멀티모달 생성형 AI 기술 동향

제주대학교 ▌윤여찬

1. 서 론

인공지능 분야는 생성형 AI의 등장과 성공으로 패러다임의 변화를 목격했으며, 특히 OpenAI의 ChatGPT, Google의 Bard와 같은 모델이 널리 사용되면서 더욱 주목을 받고 있다. 또한 Llama, Vicuna 및 Polyglot 등 상대적으로 소형의 대규모 언어 모델(LLM)이 오픈소스로 공개되면서 기업, 대학 등에서도 연구에 박차를 가하고 있으며 보안, 도메인 적응이 필요한 부분에서 활용되며 이에 기반한 다양한 서비스가 출시되고 있다. 많은 성공과 관심을 기반으로 집중된 연구개발 투자는 빠른 시간안에 기존의 생성형 AI를 다양한 형태로 진화시키고 있으며, 그 중에서도 멀티모달 대규모 언어모델은 최근 빠르게 진화되며 연구되고 있는 분야 중 하나이다.

멀티모달 생성형 AI는 여러 데이터 유형의 정보를 입력으로 받아 처리하는 획기적인 AI 모델 유형으로 주목받고 있다. 이러한 모델은 기존의 텍스트 중심 접근방식을 넘어 텍스트, 이미지, 오디오 등 다양한 양식의 데이터를 통합하고 해석한다. 이러한 멀티모달 접근방식은 복잡한 상황 및 정보를 보다 총체적으로 이해할 수 있게 해주며, 다양한 형태의 정보를 해석하는 인간의 인지능력과 가까운 해석 능력을 보인다. 예를 들어서 타인의 감성을 분석하는 태스크의 경우 발화 텍스트, 표정 이미지 등 하나의 모달리티를 해석하는 연구에서 진화하여 표정, 음성, 발화 내용을 동시에 고려하면 인간의 복잡한 감성을 더 종합적으로 판단하여 정확한 감성을 인식하는 데 도움이 된다.

텍스트 기반 LLM으로 출발한 ChatGPT는 발표된 지 일 년 남짓한 기간 동안 산업계의 혁신을 주도하며 어마어마한 성공을 이루어 냈다. 최근에는 ChatGPT 또한 업그레이드를 통하여 이미지를 입력으로 받아 처리하는 멀티모달 LLM으로서 한 단계 진화하였으며, 최근 발표된 구글의 제미나이(Gemini)는 시연 동영상에서 비디오를 입력으로 받아 처리하는 것을 확인할 수 있다. 하지만 현재도 Bard, Clova X 등의 LLM은 텍스트 기반의 LLM으로 구성되어 있으며 이 경우에 텍스트 콘텐츠를 이해하고 생성하는 데 탁월하지만, 다양한 모달리티를 고려하여 종합적인 판단을 하는 데는 한계점이 명확하다. 멀티 텍스트, 이미지, 사운드의 정보를 통합하고 합성할수록 더욱 풍부하고 맥락에 맞는 이해를 제공할 수 있으며, 이러한 기능은 시각적 및 청각적 단서에서 감정을 해석하거나 다양한 데이터 유형의 조합이 포함된 시나리오를 이해하는 등 세상을 종합적으로 파악해야 하는 작업에 매우 중요하다.

멀티모달 생성형 AI는 최근 많은 관심을 받고 발전되어 가고 있다. OpenAI의 CLIP과 Google의 멀티모달 트랜스포머는 이미지 캡션과 시각적 질의응답 등 여러 모달리티를 아우르는 작업에서 이러한 모델의 잠재력을 보여주었다. 이러한 발전은 멀티모달 LLM의 기술력을 보여줄 뿐만 아니라 AI가 인간의 인지와 유사한 방식으로 세상과 상호 작용하고 이해할 수 있는 미래를 암시한다고 할 수 있다.

2. 멀티모달 생성형 AI 연구 동향

생성형 AI 분야는 최근 몇 년 동안 다양한 전문 분야로 연구가 다각화되면서 상당한 발전을 거듭하고 있다. 이 부문에서는 기존의 자연어처리 기반의 LLM 동향에서 시작하여 비디오, 오디오 기반의 최신 생성형 AI 연구 동향과 함께 생체신호 등 다양한 모달리티를 활용한 최신 연구 동향을 살펴보도록 한다.

2.1 자연어처리 기반의 생성형 AI

자연어 중심의 대규모 언어 모델(LLM)은 현재 AI 연구의 최전선에 있다고 해도 과언이 아니다. 이 모델들은 기본적인 텍스트 처리에서부터 복잡한 구조의 문장과 문맥을 이해하고 답변을 생성하며, 사용자의 정보를 기억하며 적절한 대화를 진행하는 것이 가능하다. Tack[1]은 Blender와 GPT-3와 같은 최신 생성

모델이 교육을 목적으로 한 대화에서 효과적인 AI 교사가 될 수 있는지를 분석하였다. 이 연구는 학생을 이해하고 도와주는 능력 측면에서 실제 교사와 이 AI 모델들을 비교하는 방법을 소개하였고, 대화를 위한 생성형 AI가 교육목적으로 활용 가능성이 높음을 시사하였다. Han[2]은 의료 도메인에 생성형 AI 모델을 학습시킨 엔진과 데이터셋을 발표하였다. 의료분야에서 대화형으로 동작 가능한 AI 시스템 개발을 위하여 160,000개의 데이터셋을 구축하였고 의료 교육과 환자 관리 분야 활용을 염두에 두었다. Meyer[3]는 대화형 AI를 GPT를 통해 학습할 때, 사람이 직접 생성한 데이터를 대량 구축하는 대신, GPT로 자동으로 데이터를 생성하는 방식을 활용하였다. 이는 최근 GPT를 활용한 상당수의 연구에서 활용되는 방식으로 인공지능이 만든 데이터가 사람이 만든 데이터만큼 품질이 높다는 것을 시사한다. 산업을 위한 대화형 AI에 관한 연구도 LLM을 기반으로 이루어지고 있다. Singh[4]는 드릴링 및 생산 모니터링과 관련된 질문에 답변하고, 진단 분석을 수행하며, 운영 개선을 위한 권장사항을 생성하는 대화형 생성 AI 챗봇을 구축하기 위한 방법론을 제안하였다. Colabianchi[5]는 제조 분야에서 음성 대화 시스템(SDS)의 포괄적인 탐색을 다룬다. 개념적 아키텍처와 분류를 설정하여 SDS 요소의 설계 및 선택을 안내하고, 실제 사례 응용 프로그램을 통해 연구 결과를 검증하며 개선 영역을 강조하였다.

LLM은 창의적인 글쓰기 능력을 발휘할 수 있도록 진화해왔다. Rivero[6]는 자유롭게 작성된 글을 공무문서 등의 Formal 형태의 문서로 스타일 변환하는 방법에 대하여 발표하였다. GPT2를 사용하였고, 외부에 제출해야 하는 형식화된 문서를 작성할 때 유용하게 활용할 수 있다. Meroño-Peñuela[7]는 GPT2를 이용하여 논문 작성을 도울 방법을 제안하였다. 이 연구는 2002년부터 2019년까지의 국제 시맨틱웹 컨퍼런스에 발표된 논문을 학습 셋으로 사용하였다. Lee[8]와 Athanassopoulos[9]는 각각 ChatGPT를 활용한 논문 작성 방법과 외국어학습 방법을 제시하다. 어려울 수 있는 글쓰기를 생성형 AI를 도구로 사용하여 보조받을 수 있는 방법에 대하여 제안하고 있다.

2.2 이미지 및 비디오 기반 생성형 AI

자연어처리를 위해 개발되었는 GPT 기술은 이미지 및 비디오 분야로 확대되어 높은 성능을 보여주고 있다. 비디오 기반의 생성형 AI는 크게 비디오 혹은 이미지를 생성하는 모델과 비디오 혹은 이미지를 분석하여 다양한 작업을 처리해주는 태스크로 분류할 수 있다. VideoGPT[10]는 생성형 딥러닝 기술로 빈번하게 활용되는 Variational Auto Encoder(VAE)와 GPT 아키텍처를 기반으로 하는 새로운 비디오를 생성해준다. VideoGPT는 이미지 데이터셋인 BAIR Robot 데이터셋과 UCF-101, Tumblr GIF 데이터셋에서 높은 품

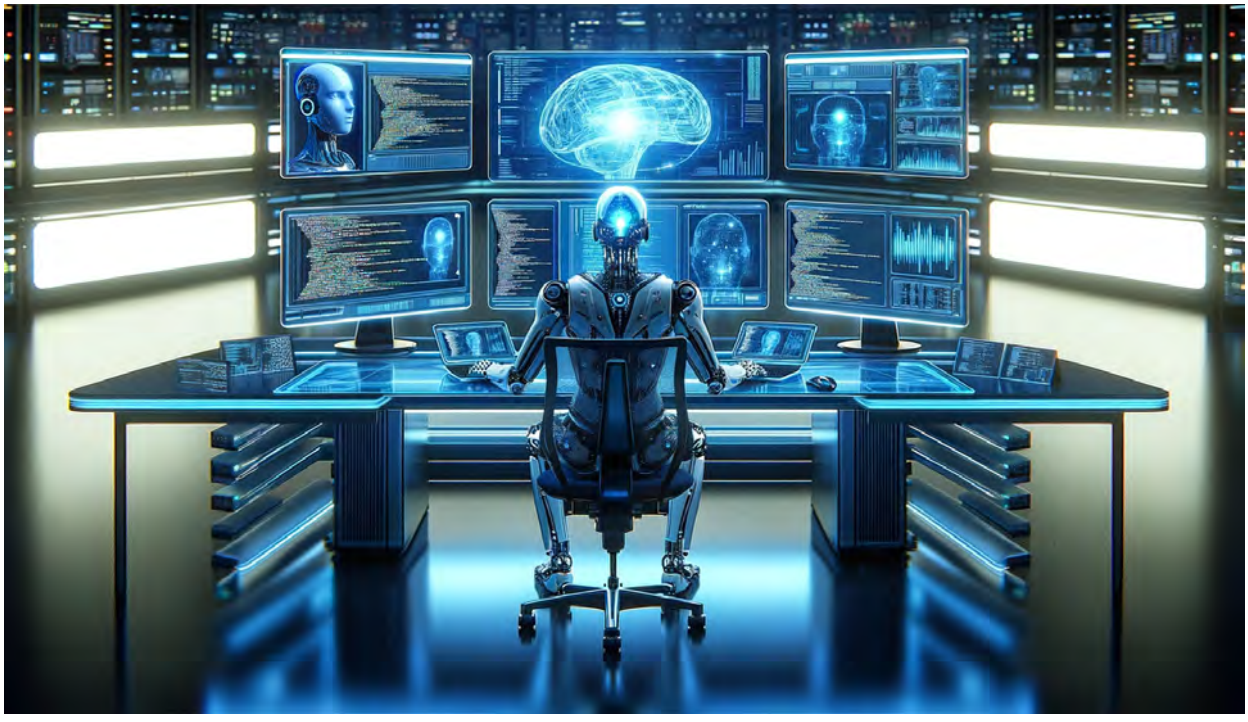

그림 1 생성형 AI DALL-E로 생성한 이미지, 키워드: 멀티모달 AI

질의 자연스러운 비디오를 생성할 수 있으며, GAN 기반의 비디오 생성 모델에 비해 높은 성능을 보여주었다. CogVideo[11]는 90억 개의 매개변수를 가진 트랜스포머 모델로, 입력된 텍스트를 이용하여 비디오를 생성할 수 있도록 대규모 사전 학습을 수행하였다. 이를 통하여 복잡한 움직임에 대한 정확률을 높였고, 생성 시에 변화의 크기를 제어할 수 있도록 하였다. DALL-E[12]는 Open AI의 텍스트 기반 이미지 생성 AI로 ChatGPT 사이트에서 이용할 수 있다. DALL-E는 대규모 생성 모델을 활용하여 텍스트-이미지 생성 작업을 개선하는 접근방식을 제안하였다. 12조 개의 매개변수를 가진 트랜스포머를 인터넷에서 수집한 2억 5000만 개의 이미지-텍스트 쌍으로 학습하여 고품질 이미지 생성 모델을 개발하였다. 이 모델은 인간이 평가하였을 때 90%의 선호를 가진 이미지를 생성하며 높은 품질로 컨셉아트 생성, 이모티콘 생성 등 다양한 분야에서 활용할 수 있다.

그림 1은 DALL-E로 생성한 그림 이미지이다. 그림 1과 같이 간단한 키워드 입력으로 높은 품질의 이미지를 생성할 수 있다. Stable-Diffusion[13]은 고해상도 이미지 합성을 위한 Latent Diffusion Models (LDM)을 사용하였다. LDM은 고해상도 이미지 생성, 이미지 인 페인팅, 클래스 조건부 이미지 합성에 있어 새로운 최고 기준을 달성하였으며, 높은 품질의 이미지를 생성한다. 이미지 생성을 위한 AI 기술이 보급됨에 따라 Oppenlaender[14]은 생성형 AI를 이용한 디지털 이미지 및 예술 작품 생성에 대하여 분석하였다. 이 논문은 텍스트-이미지 생성은 'AI 예술'이라 지칭하였고, '프롬프트 엔지니어링'을 통해 생성되는 이미지의 품질이 획기적으로 개선될 수 있음을 보여주었다. NExT-GPT[15]는 다양한 모달리티(텍스트, 이미지, 비디오, 오디오)를 이해하고 생성할 수 있는 멀티모달 대규모 언어모델이다. 이 모델은 기존의 멀티모달 언어모델이 텍스트 기반 입력에만 초점을 맞춘 것과 달리, 다양한 모달리티의 입력과 출력을 모두 처리할 수 있다. NExT-GPT는 세 단계로 구성되어 있으며, 입력을 인코딩하고, 중앙의 LLM을 통해 의미를 이해하고 추론하며, 다양한 모달리티의 콘텐츠를 생성한다. 이 모델은 미리 학습된 인코더와 디코더를 사용하여, 적은 비용으로 효과적인 학습과 확장을 할 수 있다.

이미지 혹은 비디오를 분석하여 활용하는 기술 역시 많은 관심을 받으며 연구가 진행되고 있다. BLIP: Bootstrapping Language-Image Pre-training[16] 모델은 비전-언어 태스크에서의 이해 및 생성 능력을 향상하기 위해 개발되었다. 이 모델은 이미지와 텍스트 쌍을 사용하여 사전 학습되며, 이를 통해 더 나은 이미지-텍스트 검색, 이미지 캡셔닝, 이미지 기반 질의응답이 가능하다. BLIP은 노이즈가 많은 웹 데이터를 활용하여 이미지에 대한 캡션을 생성하고, 부정확한 캡션을 제거하여 데이터의 질을 향상하고 모델의 성능을 극대화하였다. 또한 다양한 비전-언어 태스크에서 최신 기술 수준을 달성하였다. Visual ChatGPT[17]는 대화형 AI(ChatGPT)와 Stable Diffusion, BLIP과 같은 대형 비전 모델을 결합한 시스템으로, 언어뿐만 아니라 이미지도 처리하고 생성할 수 있게 한다. 이 시스템은 사용자의 다양한 질문이나 이미지편집 요청을 처리하며, 여러 AI 모델을 활용한 다단계 파이프라인을 구성하여 작동한다. Visual ChatGPT는 22개의 다른 대형 비전 모델을 포함하며, 이들 간의 상호 작용을 위한 '프롬프트 매니저'를 설계하여 각 모델의 기능과 입력-출력 형식을 정한다. 또한 ChatGPT가 복잡한 시각적 작업을 처리할 수 있도록 지원한다. 그림 2는 VisualChatGPT를 활용하여 배경을 합성한 예제를 보여준다.

Video-LLaMA[18] 는 비디오의 시각 및 청각적 내용을 이해할 수 있는 대규모 언어모델을 제공하는 멀티모달 프레임워크이다. 이는 기존 LLM이 시각 혹은 오디오 신호만 처리하는 것과 달리, 시간적 변

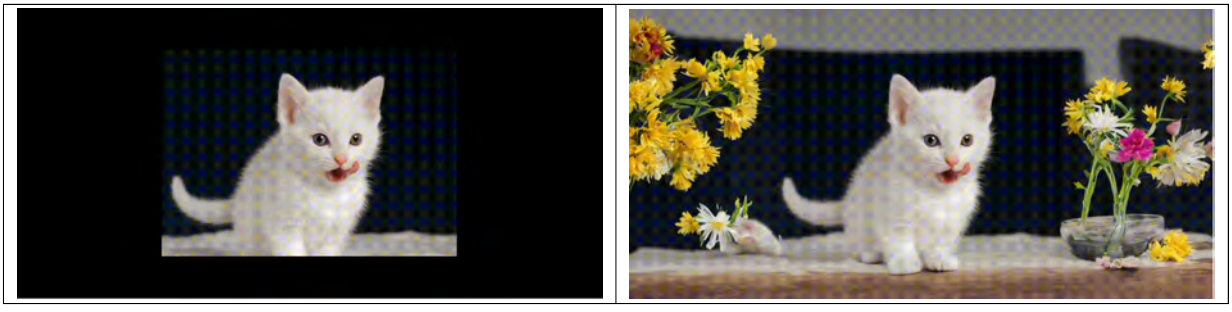

그림 2 VisualChatGPT를 이용한 배경합성
출처 : https://github.com/moymix/TaskMatrix?tab=readme-ov-file

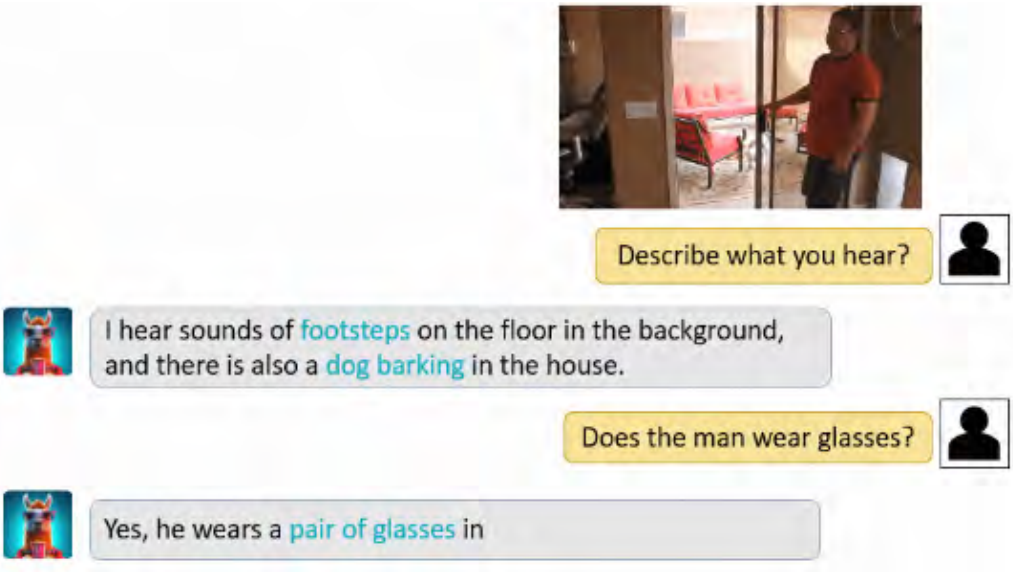

그림 3 VideoLLaMa 기반의 비디오 질의응답 예제
출처 : https://github.com/DAMO-NLP-SG/Video-LLaMA

화를 캡쳐하고 오디오-비주얼 신호를 통합함으로써 비디오를 분석하고 이해할 수 있게 한다. 그림 3은 Video-LLaMa를 사용한 비디오 분석을 보여준다. 음성 신호와 시각적신호를 분석하여 사용자의 질의에 대답하는 것을 확인할 수 있다. Video-LLaMA는 미리 학습된 이미지 인코더와 오디오 인코더를 사용하여 비디오와 언어 간의 상호 작용을 학습한다. 이 모델은 멀티모달 입력을 처리하고 비디오 콘텐츠를 분석하여 사용자의 요청에 응답할 수 있다.

2.3 그 밖의 모달리티를 활용한 생성형 AI

비디오와 이미지, 텍스트 모달리티 이외에도 오디오와 인간의 생체신호 등을 활용한 생성형 AI에 관한 연구도 이루어지고 있다. 대표적인 분야로 오디오 생성을 꼽을 수 있다. AUDIOGEN[19]은 생성형 AI 기술을 이용하여 입력된 텍스트를 기반으로 오디오를 생성하는 연구를 수행하였다. Ghosal[20]은 지시 데이터셋을 활용하여 텍스트-오디오 변환(Text-to-Audio, TTA) 작업을 생성형 AI로 수행하였다. 최근 대규모 언어 모델(Large Language Models, LLM)의 발전은 지시(instruction) 및 사고 과정(chain-of-thought) 기반의 미세 조정을 가능하게 하여 다양한 자연어 처리(NLP) 과제에서 탁월한 성능을 보여주었고 Ghosal은 LLM과 지시데이터를 이용하여 오디오 생성의 성능을 획기적으로 증가시켰다. Zhang[21]은 대규모 언어 모델(Large Language Models, LLMs)을 활용하여 인간의 신경 상태와 언어 이해 능력을 분석하였다. 특히, 텍스트 내 특정 키워드와 관련된 단어들의 중요도에 따라 뇌가 어떻게 단어를 처리하는지에 집중하였다. 이를 위해 LLM, 눈 움직임 추적(eye-gaze), 그리고 뇌파(EEG) 데이터를 함께 분석하는 방법을 제안하였으며, 키워드와 높은 관련성을 가진 단어들은 더 많은 눈 움직임을 유발한 것을 확인하였다.

3. 멀티모달 생성형 AI 응용 분야

2장에서는 멀티모달 생성형 AI 기반의 다양한 연구를 소개하였다. 이 장에서는 가능한 활용방안에 대하여 설명하고자 한다. 앞 장에서 기술하였다시피 멀티모달 생성형 인공지능(AI)은 다양한 형태의 데이터를 처리하고 생성하는 능력을 갖춘 혁신적인 기술이다. 이 기술은 교육 분야에서 맞춤형 콘텐츠를 제작하여 개인화된 학습 경험을 제공하는 데 활용할 수 있다. 예를 들어 인터넷 강의 영상을 보고 학생의 질문에 생성형 AI가 즉각적으로 해답을 제공해 주는 것이 가능하다. 건강 관리에서는 환자의 의료 기록, 영상, 음성 데이터를 분석해 진단과 치료 계획을 지원할 수 있다. 치료, 진단, 환자와의 소통에 의료인의 역할을 생성형 AI가 보조할 수 있다. 엔터테인먼트 산업에서는 영화, 음악, 게임 제작에 창의적인 기여를 할 것으로 기대된다. VisualChatGPT와 같은 기술을 통해 빠르고 효율적으로 영상 편집하는 것이 가능해질 것으로 예측된다. 자동차 산업에서는 자율주행 차량이 다양한 센서 데이터를 통합해 주변 환경을 인식하고 반응하는 데 생성형 AI가 중요한 역할을 할 것으로 기대된다. 마케팅 분야에서는 소비자 데이터를 분석하

여 맞춤형 광고와 전략을 개발하고, 보안 분야에서는 감시 카메라의 영상과 오디오 데이터를 분석해 보안 위협을 식별하는데 생성형 AI가 활용될 것이다. 로보틱스에서는 로봇이 다양한 데이터를 해석하여 더욱 정교한 작업을 수행하게 하고, 고객 서비스 분야에서는 음성 인식과 이미지 분석을 통해 고객 문의에 효율적으로 대응하는 것이 가능해질 것이고, 디자인과 예술 분야에서는 새로운 작품 창조나 아이디어 시각화를 돕고, 과학 연구에서는 실험 데이터를 분석하여 연구 결과를 도출할 수 있도록 생성형 AI가 활용될 것이다. 언어 번역에서는 시각적 정보까지 활용하여 다양한 언어의 텍스트와 음성 데이터를 이해하고 번역하며, 재난 관리에서는 위성 이미지와 현장 데이터를 분석해 즉각적이고 효율적으로 재난 대응을 지원할 수 있게 될 것이다. 마지막으로 스마트 홈 기술에서는 가정 내 센서 데이터를 분석해 기존의 스마트 홈 기술을 획기적으로 진보시킬 것이다. 이처럼 멀티모달 생성형 AI는 그 응용 범위가 매우 넓고 다양한 분야에서 혁신을 가져올 것으로 예측된다.

4. 전망 및 결론

본 원고에서는 멀티모달 생성형 AI의 현재 연구 동향과 응용 가능한 분야에 대하여 살펴보았다. 최근 생성형 AI는 4차 산업혁명의 핵심 기술로 떠오르고 있으며, 집중적으로 기술투자가 이루어지고 있는 분야 중 하나이다. 최근 발표한 구글의 제미니는 생성형 AI 기술의 한계를 가늠할 수 없게 하였고, 새로운 시대가 오고 있음을 알려주고 있다. 앞으로 생성형 AI의 주도권을 가진 기업가 국가가 혁신의 주역으로 떠올 것이 확실시되고 있어, 산학연계의 많은 관심과 투자가 절실하다

참고문헌

[1] Tack, Anaïs, and Chris Piech. “The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues.” arXiv preprint arXiv:2205.07540 (2022).

[2] Han, Tianyu, et al. “MedAlpaca--An Open-Source Collection of Medical Conversational AI Models and Training Data.” arXiv preprint arXiv:2304.08247 (2023).

[3] Meyer, Selina, et al. “Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai.” Proceedings of the 4th Conference on Conversational User Interfaces. 2022.

[4] Singh, Ajay, Tianxia Jia, and Varun Nalagatla. “Generative AI Enabled Conversational Chatbot for Drilling and Production Analytics.” Abu Dhabi International Petroleum Exhibition and Conference. SPE, 2023.

[5] Colabianchi, Silvia. “Advancing Spoken Dialog Systems for Manufacturing: From Conceptual Architecture and Taxonomy to Real Case Applications and Future Directions.” Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems. 2023.

[6] de Rivero, Mariano, Cristhiam Tirado, and Willy Ugarte. “FormalStyler: GPT based Model for Formal Style Transfer based on Formality and Meaning Preservation.” KDIR. 2021.

[7] Meroño-Peñuela, Albert, Dayana Spagnuelo, and GPT-2. “Can a Transformer Assist in Scientific Writing? Generating Semantic Web Paper Snippets with GPT-2.” The Semantic Web: ESWC 2020 Satellite Events: ESWC 2020 Satellite Events, Heraklion, Crete, Greece, May 31 – June 4, 2020, Revised Selected Papers 17. Springer International Publishing, 2020.

[8] Lee, Ping Yein, et al. “Use of ChatGPT in medical research and scientific writing.” Malaysian Family Physician: the Official Journal of the Academy of Family Physicians of Malaysia 18 (2023): 58.

[9] Athanassopoulos, Stavros, et al. “The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom.” Advances in Mobile Learning Educational Research 3.2 (2023): 818-824.

[10] Yan, Wilson, et al. “Videogpt: Video generation using vq-vae and transformers.” arXiv preprint arXiv:2104.10157 (2021).

[11] Hong, Wenyi, et al. “Cogvideo: Large-scale pretraining for text-to-video generation via transformers.” arXiv preprint arXiv:2205.15868 (2022).

[12] Ramesh, Aditya, et al. “Zero-shot text-to-image generation.” International Conference on Machine Learning. PMLR, 2021.

[13] Rombach, Robin, et al. “High-resolution image synthesis with latent diffusion models.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[14] Oppenlaender, Jonas. “The creativity of text-to-image

generation.” Proceedings of the 25th International Academic Mindtrek Conference. 2022.
[15] Wu, Shengqiong, et al. “Next-gpt: Any-to-any multimodal llm.” arXiv preprint arXiv:2309.05519 (2023).
[16] Li, Junnan, et al. “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.” International Conference on Machine Learning. PMLR, 2022.
[17] Wu, Chenfei, et al. “Visual chatgpt: Talking, drawing and editing with visual foundation models.” arXiv preprint arXiv:2303.04671 (2023).
[18] Zhang, Hang, Xin Li, and Lidong Bing. “Video-llama: An instruction-tuned audio-visual language model for video understanding.” arXiv preprint arXiv:2306.02858 (2023).
[19] Kreuk, Felix, et al. “Audiogen: Textually guided audio generation.” arXiv preprint arXiv:2209.15352 (2022).
[20] Ghosal, Deepanway, et al. “Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model.” arXiv preprint arXiv:2304.13731 (2023).
[21] Zhang, Yuhong, et al. “Integrating LLM, EEG, and Eye-Tracking Biomarker Analysis for Word-Level Neural State Classification in Semantic Inference Reading Comprehension.” arXiv preprint arXiv:2309.15714 (2023).

약 력

윤 여 찬

2004 고려대학교 컴퓨터학과 졸업(학사)
2007 고려대학교 컴퓨터학과 졸업(석사)
2020 고려대학교 컴퓨터학과 졸업(박사)
2007~2022 한국전자통신연구원 책임연구원
2022~현재 제주대학교 인공지능전공 조교수
관심분야 : 자연어처리, 멀티모달 딥러닝
Email : ycyoon@jejunu.ac.kr