JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 26, No. 3, pp. 763-775, Mar. 2025

Check for updates

언어모델을 이용한 반전이 있는 스토리 생성 및 평가

박 정 윤[1] · 배 병 철[2*]
[1]홍익대학교 게임학부 석사과정
[2]홍익대학교 게임학부 조교수

Generation and Evaluation of Twisted-Plot Stories Using Large Language Models

Jeongyoon Park[1] · Byung-Chull Bae[2*]

[1]Master's Course, School of Games, Hongik University, Seoul 04066, Korea
[2]Assistant Professor, School of Games, Hongik University, Seoul 04066, Korea

[요 약]

급전(Peripeteia)은 아리스토텔레스가 시학에서 제시한 개념으로, 반전의 원형으로 간주되며 서사의 긴장감을 고조시키고 독자의 몰입을 유도하는 중요한 스토리텔링 요소이다. 다양한 매체에서 반전은 서사의 깊이와 의미를 확장하는 데 기여하지만, 언어 모델을 활용한 반전 생성 연구는 여전히 제한적이다. 본 연구는 대규모 언어 모델(LLMs)을 활용하여 반전이 포함된 스토리를 생성하며 이를 위한 효과적인 프롬프트 엔지니어링 기법을 탐구한다. 반전의 퀄리티, 일관성, 문법적 정확도, 핍진성, 흥미도를 기준으로 GPT-4o-mini, Llama-3-8B, Claude-3-haiku 모델을 활용해 평가했다. 평가 결과 서사 요소의 특징 및 목적에 맞는 프롬프트 설계가 반전 생성의 품질에 영향을 미치는 것으로 나타났다. 본 연구는 언어 모델을 이용해 반전 서사 생성 및 평가를 진행하며 언어 모델의 스토리 생성 능력을 향상시키기 위한 방향성을 제시한다.

[Abstract]

Peripeteia, introduced by Aristotle in Poetics, is considered the archetype of plot twists, playing a critical role in heightening narrative tension and engaging readers. While plot twists contribute to the depth and meaning of narratives across various media, research on generating plot twists using language models remains limited. This study explores effective prompt engineering techniques for generating plot-twist narratives using Large Language Models (LLMs). Narratives were evaluated based on twist quality, coherence, grammatical accuracy, verisimilitude, and enjoyment, utilizing GPT-4o-mini, Llama-3-8B, and Claude-3-Haiku. The results indicate that prompt designs tailored to the characteristics and objectives of narrative elements significantly influence the quality of generated twists. By generating and evaluating plot-twist narratives using language models, this study provides insights into enhancing the narrative generation capabilities of LLMs.

색인어 : 스토리 생성, 스토리 평가, 프롬프트 엔지니어링, AI 생성 콘텐츠, 자연어 처리
Keyword : AI-Generated Content, Natural Language Processing, Prompt Engineering, Story Evaluation, Story Generation

http://dx.doi.org/10.9728/dcs.2025.26.3.763

Received 16 January 2025; **Revised** 06 March 2025
Accepted 18 March 2025

***Corresponding Author; Byung-Chull Bae**

Tel: +82-44-860-2074
E-mail: byuc@hongik.ac.kr

Ⅰ. 서 론

급전(peripeteia)은 아리스토텔레스가 시학에서 서사적 전환의 중요성을 논하며 제시한 개념[1]으로, 오늘날 반전의 원형으로 이해할 수 있다. 급전과 반전은 모두 독자의 기대를 전복하고 새로운 시각을 열어줌으로써 서사의 긴장감을 고조시키는 공통점을 갖으며, 반전에서의 예측하지 못한 놀라움의 제시를 통해 이야기의 즐거움(enjoyment)과 재미를 증대시키는 데 큰 역할을 한다[2]. 반전은 TV 시리즈[3]와 소설[4]을 넘어 광고[5] 등 서사가 사용되는 모든 곳에서 스토리텔링의 중요한 구성요소로 간주되지만 언어모델(Large Language Model; LLM)을 이용해 반전을 생성하는 시도는 아직 제한적이다.

기존의 스토리 생성 연구는 주로 서사의 구조적 일관성[6]-[9] 문제를 해결하는데에 초점이 맞추어져 있었다. 또한 더 높은 품질의 스토리 생성을 위해 캐릭터의 행동[7], 캐릭터의 감정[10], 이벤트[11]-[13], 특정 장르[14],[15], 스토리의 길이[8] 등 다양한 요소에 초점을 맞추어 강화학습[6],[10], 캐릭터 임베딩[7], 동적 플롯 상태 추적[8], 반지도학습[15] 등의 방법과 그래프 기반 구조[12],[13] 등 기술적 방법을 접목한 연구가 진행되어 왔다.

이처럼 스토리 생성 연구는 스토리 요소와 자연어처리(NLP; Natural Language Processing)기술을 접목해 활발하게 이루어지고 있다. 하지만 독자의 기대를 전복하고 긴장감을 고조시키는 반전 서사에 초점을 맞춘 연구는 상대적으로 부족한 실정이다. 이러한 연구 공백을 해결하기 위해, 본 연구는 언어 모델 및 다양한 프롬프트 엔지니어링을 활용한 반전이 있는 스토리를 생성한다.

텍스트를 생성하는 대부분의 연구는 주로 자동 평가와 사람 평가를 통해 평가한다. 자동 평가는 정량적인 지표를 기반으로 텍스트의 품질을 분석하며, 평가 기준의 일관성과 객관성을 제공한다. 대표적인 평가 지표로는 BLEU[16], ROUGE score[17] 등이 있으며 self-BLEU를 이용해 텍스트의 다양성을 정량적으로 평가할 수 있다. 사람 평가는 자동 평가로 평가하기 어려운 흥미, 창의성, 몰입도와 같은 주관적 요소를 평가할 수 있다. 많은 연구에서 자동 평가와 사람 평가를 함께 진행해 평가의 다양성을 추구한다.

하지만 사람 평가의 경우 평가 참여자에게 적정한 보상을 지급해야 하며, 평가 참여자를 관리하는 과정에서 상당한 시간과 비용이 소모된다. 또한 예산과 시간 문제로 많은 평가 참여자를 고용하기 어려운 문제가 있다. 이에 따라 스토리의 비정량적 요소를 평가하기 위한 자동 평가의 필요성이 대두되고 있다. 언어 모델을 이용한 자동 평가는 질문 응답(QA), 추론, 텍스트 요약[18], 대화 평가[19], 에세이[20] 등의 다양한 도메인에서 활발한 연구가 진행되고 있으며 동시에 스토리와 같이 복잡한 서사 구조를 가진 텍스트를 평가하는 연구 또한 시도되고 있다. 선행 연구로서 도입부, 중반부, 결말 등으로 스토리의 구조를 평가하거나 캐릭터 및 장면의 묘사를 평가[21]한 경우는 있으나 스토리가 얼마나 그럴 듯 하게 느껴지는 지를 평가하기 위한 핍진성과 반전의 퀄리티를 평가 기준으로 삼는 시도는 없었다. 본 연구에서는 반전의 퀄리티, 일관성, 문법적 정확도, 핍진성, 흥미도를 기준으로 사람 평가와 언어 모델 기반 자동 평가를 진행했다.

또한 선행 연구[22]에서는 다양한 언어 모델을 사용해 언어모델별 스토리 생성 특성과 한계를 분석하였다. 그 결과, GPT 계열 모델인 ChatGPT는 타 모델에 비해 흥미도가 낮은 스토리를 생성하는 경향이 있음을 확인했으나 이를 극복하기 위한 프롬프트 엔지니어링 기법에 대한 연구는 수행되지 않았다.

이에 본 논문에서는 선행 연구의 연구 공백을 보완하고자 다양한 프롬프트를 설계하여 흥미로운 스토리를 위해 대중적으로 사용되는 서술 기법 중 하나인 반전을 활용한 스토리를 생성한다. 언어 모델을 이용한 자동 평가 및 사람 평가를 통해 설계한 프롬프트의 스토리 생성 성능을 분석한다. 최종적으로 실험 및 평가 결과를 바탕으로 반전이 있는 스토리 생성을 위한 효과적인 프롬프트 엔지니어링 기법을 제안하는 것을 목표로 한다.

본 연구는 네 가지 기여점을 갖는다. 첫 번째, 다양한 프롬프트 설계 기법을 비교하여 반전 서사 생성의 최적 전략을 탐색하였다. 두 번째, 자동 평가와 사람 평가를 병행하여 평가 방식 간의 차이를 분석하고, 평가의 신뢰성을 높이는 방향을 제시하였다. 세 번째, 언어 모델 별 평가 경향성을 비교하여, 특정 모델이 평가에서 가지는 편향성을 확인하고 이를 보완하기 위한 방향성을 제안하였다.네 번째, 반전의 질과 핍진성 평가를 도입하여 기존 연구보다 서사적 완성도를 보다 정밀하게 평가하였다.

Ⅱ. 관련 연구

2-1 스토리 생성

언어 모델은 트랜스포머 기반 아키텍처[23]와 대규모 사전 학습을 통해 빠르게 발전하고 있으며, 자연어 처리의 다양한 응용 분야에서 혁신적인 성과를 이뤄내고 있다. 2020년 6월 공개된 GPT-3, 2022년 11월 공개된 GPT-3.5에 이어 2023년 3월에 공개된 GPT-4와 2024년 5월 공개된 GPT-4o까지 언어 모델의 빠른 발전은 자연어 처리 연구와 응용에서 새로운 지평을 열고 있다.

스토리 생성은 언어 모델에 창의성과 논리적 사고를 동시에 요구하는 복합적인 태스크이다. 논리적 일관성과 창의성 요구를 동시에 충족하는 언어 모델의 스토리 생성 능력을 더욱 극대화하기 위한 연구가 이루어지고 있다[24].

스토리 생성 연구에서는 텍스트 생성 과정에서 흔히 나타

나는 논리적 오류, 중복 생성 그리고 텍스트가 길어질수록 일관성이 유지되지 않는 문제가 발생하였고, 이를 주요 개선 과제로 삼는 연구들이 제시되고 있다[12],[25]. 또한, 언어 모델이 생성한 이야기와 사람이 작성한 이야기를 비교했을 때, 서스펜스와 긴장감이 부족하며 서사 구조가 단조로운 문제점이 제시된다[26]. 스토리 생성 연구는 단순히 독자가 읽기에 문법적, 논리적 오류가 없는 자연스러운 문장을 생성하는 데에 그치지 않고, 장르[14],[15], 캐릭터[7],[10], 이벤트[11]-[13]와 같은 서사적 요소를 자연어 처리 기술과 접목해 논리적 일관성과 서사 구조를 강화하는 시도들이 계속되어 왔다.

선행 연구에서는 긴 텍스트 생성 시 계층적 생성 모델을 통해 플롯 전개를 제어하는 모델을 제안하여 스토리의 창의성과 일관성을 높이는 시도가 있으며[27], 일관성을 위해 캐릭터의 성격과 행동을 기반으로 서사의 일관성을 유지하는 방법론을 탐구하였고[7], 이벤트로 주어진 구조적 정보의 인과관계를 반영해 이벤트 중심의 이야기를 생성하여 서사적 논리성을 강화하려는 시도도 있었다[28].

이러한 연구들은 언어 모델이 단순히 스토리를 위한 텍스트 생성을 넘어, 보다 구체적이고 정교한 스토리 요소를 통합해 보다 일관성 있는 스토리를 생성하는 데에 기여하고 있다. 하지만 선행 연구에서는 반전이 있는 스토리 생성과 같이 독자의 기대를 전복하고 새로운 시각을 제시하는 스토리텔링 기법인 반전을 접목한 스토리 생성 연구는 아직 충분히 이루어지지 않았다.

또한 선행 연구에서는 다양한 언어 모델을 이용해 한국어로 SF와 로맨스 장르 소설 생성 및 평가를 통해 모델 별 특징과 한계점을 분석하는 선행 연구를 진행한 바 있다[22]. 그러나 장르 당 5개라는 적은 수의 스토리 생성 및 특정 장르(SF와 로맨스)에 국한되어 실험을 진행했다는 한계가 있다.

이를 보완하기 위해 다양한 프롬프트를 설계하여 프롬프트당 250개의 스토리를 생성해 성능의 신뢰도를 높이고자하며, 장르에 국한되지 않고 대중적으로 사용되는 서사 장치인 반전이 있는 스토리를 생성한다.

본 연구는 스토리 생성 도메인에서 언어 모델의 활용 가능성을 확장하기 위해 반전이 있는 스토리를 생성하고, 이를 가능하게 하는 효과적인 프롬프트 설계 방법을 탐구한다. 나아가, 사람 평가와 함께 언어 모델 기반 스토리 자동 평가를 통해 생성한 반전이 있는 스토리의 품질을 평가하고 분석한다.

2-2 스토리 평가

텍스트를 평가하는 방법에 대한 연구는 자연어 처리 분야에서 텍스트 생성 연구와 함께 발전해왔다. 텍스트 생성 관련 평가는 정량적 평가(Quantitative Evaluation)와 정성적 평가(Qualitative Evaluation)로 구분된다. 정량적 평가는 BLEU, ROUGE와 같은 지표를 사용하여 생성된 텍스트와 참조 텍스트 간의 유사성을 측정하며, 일부 평가에서는 텍스트의 다양성이나 문법적 정확성을 고려하기도 한다. 정성적 평가는 사람 평가자가 생성 텍스트의 품질을 창의성, 일관성, 가독성 등의 기준으로 정량적 평가지표로 평가하기 어려운 텍스트의 완성도를 평가한다. 언어모델을 사용해 생성한 텍스트의 종류에 따라 사용되는 평가 방법이 달라지며, 본 절에서는 스토리 텍스트의 평가 방법에 대해 주로 논의하고자 한다.

먼저, 언어모델을 이용해 생성된 스토리를 평가할 때 사용되는 정량적 평가에 대해서 이야기한다. 정량적 평가는 대표적으로 BLEU[16]와 ROUGE[17]와 같은 n-gram 기반 지표를 활용하여 생성된 텍스트와 참조 텍스트 간의 유사성을 측정하는 데 중점을 둔다. BLEU는 기계 번역의 정확성을 평가하기 위해 개발되었으며, 생성된 텍스트의 n-gram 정밀도를 계산한다. 많은 스토리 생성 연구에서 평가 방법으로 사용한다[8]-[11],[13],[26]. ROUGE는 주로 요약 작업에서 사용되며, 참조 텍스트와의 n-gram 중복을 기반으로 재현율을 측정하며 많은 스토리 생성 연구에서 평가 기준으로 사용되고 있다[8],[9],[13].

또한, BLEU를 변형한 Self-BLEU를 통해 생성된 텍스트 간의 유사성 및 다양성을 평가한다. Self-BLEU는 생성된 텍스트에서 샘플 텍스트와 다른 샘플 텍스트의 유사성을 BLEU의 방법을 이용해 측정하는 방식이다. 이는 스토리 생성에서 반복적인 패턴이나 내용 중복을 평가하는 지표로 사용될 수 있으며 다양한 연구에서 평가 방법으로 사용되었다[8]-[10]. 그러나 이러한 정량적 평가는 텍스트의 표면적 유사성에 초점을 맞추므로 창의성, 흥미, 서사적 일관성과 같은 스토리의 질적 특성을 평가하기 어려운 문제가 있다.

정성적 평가는 주로 사람 평가(Human Evaluation)를 통해 스토리의 질적 요소를 측정한다. 많은 연구에서 사용하고 있는 평가기준으로는 일관성[9],[24], 유창성[8], 반복성[8], 창의성[8],[9],[24], 문법적 정확도[26], 몰입도[8], 주인공의 감정이 원하는 감정 곡선을 따르는 감정 충실성[10], 스토리의 품질[10], 관련성[12], 논리성[12],[25], 흥미[24], 연관성[24] 등이 있다. 사람 평가를 통해 스토리의 세부 요소 및 완성도를 평가할 수 있으나 사람 평가에는 시간과 비용이 많이 소요되는 단점이 있다.

이러한 한계를 보완하기 위해, 최근에는 언어 모델을 활용한 평가 방법이 제안되고 있다. 언어모델을 이용해 언어모델이 생성한 텍스트가 언어모델이 제시한 출처의 콘텐츠 간의 일치도를 측정해 언어모델의 신뢰성과 정확성을 측정하는 시도가 있었다[18]. StoryER에서는 사람의 선호도(preference)를 반영해 사람 평가와 상관성 있는 스토리 평가 방법을 제안했다. 이 연구에서는 스토리 평가 시 두 스토리 중 더 선호하는 것을 선택하고, 스토리의 플롯, 캐릭터 등을 점수로 매기며 평가 점수에 대해 설명할 것을 요구해 사람의 선호도를 반영한 스토리를 위한 언어 모델 자동 평가를 진행했다[21].

이처럼 언어 모델을 이용한 자동 평가에 대한 연구에서는 언어 모델을 이용한 평가와 사람 평가의 상관관계를 분석하

는 연구가 진행되며 사람 평가 결과와 언어 모델을 이용한 평가 결과의 상관 관계를 높이는 시도가 있었다. 또한 프롬프트 설계 시 단순한 프롬프트만으로도 충분히 좋은 평가 결과를 낼 수 있으며 복잡한 프롬프트의 경우 오히려 성능을 저하시킬 수 있음을 제시했다[29].

선행 연구에서는 언어 모델로 생성한 스토리에 대해 사람 평가만 진행하였으나[22], 본 연구에서는 반전이 있는 스토리 평가를 위해 사람 평가와 함께 여러 언어 모델을 이용한 스토리 자동 평가를 진행한다. 또한 사람 평가 결과와 언어 모델을 이용한 평가 결과를 비교하여 두 평가 방법의 상관성을 분석하고, 평가 결과 분석을 통해 평가 시 사용된 언어 모델의 스토리 평가 특징을 제시한다.

2-3 프롬프트 엔지니어링

언어 모델이 발전함에 따라 프롬프트 엔지니어링(PE; Prompt Engineering)은 모델의 성능을 효과적으로 활용하기 위한 방법으로 주목받고 있다. 대표적인 프롬프트 엔지니어링 기법으로는 제로샷(Zero-shot) 프롬프팅과 퓨샷(Few-shot)프롬프팅이 있다. 제로샷 프롬프팅은 모델에게 별도의 예제나 추가 정보 없이 작업 지시만을 제공하여 출력 결과를 생성하도록 하는 방법이며, 퓨샷 프롬프팅은 몇 가지 예제를 포함한 프롬프트를 제시함으로써 모델이 작업의 맥락을 학습하고 더 정밀한 출력을 생성하도록 유도하는 방법이다.

퓨샷 프롬프팅에도 불구하고 복잡한 연산 혹은 추론 작업 시 모델이 잘못된 대답을 생성하는 문제가 있다. 이에 따라 모델의 정확도를 높이기 위해 제안된 방법으로 프롬프팅 기법으로 CoT (Chain-of-Thought)프롬프팅[30]을 이야기할 수 있다. CoT 프롬프팅은 복잡한 문제 해결 과정에서 모델에게 중간 사고 단계를 생성할 수 있도록 프롬프팅해 모델이 중간 사고 과정을 통해 올바른 답을 도출하도록 유도하는 방법이다. 하지만 CoT 프롬프팅은 사고 단계를 명시적으로 제시해야 하므로 이를 자동화하기 위해 Auto-CoT 프롬프팅[31]이 제안되었다. Auto-CoT 프롬프팅은 질문을 클러스터링하여 대표 질문을 추출하고, 모델이 Zero-Shot-CoT 방식을 통해 자동으로 사고 단계를 생성하도록 설계된 기법이다. 이를 통해 프롬프트 설계 과정에서 사람의 개입을 최소화하며 시간과 인력을 절감할 수 있다. Auto-CoT와 같은 자동화된 프롬프팅 기술은 기존 수작업 방식보다 더 효율적이고 정밀한 프롬프트 설계 가능성을 열어준다. 예를 들어, Few-shot Auto-CoT는 클러스터별 대표 질문에서 추론 과정을 자동으로 생성하여 더욱 정교한 사고 단계를 포함한 프롬프트를 설계한다.

이처럼 프롬프트 엔지니어링을 통해 언어모델의 성능을 극대화하기 위한 연구가 활발히 진행되고 있으며, 스토리 생성을 위한 프롬프트 엔지니어링 관련 연구 또한 활발히 제시되고 있다.

선행 연구에서는 인터랙티브 스토리텔링을 위한 내러티브 선택 생성을 프롬프팅을 이용해 연구했다. 해당 연구에서는 플롯을 구분해 텍스트를 나누어 구조화를 진행했으며, 분기점에서 다음 선택, 결과 등을 서술하도록 언어모델에 요청했다. 또한 프롬프트를 질문형으로 설계했다[32]. 프롬프팅을 이용해 비주얼 노벨 스토리를 생성하고 평가한 연구가 있었다[33]. 하지만 선행 연구에서는 반전이 있는 스토리 생성을 위한 프롬프트의 성능을 비교하는 실험은 진행되지 않았다. 이 밖에도 선행 연구에서는 언어모델과 프롬프팅을 이용해 장르 소설을 생성했으나[22], 하나의 프롬프트만 사용하여 스토리를 생성하였으며 프롬프트 엔지니어링을 통해 언어 모델의 스토리 생성 성능을 향상시키기 위한 프롬프팅은 시도하지 않았다는 한계가 있다. 이러한 연구 공백을 채우기 위해, 본 연구는 반전이 있는 스토리 생성을 위한 다양한 프롬프트를 설계하여 스토리를 생성하고 평가한다.

본 논문은 생성과 평가 시 모두 프롬프트 엔지니어링을 통해 언어 모델의 성능 활용을 위한 프롬프트를 탐색한다. 또한 평가 시 동일한 프롬프트를 다양한 언어 모델에 사용하여 언어 모델의 스토리 평가 특성에 대해서 분석한다. 이를 통해 프롬프트 설계가 언어 모델의 스토리 생성 성능을 실질적으로 향상시키는 도구로 활용될 수 있음을 입증하고, 반전 서사 생성을 위한 프롬프팅 설계 원칙을 제안한다.

Ⅲ. 스토리 생성

3-1 사용 모델 및 하이퍼파라미터

본 실험에서 반전이 있는 스토리 생성 시 사용한 모델은 OpenAI의 GPT-4o-mini-2024-07-18 모델이다. GPT-4o-mini-2024-07-18 모델을 선택한 이유는 비용 효율적이면서도 성능이 보장되는 모델이기 때문이다. 사용한 하이퍼파라미터(Hyper-parameter)는 max_tokens은 16384, n은 1, stop은 none, temperature는 1.0이다.

하이퍼 파라미터 중 max_tokens을 16384로 설정한 이유는 스토리 생성에서 스토리의 길이가 길어질수록 일관성 있는 스토리를 생성하기 어려우므로 최대 토큰 수로 설정해 스토리를 생성하도록 했다.

Temperature 파라미터는 확률적으로 다음 텍스트를 출력하는 언어 모델의 특성을 조절하는 파라미터로 해당 파라미터를 통해 언어 모델의 출력의 다양성과 무작위성을 조절한다. 0에 가까울수록 일관성 있고 높은 정확도를 가진 답변을 생성하므로 정답이 있거나 높은 정확도를 요구하는 태스크인 번역, 수학문제 풀이, 추론 문제 풀이 등에 적합한다. 반면 1에 가까울수록 확률이 낮은 단어가 종종 선택하며 적당한 다양성을 가진 답변을 생성하므로 일반적인 대화와 내러티브를 생성하는 태스크에 적합하다. 본 연구에서는 temperature를

0.7과 1.0 두 가지를 이용해 반전이 있는 스토리를 생성했으나, 평가 결과 1.0을 사용하는 것이 더 적합하다고 판단되어 1.0을 사용하였다.

3-2 사용 프롬프트

본 논문에서는 다양한 프롬프트 설계를 활용해 언어 모델을 이용한 반전이 있는 스토리를 생성 및 평가를 통해 효과적인 프롬프트 설계를 제안하고자 한다. 이를 위해 다양한 프롬프트를 설계하여 GPT-4o-mini-2024-07-18 모델과 영어로 작성된 프롬프트를 이용해 반전이 있는 영어 스토리를 생성했다. 사용한 프롬프트는 총 5가지로 표 1과 같다.

표 1. 반전이 있는 스토리 생성 시 사용한 프롬프트
Table 1. Used prompt for generating story with a twist

Num	Prompt
1	“Write a story with a twist”
2	“Write a story with a twist that effectively utilizes flashbacks. The narrative should unfold in a way that the flashbacks gradually reveal crucial details about the characters or events, ultimately leading to an unexpected twist at the end. Ensure the story maintains grammatical accuracy, consistent flow, and a believable plot while engaging the reader. The flashbacks should be seamlessly integrated into the main storyline, enhancing the emotional depth and impact of the twist”
3	“"Write a story with a twist using flashbacks to reveal critical details that lead to the unexpected ending."”
4	"Write a story about a clerk at a police station who discovers a serial killer’s notebook and finds their own name listed as the next victim. The story should begin with the protagonist living an ordinary life but gradually escalate in tension as they uncover the truth behind the notebook. Build suspense throughout the narrative, and include a shocking twist at the end, revealing an unexpected truth about the protagonist or their situation that leaves a lasting impression on the reader."
5	"Write a suspenseful story about a police clerk who discovers a serial killer's notebook listing them as the next victim. Begin with their ordinary life, build tension as they investigate, and conclude with an unexpected twist that reveals a shocking truth about the protagonist or their situation."

반전이 있는 스토리 생성을 위한 프롬프트 설계시 추가 정보를 제공하지 않는 제로샷 프롬프팅과 플래시백, 로그라인, 스토리 구조에 따른 전개 등 다양한 스토리 컨텍스트를 제공하는 방법을 활용한 퓨샷 프롬프트를 설계하였다. 이를 통해 다양한 프롬프팅 방식의 반전이 있는 스토리 생성을 분석해 효과적인 프롬프팅 방법을 제시하고자 한다.

1번부터 3번 프롬프트는 소설 생성 시의 인물, 사건, 배경 등에 대한 컨텍스트를 제공하지 않고 제로샷 프롬프트를 다양하게 설계하여 성능을 비교하고 분석했다. 1번 프롬프트는 언어 모델의 기본적인 반전이 있는 스토리 생성 능력을 타 프롬프트와 비교하기 위한 베이스 라인(baseline)으로, 추가적인 조건을 제시하지 않고 오직 반전이 있는 스토리를 생성할 것을 요청했다. 2번 프롬프트는 반전이 있는 스토리를 평가하는 평가 기준인 반전의 퀄리티, 일관성, 문법적 정확도, 흥미, 핍진성에 부합한 스토리를 생성할 것을 요청했다. 3번 프롬프트에서는 가장 대중적인 반전을 표현하는 기법 중 하나인 플래시백 기법(flashback)을 사용해 반전이 있는 스토리를 생성할 것을 요청했다.

4번과 5번 프롬프트에서는 퓨샷 프롬프팅을 통해 언어 모델이 보다 구체적인 컨텍스트를 제공하는 프롬프트를 설계하였다. 이를 위해, GPT-4o를 통해 반전이 있는 로그라인 5개를 생성하도록 요청한 후, 가장 흥미로운 로그라인을 선택하는 과정을 거쳤다.

로그라인 생성 시 사용한 프롬프트는 “Write five log lines of interesting stories with twists”이며 답변은 표 2와 같다. 언어모델이 선정한 가장 흥미로운 로그라인은 “The Serial Killer's Notebook”으로, 해당 로그라인을 스토리 생성 시 사용했다.

4번 프롬프트는 로그라인과 함께 스토리의 시작, 중간, 끝의 스토리 구조를 구체적으로 설명했으며, 일반적인 삶과 반전의 대비를 강조할 것을 요청하는 등 반전을 구성하는 요소에 대해 세세한 지시를 포함했다. 스토리의 구조를 활용하되 이야기의 전개를 비교적 간결하게 지시한 5번 프롬프트를 작성해 스토리를 생성 후 비교 및 분석을 진행한다.

각 프롬프트별로 반전이 있는 스토리를 250개씩 생성했으며 하이퍼 파라미터는 모두 동일하게 설정했다.

표 2. 언어모델이 생성한 반전이 있는 로그라인
Table 2. Log line with a twist generated by language model

Title	Log line
The Wish-Granting Mirror	A high school girl starts making wishes to an old mirror, and everything she desires comes true, only to realize that her happiness comes at the cost of others' misfortune.
The Mysterious Box	A regular delivery man discovers an old photograph with his name on it inside an opened package, only to find that the photo is linked to his younger sibling who went missing ten years ago.
The True Face of a Friend	A world-famous bestselling author owes their success to advice from an old friend, only to discover that the friend is a manifestation of a rejected version of their past self.
The Serial Killer's Notebook	A clerk working at a police station stumbles upon the notebook of a serial killer, only to find their own name on the list of the next victims.
The Price of Time Leaps	A scientist researching time travel goes back in time to save a loved one, only to realize that the technology they created will lead to humanity's destruction in the future.

3-3 생성 결과 및 분석

표 1에 제시된 언어 모델을 사용하여 각 프롬프트당 250개의 반전이 있는 스토리를 생성했다. 생성된 스토리의 평균

토큰 수에 대한 통계는 표 3에 제시되어 있으며, 평균 및 표준 편차는 소수점 셋째 자리에서 반올림하여 작성되었다. 표에서 괄호 밖의 수치는 평균을, 괄호 안의 수치는 표준 편차를 의미한다.

모든 프롬프트에서 max_tokens을 16,384로 설정했으나, 해당 최대 토큰 수를 충족한 스토리는 단 한 개도 없었다. 평균적으로 가장 많은 토큰을 생성한 프롬프트는 4번으로, 생성된 스토리의 평균 길이는 1,056.79 토큰이었다. 반면, 평균적으로 가장 짧은 스토리를 생성한 프롬프트는 가장 간략하게 작성된 1번 프롬프트로, 평균 916.20개의 토큰을 생성했다.

표 3. 프롬프트 별 생성 토큰 수 평균
Table 3. Average token count per prompt

Num	Average Token Count
1	916.20(109.92)
2	1004.90(109.50)
3	993.06(110.76)
4	**1056.79(87.79)**
5	1020.71(87.19)

생성된 스토리의 평균 토큰 수를 비교한 결과, 프롬프트의 상세한 서술 정도가 스토리 길이에 영향을 미치는 것으로 나타났다. 가장 긴 스토리를 생성한 4번 프롬프트는 비교적 가장 구체적인 지침을 포함하고 있었으며, 이는 모델이 더 풍부한 내용을 생성하도록 유도했을 가능성이 크다.

반면, 1번 프롬프트는 간략한 서술로 구성되어 있어 상대적으로 짧은 스토리를 생성하는 경향을 보였다. 또한, max_tokens을 16,384로 설정했음에도 불구하고, 생성된 스토리 중 최대 토큰 수에 도달한 사례가 없었다는 점은 모델이 특정 조건 내에서 자연스럽게 스토리를 마무리하는 경향을 보인다는 것을 시사한다. 이는 프롬프트의 구조뿐만 아니라 모델의 내부 학습 방식과도 관련이 있을 가능성이 있으며, 추가적인 실험을 통해 정량적인 분석이 필요할 것으로 보인다. 따라서, 추후 연구에서는 생성해야 할 스토리의 길이를 max_tokens으로 설정하고, 이를 명시적으로 프롬프트에 제공하여 다양한 길이의 스토리를 생성하고 분석할 예정이다.

Ⅳ. 평가 및 결과분석

4-1 평가 방법

본 논문에서는 언어모델을 사용해 반전이 있는 스토리를 생성했으며 언어모델을 이용한 자동 평가와 사람 평가를 진행했다. 언어 모델을 이용한 자동 평가 시 사용한 언어모델은 스토리 생성 시 사용한 언어모델과 같은 OpenAI 사의 GPT-4o-mini-2024-07-18 모델, Meta의 Llama-3-8B-8192 모델과 Anthropic의 Claude-3-haiku-20240307 모델 총 3가지 모델을 사용했다. GPT-4o-mini의 경우에는 생성에 사용했으며 OpenAI 사의 모델 중 가장 비용 효율적인 모델이므로 평가 모델로 선정했다. 또한 Meta의 오픈소스 모델인 Llama를 사용해 오픈소스 모델과 상업용 소스모델의 특성을 비교하였으며, 추론과 수학 등의 태스크에서 높은 정확도의 출력을 생성하는 Claude-3-haiku를 사용해 다양한 모델 간의 차이를 분석했다.

생성 시와 마찬가지로 평가 시에도 영어로 된 프롬프트를 이용해 평가를 진행했으며, 출력 텍스트 또한 영어를 사용했다. 입력과 출력을 영어로 진행한 이유는 다른 언어에 비해 상대적으로 높은 성능을 낼 수 있기 때문이다.

평가 시 사용한 하이퍼파라미터는 temperature=0.5, max_tokens=1024, top_p=1, stop=none이다.

표 4. 언어 모델을 이용한 스토리 평가 프롬프트 예시
Table 4. Example of a story evaluation prompt using a language model

Story evaluation prompt of GPT-4o-mini

```
messages = [
    {
        "role": "system",
        "content": "You are a strict evaluator."
    },
    {
        "role": "user",
        "content": (
            "Evaluate the following story strictly based on the criteria below. "
            "Provide critical and detailed feedback, and do not hesitate to assign lower scores if necessary:₩n"
            "1. Plot Twist Quality (1-5): 1 - not surprising, 5 - completely unexpected and impactful.₩n"
            "2. Coherence (1-5): 1 - illogical, 5 - very consistent and logical.₩n"
            "3. Grammar Accuracy (1-5): 1 - many errors, 5 - perfect.₩n"
            "4. Verisimilitude (1-5): 1 - highly unrealistic or inconsistent, 5 - very realistic or internally consistent.₩n"
            "5. Enjoyment (1-5): 1 - very boring, 5 - highly engaging and memorable.₩n"
            "Provide scores and explanations in the format:₩n"
            "- Plot Twist Quality: X (Reason: ...)₩n"
            "- Coherence: X (Reason: ...)₩n"
            "- Grammar Accuracy: X (Reason: ...)₩n"
            "- Verisimilitude: X (Reason: ...)₩n"
            "- Enjoyment: X (Reason: ...)₩n₩n"
            f"Story: {story}"
        )
    }
]
```

각각의 언어모델에 대해 이용해 5개의 프롬프트를 이용해 250개씩 생성한 반전이 있는 스토리를 반전의 퀄리티, 일관성, 문법적 정확도, 핍진성, 흥미를 기준으로 1점부터 5점 사이의 리커트 척도로 평가했다. 평가 기준의 경우 선행 연

구에서 스토리를 평가할 때 사용한 기준들과 함께 반전이 있는 스토리 평가에 적합하다고 생각하는 흥미, 핍진성을 추가했다. 평가 프롬프트는 선행 연구[29]를 참고하여 단순히 점수만 생성하는 것이 아닌 평가한 이유를 함께 생성하도록 했다.

언어모델을 이용한 평가 시 단순히 평가 기준을 제시하며 점수와 이유를 요청하면 대부분의 스토리를 모든 항목에서 5점과 4점을 주로 주게 되어 평가의 변별력 문제가 생기므로 평가 시 사용한 언어 모델 GPT-4o-mini와 Llama-3-8B, 그리고 Claude-3-Haiku 모두 'You are a evaluator'에서 'strict'라는 단어를 추가해 'You are a strict evaluator'로 엄격한 평가를 요구하는 프롬프팅을 진행했다. 언어 모델 평가 시 사용한 프롬프트 예시는 표 4와 같다.

언어 모델을 이용한 자동 평가 결과의 신뢰성을 검증하기 위해 사람 평가를 함께 진행하였다. 평가 참여자는 총 8명으로 여성 2명, 남성 6명이다. 나이는 19-24세 3명, 25세-29세 5명이며, 모두 공학(engineering) 혹은 전산학(computer science)을 전공하였다.

사람 평가와 자동 평가는 동일한 평가 기준(반전의 퀄리티, 일관성, 문법적 정확도, 핍진성, 흥미도)을 이용했으며, 자동 평가와 마찬가지로 평가 참여자는 1점부터 5점 사이의 리커트 척도로 평가함과 동시에 평가 이유를 함께 서술하도록 요청받았다. 평가 참여자에게는 소정의 기프티콘을 평가 참여에 대한 보상으로 지급하였다.

언어 모델을 이용한 자동 평가와 사람 평가 시 차이점은 두 가지이다. 첫 번째로, 예산과 시간 관계상 언어 모델이 생성한 모든 스토리를 평가하기 어려우므로 각 프롬프트를 이용해 생성된 스토리 중 자동 평가 결과 가장 높은 점수를 받은 스토리 중 하나를 샘플링해 평가에 사용하였다. 두 번째로, 사람 평가 시에는 영어로 작성된 이야기를 GPT-4o를 이용해 번역한 번역본과 원본을 함께 제공하였으며, 참가자는 제공된 원본과 번역본 중 하나 혹은 두 개를 선택하여 읽고 평가를 진행했다.

대다수의 참가자는 번역본과 영어로 작성된 원본을 모두 읽고 평가하였으며, 일부 참가자는 번역본만 읽고 평가를 진행했다. 번역본만 읽은 참가자와 원본과 번역본을 함께 읽은 참가자 간의 점수 분포 및 평가 항목별 경향성에서 유의미한 차이가 나타나지 않았다. 이는 번역의 질이 평가 과정에 미치는 영향을 최소화하였음을 시사하며, 번역본을 제공하는 방식이 평가의 일관성을 유지하는 데 효과적일 수 있음을 의미한다.

4-2 언어 모델 기반 평가 결과

표 1의 다섯 가지 프롬프트를 이용해 각 프롬프트 별로 250개씩 생성한 반전이 있는 스토리에 대해 언어모델을 이용한 자동 평가 결과를 언어모델 별 프롬프트 평가 결과 평균과 표준 편차를 표 5-7에 작성하였다.

모든 평가 결과 평균과 표준 편차는 소수점 셋째 자리에서 반올림해 작성되었다. 괄호 밖에 작성된 수치는 평가 결과의 평균이며, 괄호 안에 작성된 수치는 표준 편차이다.

표 5. GPT-4o-mini를 이용한 평가 결과 평균과 표준편차
Table 5. Mean and standard deviation of evaluation results using GPT-4o-mini

	Quality of plot twist	Coherence	Grammatical Accuracy	Verisimilitude	Enjoyment
1	3.90 (0.29)	**4.54 (0.53)**	**4.96 (0.18)**	3.73 (0.44)	4.29 (0.45)
2	3.99 (0.15)	4.51 (0.54)	4.94 (0.23)	**3.75 (0.43)**	**4.35 (0.47)**
3	3.95 (0.21)	4.34 (0.51)	4.90 (0.28)	3.71 (0.45)	4.20 (0.40)
4	**4.04 (0.20)**	3.99 (0.51)	4.84 (0.36)	3.28 (0.45)	4.14 (0.35)
5	4.02 (0.17)	4.01 (0.46)	4.86 (0.34)	3.26 (0.35)	4.16 (0.44)

표 6. Llama-3-8B를 이용한 평가 결과 평균과 표준편차
Table 6. Mean and standard deviation of evaluation results using Llama-3-8B

	Quality of plot twist	Coherence	Grammatical Accuracy	Verisimilitude	Enjoyment
1	4.00 (0.06)	**4.47 (0.50)**	**4.98 (0.10)**	**4.01 (0.22)**	**4.46 (0.51)**
2	4.00 (0.10)	4.23 (0.43)	4.88 (0.35)	3.99 (0.21)	4.16 (0.38)
3	3.99 (0.06)	4.18 (0.40)	4.78 (0.44)	3.95 (0.24)	4.12 (0.35)
4	**4.02 (0.12)**	4.15 (0.37)	4.82 (0.39)	3.96 (0.21)	4.02 (0.20)
5	4.00 (0.12)	4.20 (0.40)	4.84 (0.39)	3.92 (0.32)	4.04 (0.21)

표 7. Claude-3-Haiku를 이용한 평가 결과 평균과 표준편차
Table 7. Mean and standard deviation of evaluation results using Claude-3-Haiku

	Quality of plot twist	Coherence	Grammatical Accuracy	Verisimilitude	Enjoyment
1	4.24 (1.06)	4.62 (0.75)	4.84 (0.42)	4.52 (0.86)	4.37 (1.07)
2	4.55 (0.96)	4.72 (0.76)	4.86 (0.55)	4.69 (0.81)	4.60 (0.96)
3	4.58 (0.73)	4.76 (0.53)	4.92 (0.29)	4.74 (0.55)	4.60 (0.76)
4	**4.94 (0.28)**	**4.98 (0.16)**	**4.99 (0.06)**	**4.98 (0.15)**	**4.96 (0.27)**
5	4.82 (0.59)	4.88 (0.44)	4.96 (0.23)	4.88 (0.45)	4.84 (0.60)

평가 결과에 따르면, 가장 간단한 프롬프트인 "Write a story with a twist"(1번 프롬프트)는 모든 모델에서 문법적으로 높은 평가를 받았다. 반면 반전의 질은 낮은 평가를 주

로 받았다. 특히 Claude-3-Haiku 모델은 해당 프롬프트의 반전의 질 품질을 타 프롬프트에 비해 현저히 낮은 점수로 평가했다. 해당 모델이 반전의 질 항목에서 1점을 부여한 이유 중 하나로는 “The so-called "twist" is as predictable as the rising and setting of the sun. Any halfway competent reader could see it coming from a mile away. It lacks any semblance of originality or impact.”(소위 말하는 ‘반전’은 해가 뜨고 지는 것만큼 예측 가능하다. 어느 정도 독해 능력을 가진 독자라면 누구나 그 결과를 쉽게 예상할 수 있었을 것이다. 반전에는 독창성이나 강렬함이 전혀 없다.)와 같이 스토리가 예측 가능하며 독창성이 없음을 꼽았다. 반면 같은 스토리를 Llama-3-8B는 “The story has a few unexpected turns, such as the discovery of the well, the appearance of Amelia's ancestors, and the revelation of the curse. However, the final twist, where Amelia breaks the curse by making a selfless wish, is somewhat predictable and doesn't feel completely unexpected.”(이 스토리에는 우물의 발견, 아멜리아의 조상 등장, 저주의 비밀이 드러나는 등의 몇 가지 예기치 못한 전개가 포함되어 있다. 그러나 아멜리아가 이타적인 소원을 빌어 저주를 깨는 최종 반전은 다소 예측 가능하며 전혀 뜻밖이라는 느낌이 들지 않는다.)로 스토리가 예측 가능함을 꼽으며 다소 비판적인 평가를 했으나 4점을 부여했다.

2번 프롬프트의 경우 평가 기준에 맞게 반전이 있는 스토리를 생성할 것을 요구했으며, GPT-4o-mini와 Llama-3-8B의 경우 문법적으로 강점이 있음을 평가했다. 특징적으로 Claude-3-Haiku에서는 핍진성이 약하다고 했으나 타 모델들에서는 핍진성 부분에서 높은 평가를 받는 축에 속하는 상반된 결과를 보였다. GPT-4o-mini에서는 “While the concept of a magical mirror is a common fantasy trope, the internal consistency of the characters' motivations and the mirror's powers is somewhat lacking. The emotional connection Eleanor feels with Miranda is compelling, but the mechanics of how this connection operates could be better explored to enhance believability. Additionally, the portrayal of the historian feels somewhat underdeveloped.”(마법 거울이라는 개념은 흔한 판타지 장치이지만, 주인공들의 동기와 거울의 능력 사이의 내부적인 일관성이 다소 부족하다. 엘리너가 미란다와 느끼는 감정적 연결은 매력적이지만, 이러한 연결이 어떻게 작동하는지에 대한 기제가 더 잘 탐구되었더라면 설득력이 더 높아졌을 것이다. 또한, 역사가의 묘사는 다소 미흡하게 느껴진다.)라는 평은 주인공의 동기와 소재의 일관성이 부족하며 묘사가 미숙한 점을 이유로 1점을 부여했으나 Claude-3-Haiku의 경우 “The story is gripping, thought-provoking, and emotionally resonant, leaving a lasting impact on the reader and making it highly engaging and memorable.”로 흥미진진하고, 감정적으로 공감을 불러 일으켜 독자에게 여운을 오래 남길 수 있다는 이유로 5점을 부여하는 상반되는 결과를 보였다.

3번 프롬프트의 경우 반전의 종류 중 하나인 플래시백을 이용해 반전이 있는 스토리를 생성할 것을 요청했으며, 1번 2번 프롬프트에 비해서는 낮은 평가를 받았으나 모든 모델의 모든 평가 지표에서 중간 혹은 그 이하의 성능을 나타내는 것으로 평가되었다.

4번 프롬프트에서는 문법적 정확도 외에는 모든 언어모델에서 공통적으로 4번 프롬프트는 반전이 있는 로그라인과 스토리의 시작 및 결말에 대해 어떻게 작성할 지에 대한 가이드라인을 입력했다. GPT-4o-mini와 Llama-3-8B에서는 각각 핍진성이 낮으며, 일관성과 재미 부분에서 약점을 나타냈다. 하지만 반전의 퀄리티 부분에서는 타 프롬프트에 비해 조금 더 높은 평가를 받았다. Claude-3-Haiku에서는 모든 지표에서 타 프롬프트에 비해 가장 높은 평가를 받는 결과를 보였다.

5번 프롬프트는 4번 프롬프트를 간략하게 변형했다. 로그라인을 제공하되 4번 프롬프트처럼 스토리의 세부요소에 대한 지시없이 로그라인과 반전을 어떻게 표현할 지에 대해 보다 간략하게 지시했다. 평가 결과 GPT-4o-mini와 Llama-3-8B에서는 4번 프롬프트에 비해서는 반전의 퀄리티와 핍진성이 낮았으나 다른 항목에서는 모두 높은 점수를 보였다.

모델별 평가 결과를 비교한 결과, GPT-4o-mini와 Llama-3-8B는 핍진성 평가에서 상대적으로 엄격한 경향을 보였다. 반면, Claude-3-Haiku는 동일한 프롬프팅에서도 후한 평가를 내리며, 특히 핍진성 측면에서 엄격한 평가가 이루어지지 않았다. 이를 개선하기 위해 추가적인 프롬프팅을 시도했으나, 다른 모델과 달리 명확히 평가 방식이 조정되지 않는 경향이 있었다.

Claude-3-Haiku의 경우 흥미로움에도 불구하고 서사적인 부족함이 있다면 1점과 2점, 크게 없다고 생각되면 4점과 5점을 부여하는 경향을 보였다. 하지만 대부분의 경우 4점과 5점으로 평가하므로 평가 결과의 변별력이 낮았으며, Claude-3-Haiku의 평가 결과 표준 편차는 모두 1에 가깝다. 이는 타 모델과 비교했을 때 Claude-3-Haiku의 평가 결과는 변동성이 높음을 알 수 있다.

반면, GPT-4o-mini와 Llama-3-8B의 경우에는 서사적인 부족함이 있음에도 불구하고 다른 요소를 통해 좋은 평가를 내리는 경향이 있었으며 두 언어 모델의 평가는 Claude-3-Haiku와 다르게 유사한 것을 알 수 있었다. 두 모델은 핍진성뿐 아니라 서사적인 완성도를 중점으로 평가하는 경향이 있었으며, 프롬프트의 변형에 따라 유연하게 바뀌는 모습을 보였다. 특히 엄격한 평가를 요하는 프롬프트를 제공 시 이전에 비해 보다 엄격하게 점수를 부여하는 경향을 보였다.

모든 언어 모델에서 4번 프롬프트를 사용했을 때, 반전의 질 항목에서 가장 높은 점수가 부여되었으며, 다른 프롬프트 대비 평가 평균이 최소 0.02에서 최대 0.70까지 높게 나타났

다. 이는 프롬프트 설계가 반전이 있는 스토리의 품질에 중요한 영향을 미친다는 점을 시사한다.

그러나 GPT-4o-mini와 Llama-3-8B의 평가 결과를 분석한 결과, 반전의 질을 높이기 위해 많은 내용을 요구하는 프롬프팅이 적용되었을 때 상대적으로 일관성, 핍진성, 흥미 등의 항목에서 낮은 점수가 부여되는 경향이 확인되었다. 이는 반전의 질을 강조하는 프롬프트가 전체적인 이야기의 균형성에 영향을 미칠 가능성이 있음을 의미한다.

4-3 사람 평가 결과

표 1의 다섯 가지 프롬프트를 이용해 각 프롬프트 별로 250개씩 생성한 반전이 있는 스토리에 대해 언어모델을 이용한 사람 평가 결과를 표 8에 작성하였다. 모든 평가 결과 평균과 표준 편차는 소수점 셋째 자리에서 반올림해 작성되었다. 괄호 밖에 작성된 수치는 평가 결과의 평균이며, 괄호 안에 작성된 수치는 표준 편차이다.

표 8. 사람 평가 결과 평균과 표준편차
Table 8. Mean and standard deviation of human evaluation results

	Quality of plot twist	Coherence	Grammatical Accuracy	Verisimilitude	Enjoyment
1	2.87 (0.99)	**3.50 (0.97)**	**4.62 (0.48)**	2.50 (0.97)	3.00 (0.75)
2	3.00 (1.34)	4.37 (1.13)	**4.62 (0.48)**	**3.62 (1.39)**	3.50 (0.97)
3	2.50 (1.27)	3.37 (0.97)	4.50 (0.48)	2.00 (1.00)	2.50 (1.27)
4	**3.62 (1.38)**	**3.50 (0.95)**	**4.62 (0.48)**	2.75 (1.06)	**3.62 (0.95)**
5	2.87 (1.46)	3.37 (1.27)	4.50 (0.48)	2.12 (0.69)	3.50 (0.97)

평가 결과 분석을 통해 문법적 정확성과 창의적 요소 간 평가의 일관성 차이를 확인할 수 있었다. 문법적 정확성의 경우, 모든 프롬프트에서 평균적으로 높은 점수를 기록했으며, 표준편차 역시 다른 평가 항목에 비해 낮은 수준을 유지하였다. 이는 평가자들이 언어 모델이 생성한 스토리의 문법적 요소를 안정적으로 인식하였으며, 모델이 문법적으로 오류가 적은 스토리를 생성했음을 시사한다.

반면, 반전의 질, 개연성, 흥미도와 같은 창의적 요소에서는 높은 변동성을 보였다. 특히 개연성 항목에서 2번 프롬프트는 1.39의 높은 표준편차를 나타내며 평가자 간 의견 차이가 두드러졌다. 평가자들의 평가 이유를 분석한 결과, 특정 맥락에서 개연성을 유지하는 데에 어려움을 겪었으며, 평가자별로 개연성에 대한 기대치가 상이했음이 드러났다. 이는 일부 평가자가 사건 전개의 자연스러움을 중요하게 여긴 반면, 다른 평가자들은 극적인 서사적 장치를 더 선호했기 때문임을 평가 이유 분석을 통해 알 수 있었다.

또한, 일관성 항목에서도 표준편차가 0.95~1.27 사이로 다소 높은 수준으로 나타났으며, 가장 낮은 점수를 받은 3번과 5번 프롬프트에서는 서사적 연결이 부족하다는 평가가 타 프롬프트에 비해 다수 존재했다.

반전의 질 항목에서 가장 높은 점수를 받은 프롬프트는 4번이었다. 해당 프롬프트는 "진실이 밝혀짐과 동시에 또 다른 인물의 동조로 의외성이 강화됨"과 같은 평가를 받았으며, 반전이 효과적으로 구현되었다고 평가받았다. 그러나 "이해가 되지 않는 반전"이라는 이유로 1점을 부여한 평가도 존재하여, 이는 앞서 개연성과 같이 반전에 대한 해석이 평가자마다 상이함을 보여준다. 반면, 가장 낮은 평가를 받은 3번 프롬프트는 "반전이 잘 느껴지지 않음", "이야기가 갑작스럽게 마무리됨" 등의 이유로 부정적인 평가를 받았다.

일관성 항목에서는 2번 프롬프트가 가장 높은 평가를 받았으며, "이야기의 전개가 논리적이며 캐릭터 설정이 일관됨"과 같은 피드백이 주어졌다. 반면, 3번 프롬프트는 "주변 인물 및 주인공의 감정 변화에 대한 서술 없이 급격한 전개가 이루어짐"이라는 이유로 낮은 점수를 받았다. 이는 3번 프롬프트로 생성된 스토리가 사건의 발단과 전개 과정이 설득력을 갖추지 못했음을 의미한다.

핍진성 항목에서는 가장 높은 점수를 받은 프롬프트는 2번이다. "주인공의 심리 변화와 감정적 갈등이 현실적으로 그려짐", "개연성이 좋았음"과 같은 피드백을 받았다. 핍진성 항목에서 가장 낮은 점수를 받은 3번 프롬프트는 "주변 인물에 대한 설명, 주인공의 감정 변화에 대한 묘사 없이 이야기가 갑자기 진행되고 종결됨", "이야기에서 주장하는 사건의 발단이 사건의 트리거로 충분하지 않음"과 같은 피드백을 받았다.

흥미도 항목에서는 4번 프롬프트가 가장 높은 평가를 받았으며, "서스펜스와 심리적 긴장감이 강하게 작용함"이라는 점이 긍정적으로 평가되었다. 반면, 3번 프롬프트는 "초반에는 기대감을 유발했으나, 과도한 과거 회상과 갑작스러운 결말로 인해 몰입감이 저하됨"이라는 피드백을 받으며 가장 낮은 평가를 받았다.

이러한 평가 결과를 종합 시 향후 프롬프트 설계에서는 스토리의 개연성과 몰입감을 강화하기 위한 서사적 조정이 필요함을 알 수 있다. 또한, 반전의 질과 같은 창의적 요소는 평가자의 주관적 해석에 따라 큰 차이를 보이므로, 보다 효과적인 평가 기준을 수립하는 것이 중요함을 알 수 있다.

이를 위해 평가 지침을 명확하게 정의하고, 평가자 간 해석의 일관성을 높이기 위한 구체적인 가이드라인을 만드는 것을 향후 연구에서 적극 고려할 예정이다. 한편, 문법적 정확성이 안정적으로 유지된 점을 고려할 때, 향후 연구에서는 더욱 창의적 요소의 개선에 초점을 맞춘 실험 설계가 필요할 것으로 판단된다.

4-4 언어 모델 평가와 사람 평가의 상관관계 분석

언어 모델이 수행한 자동 평가 결과와 사람 평가 결과 간의 상관관계를 분석하여, 각 언어 모델이 사람 평가자들의 평가 방식과 얼마나 유사한지를 파악하였다. 또한 단순히 상관계수 값만으로 언어 모델과 사람 평가 간의 유의미한 관계를 단정할 수 없으므로, t-검정을 수행하여 상관계수의 통계적 유의성을 검정하였으며 결과를 표 9에 작성했다.

표 9. 언어 모델 자동 평가와 사람 평가의 상관관계 분석 및 p-value 검정 결과

Table 9. Correlation analysis between language model and human evaluation with p-value significance test

Modelzz	Pearson correlation coefficient (r)	p-value
GPT-4o-mini	0.88	0.0490
Llama-8B	**0.92**	**0.0268**
Claude-3-haiku	0.86	0.0615

그 결과, Llama-3-8B는 사람 평가와 가장 높은 상관관계(r=0.92)를 보였으며, p-value가 0.0268로 검정 결과 통계적으로 유의미한 상관관계를 가짐을 확인할 수 있었다. 이는 Llama-3-8B의 평가 점수가 Human 평가 점수와 높은 연관성을 가지며, 우연에 의해 발생한 결과가 아닐 가능성이 크다는 것을 의미한다.

반면, 언어 모델 중 후한 평가를 하는 경향을 보였던 Claude-3-Haiku는 사람 평가 점수와의 상관계수는 상대적으로 낮았으며(r =0.86), p-value가 0.0615로 나타났다. 즉, Claude-3-Haiku의 평가와 사람 평가자들의 평가 방법 및 기준에 다소 차이가 있음을 시사한다. GPT-4o-mini는 사람 평가와의 상관계수가 0.88로 높은 편이었지만, p-value가 0.0490으로 경계선에 위치하여 사람 평가와의 상관성을 확신할 수 있는 수준은 아님을 알 수 있었다.

결론적으로 Llama-3-8B는 다른 언어 모델에 비해 사람 평가와 가장 유사한 평가 경향을 보였으며, 언어 모델 중 사람 평가자들의 평가와 가장 유사한 평가 결과를 생성하였다.

4-5 결과 분석

본 절에서는 자동 평가 결과와 사람 평가 결과를 통해 효과적인 프롬프트 설계를 위한 결론 도출 및 언어 모델별 평가 경향성을 제시하고자 한다.

우선 본 연구에서는 프롬프트의 서술 방식이 생성된 스토리의 길이에 영향을 미친다는 점을 확인하였다. 구체적인 지침을 포함한 4번 프롬프트는 평균적으로 가장 긴 스토리를 생성한 반면, 간략하게 작성된 1번 프롬프트는 가장 짧은 스토리를 생성하는 경향을 보였다. 또한, max_tokens을 16,384로 설정했음에도 불구하고 최대 토큰 수에 도달한 스토리는 없었으며, 이는 모델이 특정 조건 내에서 자연스럽게 스토리를 마무리하는 경향이 있음을 시사한다. 이러한 결과는 프롬프트의 구조뿐만 아니라 모델의 내부 학습 방식과도 관련될 가능성이 있으며, 향후 연구에서는 보다 정밀한 실험을 통해 프롬프트 설계와 생성 결과 간의 관계를 정량적으로 분석할 필요가 있다.

자동 평가와 사람 평가 모두에서 로그라인을 통해 스토리의 컨텍스트를 제시하고, 스토리의 구조(처음, 중간, 끝)를 활용하여 구체적인 지침을 제공한 4번 프롬프트가 반전의 질에서 가장 우수한 평가를 받았다. 이는 반전 요소를 명확히 구조화하여 제시하는 방식이 언어 모델이 보다 완성도 높은 반전을 생성하는 데 효과적임을 시사한다. 그러나 4번 프롬프트는 다른 평가 항목(일관성, 문법적 정확도, 핍진성, 흥미도)에서는 상대적으로 낮은 점수를 받는 경향을 보였다. 이는 반전의 질을 높이기 위해 많은 정보를 모델에 제공하면서도, 그 과정에서 모델이 이야기의 자연스러운 흐름을 형성할 수 있는 여지가 줄어들었기 때문으로 해석할 수 있다.

이를 통해, 반전이 있는 스토리를 생성할 때 스토리 진행에 따라 반전의 핵심 요소를 구체적으로 기술하는 방식이 반전의 질을 향상시키는 데 효과적임을 확인할 수 있다. 그러나 이러한 방식은 언어 모델의 창의적 서술을 제한할 가능성이 있으며, 그 결과 일관성과 핍진성, 흥미도 측면에서 완성도가 저하되는 경향을 보인다. 따라서 반전의 질을 유지하면서도 이야기의 자연스러움을 확보할 수 있는 최적의 프롬프트 설계가 필요성을 제시한다. 이러한 결과는 언어 모델별 평가 기준이 다를 수 있으며, 모델별로 최적화된 프롬프팅 기법이 필요함을 의미한다.

또한, 각 언어 모델의 평가 경향성에 차이가 존재함을 확인할 수 있었다. GPT-4o-mini와 Llama-3-8B는 핍진성과 일관성을 보다 엄격하게 평가하는 반면, Claude-3-Haiku는 상대적으로 후한 점수를 부여하는 경향을 보였다. 특히 Claude-3-Haiku는 같은 프롬프팅 전략을 사용하더라도 평가의 변별력이 낮고, 타 모델과 비교했을 때 핍진성과 서사적 완성도 평가에서 일관되게 관대한 평가를 부여하는 경향이 확인되었다.

한편, 자동 평가와 사람 평가 간의 상관관계를 분석한 결과, Llama-3-8B의 평가 결과가 사람 평가와 가장 유사한 패턴을 보였다. 자동 평가는 문법적 정확성, 논리적 일관성, 텍스트의 구조적 측면과 같은 정량적 요소를 측정하는 데 강점을 가지는 반면, 사람 평가는 서사적 완성도, 감성적 반응과 같은 정성적 요소를 고려하기 때문에 두 평가 방식 간 차이가 발생할 가능성이 크다. 이로 인해 자동 평가와 사람 평가 간의 차이가 발생할 가능성이 높으며, 향후 연구에서는 자동 평가 지표와 사람 평가 지표 간의 정합성을 높이는 방법을 탐색하여 보다 신뢰도 높은 평가 체계를 구축할 필요가 있음이 확인되었다.

또한, 사람 평가의 경우 평가자 간 편차가 크게 발생하는 문제가 확인되었다. 이는 스토리에 대한 개별 평가자의 주관적 해석 차이에서 기인하는 것으로 보이며, 특히 창의적 요소와 감성적 반응을 평가할 때 편차가 더욱 두드러졌다. 이러한

문제를 해결하기 위해서는 평가 기준을 보다 구체적으로 정리한 세분화된 가이드라인을 제공하고, 평가자의 사전 교육을 강화하는 방법이 필요할 것이다. 향후 연구에서는 평가 과정에서 일관된 기준을 유지할 수 있도록 평가 프레임워크를 정교화하고, 평가자의 주관적 편차를 줄이기 위한 방법을 적용하는 연구가 필요하다.

연구 결과, 프롬프트의 설계 방식이 반전의 질뿐만 아니라 창의성과 일관성에도 중요한 영향을 미친다는 점을 시사한다. 특히, 구조적 지침이 강화된 프롬프트는 반전의 질을 높이는 데 효과적이지만, 모델의 자유로운 해석을 제한하여 창의성과 일관성 면에서 성능의 손실을 초래할 가능성이 있다. 따라서 향후 연구에서는 프롬프트의 구체성과 모델의 창의성 사이의 균형을 분석하고, 언어 모델을 활용한 자동 평가의 신뢰성을 검증함으로써, 반전 서사의 생성 및 평가 체계를 보다 정교하게 개선할 필요가 있다.

V. 결 론

본 연구는 언어 모델을 활용하여 반전이 있는 스토리를 생성하기 위한 효과적인 프롬프트 엔지니어링 기법을 탐색하였다. 이를 위해 5가지 프롬프트를 설계하여 각 250개의 반전이 있는 스토리를 생성하고, 3가지 언어 모델(GPT-4o-mini, Llama-3-8B, Claude-3-Haiku)을 활용한 자동 평가 및 사람 평가를 병행하여 분석하였다. 또한, 자동 평가와 사람 평가 간의 상관관계를 분석하여 평가 방식 간의 차이를 규명하였다.

향후 연구에서는 반전의 질과 스토리의 일관성을 동시에 보장하기 위한 하이브리드 프롬프트 설계 기법을 도입할 필요가 있다. 예를 들어, 반전의 핵심 요소를 명확히 제공하되, 세부적인 서사 진행 방식은 언어 모델의 자율성을 보장하는 방식이 가능할 것이다. 또한, 모델이 생성한 결과물을 사후적으로 평가하여 보정하는 후처리 기반 평가 보정 기법을 적용하는 것도 해결책이 될 수 있을 것으로 보인다.

또한, 모델별 평가 경향성을 고려한 최적화 전략이 필요함이 확인되었다. GPT-4o-mini와 Llama-3-8B 모델은 일관성과 핍진성을 엄격하게 평가하는 반면, Claude-3-haiku 모델은 후한 평가를 내리는 경향을 보인다. 이러한 차이가 발생하는 원인은 모델의 학습 방식, 사전 훈련 데이터의 구성, 평가 태스크 수행 방식 등에 기인할 가능성이 크며, 향후 연구에서는 모델별 특성을 반영한 맞춤형 프롬프팅 기법을 탐색하여 특정 모델이 가진 평가 편향성을 보완할 필요가 있다.

결론적으로, 본 연구는 반전이 있는 서사의 효과적인 생성 및 평가를 위한 프롬프트 엔지니어링 기법을 탐색하였으며, 언어 모델을 활용한 서사 생성 연구의 확장을 위한 기초적인 방향성을 제시하였다. 향후 연구에서는 프롬프트와 언어 모델 간의 상호작용을 보다 정밀하게 분석하고, 언어 모델을 활용한 자동 평가의 신뢰성을 높이기 위한 연구를 지속적으로 진행함으로써, 반전이 있는 서사의 생성 및 평가 체계를 더욱 발전시킬 수 있을 것으로 기대된다.

감사의 글

이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2021R1A2C1012377).

참고문헌

[1] Aristotle, *Poetics*, Oxford, UK: Clarendon Press, 1968. https://doi.org/10.1093/actrade/9780198140245.book.1

[2] C. Chun, “Structural Analysis of Cognitive and Emotional Responses of Story Users: Focusing on Suspense, Curiosity, and Surprise,” *Journal of Culture Industry,* Vol. 24, No. 3, pp. 179-187, September 2024. https://doi.org/10.35174/JKCI.2024.09.24.3.179

[3] H. J. Pérez, “The Plot Twist in TV Serial Narratives,” *Projections*, Vol. 14, No. 1, pp. 58-74, March 2020. https://doi.org/10.3167/proj.2020.140105

[4] M. Terlunen, All Along…! The Pre-History of the Plot Twist in Nineteenth-Century Fiction, Ph.D. Dissertation, Columbia University, New York, NY, 2022.

[5] C. M. Sung and H. J. Kim., “Exploring the Narrative Persuasion on the Story Plot: Focusing on the Plot-Twist Advertising,” *Advertising Research*, No. 130, pp. 75-105, September 2021. https://doi.org/10.16914/ar.2021.130.75

[6] J. Xu, X. Ren, Y. Zhang, Q. Zeng, X. Cai, and X. Sun, “A Skeleton-Based Model for Promoting Coherence Among Sentences in Narrative Story Generation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, pp. 4306-4315, October-November 2018. https://doi.org/10.18653/v1/D18-1462

[7] D. Liu, J. Li, M.-H. Yu, Z. Huang, G. Liu, D. Zhao, and R. Yan, “A Character-Centric Neural Model for Automated Story Generation,” in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20),* New York: NY, pp. 1725-1732, February 2020. https://doi.org/10.1609/aaai.v34i02.5536

[8] H. Rashkin, A. Celikyilmaz, Y. Choi, and J. Gao, “PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Online, pp. 4274-4295, November

2020. https://doi.org/10.18653/v1/2020.emnlp-main.349

[9] K. Park, N. Yang, and K. Jung, "LongStory: Coherent, Complete and Length Controlled Long Story Generation," in *Proceedings of the 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2024)*, Taipei, Taiwan, pp. 184-196, May 2024. https://doi.org/10.1007/978-981-97-2253-2_15

[10] F. Brahman and S. Chaturvedi, "Modeling Protagonist Emotions for Emotion-Aware Storytelling," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Online, pp. 5277-5294, November 2020. https://doi.org/10.18653/v1/2020.emnlp-main.426

[11] L. J. Martin, P. Ammanabrolu, X. Wang, W. Hancock, S. Singh, B. Harrison, and M. O. Riedl, "Event Representations for Automated Story Generation with Deep Neural Nets," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans: LA, pp. 868-875, February 2018. https://doi.org/10.1609/aaai.v32i1.11430

[12] H. Chen, R. Shu, H. Takamura, and H. Nakayama, "GraphPlan: Story Generation by Planning with Event Graph," in *Proceedings of the 14th International Conference on Natural Language Generation (INLG 2021)*, Aberdeen, UK, pp. 377-386, August 2021. https://doi.org/10.18653/v1/2021.inlg-1.42

[13] C. Tang, Z. Zhang, T. Loakman, C. Lin, and F. Guerin, "NGEP: A Graph-Based Event Planning Framework for Story Generation," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2022)*, Online, pp. 186-193, November 2022. https://doi.org/10.18653/v1/2022.aacl-short.24

[14] P. W. Mirowski, K. W. Mathewson, J. Pittman, and R. Evans, "Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany, 355, April 2023. https://doi.org/10.1145/3544548.3581225

[15] J. U. Cho, M. S. Jeong, J. Y. Bak, and Y.-G. Cheong, "Genre-Controllable Story Generation via Supervised Contrastive Learning," in *Proceedings of the ACM Web Conference 2022 (WWW '22)*, Lyon, France, pp. 2839-2849, April 2022. https://doi.org/10.1145/3485447.3512004

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, Philadelphia: PA, pp. 311-318, July 2002. https://doi.org/10.3115/1073083.1073135

[17] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proceedings of the ACL-04 Workshop*, Barcelona, Spain, pp. 74-81, July 2004.

[18] X. Yue, B. Wang, Z. Chen, K. Zhang, Y. Su, and H. Sun, "Automatic Evaluation of Attribution by Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023),* Singapore, pp. 4615-4635, December 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.307

[19] C.-H. Chiang and H.-Y. Lee, "A Closer Look into Automatic Evaluation Using Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023),* Singapore, pp. 8928-8942, December 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.599

[20] W. Xia, S. Mao, and C. Zheng, "Empirical Study of Large Language Models as Automated Essay Scoring Tools in English Composition——Taking TOEFL Independent Writing Task for Example," arXiv:2401.03401, January 2024. https://doi.org/10.48550/arXiv.2401.03401

[21] H. Chen, D. M. Vo, H. Takamura, Y. Miyao, and H. Nakayama, "StoryER: Automatic Story Evaluation via Ranking, Rating and Reasoning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE, pp. 1739-1753, December 2022. https://doi.org/10.18653/v1/2022.emnlp-main.114

[22] J. Y. Park, G. Y. Kim, and B. C. Bae, "Generation and Analysis of Short Novels by Genre Using Language Models," in *Proceedings of the Korean Software Congress (KSC 2023)*, Busan, pp. 865-867, December 2023.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, Long Beach: CA, pp. 6000-6010, December 2017. https://doi.org/10.48550/arXiv.1706.03762

[24] M. Bae and H. Kim, "Collective Critics for Creative Story Generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami: FL, pp. 18784-18819, November 2024. https://doi.org/10.18653/v1/2024.emnlp-main.1046

[25] J. Guan, F. Huang, Z. Zhao, X. Zhu, and M. Huang, "A

Knowledge-Enhanced Pretraining Model for Commonsense Story Generation," *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 93-108, 2020. https://doi.org/10.1162/tacl_a_00302

[26] Y. Tian, T. Huang, M. Liu, D. Jiang, A. Spangher, M. Chen, ... and N. Peng, "Are Large Language Models Capable of Generating Human-Level Narratives?," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami: FL, pp. 17659-17681, November 2024. https://doi.org/10.18653/v1/2024.emnlp-main.978

[27] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical Neural Story Generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, pp. 889-898, July 2018. https://doi.org/10.18653/v1/P18-1082

[28] P. Ammanabrolu, E. Tien, W. Cheung, Z. Luo, W. Ma, L. J. Martin, and M. O. Riedl, "Story Realization: Expanding Plot Events into Sentences," in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI-20)*, New York: NY, pp. 7375-7382, February 2020. https://doi.org/10.1609/aaai.v34i05.6232

[29] C. Chhun, F. M. Suchanek, and C. Clavel, "Do Language Models Enjoy Their Own Stories? Prompting Large Language Models for Automatic Story Evaluation," *Transactions of the Association for Computational Linguistics*, Vol. 12, pp. 1122-1142, 2024. https://doi.org/10.1162/tacl_a_00689

[30] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, ... and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, New Orleans: LA, pp. 24824-24837, November-December 2022. https://doi.org/10.48550/arXiv.2201.11903

[31] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic Chain of Thought Prompting in Large Language Models," arXiv:2210.03493, October 2022. https://doi.org/10.48550/arXiv.2210.03493

[32] S. Harmon and S. Rutman, "Prompt Engineering for Narrative Choice Generation," in *Proceedings of the 16th International Conference on Interactive Digital Storytelling (ICIDS 2023)*, Kobe, Japan, pp. 208-225, November 2023. https://doi.org/10.1007/978-3-031-47655-6_13

[33] M. C. Gursesli, P. Taveekitworachai, F. Abdullah, M. F. Dewantoro, A. Lanata, A. Guazzini, ... and R. Thawonmas, "The Chronicles of ChatGPT: Generating and Evaluating Visual Novel Narratives on Climate Change Through ChatGPT," in *Proceedings of the 16th International Conference on Interactive Digital Storytelling (ICIDS 2023)*, Kobe, Japan, pp. 181-194, November 2023. https://doi.org/10.1007/978-3-031-47658-7_16

박정윤(Jeongyoon Park)

2019년～2023년: 홍익대학교 게임소프트웨어전공(학사)
2023년～현 재: 홍익대학교 게임학과 공학계열 석사과정
※관심분야 : 자연어 처리(NLP), 자연어 생성(NLG), 스토리 생성, 인터랙티브 스토리텔링

배병철(Byung-Chull Bae)

2009년 : 노스캐롤라이나주립대학교 컴퓨터학과 (공학박사)
2009년～2011년: 삼성전자 종합기술원
2011년～2012년: 코펜하겐IT대학 방문연구원
2013년～2014년: 코펜하겐IT대학 시간 강사
2014년～2015년: 성균관대학교 BK연구교수
2015년～현 재: 홍익대학교 게임학부 조교수
※관심분야 : 인터래티브 스토리텔링, 게임 인공지능, HCI