

창의적 콘텐츠 생성에서 GPT-4의 성능 평가: 인간 작성물과의 비교 연구

김성은¹ · 김자경¹ · 오동석^{2*}

¹경북대학교 영어영문학과 박사수료

²경북대학교 영어영문학과 조교수

Performance Evaluation of GPT-4 in Creative Content Generation: A Comparative Study with Human Outputs

Sungeun Kim¹ · Jakyoun Kim¹ · Dongsuk Oh^{2*}

¹Ph.D. Candidate, Department of English Language and Literature, Kyungpook National University, Daegu 41566, Korea

²Assistant Professor, Department of English Language and Literature, Kyungpook National University, Daegu 41566, Korea

[요약]

대규모 언어 모델 (LLMs)은 다양한 분야에서 경쟁력 있는 성능을 보여주었으며, 특히 창의성을 요구하는 작업에서 광범위한 응용 서비스를 제공한다. 본 논문은 GPT-4와 같은 대규모 언어 모델이 창의적 글쓰기를 생성하는 데 있어 인간의 텍스트 작성물과 비교하여 어떤 성과를 내는지 평가한다. 본 연구는 GPT-4가 생성한 100개의 텍스트와 인간이 작성한 100개의 텍스트를 유창성, 유연성, 독창성, 구체성, 유용성 및 특정 창의적 전략과 같은 벤치마크를 사용하여 분석한다. 연구 결과에서 GPT-4는 높은 품질의 창의적 콘텐츠를 생성하며 인간의 성과에 근접하는 것으로 나타난다. 평가자들 간의 일관성 부족에도 불구하고, GPT-4는 인간의 창의력을 향상시키고 창의적 글쓰기의 질을 개선할 수 있는 상당한 잠재력을 보여준다. 그러나 GPT-4는 사실 정보와 역사적 배경에 의존하는 훈련 데이터로 인해 인간 창의성과의 중요한 차이를 나타낸다.

[Abstract]

Large language models (LLMs) have demonstrated competitive performance across various domains, particularly in tasks requiring creativity, and thus offer a wide range of applications. This study evaluates the performance of large language models (LLMs), such as GPT-4, in generating creative writing through a comparison with text authored by humans. This study evaluates 100 creative writing pieces generated by GPT-4 and 100 texts authored by humans, using parameters such as fluency, flexibility, originality, elaboration, usefulness, and specific creative strategies. The findings reveal that GPT-4 is able to emulate the performance of human authors closely, producing high-quality and creative content. Despite the inconsistencies among the evaluators, GPT-4 demonstrates the significant potential to enhance human creativity and improve the quality of creative writing. However, the limitations inherent to the training data of GPT-4, including its dependence on factual and historical background information, indicate critical differences from human creativity.

색인어 : 챗GPT 4, 대규모 언어 모델, 창의적 글쓰기, 창의적 콘텐츠, 창의성 평가

Keyword : GPT-4, Large Language Models (LLMs), Creative Writing, Creative Contents, Creativity Evaluation

<http://dx.doi.org/10.9728/dcs.2024.25.9.2479>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 30 July 2024; Revised 04 September 2024

Accepted 23 September 2024

*Corresponding Author; Dongsuk Oh

Tel: +82-53-950-7155

E-mail: inow3555@knu.ac.kr

1. 서론

인공지능(AI: Artificial Intelligence)의 발전은 데이터 분석부터 창의적 콘텐츠 생성에 이르기까지 다양한 분야에서 혁신을 일으켰다[1]. 이러한 기술적 패러다임의 변화에는 GPT-4, Gemini, Llama와 같은 대규모 언어 모델(LLMs; Large Language Models)이 중요한 역할을 하고 있다. 대규모 언어 모델(LLMs)은 방대한 데이터와 정교한 알고리즘을 통해 인간과 유사한 텍스트를 생성하는 데 탁월하다[2],[3]. 특히 GPT-4와 Gemini 같은 폐쇄형 모델은 언어의 정밀성과 콘텐츠 생성의 복잡성에서 뛰어난 성능을 보여준다[4],[5].

폐쇄형 모델은 기업이나 기관에서 개발하고 유지 관리하며, 접근 권한이 제한된 모델을 의미한다. 이들은 독점 데이터와 고도의 컴퓨팅 자원을 활용하여 개발되며, 높은 정확도와 정밀도를 자랑한다[4],[5]. 반면, Meta의 LLaMA (Large Language Model Meta AI)와 같은 개방형 모델은 오픈소스 커뮤니티에서 개발되고 공개적으로 접근할 수 있어 다양한 사용자와 연구자가 쉽게 활용할 수 있다[6],[7]. 그러나, 폐쇄형 모델은 대규모 언어모델 벤치마크에서 개방형 모델보다 높은 성능을 보이고 있다[8].

폐쇄형 모델 중 GPT-4는 다양한 언어 작업에서 뛰어난 성과를 보이며, 인간이 작성한 것과 비교할 만한 고품질의 콘텐츠를 생성할 수 있다[4]. 이러한 이유로 본 연구에서는 GPT-4를 사용하여 대규모 언어 모델의 창의성을 조사하고, 인간 전문가의 평가에서 AI가 어떻게 수행하는지 탐구한다.

기존 연구들은 주로 AI가 생성한 콘텐츠의 완성도와 논리적 일관성, 예를 들어 문법성, 일관성, 관련성 등을 평가하였다[9]. 그러나 본 연구는 창의적 글쓰기에서의 창의성에 중점을 두고, GPT-4와 인간이 생성한 텍스트를 비교하여 유창성, 유연성, 독창성, 구체성, 유용성, 창의적 전략의 여섯 가지 창의성 기준에 따라 평가한다[10]. 인간이 작성한 글 100개 및 GPT-4가 생성한 100개 (총 200개)의 텍스트를 평가하였으며, 영어영문학 석사 학위 이상을 보유하고 있는 8명의 교육자가 엄격하게 검토하였다.

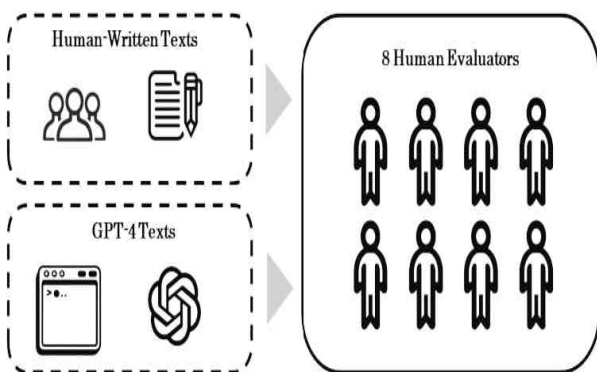


그림 1. 창의적 텍스트를 평가하는 과정
 Fig. 1. Process for evaluating creative texts

본 연구의 결과와 기여는 다음과 같다.

- 창의성 벤치마크를 통해 평가된 AI 생성 콘텐츠가 인간이 작성한 글과 비교할 때 유사한 결과를 생성할 수 있음을 보여주며, 창의적 글쓰기에서 AI의 잠재력을 검증한다. 본 연구는 창의적 글쓰기 평가에서 대규모 언어 모델을 사용하는 것이 효과적임을 제안하고 입증한다.

- 인간이 쓴 글과 AI 생성 콘텐츠를 비교 및 평가하여 창의적 글쓰기 과제에서 대규모 언어 모델(LLMs) 사용의 장점과 한계를 분석한다. 이를 통해 AI의 창의성에 미치는 영향을 탐구하고, AI가 창의적 작업에서 어떤 역할을 할 수 있는지에 대한 새로운 통찰을 제공하고자 한다.

II. 관련 연구

2-1 대규모 언어 모델

대규모 언어 모델(LLMs)은 최근 자연어 처리(NLP; Natural Language Process) 분야에 혁신적인 변화를 가져왔다. 트랜스포머 아키텍처의 도입과 함께 다양한 사전 학습 모델이 개발되었으며, OpenAI의 GPT-1은 대규모 데이터셋을 활용해 일관된 텍스트를 생성하는 첫 단계를 밟았다[11]. GPT-1은 1억 1700만 개의 파라미터로 구성되었으며, 반지도 학습과 지도 학습을 결합해 다양한 작업에 적응할 수 있는 범용 표현을 학습하였다[11].

BERT (Bidirectional Encoder Representations from Transformers)는 트랜스포머의 인코더를 사용하여 문맥을 양방향으로 이해하며, MLM (Masked Language Modeling) 과 NSP (Next Sentence Prediction)을 통해 자연어 이해 능력을 크게 향상시켰다[12]. RoBERTa (Robustly optimized BERT approach)는 더 많은 데이터, 더 긴 학습 시간, 하이퍼 파라미터 튜닝을 통해 BERT의 성능을 극대화하였으며, Dynamic Masking, Full Sentence without NSP (Next Sentence Prediction), Large Mini-batches, Larger Byte-Level BPE (Byte Pair Encoding) 등의 기법을 적용하였다 [13]. ALBERT (A Lite BERT)는 파라미터 공유와 임베딩 파라미터 분해를 통해 모델의 메모리 사용 효율성을 높였으며, 문장 순서 예측(SOP; Sentence Order Prediction)을 통해 문장 간 일관성을 향상시켰다. 이로 인해 ALBERT는 GLUE 및 SQuAD 벤치마크에서 뛰어난 성능을 보였다 [14]. GPT-2는 15억 개의 파라미터로 구성된 모델로, 일관된 장문의 텍스트를 생성하는 능력을 보여주었다 [15].

GPT-2는 트랜스포머 디코더를 사용하여 텍스트를 순차적으로 생성하며, Fine-Tuning 없이 Zero-shot 능력을 통해 Task를 수행할 수 있는 가능성을 보여주었다. 그러나, 여전히 정확하게 과제를 수행하기 위해서는 Fine-Tuning 과정을 수행해야 한다[15]. BART (Bidirectional and Auto-Regressive Transformers)는 인코더-디코더 구조를

결합한 디노이징 오토인코더를 사용하여 텍스트 생성, 번역, 요약에서 뛰어난 성능을 보였다[16]. 또한, 다양한 노이징 방법론을 통해 입력 텍스트를 재구성하고, seq2seq 모델을 사용하여 원래 텍스트를 복원하는 방식을 학습하여 다운스트림 작업에 적용할 수 있고, SQuAD와 GLUE 벤치마크에서 뛰어난 성과를 기록하며, 특히 번역, 대화 생성, 질문 응답, 요약 작업에서 이전 기술보다 높은 성능을 보여주었다[16]. T5 (Text-to-Text Transfer Transformer)는 모든 자연어 처리 (NLP) 작업을 Text-to-Text 문제로 변환하여 전이 학습의 효율성을 크게 향상시켰다[17]. T5의 모델 아키텍처는 기본 트랜스포머 구조를 그대로 사용하면서도, 모든 자연어 처리 (NLP) 작업을 하나의 통일된 프레임워크로 처리할 수 있어 높은 성능을 발휘한다[17].

GPT-3는 1750억 개의 파라미터를 통해 다양한 작업 수행 능력을 크게 향상시켰으며, zero-shot, one-shot, few-shot 학습 능력이 파라미터 수가 증가함에 따라 향상됨을 보여주었다[18]. GPT-1 (1.17억 파라미터)과 GPT-2 (15억 파라미터)와 비교했을 때, GPT-3는 대규모 파라미터를 보유하기 때문에 보다 복잡하고 다양한 언어 모델링 작업을 수행할 수 있다[18]. GPT-3.5는 InstructGPT의 RLHF (Reinforcement Learning with Human Feedback)와 동일하게 학습되었으며, 기존의 InstructGPT보다 대화형 응용 프로그램에 최적화되었으며, 안전성과 대화의 맥락 및 일관성을 유지하는 능력을 개선시켰다[19]. 이를 통해 GPT-3.5는 GPT-3보다 더욱 향상된 대화 성능을 제공한다[19].

GPT-4는 GPT-3.5와 다르게 다양한 형태의 이미지를 입력으로 받을 수 있고, 해당 이미지와 Prompt에 따라 답변을 도출할 수 있는 능력이 추가되었다[4],[20]. 그리고, 방대한 파라미터와 긴 텍스트를 통해 복잡한 의존성을 처리하는 능력이 뛰어나며, 다양한 대규모 언어모델을 평가하는 평가하는 리더보드에서 이전 모델들을 능가하는 성과를 보였다 [4],[8]. 이에 본 연구에서 대규모 언어 모델 중 GPT-4를 사용하여 인간과의 글을 비교하여 평가하였다.

2-2 창의성 평가 유형

창의성 평가 방법은 창의적 콘텐츠를 평가하기 위해 다양하고 체계적인 방법을 제공한다. 본 절에서는 창의성 평가 유형들을 설명하고, 창의성의 다양한 측면을 분석하여 창의적 결과물의 포괄적인 평가를 제공하는 방식을 논의한다.

Boden은 AI와 창의성에 관한 세 가지 유형의 창의성을 제시한다. 조합적, 탐구적, 변혁적 창의성으로 구분되며, 이는 AI를 사용하여 모델링하고 평가할 수 있다[21]. 조합적 창의성은 시적 이미지 혹은 비유와 같은 친숙한 아이디어의 새로운 조합을 생성하는 것을 포함한다[21]. 탐구적 창의성은 기존 스타일 내에서 새로운 아이디어를 생성하기 위해 구조화된 개념 공간을 탐구하는 것이다. 변혁적 창의성은 개념 공간의 하나 이상의 차원을 변경하여 이전에는 불가능했던 아이

디어를 생성하는 가장 심오한 유형이다[21].

Dietrich는 창의성의 인지 및 신경 메커니즘을 이해하기 위한 새로운 이론적 틀을 제안하며 이를 의도적 모드, 자발적 모드, 몰입 모드로 구분하였다[22]. 의도적 모드는 작업 기억 혹은 계획과 같은 실행 기능과 관련된 전두엽 피질에 의해 주도되는 의식적이고 노력적인 처리를 평가한다. 자발적 모드는 전두엽 피질의 개입이 적고 종종 갑작스러운 통찰을 이끌어내는 무의식적이고 하향식(bottom-up) 처리를 평가하는 데 초점을 맞춘다[22]. 몰입 모드는 암묵적 시스템과 운동 효율성에 의해 주도되는 고도로 집중되고 노력 없는 활동 동안의 창의성을 평가한다. 이러한 다양한 모드를 통해 창의성 평가의 복잡성을 포착하고, 다양한 창의적 과정의 본질을 더 잘 이해하고 평가할 수 있다[22].

Mozaffari의 연구에서는 창의적 글쓰기 지도 평가 방법으로 창의적 언어의 주요 특성인 이미지, 캐릭터화, 목소리, 이야기 등을 포함하는 분석적 평가 기준표를 사용하여 글을 평가하였다[23]. 이미지는 구체적이고 중요한 세부 사항과 문학적 수사를 통해 독자의 감각을 자극하며, 목소리는 작가의 감수성과 세계관을 반영하는 독특한 서술 스타일로 나타난다 [23]. 캐릭터화는 직접적 및 간접적 방법을 통해 캐릭터를 설득력 있게 개발하고, 이야기는 독자가 내러티브를 통해 자신의 결론을 도출할 수 있도록 정보를 제공한다[23]. 각 글은 우수, 양호, 보통, 미흡의 네 가지 달성 수준으로 평가되며, 이러한 평가지표는 신뢰성과 타당성을 통해 평가된다[23].

D'Souza는 서사적 글쓰기에 나타나는 창의성을 평가하기 위한 방법을 체계적인 문헌 검색을 통해 설명하고 있다[24]. 이 연구는 창의성 평가를 위해 사용된 다양한 설계와 결과를 통해 창의성 평가를 더욱 효과적으로 만들기 위한 방법을 탐구한다[24]. 연구 결과에 따르면, 창의성을 평가하기 위해 사용된 주요 기술로는 합의 평가 기법(CAT; Consensual Assessment Technique), 평가지표, 동료 평가, 그리고 컴퓨터 언어 분석이 있다[24]. 각 기술은 고유한 장점과 한계를 가지고 있으며, 창의적 글쓰기를 평가하는 데 있어 종합적인 접근이 필요함을 시사한다[24].

본 연구는 창의성 평가의 다양한 접근 방식을 바탕으로 Shively 등이 연구한 평가 기준표를 사용하였다. 이 기준은 Kaufman과 Beghetto의 창의성 평가 정의를 포함하고, Guilford가 제시한 유창성(아이디어의 양), 유연성(다양한 유형의 아이디어), 구체성(아이디어의 세분화), 독창성(아이디어의 독창성) 기준을 적용하였다[10],[25]. 또한, 창의성이 사회적 맥락 내에서 목적을 가져야 한다는 점을 강조한 Runco와 Jaeger가 연구한 유용성(아이디어의 실제 적용 가능성)과 특정 창의적 전략(특정 전략을 선택하여 적용하는 능력)도 포함하였다[10],[26]. 이러한 기준을 통해 인간과 AI, 특히 GPT-4의 창의적 결과물을 구체적인 기준에 따라 명확하고 편향되지 않게 비교하고자 한다.

III. 데이터 실험 설계

3-1 작업 데이터 구성

1) 데이터

창의적 표현을 자극하는 다양한 프롬프트를 기반으로 콘텐츠를 작성하는 프롬프트는 Kaggle의 WritingPrompts 데이터셋에서 선택되었으며, 이 데이터는 모두 영어로 제시되었다. 유창성, 유연성, 독창성, 구체성, 유용성 및 특정 창의적 전략의 여섯 가지 기준에 따라 이야기를 평가했다[27].

2) 프롬프트

프롬프트는 다양한 유형의 글쓰기를 포함한다. WritingPrompts 데이터셋에서 무작위로 100개의 프롬프트를 선택하여 주제의 다양성을 확보하였고, 동일한 주제의 프롬프트는 제외되었다. 프롬프트 유형 중 주요 프롬프트는 다음과 같이 분류된다[28].

- WP (Writing Prompt): 시나리오나 아이디어를 제공하여 작성자가 이를 기반으로 이야기를 작성하게 한다.
- CW (Constrained Writing): 특정 글자를 사용하지 않거나 엄격한 단어 제한 내에서 쓰는 것과 같이 특정 글쓰기 제약을 부여한다.
- EU (Established Universe): 기존의 허구적 세계를 확장하여 새로운 이야기, 캐릭터, 사건 등을 창작하게 한다.
- TT (Theme Thursday): 특정 주제에 관한 프롬프트로, 목요일마다 공개되어 해당 주제를 탐구하게 한다.
- PM (Prompt Me): 일반적 글쓰기와 반대로 작성자가 다른 사람들에게 특정 주제에 대한 글을 쓰도록 요청하는 방식으로 새로운 반전, 결과, 관점을 제시하게 한다.
- IP (Image Prompt): 이미지를 기반으로 이야기를 작성하게 한다.

3) 인간 글쓰기 인터페이스

참여자는 영어영문학을 전공하는 석사 이상 학위 보유자 한국인 10명으로 구성되었다. 각 참여자는 WritingPrompts 데이터셋에서 선정된 총 100개의 이야기 중 다른 주제로 10개씩 작성하였으며, 주제별 이야기는 400단어 이내로 작성되었다.

4) AI 글쓰기 인터페이스

GPT-4는 인간 참여자와 동일한 프롬프트와 조건을 사용하여 이야기를 생성하였다. 콘텐츠 생성은 OpenAI의 API를 사용하여 자동으로 수행되었다. GPT-4의 창의적 글쓰기를 위한 프롬프트는 다음과 같다.

“100개의 주제로 100개의 프롬프트가 제시되었다. 각 주제에 대해 창의적인 콘텐츠를 작성해야 한다. 글쓰기에는 여러 유형이 있다. 프롬프트에 [WP]가 포함된 경우, 창의적 아이디어를 제공하여 작성자가 이를 기반으로 콘텐츠를 작성하게 한다. [CW]의 경우, 특정 단어나 특정 스타일을 사용하는 등의 제약을 두고 글쓰기를 해야 한다. [EU]는 기존의 허구

적 세계를 확장하여 새로운 이야기, 캐릭터, 사건 등을 창작하게 해야 한다. [TT]는 특정 주제나 글을 쓰는 이유에 대해 글을 작성하고, 이를 이야기를 통해 나타내야 한다. [PM]은 새로운 반전, 결과, 관점을 중점으로 창의적 글쓰기를 해야 한다. [IP]의 경우 이미지를 기반으로 이야기를 작성하게 한다. 이 범주에 포함되지 않는 프롬프트의 경우, 일반적으로 창의적인 콘텐츠를 작성해야 한다. 글은 400단어 이내로 작성하고, 문장의 끝맺음은 마침표(.)로 끝내고 쉼표(,)로 끝내지 않도록 주의해야 한다. 지시 사항을 주의 깊게 읽고, 과제를 수행해야 한다.”

3-2 평가 기준

1) 창의성 평가 기준

본 연구는 Shively 등의 창의성 평가 연구에서 개발한 기준을 바탕으로 여섯 가지 창의성 측면을 평가하도록 설계되었다[10]. 창의성 평가는 유창성, 유연성, 구체성, 독창성, 유용성, 특정 창의적 전략의 기준 총 여섯 가지 항목을 구분하여 평가하였다. 이러한 다차원적 접근 방식을 통해 창의적 사고의 다양한 특성을 종합적으로 반영하므로, 공정하고 편향되지 않는 엄격한 평가를 하고자 하였다. 이를 통해 인간의 글과 AI의 글을 비교하여 창의적 사고 능력을 종합적으로 평가하고, 발전시킬 기회를 제공하고자 한다. 각 기준은 창의성의 고유한 측면을 포착하여 창의적 글쓰기 결과물을 포괄적이고 정량적으로 측정한다. 평가 척도의 경우 1점, 2점, 3점으로 나타내며, 1점은 초급 수준 (Novice), 2점은 중급 수준 (Developing), 3점은 고급 수준 (Expert) 단계로 구분하였다. 각 단계에 대한 구체적인 기준들과 평가 내용은 다음 실험 결과 표 1에 담고 있다[10].

- 유창성(Fluency): 아이디어의 양을 평가하며, 다수의 아이디어를 생성하는 능력을 의미한다. 이는 창의적 사고의 기초가 되는 능력으로, 다양한 해결책을 탐색하는 데 중요한 역할을 한다.
- 유연성(Flexibility): 다양한 유형의 아이디어를 고려하는 능력을 평가한다. 이는 문제 해결을 위해 다양한 접근 방식을 탐색하고, 고정된 사고방식에서 벗어나도록 돕는다.
- 독창성(Originality): 참신하고 새로운 아이디어를 생성하는 능력을 평가한다. 이는 창의성의 핵심 요소로, 기존의 아이디어를 뛰어넘는 혁신적인 해결책을 도출하는 데 기여한다.
- 구체성(Elaboration): 아이디어를 세부적으로 발전시키는 능력을 평가한다. 이는 아이디어의 실행 가능성을 높이고, 구체적이고 실현 가능한 해결책을 개발하는 데 도움을 준다.
- 유용성(Usefulness): 아이디어가 실제로 유용한지, 사용자에게 가치를 제공하는지를 평가한다. 이는 창의적인 아이디어가 실제 상황에서 적용 가능한지 판단하는 데 중요하다.

- 특정 창의적 전략(Specific Creativity Strategy): 창의적 사고를 향상하기 위해 특정 전략을 선택하고 적용하는 능력을 평가한다. 이는 창의적 문제 해결 과정에서 전략적 사고의 중요성을 강조한다.

2) 인간 평가 인터페이스

평가는 Google Forms를 사용하여 수행되었으며, 콘텐츠, 평가 지침 및 기준(유창성, 유연성, 독창성, 구체성, 유용성,

특정 창의적 전략)을 포함하였다[10]. 그림 2는 인간 평가 양식의 예를 보여준다. 각 항목은 1에서 3점까지의 척도로 평가되었으며, 1점은 초급 수준(Novice), 2점은 중급 수준(Developing), 3점은 고급 수준(Expert)을 나타낸다. 평가자들은 8명으로 모두 영어영문학과 석사 이상의 학위를 가진 한국인 영어 관련 교육자로 구성하였다. 평가자들은 텍스트의 출처(인간 또는 AI)에 대해 정보가 제공되지 않아 편향 없는 평가가 가능하도록 하였다.

표 1. 창의성 평가 척도

Table 1. Creativity evaluation results

	Novice	Developing	Expert
Fluency	Writers considered one idea.	Writers considered several ideas.	Writers considered many ideas.
Flexibility	Writers considered one type of idea.	Writers considered several types of ideas.	Writers considered many types of ideas.
Originality	Writers developed a common idea that many other writers would have suggested and/or replicated an existing idea.	Writers developed an interesting idea that several other writers would have suggested and/or minimally added onto an existing idea.	Writers developed a unique idea that few other writers suggested and/or substantially built upon an existing idea in a unique way.
Elaboration	Writers added minimal details and improvements to their ideas.	Writers added a few details and improvements to their ideas.	Writers added many significant details and improvements to their ideas.
Usefulness	Writers proposed ideas that may meet the end-user's needs in certain conditions.	Writers proposed ideas that would meet the end-user's needs.	Writers proposed ideas that would meet the end-user's needs and significantly add value to their lives.
Specific Creativity Strategy	Writers randomly selected and implemented a creative thinking strategy, and/or they were unable to leverage the strategy to improve their ideas.	Writers selected and implemented a creative thinking strategy to develop their ideas. They explained how the strategy supported their creativity.	Writers deliberately selected and implemented a creative thinking strategy to develop their ideas. They explained how the strategy supported their creativity.

Evaluation of Human writing and GPT-4 writing for the Quality of creativity

This is a survey comparing qualities of Texts between Human and GPT-4. Please rate us using the instructions below.

•Creativity evaluation criteria items

- 1.Fluency
2. Flexibility
3. Originality
- 4.Elaboration
- 5.Usefulness
6. Specific creativity strategy

•Scale

1. Novice > 1 point
2. Developing > 2 points
3. Expert > 3 points

Please rate each item as 1 point for novice, 2 points for developing, and 3 points for expert according to the standard description. (The higher the score, the higher the level)

1. In the case of Fluency items, novice writers considered one idea, and in the Developing 6. In the case of Specific creativity strategy items, novice is writers randomly selected and implemented a creative thinking strategy, and/or they were unable to leverage the strategy to improve their ideas. In the Developing stage, writers considered several types of ideas at the developing stage, and writers considered many type of ideas at the expert stage.
2. Flexibility has the criteria of writers considered one type of idea at the novice stage, writers considered several types of ideas at the developing stage, and writers considered many type of ideas at the expert stage.

story 27-1

On my first day at a new job in a large office building, I was thrilling with nerves and excitement, eager to make a good impression. Little did I know, I was about to face a challenge far beyond any typical workplace hurdle. As I was getting acquainted with my desk, a loud bell rang throughout the office at exactly twelve o'clock. Suddenly, the atmosphere shifted dramatically. My coworkers named Emily and Chris, who were friendly moments ago, transformed before my eyes. Desks became barricades, and office supplies turned into makeshift weapons. I stood there, stunned, as an announcement echoed through the building, revealing a shocking daily ritual: a purge-style battle for survival. For one hour, employees engaged in a fierce competition, with survivors earning not only the chance to return the next day but also a significant pay raise.

Adrenaline coursed through my veins as I quickly realized the seriousness of the situation. Survival instincts kicked in: I found myself ducking and weaving through the chaos, using my wits and whatever tools I could find to evade confrontation. It wasn't about fighting; it was about surviving until the bell rang again, signaling the end of the madness.

When it was over, the office slowly returned to its usual state of calm, as if nothing had happened. Those who made it through resumed their tasks, slightly richer and ready to face another day. I was among them, bewildered but unharmed, now initiated into the bizarre reality of my new workplace. This wasn't just a job; it was a daily test of survival, where the rewards extended beyond the paycheck.

	1	2	3
1. Fluency	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. Flexibility	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. Originality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4.Elaboration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. Usefulness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. Specific creativity stra...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Evaluation of Human writing and GPT-4 writing for the Quality of creativity

3. Originality is the novice stage where the writer developed a common idea that many other writers would have suggested and/or replicated an existing idea. In the Developing stage, the writer developed an interesting idea that several other writers would have suggested and/or minimally added onto an existing idea. In the Expert stage, the writer developed a unique idea that few other writers suggested and/or substantially built upon an existing idea in a unique way.

4 Elaboration has the following standards at the novice stage: writers added minimal details and improvements to their ideas. at the developing stage, writers added a few details and improvements to their ideas. and at the expert stage, writers added many significant details and improvements to their ideas.

5. In the case of Usefulness, novice refers to writers proposed ideas that may meet the end-user's needs in certain conditions. In the Developing stage, writers proposed ideas that would meet the end-user's needs. In the Expert step, writers proposed ideas that would meet. It is a standard of the end-user's needs and significantly add value to their lives.

6. In the case of Specific creativity strategy items, novice is writers randomly selected and implemented a creative thinking strategy, and/or they were unable to leverage the strategy to improve their ideas. In the Developing stage, writers selected and implemented a creative thinking strategy to develop their ideas. They explained how the strategy supported their creativity. Expert's steps were writers deliberately selected and implemented a creative thinking strategy to develop their ideas. They explained how the strategy supported their creativity.

story 27-2

As the clock struck twelve, a deafening bell echoed through the vast office building. I had just started my new job and was still getting acquainted with the labyrinthine corridors and cubicles. Suddenly, the humdrum of keyboards and printers ceased, replaced by an eerie silence. The once bustling office turned into a ghost town within seconds.

I looked around, puzzled. My colleagues had vanished, their desks abandoned. A chill ran down my spine as I noticed the sharpened pencils, staplers, and letter openers left behind. It was then that I saw it—a post-it note stuck to my computer screen: "Survive till 1 PM, get a pay raise." It read:

My heart pounded in my chest as realization dawned on me. This wasn't just an office; it was a battlefield. Every day at noon, a deadly game unfolded within these walls. Survival meant not just keeping your job, but also earning a significant pay raise.

The rules were simple: survive or perish. As the adrenaline coursed through my veins, I knew I had to adapt quickly. I grabbed a letter opener, ready to face whatever came my way. The office was no longer just a place of work; it was a war zone where survival was the only option.

	1	2	3
1. Fluency	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2. Flexibility	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3. Originality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4.Elaboration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. Usefulness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. Specific creativity stra...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

그림 2. 인간 평가 양식

Fig. 2. Human evaluation form

IV. 연구 결과

4-1 측정지표

본 연구의 정량적 평가는 평균(Mean), 표준 편차(Standard Deviation), 그리고 상호 평가자 일치도(IAA; Inter-Annotator Agreement)로 구성하였고, 자세한 설명은 다음과 같다.

평균은 창의성 평가 척도로 나타난 결과값들에 대한 평균 점수이다. 이는 각 텍스트의 점수를 합산한 후 텍스트의 총 개수로 나눈 값으로 계산된다. μ 는 평균, N 은 텍스트의 총 개수, x_i 는 각 텍스트의 점수를 의미한다. 평균 점수가 높을 경우, 창의성이 전반적으로 우수하게 평가되었음을 나타내며, 낮을 경우는 반대의 의미를 가진다[29].

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \tag{1}$$

표준 편차는 평균 주변의 점수의 분산 또는 변동성을 측정한다. 이는 데이터가 평균으로부터 얼마나 떨어져 있는지를 나타내며, 점수의 일관성을 평가하는 데 중요한 역할을 한다. σ 는 표준 편차, n 은 점수의 총 개수, \bar{x} 은 평균점수, x_i 는 각 점수를 의미한다. 표준 편차가 낮을 경우, 점수들이 평균에 가까이 몰려 있음을 의미하며, 높을 경우 점수들이 넓게 분포되어 있음을 의미한다[30].

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (\bar{x} - x_i)^2}{n - 1}} \tag{2}$$

상호 평가자 일치도는 여러 평가자들이 부여한 점수의 일관성을 평가한다. 이는 평가자 간의 신뢰성을 측정하는 데 사용되며, Krippendorff의 알파(α)를 사용하여 계산된다. D_o 는 관측된 불일치도, D_e 는 기대 불일치도, n 은 평가자 수, c 는 항목의 인덱스, k 는 범주의 수, δ_{ck}^2 는 c 와 k 사이의 평가 차이를 제공한 값, p_k 는 범주 k 의 비율, n_k 는 범주 k 에 속한 항목 수를 의미한다. IAA는 0과 1의 사이로 측정되며, 1에 가까울수록 평가자들 간의 점수 일치도가 높음을 나타내며, 0에 가까울수록 일치도가 낮음을 의미한다[31].

표 2. 창의성 평가 결과

Table 2. Creativity evaluation results

	Fluency		Flexibility		Originality		Elaboration		Usefulness		Specific Creativity Strategy	
	Mean/SD	IAA/%	Mean/SD	IAA/%	Mean/SD	IAA/%	Mean/SD	IAA/%	Mean/SD	IAA/%	Mean/SD	IAA/%
Human	2.10/ 0.67	0.14/ 0.99%	2.05/ 0.75	0.20/ 0%	2.06/ 0.76	0.16/ 0%	2.07/ 0.75	0.21/ 0%	2.02/ 0.67	0.16/ 0%	1.83/ 0.67	0.12/ 0%
GPT-4	1.87/ 0.68	0.15/ 0%	2.25/ 0.76	0.28/ 0%	2.23/ 0.72	0.16/ 0%	2.29/ 0.72	0.21/ 1%	2.19/ 0.70	0.28/ 0%	1.93/ 0.69	0.13/ 0%

$$\alpha = 1 - \frac{D_o}{D_e} \tag{3}$$

$$D_o = \frac{1}{n} \sum_c \sum_k \delta_{ck}^2 \tag{4}$$

$$D_e = \frac{1}{n} \sum_c \sum_k \delta_{ck}^2 = \frac{1}{n(n-1)} \sum_c n_c \sum_k n_{kmetric} \delta_{ck}^2 \tag{5}$$

4-2 주요 결과

인간과 GPT-4가 생성한 글의 창의성 평가에서 여섯 가지 기준에 대한 상세한 분석은 표 2에 제시되어 있다. 본 연구의 결과는 표 2와 같이 GPT-4가 창의적 콘텐츠 생성 작업에서 각 항목별로 인간 저자에 비해 어떤 강점과 한계를 가지고 있는지를 정량적으로 보여준다.

먼저, 유창성(Fluency)에서 인간이 작성한 글의 평균 점수는 2.10, GPT-4는 1.87로 나타났다. 두 그룹 표준 편차가 각각 0.67, 0.68로 거의 비슷한 분산도를 보였다. 상호 평가자 일치도(IAA)는 인간이 0.14, GPT-4가 0.15로 나타났다. 이는 유창성 평가에서 일관성이 낮음을 시사하며, 평가자 간의 주관적 차이가 있음을 반영한다. GPT-4는 아이디어의 양에서 인간과 비슷한 성능을 보였지만, 여전히 인간보다 약간 낮은 점수를 기록했다. 이는 GPT-4가 다양한 아이디어를 생성하는 데 있어 인간과 유사한 능력을 보이나, 완벽하지 않음을 시사한다.

유연성(Flexibility)에서는 GPT-4가 평균 2.25로 인간의 2.05를 앞선다. GPT-4의 상호 평가자 일치도(IAA)는 0.28로 인간의 0.20보다 높아, 텍스트 유연성에 대한 평가에서 약간 더 높은 일관성을 보였다. GPT-4는 다양한 문체와 어조를 자유자재로 사용하며, 여러 유형의 아이디어를 통합하는 능력이 뛰어나다는 점에서 인간보다 조금 더 높은 평가를 받았다. 이는 GPT-4가 다양한 문맥에서 효과적으로 스타일을 적용시키는 언어적 유연성을 보여줌을 의미한다.

독창성(Originality) 면에서 GPT-4는 평균 2.23으로 인간의 평균인 2.06을 약간 앞선다. 표준 편차는 인간의 글이 0.76, GPT-4는 0.72로 나타났다. 그리고 상호 평가자 일치도(IAA)의 결과값이 0.16으로 동일하여 독창성에 대한 평가자 간 일치도가 유사함을 보여준다. GPT-4는 특히 기존 아이디어를 변형하여 새로운 시각을 제시하거나, 전혀 새로운 아이디어를 창출하는 데 능숙함을 보였다.

구체성(Elaboration) 기준에서 GPT-4는 평균 2.29로 인간의 평균인 2.07보다 높다. 표준 편차는 인간의 글이 0.75, GPT-4는 0.72로 나타났다. 그리고 두 그룹 모두 0.21의 같은 상호 평가자 일치도(IAA) 결과를 보여주었다. 이는 GPT-4가 상세하고 확장된 콘텐츠를 제공하는 데 능숙함을 시사하고, GPT-4는 콘텐츠의 전체적인 흐름과 완성도에서 뛰어난 성능을 보였고, 텍스트 내의 세부 사항과 추가 정보를 통해 이야기를 더욱 풍부하게 만드는 데 강점을 보였다.

유용성(Usefulness) 부문에서 GPT-4는 평균이 2.19, 표준 편차 0.70으로 인간의 평균 2.02, 표준 편차 0.67보다 높다. GPT-4의 상호 평가자 일치도(IAA)는 0.28로 인간의 0.16보다 높다. 이는 GPT-4의 콘텐츠가 실제 상황에서 더 유용하거나 관련성이 있는 것으로 인식됨을 나타낸다. 또한, GPT-4는 독자가 쉽게 이해하고 활용할 수 있는 명확한 해결책과 실질적인 정보를 제공하는 데 강점을 보였다.

특정 창의적 전략(Specific Creativity Strategy)에서는 GPT-4가 평균 1.93으로 인간의 평균인 1.83을 앞선다. 표준 편차는 인간의 경우 0.67, GPT-4는 0.69의 값을 보여주었다. 상호 평가자 일치도(IAA)는 인간이 0.12, GPT-4가 0.13의 값을 나타내며, GPT-4가 창의적 사고 전략을 활용하는 데 능숙함을 시사한다. GPT-4는 다양한 창의적 전략을 효과적으로 적용하여 독창적이고 혁신적인 아이디어를 제시하는 잠재력을 보여준다.

종합적으로, 본 연구의 주요 결과는 GPT-4가 창의적 콘텐츠 생성의 여러 측면에서 인간 저자와 유사하거나 때로는 더 뛰어난 성능을 보였음을 나타내며, 유창성, 유연성, 독창성, 구체성, 유용성 및 특정 창의적 전략에서 높은 완성도의 창의적 콘텐츠를 생성할 수 있는 능력을 강조한다. 그러나 표준화된 분석 프레임워크를 사용하더라도 창의적 글쓰기 평가에는 본질적으로 주관성이 내포되어 있다. 상호 평가자 일치도(IAA)의 비율(%)의 의미는 8명의 평가자가 항목별로 정확하게 동일한 평가를 내린 평가 비율을 나타내며, 유창성(Fluency)에서 인간의 글은 0.99%, 구체성(Elaboration)에

서 GPT-4가 1%를 제외하면, 나머지 항목에서는 0%로 동일하게 판단되는 경우가 거의 없었다. 이는 평가자들 간의 주관적 차이가 크다는 것을 의미한다. 인간 평가자는 창의성의 다른 측면을 강조할 수 있으며, 개인적 편견이 점수에 영향을 미칠 수 있다. 이는 일관된 평가를 보장하기 위해 더 표준화되고 객관적인 평가 프레임워크의 필요성을 강조한다. GPT-4는 창의적 글쓰기의 여러 측면에서 인간 저자와 유사한 성능을 보였으며, 이는 GPT-4가 창의적 작업에서 중요한 도구가 될 수 있음을 시사하며, 인간의 창의성을 보완하고 증진시킬 수 있는 잠재력을 강조한다. 그러나 창의성 평가의 주관성을 고려할 때, 평가의 일관성과 객관성을 유지하기 위한 추가 연구가 필요하다. 이는 특히 창의적 글쓰기와 같은 창의성 평가에서 더욱 중요하다. GPT-4의 성능을 더욱 향상시키기 위해서는 다양한 문화적 배경과 장르를 포함한 데이터셋의 다양화와 표준화된 평가 방법의 개발이 필요하다.

4-3 사례 연구

정성적 평가는 GPT-4이 생성한 글을 인간이 쓴 글과 비교하여 창의성 평가에서 두드러진 강점 또는 약점을 보인 특정 사례를 조사한다. 이 분석은 유창성, 유연성, 독창성, 구체성, 유용성 및 특정 창의적 전략의 여섯 가지 창의성 기준과 특이점을 포함하여 분석하였다. 이러한 기준을 통해 예시들을 검토함으로써, 본 연구는 연구 결과에 대한 원인을 분석하고, GPT-4와 같은 대규모 언어 모델의 텍스트 생성에서의 창의성 향상을 목표로 한다.

1) 인공지능 글의 장점

인간이 쓴 글과 GPT-4가 생성한 콘텐츠 사이의 놀라운 유사성을 강조하며, 비슷한 맥락과 배경지식 그리고 어휘를 나타낸다. 특히 이야기 구조와 서사 방식 면에서 AI가 복잡한 시나리오에서 인간의 스토리텔링을 재현할 수 있는 능력을 보여준다. 이에 대한 예시는 표 3과 표 4와 같다.

표 3. 일관된 문맥과 전개

Table 3. Consistent context and development

[WP] Leonardo DiCaprio in a fit of rage begins to torpedo his own career by deliberately acting poorly and taking on bad films. He finally wins an oscar for starring in Paul Blart : Mall Cop 3.		
	GPT-4	Human
Contents	In a shocking turn of events, <i>Hollywood's golden boy</i> , Leonardo DiCaprio, has embarked on a self-destructive journey to torpedo his own illustrious career. The actor, known for his impeccable performances and choice of critically acclaimed films, has suddenly started delivering subpar performances and choosing questionable scripts. It's as if he's deliberately trying to tarnish his own reputation. <i>The first sign of this bizarre behavior was when he signed on for the third installment of the widely panned 'Paul Blart: Mall Cop' series.</i> Fans and critics alike were left scratching their heads at this unexpected decision. But the real shock came when DiCaprio, in an apparent fit of rage, began to act poorly.	A famous actor, Leonardo DiCaprio intentionally damages his career while he tries to win an Oscar. Interestingly, he chooses to star in terrible films and fails in them. The film <i>"Paul Blart: Mall Cop 3,"</i> which no one thought would be helpful for his reputation, was his choice to star in. However, his role of a mall cop going through a difficult time in this film manages to touch the hearts of all viewers. Award season, which includes the Oscars, is among us, and rumors exist that Leonardo DiCaprio might take home an award for this performance. To top it all off, he performs. His Oscar-winning performance was in a film that was meant to be a farce.

표 4. 일관된 문맥과 전개 유사한 구성과 어휘

Table 4. Similar composition and vocabulary

[WP] The most saddest love story you can ever think of.		
	GPT-4	Human
Contents	<i>In the quaint town of Verona, lived two souls, Amelia and Ethan, bound by an unspoken love. Their hearts beat in sync, yet their words never conveyed their feelings. Amelia, a painter, expressed her love through vibrant strokes on canvas, while Ethan, a poet, poured his heart into verses only Amelia could comprehend. One day, Amelia was diagnosed with a rare eye disease, gradually stealing her sight. The world she painted so vividly started fading into darkness. Yet, her love for Ethan remained undimmed. She continued to paint, guided by the memory of colors and the warmth of Ethan's voice. Ethan, on the other hand, was devastated. He wrote poems filled with despair, each word echoing his fear of losing Amelia to the encroaching darkness. His verses became his solace, his silent confession of love.</i>	<i>In a quaint village where seasons painted vivid tales, Emily and Kevin's romance blossomed under the warmth of summer suns. Their love, tender and fierce, promised a lifetime of shared dreams. Yet, destiny had woven a different path. Kevin, an artist whose soul found expression in his canvases, immortalized Emily with strokes that spoke of undying love. They envisioned a future together, building a life on hopes and whispers of forever. But a shadow fell upon their dreams as Kevin was stricken by an illness, its rarity matched only by the depth of their love. Emily became his unwavering sentinel, her presence a beacon through his darkest hours. Kevin, even as his strength waned, continued to create, his final masterpiece a tribute to their profound bond.</i>

표 3은 레오나르도 디카프리오에 대한 인간의 글과 GPT-4이 생성한 콘텐츠 일부분을 발췌한 내용을 담고 있다. 디카프리오의 분노에 사로잡혀 일부러 연기를 엉망으로 하고 나쁜 영화를 선택함으로써 자신의 경력을 망치기 시작하지만, 그는 “Paul Blart: Mall Cop 3에서 주연을 맡아 오스카상을 받는다는 관한 주제이다. 레오나르도 디카프리오가 오스카상을 받기 전 “Paul Blart: Mall Cop 3”이라는 작품 선택에 관한 이야기는 인간과 GPT-4 모두에서 유사한 맥락을 보여준다. 인간의 이야기는 디카프리오의 예상치 못한 경력 선택과 대중 및 비평가의 혼란을 상세히 묘사하며, 결국 오스카상 수상으로 결말을 맺는다. GPT-4도 유사한 방식으로 그의 명성과 대조되는 영화 선택, 오스카 수상, 그리고 성공의 아이러니를 강조한다. 두 콘텐츠 모두 비슷한 구성과 어휘를 사용하여 디카프리오의 행동의 극적이고 역설적인 특성을 강조한다. 이는 GPT-4가 인간의 창의성을 모방하여 흥미롭고 캐릭터 중심의 이야기를 생성할 수 있음을 보여준다. 또한, GPT-4가 창의적 글쓰기에서 여섯 가지 창의성 기준을 얼마나 잘 충족하는지를 보여준다. 인간과 AI가 생성한 이야기 모두 독자들을 끌어들이는 뛰어난 유창성을 가지고 있으며, GPT-4는 다양한 줄거리 전개를 처리하는 데 유연성을 발휘한다. 두 이야기 모두 독창성이 두드러지며, 혁신적 사고를 반영하는 독특하고 매력적인 주인공 특성을 표현한다. 풍부한 세부 묘사는 이야기의 전반적인 완성도를 높이고, 주인공의 성공과 반전 상황 같은 의미 있는 주제를 흥미롭게 전달한다. GPT-4는 글의 반전과 세밀한 캐릭터 개발과 같은 창의적 전략을 능숙하게 활용하여 이러한 주제를 효과적으로 전달한다.

슬픈 사랑에 관한 주제를 담고 있는 표 4의 콘텐츠는 인간과 GPT-4 모두 사랑이 예술을 통해 표현되고, 유사한 구성과 어휘를 사용한다. 인간의 글은 에밀리와 케빈의 사랑 이야기를 조용한 마을을 배경으로 펼쳐지며, 질병으로 인해 시험에 들게 한다. GPT-4는 베로나에서 아멜리아와 이든의 사랑 이야기를 시와 그림을 통해 표현하며, 아멜리아의 시력 저하에도 불구하고 사랑을 나누는 이야기가 펼쳐진다. 두 이야기 모

두 예술을 통해 캐릭터를 소개하며, 비극적 질병이 관계에 미치는 영향을 강조한다. 이는 GPT-4가 극적이고 로맨틱한 요소를 효과적으로 활용하여 인간의 스토리텔링을 잘 모방함을 보여준다. 이는 GPT-4가 창의적 글쓰기에서 여섯 가지 창의성 기준을 잘 충족하는지를 보여준다. 두 이야기 모두 유창성 면에서 뛰어나며, GPT-4는 로맨틱하고 극적인 문맥에 맞춰 스타일을 유연하게 적용한다. AI의 결과물은 사랑과 비극의 본질을 극적으로 포착하여 새로운 시각을 제공하는 독창성을 보인다. 상세한 묘사는 감정적 영향을 강화하며 이야기의 깊이와 풍부함을 더하고, 사랑이라는 의미 있는 주제를 전달하여, 감정적으로 청중과 공감대를 형성한다. 또한, GPT-4는 생생한 이미지와 감정적 깊이와 같은 특정 창의적 전략을 능숙하게 활용한다.

2) 인공지능 글의 단점

GPT-4는 학습된 데이터로 인해 새로운 텍스트를 창조하는 것이 어렵다. 이는 창의적 글쓰기의 다양하고 참신한 아이디어를 창출하는 데 GPT-4의 한계를 강조한다. 이에 대한 예시는 표 5와 표 6과 같다.

표 5는 레드잇 (Reddit)의 달 구매 이야기를 다른 내용에서 인간의 글과 GPT-4가 생성한 콘텐츠의 일부분을 보여준다. GPT-4는 창의적인 글을 생성하는데 어려움을 보여준다. GPT-4의 이야기는 레드잇의 CEO부터 출현하여 회사에서 어떻게 달을 구매하는 지에 대해 흥미롭게 묘사하지만, 이 이야기는 레드잇에 대한 기존 사실에 크게 의존한다. 레드잇을 설명할 때 이미 레드잇이라는 회사가 소셜미디어 플랫폼이라는 사실을 직접적으로 표현하였다. 이는 GPT-4는 새로운 콘텐츠를 생성하는 능력이 제한되어 있다고 할 수 있다. 반면, 인간 저자는 더 창의적이고 새로운 이야기를 쓸 수 있다. 또한, GPT-4가 생성한 글은 유창성 면에서 논리적 일관성의 부족으로 이야기의 흐름이 방해된다. 유연성 면에서는 AI가 학습 데이터의 한계를 넘어서지 못해 창의적으로 적응하는 데 어려움을 겪는다. 독창성은 기존 사실 정보에 대한 의존으로 제한되어

표 5. 사실 정보에 대한 의존

Table 5. Dependence on factual information

[WP] Reddit buys the moon.		
	GPT-4	Human
Contents	In an unprecedented move, <i>Reddit</i> , the popular social media platform, has purchased the moon. Yes, you read that right. The celestial body that has inspired poets and guided sailors for centuries is now under the ownership of a digital platform. The news broke out when Reddit's CEO casually mentioned it during a live AMA (Ask Me Anything) session. "Oh, by the way, we bought the moon," he said nonchalantly, sending the entire internet into a frenzy.	One day, <i>Emily</i> made a company called <i>Reddit</i> did something really big: they bought the moon! It sounds like a story from a book, but let's pretend it happened. After <i>Reddit</i> owned the moon, Everyone got to share their ideas for ideas of name and process. Some people wanted to put the <i>Reddit</i> logo on the moon. Others thought it would be cool to send messages to outer space. So, <i>Reddit</i> started to work on these ideas. They sent robots to the moon to draw their big logo.

표 6. 역사적 배경에 대한 의존

Table 6. Dependence on historical background

[WP] Narrate a game of Civilization as though it were excerpted from a history textbook.		
	GPT-4	Human
Contents	<i>Chapter 12: The Rise and Fall of the Great Empires</i> In the year 4000 BC, a civilization emerged from the fertile banks of the Nile. This was the birth of the <i>Egyptian Empire, led by the wise Pharaoh Ramesses II</i> . The Egyptians, known for their architectural prowess, constructed grand pyramids and temples that stood as testaments to their advanced knowledge in engineering and mathematics. Meanwhile, on the other side of the world, <i>the Chinese Dynasty under Emperor Qin Shi Huang</i> began its journey towards greatness. They built the Great Wall, a monumental feat of human endeavor, protecting their lands from barbarian invasions.	Once upon a time, a game called " <i>Civilization</i> ," people from all over the world came to build their own countries from the very beginning. It was like playing with the biggest set of building blocks ever! In this game, everyone started with just a tiny village. Imagine a few huts and farms, with people trying to figure out how to live together. They had to learn how to farm, build things, and protect themselves from others who might want to take what they had. As time went by, these tiny villages grew into big cities. The people learned how to write, make laws, and trade with their neighbors.

고유한 콘텐츠 생성을 방해한다. 구체성 부분에서 글을 작성하는 섬세한 능력은 부족하여 이야기에 새로움과 확장을 제공하지 못한다. 이야기의 유용성은 사실 정보를 기반으로 작성하여 실용성을 강조하는 것처럼 보이지만, 창의적 글쓰기와 관련성이 부족하다. 특정 창의적 전략에서는 상상력을 발휘하여 독창적이고 창의적인 이야기를 만들어내기보다는, 이미 존재하는 패턴과 구조를 따르는 경향이 있다.

이는 AI가 기대되는 창의적 감각과 독창성을 발휘하지 못하게 하며, 생성된 콘텐츠는 인간의 창의적 작업에 비해 단조롭고 예측 가능하게 된다. 결과적으로 GPT-4는 기존의 정보를 재구성하거나 약간 변형하는 데는 능숙하지만, 완전히 새로운 상황이나 개념을 창출하는 데는 부족하다는 것을 볼 수 있다.

표 6은 문명(Civilization) 게임의 한 장면을 역사 교과서에서 발췌한 것처럼 서술하라는 주제에 관해 인간의 글과 GPT-4가 생성한 콘텐츠를 구분하여 일부분 발췌하였다. GPT-4는 방대한 지식 기반을 활용하여 문명 게임을 설명하는 글을 생성했지만, 사전 훈련 데이터에 대한 의존성을 보여준다. 예를 들어, GPT-4는 이집트의 람세스와 중국 황제 진시황 같은 역사적 인물들을 등장시켜 글을 작성하였다. 이는 AI가 기존의 역사적 배경과 사실에 기초한 데이터를 활용하여 글을 생성하는 경향을 보여준다.

이러한 접근 방식은 역사적 정확성을 보장할 수는 있지만, 창의적 글쓰기에 있어서는 한계를 보인다. 반면, 인간 저자는 역사적 배경에 구애받지 않고 더 창의적인 요소를 추가하여

이야기를 풍부하게 만든다. 예를 들어, 인간 저자는 실제 역사적 인물과 배경과는 무관한 이야기를 창조하여 독창성과 상상력을 나타낸다. 표 6은 여섯 가지 창의성 기준에서 GPT-4의 약점을 잘 보여준다. 유창성 면에서 GPT-4는 부드러운 흐름을 유지하지만, 창의적 유동성이 부족하여 단조로운 느낌을 준다. AI는 기존의 패턴을 따라가는 경향이 강해, 예상 가능한 전개로 이어지는 경우가 많다. 유연성 면에서는 학습 데이터의 한계를 넘어서지 못하고 새로운 상황이나 전개에 적응하는 데 어려움을 겪는다. 구체성 부분에서는 역사적 세부 사항이 풍부하게 나타나지만, 이는 창의적 글쓰기의 역량을 충분히 발휘하지 못하게 만든다. GPT-4 글에서 유용성은 역사적 사실에 의존하여 작성되므로, 창의적 글쓰기의 관점에서 볼 때 실용적이지 않다. 특정 창의적 전략에서는 GPT-4가 창의성을 발휘하지 못해 콘텐츠에 참신함과 새로운 구성이 부족하다는 것을 보여준다. 이는 GPT-4가 기존 데이터에 있는 정보를 바탕으로 글을 생성하기 때문에 인간의 상상력을 따라잡기 위해 개선이 필요함을 강조한다.

V. 결론

대규모 언어 모델 (LLMs)인 GPT-4는 방대한 데이터를 활용하여 창의적인 콘텐츠를 생성하는 데 있어 뛰어난 잠재력을 보여준다. 이 능력은 다양한 관점과 혁신적인 아이디어를 제공하며, 창의적 글쓰기 작업에서 인간의 성과에 근접한 결과를 도출한다. 기존 데이터에 대한 의존에도 불구하고,

LLMs는 고품질의 창의적 콘텐츠를 생성하는 데 효과적인 도구임을 입증한다.

본 연구의 결과는 GPT-4의 능력과 내재한 한계를 동시에 강조한다. AI는 인간과 유사한 창의성을 비슷하게 도출할 수 있지만, 그 결과물은 훈련 데이터의 범위와 특징에 의해 제한된다. 이는 AI가 생성한 창의성과 인간의 창의성 사이의 중요한 차이점을 강조하며, 인간의 창의성은 더 참신하고 독창적인 사고 과정을 포함한다. 그럼에도 불구하고, GPT-4의 창의적 콘텐츠 생성 능력은 주목할 만하며, AI가 창의적 작업에서 유용하게 사용될 수 있음을 시사한다. 이에 후속 연구는 AI 시스템을 훈련하는 데이터셋을 다양화하고, 다양한 분야의 창의적 텍스트를 포함하여 AI의 창의적 능력을 더욱 향상하는 데 중점을 두어야 한다. 또한, AI를 창의적 프로세스에서 협력 도구로 탐구하여 인간의 창의성을 보완하고 강화하는 혁신적인 도구를 개발할 수 있다. AI 기술이 발전함에 따라 AI의 창의적 능력을 평가하는 방법을 정교화하고 알고리즘을 개선하는 것이 인간의 창의성을 모방하거나 증진할 수 있는 시스템을 개발하는 데 필수적이다.

AI의 창의적 능력을 논할 때 윤리적 고려 사항도 중요하다. AI 생성 콘텐츠를 명확히 식별하고, 인간 창작자의 권리를 존중하는 것이 중요하다. 또한, AI 시스템을 투명성, 공정성, 책임성을 촉진하는 방식으로 개발하고 배포해야 한다. AI가 창의적 맥락에서 미치는 윤리적 영향을 조사하고, 진정성과 독창성에 미치는 영향을 고려하여 윤리적 사용 지침을 수립하는 것도 중요하다. 이러한 윤리적 차원을 계속 탐구하여 AI 기술이 책임감 있게 사용되어야 한다.

또한, 현재까지 연구된 창의성 평가 도구는 종종 창의성의 개념에 대한 이해 부족과 창의성의 모든 측면을 충분히 측정하지 못하며, 전통적인 평가 방법에 의존하는 한계성을 보인다[32]. 예를 들어 평가 항목 중 본 연구에서 사용된 유창성과 유연성의 개념은 다소 모호하게 설명된 면이 있다. 보다 포괄적이고 신뢰할 수 있는 평가 방법을 개발하여 창의적 능력의 전체 스펙트럼을 포착하는 것이 필요하다. 따라서, 대규모 언어 모델(LLMs)은 창의적 글쓰기에 상당한 잠재력을 가지고 있으며, 인간의 창의성을 모방할 수 있는 성능을 확인하였다. 그러나 그 능력은 지속적인 연구와 개선을 통해 달성될 수 있다. 훈련 데이터셋의 다양성을 높이고, 평가 기준을 개선하며, 윤리적 문제를 해결하는 것이 필요하다. 대규모 언어 모델(LLMs)은 일정 수준의 창의성을 지니고 있지만, 그 주된 가치는 인간의 창의성을 보완하고 협력하는 능력에 있다. 이들을 다양한 창의적 분야에서 윤리적이고 책임감 있게 사용함으로써 그 영향력과 유용성은 더욱 명확해질 것이다.

본 연구의 주요 한계 중 하나는 창의적 글쓰기 평가 과정에서 관찰된 상호 평가자 일치도(IAA)의 변동성에 있다. IAA 값은 평가자 간 합의 정도를 나타내며, 높은 결과 수치는 더 큰 일관성을 의미한다. 그러나 창의적 글쓰기와 같은 주관적 작업에서는 개인적인 편견과 관점이 평가에 크게 영향을 미치므로 높은 상호 평가자 일치도(IAA)를 달성하는 것은 어렵

다. 창의적 작품을 평가하는 데 내재된 주관성으로 인해 평가자마다 창의성의 다양한 측면을 우선시하는 경향이 있어 일관성 없는 평가가 이루어진다. 상호 평가자 일치도(IAA)는 평가자들의 일관성을 측정할 수 있는 척도를 제공하지만, 동시에 평가자들이 창의성을 어떻게 해석하고 중요시하는지에 대한 다양한 관점을 완전히 반영하지는 못한다. 따라서 본 연구에서의 평가자 일치도 결과는 인간 평가에서 비롯된 주관적 판단과 편견, 그리고 가치관의 차이에서 발생한다.

또한, 창의적 글쓰기 평가와 같은 복잡하고 해석적인 작업에서는 높은 상호 평가자 일치도(IAA)를 달성하는 것이 어렵다. 따라서 본 연구에서 관찰된 IAA의 변동성은 인간 평가자들이 해석적 판단을 많이 요구하는 창의적 글쓰기 평가의 미묘하고 다면적인 특성에서 비롯될 수 있다. 본 연구의 낮은 상호 평가자 일치도(IAA) 비율은 다양한 평가자 간의 일관성을 유지하는 데 내재한 어려움을 강조한다. 8명의 평가자가 100개의 서로 다른 주제에 대해 동일한 점수를 부여하는 것은 당연히 쉽지 않다. 이 한계는 창의적 작품에 대한 인간 평가의 주관성이 포함되어 있음을 이해하는 것의 필요성을 강조한다. 향후 연구에서는 이러한 주관적 편향을 완화하기 위해 창의적 글쓰기 평가의 신뢰성을 향상하는 방법을 계속 탐구하고, 인간 평가와 모델 평가를 모두 포함하는 방법을 향후 연구에서 제안할 예정이다.

참고문헌

- [1] J. Vinothkumar and A. Karunamurthy, "Recent Advancements in Artificial Intelligence Technology: Trends and Implications," *Quing: International Journal of Multidisciplinary Scientific Research and Development*, Vol. 2, No. 1, pp. 1-11, January-March 2023.
- [2] J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, K. O'Connor, R. Li, P.-C. Peng, ... and J. H. Moore, "ChatGPT and Large Language Models in Academia: Opportunities and Challenges," *BioData Mining*, Vol. 16, 20, July 2023. <https://doi.org/10.1186/s13040-023-00339-9>
- [3] M. M. Mijwil, K. K. Hiran, R. Doshi, M. Dadhich, A.-H. Al-Mistarehi, and I. Bala, "ChatGPT and the Future of Academic Integrity in the Artificial Intelligence Era: A New Frontier," *Al-Salam Journal for Engineering and Technology*, Vol. 2, No. 2, pp. 116-127, April 2023. <https://doi.org/10.55145/ajest.2023.02.02.015>
- [4] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774v6, March 2024. <https://doi.org/10.48550/arXiv.2303.08774>
- [5] D. Kevian, U. Syed, X. Guo, A. Havens, G. Dullerud, P. Seiler, ... and B. Hu, "Capabilities of Large Language Models in Control Engineering: A Benchmark Study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra,"

- arXiv:2404.03647v1, April 2024. <https://doi.org/10.48550/arxiv.2404.03647>
- [6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. ... and G. Lample, “LLaMA: Open and Efficient Foundation Language Models,” arXiv:2302.13971v1, February 2023. <https://doi.org/10.48550/arXiv.2302.13971>
- [7] D. Huang, Z. Hu, and Z. Wang, “Performance Analysis of Llama 2 among Other LLMs,” in *Proceedings of 2024 IEEE Conference on Artificial Intelligence (CAI)*, Singapore, pp. 1081-1085, June 2024. <https://doi.org/10.1109/CAI59869.2024.00108>
- [8] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, ... and I. Stoica, “Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference,” arXiv:2403.04132v1, March 2024. <https://doi.org/10.48550/arXiv.2403.04132>
- [9] C.-H. Chiang and H.-Y. Lee, “Can Large Language Models Be an Alternative to Human Evaluations?,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, pp. 15607-15631, July 2023. <https://doi.org/10.18653/v1/2023.acl-long.870>
- [10] K. Shively, K. Stith, and L. D. V. Rubenstein, “Measuring What Matters: Assessing Creativity, Critical Thinking, and the Design Process,” *Gifted Child Today*, Vol. 41, No. 3, pp. 149-158, July 2018. <https://doi.org/10.1177/1076217518768361>
- [11] OpenAI. Improving Language Understanding by Generative Pre-Training [Internet]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis: MN, pp. 4171-4186, June 2019. <https://doi.org/10.18653/v1/N19-1423>
- [13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, ... and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” arXiv:1907.11692v1, July 2019. <https://doi.org/10.48550/arXiv.1907.11692>
- [14] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations,” arXiv:1909.11942v6, February 2020. [50/arXiv.1909.11942](https://doi.org/10.48550/arXiv.1909.11942)
- [15] OpenAI. Language Models are Unsupervised Multitask Learners [Internet]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- [16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, ... and L. Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 7871-7880, July 2020. <https://doi.org/10.18653/v1/2020.acl-main.703>
- [17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, ... and P. J. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, Vol. 21, No. 1, 140, January 2020.
- [18] S. Kublik and S. Saboo, *GPT-3: The Ultimate Guide To Building NLP Products With OpenAI API*, Birmingham, UK: Packt Publishing, 2022.
- [19] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, ... and X. Huang, “A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models,” arXiv:2303.10420v2, December 2023. <https://doi.org/10.48550/arxiv.2303.10420>
- [20] E. Y. Chang, “Examining GPT-4’s Capabilities and Enhancement with SocraSynth,” in *Proceedings of 2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas: NV, pp. 7-14, December 2023. <https://doi.org/10.1109/CSCI62032.2023.00009>
- [21] M. A. Boden, “Creativity and Artificial Intelligence,” *Artificial Intelligence*, Vol. 103, No. 1-2, pp. 347-356, August 1998. [https://doi.org/10.1016/S0004-3702\(98\)00055-1](https://doi.org/10.1016/S0004-3702(98)00055-1)
- [22] A. Dietrich, “The Cognitive Neuroscience of Creativity,” *Psychonomic Bulletin & Review*, Vol. 11, No. 6, pp. 1011-1026, December 2004. <https://doi.org/10.3758/BF03196731>
- [23] H. Mozaffari, “An Analytical Rubric for Assessing Creativity in Creative Writing,” *Theory and Practice in Language Studies*, Vol. 3, No. 12, pp. 2214-2219, December 2023. <https://doi.org/10.4304/tpls.3.12.2214-2219>
- [24] R. D’Souza, “What Characterises Creativity in Narrative Writing, and How Do We Assess It? Research Findings from a Systematic Literature Search,” *Thinking Skills and Creativity*, Vol. 42, 100949, December 2021.

<https://doi.org/10.1016/j.tsc.2021.100949>

- [25] J. C. Kaufman and R. A. Beghetto, "Beyond Big and Little: The Four C Model of Creativity," *Review of General Psychology*, Vol. 13, No. 1, pp. 1-12, March 2009. <https://doi.org/10.1037/a0013688>
- [26] M. A. Runco and G. J. Jaeger, "The Standard Definition of Creativity," *Creativity Research Journal*, Vol. 24, No. 1, pp. 92-96, 2012. <https://doi.org/10.1080/10400419.2012.650092>
- [27] Kaggle. Writing Prompts [Internet]. Available: <http://www.kaggle.com/datasets/ratthachat/writing-prompts/>.
- [28] Reddit. r/Writing Prompts [Internet]. Available: https://www.reddit.com/r/WritingPrompts/wiki/how_to_tag_prompts/.
- [29] J. A. Rice, *Mathematical Statistics and Data Analysis*, 3rd ed. Belmont, CA: Duxbury Press, p. 201, 2006.
- [30] D. K. Lee, J. In, and S. Lee, "Standard Deviation and Standard Error of the Mean," *Korean Journal of Anesthesiology*, Vol. 68, No. 3, pp. 220-223, June 2015. <https://doi.org/10.4097/kjac.2015.68.3.220>
- [31] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 3rd ed. Thousand Oaks, CA: SAGE Publications, pp. 222-235, 2013.
- [32] J. A. Plucker, M. C. Makel, and M. Qian, Assessment of Creativity, in *The Cambridge Handbook of Creativity*, 2nd ed. Cambridge, UK: Cambridge University Press, ch. 3, pp. 44-68, 2019. <https://doi.org/10.1017/9781316979839.005>



김성은(Sungeun Kim)

2021년 : 경북대학교 대학원 (영어교육 석사)

2021년~현 재: 경북대학교 영어영문학과 영어교육 박사수료
※ 관심분야 : 인공지능 활용 교육, 영어교육, 프롬프트 엔지니어링, 자연어처리 등



김자경(Jakyong Kim)

2022년 : 경북대학교 대학원 (영어교육 석사)

2022년~현 재: 경북대학교 영어영문학과 영어교육 박사수료
※ 관심분야 : 인공지능 활용 교육, 영어교육, 프롬프트 엔지니어링, 자연어처리 등



오동석(Dongsook Oh)

2016년 : 서강대학교 대학원 컴퓨터공학과 (석사)

2020년 : 고려대학교 대학원 컴퓨터학과 (박사)

2016년~2018년: 다이퀘스트 자연어처리 연구원
2018년~2019년: NHN Ent. 자연어처리 연구원
2019년~2020년: Human-inspired AI Research 인공지능 연구원
2023년~2023년: KT Large AI Alignment Project팀 선임연구원
2023년~현 재: 경북대학교 영어영문학과 조교수
※ 관심분야 : 신뢰가능한 인공지능, 뉴로-심블릭 인공지능, 자연어처리 등