특집원고

생성형 모델의 신뢰성 향상에 관한 연구 동향

한국전자통신연구원 ▌노지현*·배경만*

1. 서 론

최근 인공지능(AI) 기술의 발전은 다양한 산업 분야에서 획기적인 발전을 가져오고 있다. 이러한 발전의 중심에는 자연어 처리(Natural Language Processing; NLP)가 있으며, 이 분야는 인간과 기계 사이의 상호작용을 더욱 풍부하게 만들어주는 중추적인 역할을 하고 있다. 언어모델링은 핵심 구성요소인 언어모델(language model; LM)을 통해 사용자의 질문을 이해하고 이에 답변하거나, 문서를 요약하고, 번역하는 등의 복잡한 작업을 수행하는 데 사용되고 있다. 특히, 트랜스포머(Transformer)에 기반한 대형 언어모델(Large Language Model; LLM)[1][2][3]은 방대한 학습데이터를 바탕으로 훈련되어 언어모델의 확장과 발전을 선도하고 있다. 이런 발전은 사람들의 업무 방식에도 큰 변화를 불러오고 있으며, 기계가 인간의 언어를 이해하고 생성하는 능력을 지속적으로 향상시키고 있다. 이러한 발전에도 불구하고 언어모델은 여전히 극복해야 할 문제들이 존재한다. 특히, 환각(hallucination) 및 사실성(factuality) 오류, 오래된 지식(outdated knowledge) 관련 문제들은 언어모델의 신뢰성을 저해하는 주요 요인으로 지목되고 있다. 이 문제들은 대부분 현재의 언어모델이 학습데이터의 조합으로 어떻게 결론에 도달하는지, 그 과정이 정확히 어떠했는지를 완전히 이해하기 어려운 블랙박스(black-box) 시스템인 특성에서 비롯된다. 이러한 어려움으로 인해 설명 가능한 인공지능 시스템 연구의 중요성이 점차 대두되었다. 설명 가능성 연구를 통해 모델의 작동 방식을 이해하고 강점과 약점을 진단하여 디버깅하는 데 도움을 줄 수 있다. 예를 들어, 모델이 설명한 내용의 일관성을 판별함으로써 환각 현상을 발견하고 개선할 수 있다.

* 정회원

† 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00369, (4세부) 전문지식 대상 판단결과의 이유/근거를 설명가능한 전문가 의사결정 지원 인공지능 기술개발)

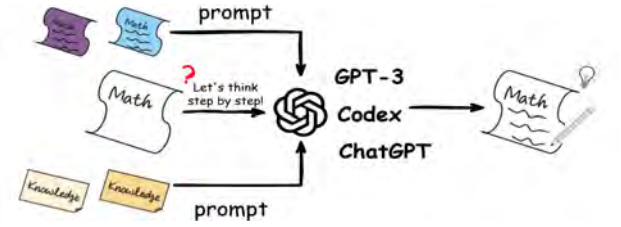

그림 1 프롬프팅을 활용한 언어모델 추론 방법[5]

신뢰할 수 있는 모델은 모델이 추론 과정을 명확하게 설명하고, 결과 생성 방식을 투명하게 제공할 수 있어야 한다. 이를 위해 그림 1과 같이 단순한 프롬프팅(Prompting)에서부터 외부 지식 (external knowledge)을 활용한 설명 방법에 이르기까지 다양한 접근방법이 연구되고 있다. 사고 사슬(Chain-of-Thought; CoT)[4]와 같은 방법은 언어모델의 추론 과정을 단계별로 설명함으로써 모델의 신뢰도를 향상하는 데 기여하고 있다. 이러한 연구는 언어모델이 더욱 정확하고 신뢰할 수 있는 결과를 제공하는 데 필수적이다.

2. 생성 모델의 사실성 강화 연구 및 환각과 사실성 검증 연구 동향

해당 부문에서는 언어모델이 당면한 한계와 도전과제, 생성형 모델의 사실성 강화 연구 동향, 그리고 환각과 사실성을 검증하는 연구 동향을 소개한다.

2.1 언어모델이 당면한 한계점과 도전과제

서론에서 언급한 바와 같이, 언어모델은 몇 가지 주요한 한계점을 보유하고 있다. 1) 최신성 반영 문제: 오래된 정보로만 학습된 모델들은 최신 정보를 파악하고 정확한 답변을 제공하는 데 어려움을 가진다. 2) 환각 및 사실성 결여 문제: 언어모델이 확률적으로 최적화된 답변을 생성하면서 진실하지 않은 정보를 포함할 가능성이 있다. 모델이 매우 그럴듯하지만, 사용자 관점에서 엉뚱한 답변을 제출하기도 한다. 3) 학습 용량 부족: 대형 언어모델로 발전되면서 상당한 양

Factual and Non-Hallucinated	Factually correct outputs.
Non-Factual and Hallucinated	Entirely fabricated outputs.
Hallucinated but Factual	1. Outputs that are unfaithful to the prompt but remain factually correct [19]. 2. Outputs that deviate from the prompt's specifics but don't touch on factuality, e.g., a prompt asking for a story about a rabbit and wolf becoming friends, but the LLM produces a tale about a rabbit and a dog befriending each other. 3. Outputs that provide additional factual details not specified in the prompt, e.g., a prompt asking about the capital of France, and the LLM responds with "Paris, which is known for the Eiffel Tower."
Non-Factual but Non-Hallucinated	1. Outputs where the LLM states"I don't know" or avoids a direct answer. 2. Outputs that are partially correct, e.g., for the question, "Who landed on the moon with Apollo 11?" If the LLM responds with just "Neil Armstrong," the answer is incomplete but not hallucinated. 3. Outputs that provide a generalized or vague response without specific details, e.g., for a question about the causes of World War II, the LLM might respond with "It was due to various political and economic factors."

그림 2 사실성 문제와 환각 문제 비교 예시[7]

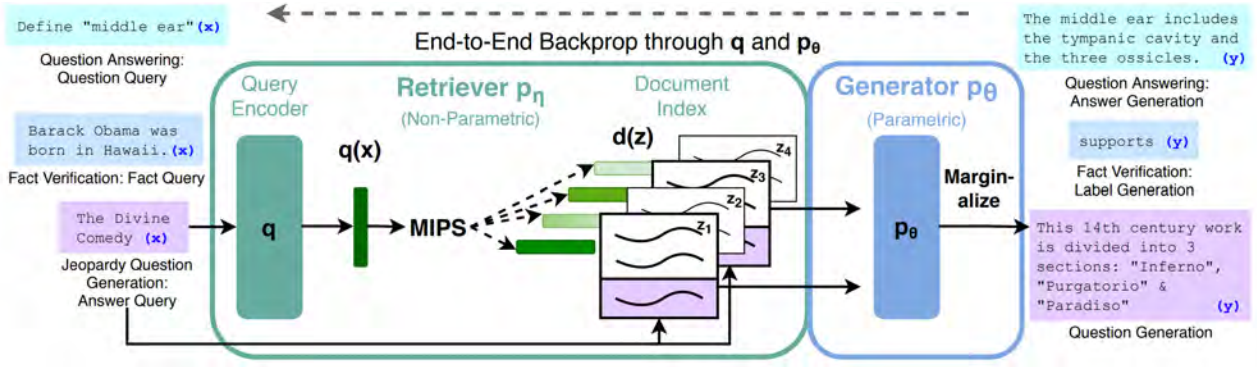

그림 3 검색 증강 생성(RAG) 기법의 초기 시스템. 검색기(Retriever)와 생성기(Generator)를 결합하여 학습[9]

의 GPU들이 필요하다는 점에서 발생한다. 4) 내부 메모리: 모델이 기본적으로 메모리 시스템을 보유하고 있지 않기 때문에 이전 문맥을 장기간 기억하는 데 어려움을 겪는다. 이전 내용을 오래 기억하기 위해서는 컨텍스트 길이가 긴 대형 언어모델을 계속 활용해야 한다.

본 장에서는 이러한 문제들 중 환각과 사실성에 초점을 맞추어 관련 연구 동향에 대해 살펴보고자 한다. 환각은 "Survey of Hallucination in Natural Language Generation" 논문[6]에서 언급한 바와 같이, 제공된 출처에 충실하지 않거나 무의미한 내용을 생성하는 것으로 정의된다. 출처는 상황에 따라 다를 수 있으며, 입력 텍스트(예, 질문)이거나 언어모델이 생성한 문단, 혹은 사실에 입각한 웹 지식을 의미할 수 있다.

LLM 활용 분야에서 사실성(Factuality)은 모델이 사실 정보를 포함하는 문맥을 생성하는 능력으로 정의될 수 있다. 여기서 사실 정보는 상식 정보나 사전, 위키 등 다양한 분야의 신뢰할 수 있는 출처에서 얻은 정보를 의미한다. 사실성 부족 문제는 관련 분야 지식의 부족, 검색 능력의 한계, 최신 정보를 업데이트하는 데의 어려움, 추론 오류 등 다양한 요인에 의해 발생할 수 있다. 환각과 사실성 부족 문제는 모두 LLM이 생성하는 데이터의 정확성, 신뢰성과 관련되어 있다. 일반적으로 환각은 사실성 문제를 포함할 수 있으나, 본 원고에서는 두 가지를 분리하여 설명하고자 한다. 환각은 사용자 입력과 모순되거나 이전에 생성된 문맥과 모순되는 경우를 포함하는 현상으로 구분하고, 사실성 문제는 현재의 사실관계와 모순되는 경우로 구분 지을 수 있다.

그림 2는 논문[7]에서 사실성 이슈와 환각 이슈를 비교한 예시이다. 예를 들어, LLM에게 "토끼와 늑대가 친구가 되는 동화를 만들어줘"라고 요청했을 때, "토끼와 개가 친구가 되는 이야기"를 생성할 경우 이는 환각 문제로 볼 수 있지만, 사실성 문제는 아니다.

2.2 사실성을 강화하는 연구 방법들

이 장에서는 언어모델의 사실성을 강화하는 여러 방법 중에서 검색 기반 접근방법을 설명하려 한다. 대규모의 데이터셋(예, Wikipedia)로 학습된 LLM은 특

정 도메인의 지식이 부족하거나 학습 이후에 발생한 지식에 대해 알지 못하는 문제를 가질 수 있다. 그리고 지속적인 추가 훈련 과정에서 기존에 학습된 지식을 잊어버리는 문제가 발생하기도 한다[8]. 이로 인해 발생하는 사실성 오류는 검색기를 활용함으로써 어느 정도 해결될 수 있다.

검색 증강 생성(Retrieval-Augmented Generation; RAG)[9]는 최신 정보 부족과 모델의 기억력 한계와 같은 문제를 해결하기 위해 채택된 방법의 하나로 현재 가장 널리 사용되고 있다. 일반적으로 RAG 방법은 외부 데이터베이스에서 데이터를 검색한 후, 검색된 데이터를 LLM의 생성 과정에 통합하는 방식으로 작동한다. 검색될 문서들은 언어모델 혹은 텍스트 임베딩 모델을 통해 미리 벡터화되어 저장소에 저장된다. 검색기는 저장소에 있는 데이터를 검색하는 데 사용되며, 입력 정보와 가장 관련성 높은 데이터를 찾아내게 된다.

검색 트랜스포머 모델인 RETRO[10]는 기존 텍스트 조각과 유사한 여러 텍스트 조각을 뉴럴 데이터베이스에서 검색해 문장을 완성하는 새로운 접근법을 제안하였다. 이 방법은 GPT-3보다 25배 더 적은 파라미터를 사용하면서도 우수한 성능을 보여주었다고 보고되었다. 논문[11]은 주어진 쿼리에 대해 구글 검색기로 웹 문서를 검색하고 이를 기반으로 few-shot 프롬프팅을 하여 언어모델의 생성 능력을 향상하는 방법을 제안하였다. LLM을 RAG 형태에 맞게 적응(조정)하는 연구들[12][13] 은 검색된 정보를 단순히 결합하는 것만으로는 능력 향상에 한계가 있다고 지적하고 있다. 즉, LLM이 검색된 데이터에 더 잘 통합되어 더 정확한 내용을 생성할 수 있도록 하는 것이 중요하다고 본다.

이러한 맥락에서, 관련 논문[14]은 Contriver [15] 검색기와 Fusion-in-Decoder [16] 를 갖춘 T5 [17] 언어모델로 구성된 통합된 아키텍처인 ATLAS를 제안하였다. ATLAS는 자연 질문에 대해 단 64개의 예시만을 사용하여 540B 모델보다 50배 적은 매개변수로 능가하는 정확도를 달성하였다. 이 외에도, 여러 연구가 진행되고 있다. 쿼리 검색에 CoT 단계를 통합하는 방법[18] 외부지식 API에 접근하는 LLM 기반 에이전트 프레임워크[19] 방법, 외부 파라매트릭 메모리에서 검색하는 방법[20] 혹은 지식 그래프에서 검색[21] 방법, 구조적 데이터(KG, 표, 데이터베이스) 검색을 통해 LLM 추론을 향상한 StructGPT[22] 방법도 있다. 이러한 방법들은 LLM의 성능을 높이고 사실성을 강화하는데 기여하면서 언어모델이 더 정확하고 신뢰할 수 있는 정보를 생성하도록 지원한다.

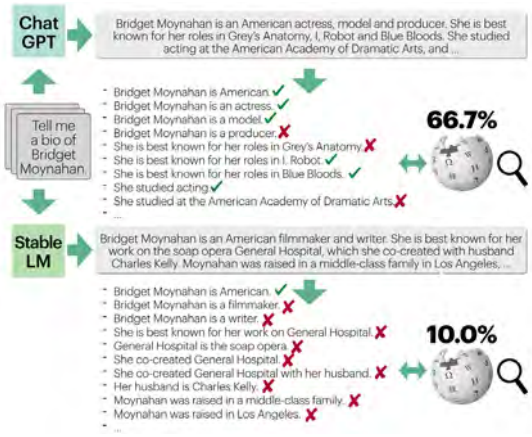

그림 4 FActScore의 개요. Atomic facts 단위로 사실성 검증 실행[28]

2.3 사실성 및 환각 검증 방법

LLM이 생성한 결과에 대한 사실성 혹은 환각을 검증하는 일은 모델의 신뢰성을 보증하는 데 필수적이다. 가장 대중적으로 사용되는 평가 지표(metric) 중 하나는 Exact Match(EM)로, 기준이 되는 텍스트와 생성된 텍스트 사이에 얼마나 비슷한 단어들이 존재하는지를 분석하여 값을 도출한다. 또한, 텍스트 간 유사성을 평가하기 위해 일반적으로 사용되는 메트릭에는 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[23], BLEU(Bilingual Evaluation Understudy)[24], METEOR(Metric for Evaluation of Translation with Explicit Ordering)[25]가 있으며, 이 방식들은 모두 n-gram을 활용한다. 이 외에도 뉴럴넷 모델을 활용한 평가 방법인 BERTScore[26]는 사전에 학습된 BERT 모델을 활용하여 정답 문장과 생성 문장 사이의 유사도를 통해 생성결과를 평가한다. 비슷하게 BART를 활용하는 BARTScore[27]도 존재한다.

최근 발표된 연구 중 FActScore[28]는 문장 단위가 아닌 더 세분화된 단위, 즉 Atomic 문장들의 사실성을 평가하는 새로운 방법을 제안하였다. 이 방법은 기존의 문장 단위 평가에서 한계점을 발견하고, 더 세밀한 분석을 가능하게 하였다. 특히 기존 방법들은 한 문장 내에서 사실과 거짓이 공존할 경우 전체를 거짓(사실이 아님)으로 분류하는 반면, 제안된 연구는 문장을 여러 개의 부분 문장으로 나누어 (논문에서는 이를 atomic facts라고 함) 각 부분 문장의 사실성을 평가하였다. 이 검증 과정은 신뢰할 수 있는 지식 소

스에 기반하였고, Wikipedia의 사람 전기(biography) 데이터를 활용해 실험하였다.

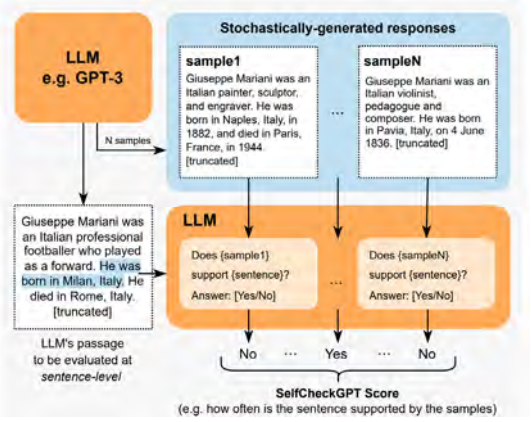

그림 5 SelfCheckGPT의 프롬프트 예시[29]

또한, LLM을 활용해 생성 텍스트의 환각성을 평가하는 SelfCheckGPT[29]도 관련 연구 분야에서 주목받고 있다. 제안된 연구도 Wikibio 데이터셋을 활용해 LLM이 생성한 문장이 환각이 있는지 없는지를 검증하였다. 이 방법은 LLM(예, GPT-3)에게 한 인물에 대한 데이터를 여러 번 생성하게 한 다음, 생성된 데이터들이 서로 유사하거나 일관되면 환각이 없음으로 평가하는 방법이다. 이 방법은 외부 데이터베이스를 전혀 사용하지 않고 적용 가능하며, 내부를 알 수 없는 블랙박스인 LLM들을 검증할 수 있는 장점을 가진다.

3. AI의 설명 가능성과 논리적 추론 연구 동향

의료 진단, 법률, 금융 등 전문지식을 요구하는 분야에서 LLM의 신뢰성은 매우 중요하다. LLM이 생성한 텍스트가 거짓된 답변이나 근거 없는 답변인 경우 심각한 비용 손실을 초래할 수 있기 때문이다. 신뢰할 수 있는 LLM은 추론 과정을 명확히 설명하고 타당한 근거를 제공할 수 있어야 한다. 앞서 설명된 검색 증강 모델(RAG)과 같은 방법은 외부 데이터베이스의 정보를 인용함으로써 사용자에게 출처를 제시할 수 있고, 사용자는 검색된 출처를 검증해서 LLM의 답변을 신뢰할지 결정할 수 있었다. ChatGPT와 같은 생성형 대화 모델은 LLM의 해석 가능성(interpretability)과 추론 방식에 대한 새로운 접근 방식을 보여주었다. 이러한 모델은 자체 추론 과정을 설명함으로써 사용자들의 신뢰를 증진할 수 있다.

추론 능력을 강화하기 위한 방법의 하나로 프롬프트 엔지니어링(Prompt Engineering)이 제시되었다. 프롬프트(Prompt)는 일반적으로 지시사항, 질문, 입력 데이터, 예시로 구성된다. 고급 프롬프트는 모델이 논리적 추론 과정을 따라 답변에 도달하도록 안내하는 CoT 프롬프팅(그림 6 (b))과 같이 더 복잡한 구조를 포함한다. CoT 프롬프팅은 “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models” 논문[4]에서 처음 제안되었으며, 필수적인 추론 단계를 LLM에게 프롬프트로 안내하여 더욱 논리적인 추론이 가능하도록 한 방식이다. 기존에는 “단계별로 생각하라(Think step-by-step)” 명령만으로도 LLM이 추론 과정을 설명하고 적절한 추론을 선보였다면 현재는 단계마다 필요한 추론 예시를 제시해 주어 더욱 복잡한 논리적 추론이 가능하게 한다. 그림 6의 (c)는 자기 일관성(Self-Consistency) CoT 방법[30]으로, CoT 과정을 통해 나온 여러 LLM의 출력 중 최고의 결과를 선택하여 기본 CoT 방법보다 더욱 일관된 추론 결과를 제공한다. 그림 6의 (d)는 사고의 나무(Tree-of-Thoughts) 방법[31]이다. 이 방법은 사고(Thought)를 나무와 같

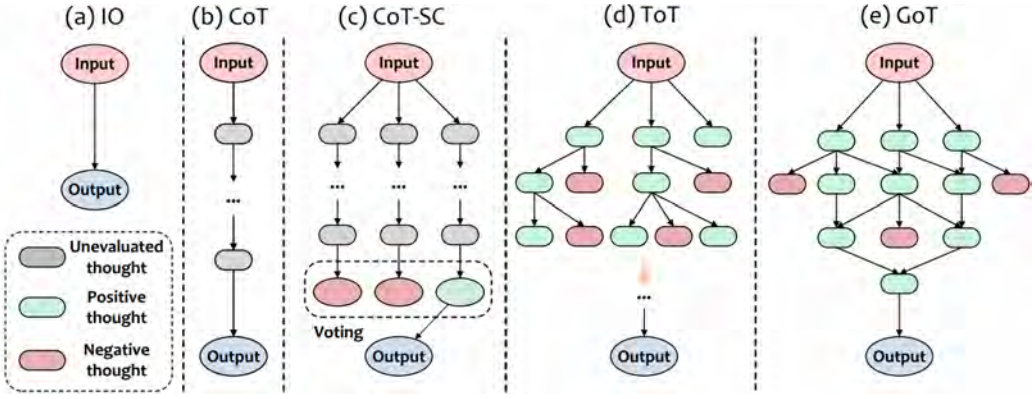

그림 6 사고 사슬(Chain-of-Thoughts) 프롬프팅 방법들. (a) Input-Output (IO) (b) Chain-of-Thought (CoT)[4] (c) Self-consistency (SC) CoT[30] (d) Tree-of-Thought (ToT)[31] (e) Graph-of-Thought (GoT)[32] (그림 참조: [33])

은 구조로 조직하고 각각의 분기가 다른 추론으로 표현된다. 이 방법은 가치판단을 하는 인간의 인지 과정과 유사하게, LLM이 다양한 가능성을 탐색할 수 있도록 하여서 다수의 시나리오 중 어떤 것이 가능성이 가장 큰지를 판단해 결정한다. 그러나 여러 번의 추론으로 발생하는 높은 계산 비용이 단점이 될 수 있다. 그림 6의 (e)는 사고의 그래프(Graph-of-Thoughts)[32]로 기존의 추론 방법을 확장하였다. 이 방법은 추론 과정을 그래프와 같이 구성하여, 중간 단계에서 사고들을 서로 묶거나 세분화함으로써 더 유연한 사고 구조가 가능하게 하였다. 이와 같은 다양한 연쇄적 추론 방법들은 모델의 추론 과정을 더욱 명확히 하여, 시스템의 전반적인 성능을 향상하는 동시에 결과에 대한 설명의 신뢰도를 높이는 데 기여한다.

4. 결 론

본 원고는 자연어처리 분야에서 언어모델의 최근 발전 방향과 신뢰도를 높이기 위한 다양한 연구 동향을 소개하였다. 최근 대형언어모델(LLM)의 발전은 인간과 기계 간 상호작용을 크게 개선하고 사용자들의 업무 수행 방식에서도 큰 변화를 가져왔다. 그렇지만 여전히 환각 문제, 최신 지식 반영의 어려움, 사실성 문제 등 아직 해결해야 할 여러 도전과제가 존재한다. 설명 가능한 AI, 검색 증강 생성(RAG), 그리고 다양한 사실성/환각 검증 메트릭 연구들은 모델의 신뢰성과 사실성을 강화하는 데 중요한 역할을 한다는 것을 확인하였다. 추론 과정의 설명 가능성과 사실성을 강화하는 연구 방향은 LLM의 활용 범위를 전문분야까지 확대하는 데 기여할 것이며, 이 기술들의 발전이 인간의 의사결정을 보조하는 데 중요한 역할을 할 것으로 기대된다.

참고문헌

[1] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt (2022).

[2] OpenAI. GPT-4 technical report. arXiv (2023).

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. Llama: Open and efficient foundation language models. arXiv, 2023

[4] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. & Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. NeuIPS, 35, 24824-24837.

[5] Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., ... & Chen, H. Reasoning with language model prompting: A survey. arXiv, 2022.

[6] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38, 2023.

[7] Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., ... & Zhang, Y. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv, 2023.

[8] Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., & Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint arXiv:2308.08747, 2023.

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33, 9459-9474, 2020.

[10] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. Improving language models by retrieving from trillions of tokens. In International conference on machine learning, PMLR, pp. 2206-2240, 2022.

[11] Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. Internet-augmented language models through few-shot prompting for open-domain question answering. arXiv preprint, 2022.

[12] Ren, R., Wang, Y., Qu, Y., Zhao, W. X., Liu, J., Tian, H., ... & Wang, H. Investigating the factual knowledge boundary of large language models with retrieval augmentation. arXiv, 2023.

[13] Wang, C., Cheng, S., Xu, Z., Ding, B., Wang, Y., & Zhang, Y. Evaluating open question answering evaluation. arXiv preprint arXiv, 2023.

[14] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. Atlas: Few-shot learning with retrieval augmented language models. arXiv, 2022.

[15] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning. arXiv, 2021.

[16] Izacard, G. and Grave, E. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proc. of Conf. of the European Chapter of the ACL: Main Volume. Association for Computational Linguistics, Online, 874 – 880, 2021.

[17] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S.,

Matena, M., ... & Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 21(140), 1-67, 2020.
[18] He, H., Zhang, H., & Roth, D. Rethinking with retrieval: Faithful large language model inference. arXiv preprint arXiv, 2022.
[19] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv, 2022.
[20] Chen, A., Pasupat, P., Singh, S., Lee, H., & Guu, K. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. arXiv, 2023.
[21] Zhang, S., Pan, L., Zhao, J., & Wang, W. Y. Mitigating language model hallucination with interactive question-knowledge alignment. arXiv, 2023.
[22] Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, W. X., & Wen, J. R. Structgpt: A general framework for large language model to reason over structured data. arXiv, 2023.
[23] Lin, C. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, 2004.
[24] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. Bleu: a method for automatic evaluation of machine translation. ACL, pp. 311-318, 2002.
[25] Banerjee, S., & Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proc. of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65-72, 2005.
[26] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. Bertscore: Evaluating text generation with bert. ICLR, 2020.
[27] Yuan, W., Neubig, G., & Liu, P. Bartscore: Evaluating generated text as text generation. NeurIPS 34, 27263-27277, 2021.
[28] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W. T., Koh, P., ... & Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP, pp. 12076-12100, 2023.
[29] Manakul, P., Liusie, A., & Gales, M. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP, pp. 9004-9017, 2023.
[30] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. Self-consistency improves chain of thought reasoning in language models. ICLR, 2023
[31] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. NeurIPS, 36, 2024.
[32] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., ... & Hoefler, T. Graph of thoughts: Solving elaborate problems with large language models. AAAI, Vol. 38, No. 16, pp. 17682-17690, 2024.
[33] Ding, R., Zhang, C., Wang, L., Xu, Y., Ma, M., Zhang, W., ... & Zhang, D. Everything of thoughts: Defying the law of penrose triangle for thought generation. arXiv, 2023.

약 력

노 지 현

2013 성균관대학교 전자전기공학부 졸업(학사)
2022 KAIST 전기및전자공학부 졸업(박사)
2022~현재 한국전자통신연구원 연구원
관심분야 : 신뢰가능한 인공지능, 소형 언어모델, 자연어처리, 자연어 생성
Email : jihyeon.roh@etri.re.kr

배 경 만

2004 동아대학교 컴퓨터공학과 졸업(학사)
2016 동아대학교 대학원 졸업(박사)
2016~현재 한국전자통신연구원 선임연구원
2022~현재 엑소브레인 과제 책임자, 전문가 의사결정지원 과제 책임자
관심분야 : 대형언어모델, 자연어처리, 설명가능한 AI, 생성형 AI
Email : kyoungman.bae@etri.re.kr