이에 대한 실험 결과는 표 2와 같다.

여기서 학습예시와 추론대상이 같은 경우에도 초거대언어모델의 성능이 100점을 출력하지 않는 이유는 초거대언어모델의 언어 이해 및 생성 능력 자체의 한계를 보여준다. 즉, 주어진 정보에 대해서도 잘 수행하지 못함을 보여주며 이는 초거대언어모델의 환각현상을 그대로 보여주는 사례이다. 또한, 다른 어체를 사용하는 경우에 추가적인 성능 하락이 있음은, 어체의 영향에 따라 언어모델의 이해 및 추론 능력이 유의미한 폭으로 하락할 수 있음을 보여준다.

이러한 분석을 기반으로 했을 때, 기존 데이터들이 이런 어체에 대한 고려 없이 구축되었다는 것은 향후 활용에 있어서 한계점을 낳을 수 있다. 즉, 언어모델의 성능을 확인하기 위한 벤치마크로서 활용되는 데이터이나, 어체수준에서의 변화로 인해 이를 객관적으로 측정하지 못하는 문제를 낳을 수 있음을 밝힌다.

5. 결론

본 연구는 초거대언어모델을 통해 한국어 작업을 수행하는 경우 발생하는 환각현상에 초점을 맞추어 진행되었다. 특히 우리는 종결어미 변환이라는 매우 단순한 작업만으로도 초거대언어모델의 성능을 크게 바꿀 수 있음을 실험을 통해 밝혔으며, 이를 통해 데이터 구축시 어체에 대한 고려가 이루어져야 함을 제안하였다. 하지만 우리의 작업은 개체명인식이라는 작업에 국한되어있으며, 검증 데이터도 100개로 비교적 적은 상황에서 진행되었기에, 본 연구의 결과가 모든 상황을 일반화하기는 어려움을 밝힌다. 본 연구를 통해 우리는 한국어에서 발생하는 어체에 따른 환각현상에 대하여, 보다 발전되고 일반화된 형태의 분석이 필요함을 제안하며, 관련한 향후 연구 방향을 계획한다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기술기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2023-2018-0-01405). 이 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2021R1A6A1A03045425). 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1A5A7026673)

참고문헌

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[2] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *arXiv preprint arXiv:2307.10169*, 2023.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry,

A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[4] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "Gpt understands, too," *AI Open*, 2023.

[5] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2339–2352, 2021.

[6] 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.

[7] Z. Liang, F. Wei, Y. Jie, Y. Qian, Z. Hao, and B. Han, "Prompts can play lottery tickets well: Achieving lifelong information extraction via lottery prompt tuning," *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 277–292, 2023.

[8] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun, "Towards understanding chain-of-thought prompting: An empirical study of what matters," *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[9] OpenAI, "Gpt-4 technical report," 2023.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] H. Alkaissi and S. I. McFarlane, "Artificial hallucinations in chatgpt: implications in scientific writing," *Cureus*, Vol. 15, No. 2, 2023.

[12] J. D. Blom, *A dictionary of hallucinations.* Springer, 2010.

[13] 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.

[14] K. Filippova, "Controlled hallucinations: Learning to generate faithfully from noisy data," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 864–870, 2020.

[15] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020.

[16] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

[17] S. Park, J. Moon, S. Kim, W. I. Cho, J. Y. Han, J. Park, C. Song, J. Kim, Y. Song, T. Oh *et al.*, "Klue: Korean language understanding evaluation," *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[18] 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, 2021.

[19] R. Lebret, D. Grangier, and M. Auli, "Neural text generation from structured data with application to the biography domain," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1203–1213, 2016.

[20] A. Parikh, X. Wang, S. Gehrmann, M. Faruqui, B. Dhingra, D. Yang, and D. Das, "Totto: A controlled table-to-text generation dataset," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, 2020.

[21] S. Wiseman, S. M. Shieber, and A. M. Rush, "Challenges in data-to-document generation," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2263, 2017.

[22] P. Parthasarathi, K. Sinha, J. Pineau, and A. Williams, "Sometimes we want ungrammatical translations," *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3205–3227, 2021.

[23] S. Agrawal, A. Anastasopoulos, L. Bentivogli, O. Bojar, C. Borg, M. Carpuat, R. Cattoni, M. Cettolo, M. Chen,

W. Chen *et al.*, "Findings of the iwslt 2023 evaluation campaign," *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pp. 1–61, 2023.

[24] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou *et al.*, "Language models are multilingual chain-of-thought reasoners," *The Eleventh International Conference on Learning Representations*, 2022.