

오픈소스 대규모 언어 모델 미세 조정을 통한 한국어 대중문화 콘텐츠 플롯 생성 연구

김 지 선¹ · 김 명 준^{2*}¹이화여자대학교 융합콘텐츠학과 석박사 통합과정²이화여자대학교 융합콘텐츠학과 교수

Fine-Tuning Open-Source LLMs for Plot Generation in Korean Popular Culture Content

Ji-Seon Kim¹ · Myoung-Jun Kim^{2*}¹Master-Doctoral Combined Course, Dept of Content Convergence, Ewha Womans University, Seoul 03760, Korea²Professor, Dept of Content Convergence, Ewha Womans University, Seoul 03760, Korea

[요 약]

본 연구는 대규모 언어 모델(Large Language Models)을 활용하여 한국 대중문화 콘텐츠의 스토리 플롯을 생성하는 새로운 방법을 제안한다. Llama-3-8B, EEVE-Korean-10.8B-v1.0, EXAONE-3.0-7.8B-Instruct 세 가지 오픈 소스 언어 모델을 3,500편의 한국 대중문화 스토리 데이터셋을 사용하여 미세 조정하였다. 모델의 성능은 BLEU, Distinct, Coverage, Order 점수를 통해 평가하였고, 서사 구조 이해와 플롯 생성의 일관성이 크게 향상되었음을 확인하였다. EXAONE 모델은 일관성과 완성도 측면에서, EEVE 모델은 다양성과 창의성 측면에서 강점을 보였다. 그러나 캐릭터 일관성 유지와 주어진 결말의 충실한 반영에는 여전히 과제가 남아있다. 이 연구는 LLM이 한국 대중문화 콘텐츠의 창의적 스토리텔링 분야에서의 잠재력을 보여주며, 각 모델의 강점과 한계에 대한 통찰을 제공한다. 또한 AI 생성 플롯에서 서사 구조의 중요성을 강조하여, AI 보조 콘텐츠 창작 분야의 향후 연구 방향을 제시한다.

[Abstract]

This paper proposes a novel method for generating Korean popular culture story plots using Large Language Models (LLMs). Three open-source models—Llama-3-8B, EEVE-Korean-10.8B-v1.0, and EXAONE-3.0-7.8B-Instruct—were fine-tuned using 3,500 Korean popular culture stories. The performances of the models were evaluated using BLEU, Distinct, Coverage, and Order scores, showing significant improvements in narrative structure understanding and plot generation. The EXAONE model demonstrated strengths in consistency and completeness, while the EEVE model excelled in diversity and creativity. However, challenges remain in maintaining character consistency and faithfully reflecting given endings. This research demonstrates the potential of LLMs in creative storytelling for Korean popular culture and provides insights into the strengths and limitations of each model. It emphasizes the importance of narrative structure in AI-generated plots, suggesting directions for future research in AI-assisted content creation.

색인어 : 대규모 언어 모델, 한국어 대중문화 콘텐츠, 인공지능 스토리텔링, 스토리 플롯 생성, 서사 구조**Keyword** : Large Language Models, Korean Pop Culture Content, Artificial Intelligence Storytelling, Story Plot Generation, Narrative Structure<http://dx.doi.org/10.9728/dcs.2024.25.11.3167>

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 13 October 2024; Revised 11 November 2024

Accepted 20 November 2024

***Corresponding Author; Myoung-Jun Kim**

Tel: +82-2-3277-3347

E-mail: mjkim@ewha.ac.kr

1. 서론

1-1 연구배경 및 목적

21세기에 접어들면서 대중문화 콘텐츠 산업은 전 세계적으로 급속도로 성장하고 있다. 한편, 인공지능 기술의 발전, 특히 자연어 처리 분야에서의 대규모 언어 모델(Large Language Models, LLMs)의 등장으로 인해 텍스트 생성 및 이해 능력이 크게 향상되었다. 이에 따라 국내외에서 LLM을 활용하여 스토리를 생성하기 위한 데이터셋을 구축하거나 시나리오의 창작 및 공연예술 등 다양한 분야에 적용하는 스토리텔링 연구가 이루어지고 있다.

Kreminski 등[1]은 Datalog 쿼리를 통한 스토리 시프팅, 캐릭터 주관성을 반영한 시뮬레이션과 플레이어 의도를 반영한 행동 제안 시스템을 결합한 AI 아키텍처를 ‘WAWLT(Why Are We Like This?)’ 게임에 적용하여 AI를 활용하는 협력적 스토리텔링의 새로운 접근 방식을 제시하였다. Akoury 등[2]은 기계 참여형 인터랙티브 스토리 생성을 위한 6천 여개의 장문 이야기와 메타데이터로 구성된 STORIUM 데이터셋을 구축, GPT-2 기반 스토리 생성 모델을 개발하였고, 이를 실제 작가들이 생성된 텍스트를 편집하는 평가 플랫폼을 제안하였다. Branch 등[3]은 GPT-3 기반 내러티브 생성 시스템을 실시간 즉흥극에 적용하여 인간과 AI의 협력적 즉흥 스토리텔링 가능성을 탐구하였다. Osone 등[4]은 비지도 다중 작업 학습을 적용한 AI TRPG 기반 대화형 스토리 공동 창작 시스템 ‘BunCho’ 시스템을 개발하여 일본어 소설 창작을 위한 AI 협업 도구를 개발하고 효과를 검증하였다. Tikhonov 등[5]은 42개 언어의 위키피디아에서 ‘줄거리’를 추출하고 장르를 태깅하여 다국어 스토리 데이터셋인 StoryDB를 구축하여 다국어 교차 언어 서사 연구를 위한 리소스를 구축하였다. Lee 등[6]은 협업 작가와 4개의 GPT-3 인스턴스 간의 글쓰기 세션을 포함하는 CoAuthor 데이터셋을 구축하여 인간-AI의 협업 글쓰기의 가능성을 분석하고 텍스트 창작 분야에서 언어 모델의 생성 능력을 탐구하였다. Mirowski 등[7]은 계층적 스토리 생성 방법론을 적용하여 Dramatron이라는 스토리 생성 AI 시스템을 개발하였고, Dramatron을 활용한 전문가와의 공동 창작 세션 수행 결과를 분석하여 AI-인간 협력 창작의 가능성을 탐구하였다. Im 등[8]은 우선 스토리 배경을 설정한 뒤, 사용자와 AI간의 릴레이 형식의 스토리 작성 및 세션 인터뷰 방식의 AI 스토리 릴레이 창작 형식을 적용하여 스토리 창작 세션을 수행하였고, 이를 통해 AI 개발자와 사용자 간 협력적 디자인 픽션 작성을 수행하고 AI 설계 고려사항을 도출하고자 하였다.

국내에서도 지속적으로 인공지능을 이용한 디지털 스토리텔링 창작 관련 연구가 이루어지고 있다.

유은순, 이오준, 김진택[9]은 전산학적 스토리 자동 생성 연구의 발전 과정과 현황을 고찰하여 AI 기반 디지털 스토리

텔링의 가능성과 한계를 제시했다. 박성준[10]은 ChatGPT의 in-context learning 능력과 시 텍스트 학습을 통한 출력 정확도에 대한 분석 연구를 진행하였다. 김태형과 박성준[11]은 이를 발전시켜 ChatGPT를 이용한 이육사 관련 문학적 대화 가능성을 검증했다. 박성준[12]의 연구에서는 더 나아가 ChatGPT의 시적 수사법 이해와 재현 능력을 분석하는 연구를 진행하였다. 안지인, 이육셋별, 장예원, 정다샘[13]은 편향성이 완화된 한국어 AI 동화 생성기를 개발하여 아동용 디지털 스토리텔링 플랫폼의 기반을 마련했다. 김용연, 남양희[14]는 챗GPT와 QuickDraw 데이터셋을 활용한 대화형 그림동화 재창작 서비스를 개발하여 독서 후 활동을 위한 새로운 디지털 스토리텔링 방식을 제안하였다. 권세미, 김다예, 정다샘[15]은 GPT-3를 이용한 회극 쓰기와 XR 기술을 결합하여 가상 인간과의 연기를 구현함으로써 새로운 형태의 디지털 스토리텔링 공연 콘텐츠 제작을 시도했다. 박정윤, 신주민, 김가연, 남지혜, 배병철[16]은 ChatGPT 3.5 API를 활용한 인터랙티브 스토리 저작 도구 프로토타입을 설계하여 AI 기반 디지털 스토리텔링 도구 개발을 시도하였다. 김지선과 김병준[17]은 영화, 드라마, 웹소설, 웹툰/만화의 스토리를 분석하고 작품 및 유닛 단위의 스토리 데이터와 메타 데이터를 구축함으로써 대중문화 스토리 콘텐츠의 창작 요소 분포를 정량적으로 분석하고 AI 스토리텔링 연구를 위한 기초 자료를 제공하였다.

이러한 연구들은 인공지능이 인간의 창작 활동을 보조하거나, 나아가 독자적인 창작물을 생성할 수 있는 가능성을 보여준다. 그러나 대부분의 연구가 영어를 대상으로 이루어지고 있어 상대적으로 한국어 대중문화 콘텐츠를 다루는 연구는 미흡한 실정이다. 한국어는 영어와 문법 구조 및 어순 등에서 큰 차이를 보이기 때문에, 영어 기반의 LLM을 그대로 적용하기에는 어려움이 있다. 따라서 LLM을 활용하여 창작 스토리텔링에 한국어의 특성을 반영한 LLM 연구가 필요하다.

이에 본 연구는 구축된 한국어 대중문화 콘텐츠 데이터셋을 활용하여 LLM 기반 스토리 플롯 생성 방법을 제안하고자 한다. 본 연구의 목적은 다음과 같다. 첫째, 한국어로 구성된 대중문화 콘텐츠 스토리 데이터를 활용하여 오픈소스 LLM을 fine-tuning하고, 한국어 스토리 생성 모델을 개발한다. 해당 데이터셋은 드라마, 영화, 웹툰, 웹소설 등 다양한 장르의 콘텐츠를 포함하며, 본 연구에서는 이를 재구성하여 학습에 활용한다. 둘째, fine-tuning된 LLM의 성능 평가를 위한 자동 평가 지표를 개발한다. 스토리 생성 태스크의 경우, 생성된 결과물의 품질을 정량적으로 측정하기 어려운 측면이 있다. 이에 본 연구에서는 선행 연구를 참고하여 한국어 스토리 생성 모델에 적합한 자동 평가 방법을 제안하고, Coverage와 Order 등 스토리 구조 평가에 특화된 지표를 확장하여 한국어 콘텐츠의 특성을 고려한 다면적 평가 체계를 구축한다. 이를 통해 모델의 성능을 객관적으로 비교·분석한다. 셋째, 제안된 방법론을 바탕으로 실제 한국어 대중문화 콘텐츠 스토리 플롯을 생성하고, 그 결과를 분석하여 제안 방법론의 실효성

을 검증한다.

본 연구를 통해 인공지능 기술과 한국 대중문화 콘텐츠의 융합 가능성을 제시하고, 창의적인 스토리텔링을 위한 새로운 패러다임을 마련하는 데 기여할 것으로 기대된다. 나아가 본 연구의 결과는 실제 대중문화 콘텐츠 제작 과정에서 창의적인 스토리텔링을 지원하는 도구로 활용될 수 있을 것이다.

1-2 연구범위 및 방법

본 연구는 LLM을 활용하여 한국어 대중문화 콘텐츠의 스토리 플롯을 생성하는 방법을 제안하는 것을 목표로 한다. 연구 대상은 드라마, 영화, 웹툰, 웹소설 등 다양한 장르의 스토리 데이터이며, 연구에 사용된 데이터셋[17]의 범위에 따라 2010년부터 2022년 사이에 제작된 작품들을 대상으로 한다. 학습된 모델은 작품의 전체 스토리를 요약한 개요 또는 시놉시스 형태의 텍스트 데이터를 생성하는 것을 목표로 하며, 세부적인 장면 묘사나 대사 등은 연구 범위에서 제외한다. LLM은 fine-tuning이 가능한 최신 오픈 소스 모델 중 한국어 처리에 적합한 모델을 선정하되, 모델의 크기 및 아키텍처는 연구 목적 및 자원을 고려하여 결정한다.

연구 방법은 다음과 같다. 첫째, 기존에 구축된 한국어 대중문화 콘텐츠 데이터셋을 활용하여, LLM 학습에 적합한 형태로 데이터를 재구성한다. 재구성된 데이터를 정제하고, LLM의 입력 형식에 맞게 전처리한 후, 학습, 검증, 테스트 세트로 분할한다. 둘째, 한국어 처리 성능이 검증된 LLM을 선정하고, 재구성된 데이터를 사용하여 LLM을 fine-tuning한다. ‘장르’, ‘주제’, ‘모티프’, ‘로그라인’ 등의 메타데이터를 활용하여 서사 단계가 명시된 플롯 형태의 스토리 텍스트 데이터를 생성하는 것을 목표로 하며, 학습 과정에서 다양한 하이퍼 파라미터를 실험하여 최적의 모델을 도출한다. 셋째, 선행 연구를 참고하여 한국어 스토리 생성 모델 평가에 적합한 자동 평가 지표를 설계하고, BLEU, Distinct, Coverage, Order 등의 다면적 평가 지표를 통해 fine-tuning된 LLM의 성능을 측정한다. 이를 통해 제안하는 방법론의 유효성을 검증하고, LLM을 활용한 한국어 대중문화 콘텐츠 창작 연구의 가능성과 한계점을 논의한다. 넷째, Fine-tuning된 LLM을 사용하여 새로운 한국어 대중문화 콘텐츠 스토리 플롯을 생성하고, 제안된 평가 지표를 통해 생성된 플롯의 품질을 객관적으로 분석한다.

본 연구는 LLM 기반 한국어 대중문화 콘텐츠 스토리 플롯 생성 기술을 개발하고, 그 효용성을 검증함으로써 인공지능 기술과 대중문화 콘텐츠 산업의 융합 가능성을 타진하고, 새로운 연구 주제를 발굴하는 데 기여하고자 한다. 나아가 본 연구의 결과는 실제 대중문화 콘텐츠 제작 과정에서 창의적인 스토리텔링을 지원하는 도구로 활용될 수 있을 것이다.

II. 관련 연구

2-1 제어 가능한 스토리 생성

스토리 생성 분야에서 제어 가능한 생성 방법에 대한 연구는 지속적으로 발전해왔다. 초기에는 Li 등[18]이 계획 기반 접근법을 제안하여 스토리의 전체적인 구조를 먼저 생성한 후 세부 내용을 채우는 방식을 도입했다. 이후 Fan 등[19]은 계층적 생성(hierarchical generation) 방식을 제안하여 스토리의 개요를 먼저 생성한 후 전체 텍스트를 생성하는 방법을 선보였다.

2020년에 들어서면서 더욱 다양한 제어 방식이 연구되기 시작하였다. 키워드 기반 제어 방식도 주목받았는데, Yao 등[20]의 Plan-and-Write 모델은 키워드 시퀀스를 먼저 생성한 후 이를 바탕으로 스토리를 작성하는 접근법을 제시하였다. Brahman과 Chaturvedi[21]는 주인공의 감정 궤적을 제어하여 스토리를 생성하는 방법을 연구하였다. 비슷한 시기에 Tambwekar 등[22]은 강화학습을 활용하여 목표 지향적인 스토리 플롯을 생성하는 방법을 개발했다. 스토리의 일관성과 논리성 향상을 위해 Guan 등[23]은 상식 지식을 사전 학습된 언어 모델에 주입하는 방법을 제안했다. 이러한 방식은 스토리의 품질을 크게 향상시켰다.

최근에는 더욱 정교한 제어 방식들이 등장했다. Rashkin 등[24]의 PLOTMACHINES는 개요를 입력으로 받아 동적 플롯 상태 추적을 통해 일관성 있는 스토리를 생성하는 프레임워크를 제안했다. Xu 등[25]의 MEGATRON-CNTRL 모델은 외부 지식을 활용하여 키워드, 스타일 등 다양한 속성으로 스토리 생성을 제어하는 방법을 활용했다. 가장 최근에는 Cho 등[26]이 장르를 제어 코드로 사용하여 특정 장르의 스토리를 생성하는 SCSC 모델을 제안했다. 이 모델은 지도 대조 학습을 통해 장르별 표현을 학습하여 효과적으로 장르에 기반한 스토리를 생성하고자 했다.

이러한 연구들은 키워드, 개요, 감정, 스타일 등 스토리의 다양한 측면을 제어하면서 일관성 있고 맥락에 맞는 스토리를 생성하는 것을 목표로 하였다. 스토리 생성의 제어 가능성, 일관성, 창의성 등 다양한 측면을 지속적으로 개선해왔다.

2-2 스토리 생성 모델의 정량 평가

스토리 생성 모델의 평가 방법은 시간이 지남에 따라 점점 더 정교해지고 다양화되었다. 초기에는 주로 BLEU, ROUGE와 같은 n-gram 중첩 기반 지표가 사용되었으나, 이러한 방식은 창의적인 텍스트 생성 평가에는 적합하지 않다는 한계가 지적되어 왔다.

Yao 등[20]이 제안한 Plan-and-write 모델의 평가에서는 BLEU-2 스코어를 사용하여 생성된 스토리의 품질을 측정하였다. 또한 생성된 텍스트의 다양성을 평가하기 위해

Distinct-4 지표를 도입하였다. 이는 생성된 텍스트에서 고유한 4-gram의 비율을 계산하는 방식으로, 높을수록 다양한 표현을 사용했다고 판단한다.

PLOTMACHINES에서는 더욱 다양한 정량적 평가 방법을 사용하였다[24]. 여기에는 Perplexity, ROUGE 스코어, 그리고 생성된 텍스트의 반복성을 측정하는 Repeat 지표가 포함된다. Repeat은 생성된 스토리에서 최소 한 번 이상 반복되는 4-gram의 비율을 계산한다. 또한 Self-BLEU 스코어를 사용하여 생성된 스토리들 간의 다양성을 평가하였다.

PLOTMACHINES에서는 Coverage 지표를 도입했다. 이 Coverage는 생성된 스토리가 주어진 개요의 요소들을 얼마나 잘 포함하고 있는지를 측정한다. PLOTMACHINES에서의 Coverage는 각 개요 문구와 생성된 텍스트 사이의 임베딩 기반 유사도 점수의 평균으로 측정된다. 한편, Guan 등[27]이 제안한 LOT (Long Outline-controlled Text generation) 벤치마크에서는 Coverage와 Order라는 새로운 평가 지표를 제시하였다. LOT의 Coverage는 PLOTMACHINES의 Coverage와 유사한 개념이지만, 계산 방식에서 차이를 보인다. LOT의 Coverage는 각 개요 문구와 생성된 텍스트 사이의 ROUGE-L 점수의 평균으로 측정된다.

LOT에서 새롭게 제안된 Order 지표는 생성된 스토리에서 개요의 요소들이 등장하는 순서가 얼마나 적절한지를 중심으로 평가된다. 이는 생성된 스토리에서 개요 요소들의 위치 순서와 원래 개요에서의 순서를 비교한다. Order 지표는 PLOTMACHINES에는 없던 새로운 평가 지표로, 스토리의 구조적 일관성을 평가한다.

EQ-Bench는 대규모 언어 모델의 감정 지능을 평가하는 벤치마크이다[28]. 이 벤치마크는 대화 속 등장인물의 감정 상태 강도를 예측하는 과제를 통해 모델의 성능을 평가한다. EQ-Bench는 생성된 텍스트의 품질을 직접적으로 평가하지는 않지만, 스토리 내의 감정적 맥락을 이해하고 생성하는 능력을 측정한다는 점에서 스토리 생성 모델의 평가에도 적용될 수 있는 가능성을 보여준다.

이러한 다양한 정량적 평가 지표는 각각 스토리 생성 모델의 다른 측면을 평가한다. BLEU와 ROUGE는 생성된 텍스트의 품질을, Distinct와 Self-BLEU는 다양성을, Coverage는 개요 반영 정도를, Order는 구조적 일관성을 평가한다. 현재는 이러한 다양한 평가 지표를 종합적으로 활용하여 생성 모델의 성능을 평가하는 것이 일반적이다. 그러나 여전히 텍스트 생성 모델의 평가는 중요한 도전 과제로 남아있으며, 특히 스토리를 평가하기 위한 정량적 평가 방법에 대한 연구가 필요하다.

III. 연구방법론

3-1 데이터셋 구성 및 전처리

대중문화 콘텐츠의 내러티브 생성을 위해서는 모델이 영화나 드라마의 구조적 특성을 충분히 학습할 수 있는 데이터가 필요하다. 본 실험에서는 한국인공지능정보사회진흥원이 공개한 대중문화 콘텐츠 스토리 데이터[17]를 목적에 맞게 재구성하여 대중문화 콘텐츠의 장르, 주제, 모티프, 로그라인 등의 메타정보와 함께, 실제 줄거리를 학습 데이터로 활용하였다.

특히 줄거리의 경우, 단순히 전체 내용을 나열하는 것이 아니라 ‘Opening Salvo’, ‘Main Character’, ‘Setting-up’ 등 미리 정의된 서사 단계에 따라 구분하여 명시적으로 재구성하였다. 미리 정의된 서사 단계는 ‘Storyhelper 15단계’와 ‘영웅의 12단계’의 2 종류로 구분된다. 줄거리에 서사 단계를 명시하는 것은 모델로 하여금 내러티브의 흐름과 구조를 보다 명확하게 학습할 수 있게 해준다. 학습 데이터의 구조는 표 1과 같다.

대중문화 콘텐츠의 플롯 창작을 위한 fine-tuning을 위한 데이터는 인공지능 학습을 위한 대중문화 콘텐츠 스토리데이터를 활용하였다. Fine-tuning을 위한 학습 데이터는 스토리데이터의 장르, 주제, 모티프와 함께 스토리데이터의 기본설정과 갈등구조를 조합하여 로그라인으로 설정하였다. 인공지능 시스템이 주어진 조건에서 플롯을 짜도록 시스템 메시지를 설정하였다. 출력 데이터는 스토리데이터 각 유닛의 서사 단계와 스토리라인을 결합하여 생성하도록 구조화하였다. 이때, 출력 플롯은 같은 서사 단계에 해당하는 유닛들의 스토리라인을 하나로 결합하지 않고, 유닛 단위로 서사 단계와 스토리라인을 결합하여 나열하였다. 즉, 출력 플롯은 스토리데이터의 유닛 개수만큼의 “서사 단계: 스토리라인” 쌍으로 이루어지도록 설정하였다. 이는 각 서사단계에 해당하는 유닛의 개수를 명시적으로 드러냄으로써 동일한 인물이 생성된 한 플롯 스토리에서 각 서사단계가 차지하는 상대적 길이를 확인하고자 하는 목적을 두었다. 각 스토리데이터의 전체 줄거리를 유닛 별 서사 단계와 함께 제공함으로써, 모델은 스토리의 구조와 흐름을 학습할 수 있게 된다.

실험에는 총 3500편의 대중문화 콘텐츠 스토리를 활용하였으며, 이 중 약 3000편을 학습에, 400편을 평가에 사용하였다.

3-2 스토리 생성 모델 구조 및 베이스 모델 선정

본 연구에서 제안하는 스토리 플롯 생성 모델의 구조는 그림 1과 같다. 장르, 주제, 모티프, 로그라인을 입력 프롬프트로 제시하면 서사단계가 명시된 구조화된 플롯을 생성하는 모델이다. 스토리데이터를 활용하여 표 1과 같은 형식의 학습

데이터를 생성하고, 이를 학습하여 베이스 LLM 모델을 fine-tuning 한다.

베이스 LLM 모델 선정에 있어 다음과 같은 기준을 적용하였다. 첫째, 한국어 처리 능력이 검증된 모델을 우선적으로 고려하였다. 둘째, 연구의 재현성과 확장성 확보를 위해 오픈소스 모델을 선택하였다. 셋째, 연구 자원의 효율적 활용과 실용적 구현 가능성을 고려하여 8B 파라미터 규모의 모델들을 주요 대상으로 삼았다.

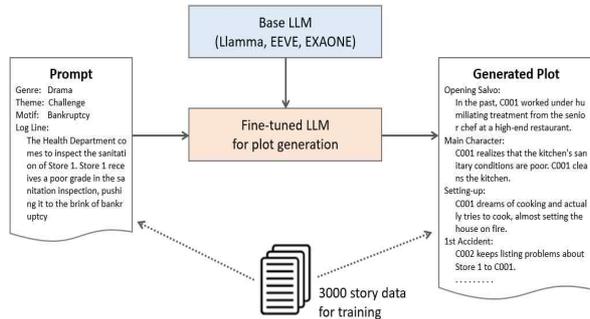


그림 1. 한국어 대중문화 콘텐츠 스토리 플롯 생성을 위한 LLM fine-tuning
 Fig. 1. LLM fine-tuning for Korean pop culture content story plot generation

이러한 기준을 바탕으로 본 연구에서는 Llama-3-8B, EEVE-Korean-10.8B-v1.0, EXAONE-3.0-7.8B-Instruct 세 가지 모델을 최종 선정하였다. Llama-3-8B는 최신 아키텍처를 적용하여 다국어 처리 능력이 향상되었으며[29], EEVE-Korean-10.8B-v1.0는 한국어 토큰을 추가하고 재학습하여 영어 모델의 성능을 유지하면서 한국어 처리 능력을 효과적으로 향상시켰다[30]. EXAONE-3.0-7.8B-Instruct는 효율적인 토큰라이저 설계와 다단계 학습 방법을 통해 영어와 한국어 모두에서 우수한 성능을 보인다[31].

Llama-3-8B는 범용성이 높고 다양한 태스크에 적용 가능하며, EEVE-Korean-10.8B-v1.0는 적은 양의 학습 데이터로도 높은 성능을 달성할 수 있는 장점이 있다. EXAONE-3.0-7.8B-Instruct는 이중언어 지원과 효율적인 토큰라이저를 통해 한국어 처리에 특화된 성능을 보인다. 이러한 특성들을 고려하여, 각 모델을 베이스 모델로 선정하고 fine-tuning을 거쳐 한국어 대중문화 콘텐츠 스토리 플롯 생성이라는 특수한 태스크에 대한 능력을 비교 분석하고자 하였다.

3-3 Fine-tuning 파라미터 선정

본 연구에서는 Llama-3-8B, EEVE-Korean-10.8B-v1.0, EXAONE-3.0-7.8B-Instruct 모델을 대상으로 한국어 대중문화 콘텐츠 스토리 플롯 생성을 위한 fine-tuning을 수행하였다. 학습 과정에서 모델은 주어진 장르, 주제, 모티프, 로그라인을

표 1. Fine-tuning 학습데이터 예시

Table 1. Fine-tuning training data example

Plot Generation Fine-tuning Training Data Sample
<pre> { "messages": [{ "role": "system", "content": "You are an AI assistant. Develop a plot from the given conditions." }, { "role": "user", "content": "Genre: Drama, Theme: Challenge Motif: Bankruptcy Log Line: The Health Department comes to inspect the sanitation of Store 1. Store 1 receives a poor grade in the sanitation inspection, pushing it to the brink of bankruptcy." }, { "role": "assistant", "content": "Opening Salvo: In the past, C001 worked under humiliating treatment from the senior chef at a high-end restaurant. Main Character: C001 realizes that the kitchen's sanitary conditions are poor. C001 cleans the kitchen. Setting-up: C001 dreams of cooking and actually tries to cook, almost setting the house on fire. 1st Accident: C002 keeps listing problems about Store 1 to C001. ... 2nd Accident: C006 comforts his daughter over the phone. C002 discovers an unexpected side of C006 in how he treats his daughter. Choice to Fight: C001 confides in his younger sibling C015 about his unusual symptoms. Resurrection: C006 finds C007's letter envelope in a corner of the auxiliary kitchen. Resurrection: C001 formally hires C002. Final Salvo: C001 realizes that the cigarettes that caused problems during the sanitation inspection were actually his. ... }] } </pre>

바탕으로 서사 단계에 따라 구조화된 완결된 플롯을 생성하도록 설계되었다. Fine-tuning은 3대의 RTX 4090 GPU를 활용하여 진행되었으며, 기본적으로 3 epoch의 학습을 수행하였다.

3-4 평가 방법 및 지표

본 연구에서는 AI 모델의 텍스트 스토리 생성 능력을 종합적으로 평가하기 위해 다양한 정량적 지표를 활용했다. 주요 평가 지표로는 PLOTMACHINES와 LOT (Long Outline-controlled Text generation) 벤치마크에서 제안된 방법론을 참고하여

BLEU 점수, Distinct 점수, Coverage 점수, 그리고 Order 점수를 사용했다.

BLEU 점수는 생성된 텍스트와 참조 텍스트 간의 표면적 유사성을 평가하는 데 사용되었다. BLEU-1부터 BLEU-4까지의 점수를 계산하여 단어 및 구문 수준에서의 일치도를 측정했다. 이를 통해 생성된 스토리가 참조 스토리와 얼마나 유사한 표현을 사용하는지 평가할 수 있었다.

Distinct 점수는 생성된 스토리의 다양성과 창의성을 평가하기 위해 도입되었다. Distinct-1부터 Distinct-4까지의 점수를 계산하여 텍스트 내 고유한 n-gram의 비율을 측정했다. 이 지표를 통해 모델이 얼마나 다양한 어휘와 표현을 사용하여 스토리를 생성하는지 분석할 수 있었다.

본 연구에서는 생성된 스토리가 입력 프롬프트의 요구사항을 얼마나 충실히 반영하는지 평가하기 위해 Coverage 지표를 도입하였다. Coverage 지표는 프롬프트에서 추출한 핵심 키워드들이 생성된 텍스트에 포함되어 있는지 여부를 기반으로 계산된다. Coverage 지표의 계산은 우선 입력 프롬프트에서 키워드를 추출하는 것으로 시작된다. 키워드는 장르(Genre), 주제(Theme), 모티프(Motif)와 같은 스토리 설정 요소뿐만 아니라, Log Line에 포함된 주요 서술어와 명사구도 포함된다. 예를 들어, 표 1의 프롬프트에서는 ‘drama’, ‘challenge’, ‘bankruptcy’, ‘health’, ‘department’, ‘inspect’, ‘sanitation’, ‘store’, ‘grade’, ‘poor’ 등이 핵심 키워드 집합으로 추출된다. 추출된 키워드는 생성된 텍스트와 대조되어 각 키워드의 출현 여부가 확인된다. 이때 특정 키워드의 중복 출현은 고려하지 않으며, 단 한 번의 출현만으로도 해당 키워드가 반영된 것으로 간주한다. Coverage 점수는 전체 키워드 중 생성된 텍스트에 포함된 키워드의 비율로 계산된다. 이는 (생성된 텍스트에 포함된 키워드 수) / (전체 키워드 수)로 표현될 수 있다. 이러한 Coverage 지표를 통해 생성 모델이 입력 프롬프트의 요구사항을 얼마나 포괄적으로 반영하는지를 정량적으로 평가할 수 있다. 특히 Log Line에서 추출된 키워드들의 포함 여부를 확인함으로써, 생성된 스토리가 의도된 플롯 요소들을 얼마나 충실히 반영하는지도 평가할 수 있다. 이는 생성된 스토리의 프롬프트 충실도(prompt fidelity)를 측정하는 객관적인 기준으로 활용될 수 있다.

또한 생성된 스토리의 서사 구조를 정량적으로 평가하기 위해 Order 점수를 도입하였다. 이는 LOT[27] 벤치마크의 순서 평가 방법론을 기반으로 하되, 본 연구의 계층적 서사 구조 특성을 반영하여 수정 및 확장하였다. Order 점수는 서사 단계(narrative stages)의 순서와 플롯 내용(plot contents)의 순서를 각각 독립적으로 평가한 후 가중 평균을 계산한다. 구체적으로, 서사 단계 순서에 60%, 플롯 내용 순서에 40%의 가중치를 부여한다. 예를 들어, “Opening Salvo”, “Main Character”, “Setting-up” 등과 같은 서사 단계의 순서가 참조 데이터와 일치하는지를 우선적으로 평가하고, 각 단계 내의 구체적인 플롯 내용의 순서도 함께 고려

한다. 순서 평가는 역전(inversion) 횟수를 기반으로 이루어진다. 먼저 생성된 텍스트와 참조 텍스트에서 공통으로 등장하는 요소들을 추출한다. 예를 들어, 표 1의 프롬프트에 대해 생성된 스토리에서 “Opening Salvo”, “Main Character”, “Setting-up” 등의 서사 단계가 추출되면, 이들의 순서를 참조 데이터의 순서와 비교한다. 순서가 뒤바뀐 쌍의 수(역전 수)를 계산하고, 이를 가능한 최대 역전 수로 정규화하여 점수를 산출한다. 이때 공통 요소가 전혀 없는 경우에는 0점을 부여하여, 서사 구조의 일관성이 완전히 결여된 경우를 명확히 구분한다. 구체적인 계산 예시를 들면, 참조 데이터에서 “Opening Salvo → Main Character → Setting-up” 순서로 나타난 서사 단계가 생성된 텍스트에서 “Main Character → Opening Salvo → Setting-up” 순서로 나타났다면, 한 쌍의 역전(Opening Salvo와 Main Character)이 발생한 것으로 계산된다. 이러한 방식으로 서사 단계의 순서 점수와 플롯 내용의 순서 점수를 각각 계산한 후, 설정된 가중치(0.6, 0.4)를 적용하여 최종 Order 점수를 산출한다. 이러한 평가 방식은 단순히 텍스트의 표면적 유사성을 넘어, 스토리의 구조적 완성도와 서사적 일관성을 정량적으로 측정할 수 있게 한다. 특히 서사 단계에 더 높은 가중치를 부여함으로써, 스토리텔링의 전체적인 흐름과 구조를 중요하게 고려하면서도 세부적인 플롯 전개 순서도 평가에 반영할 수 있다. 또한 본 연구에서는 생성된 텍스트와 참조 텍스트 간에 공통 요소가 전혀 없는 경우 0점을 부여하는 방식을 새롭게 도입하였다. 이는 서사 구조가 완전히 일치하지 않는 경우를 엄격하게 평가함으로써, 스토리의 구조적 일관성과 논리적 흐름을 더욱 정확하게 측정할 수 있게 한다.

이러한 다면적 평가 방법을 통해 AI 모델의 텍스트 스토리 생성 능력을 종합적으로 분석할 수 있었다. 각 지표는 스토리의 서로 다른 측면 - 표현의 정확성, 다양성, 주제 반영도, 구조적 일관성 - 을 평가함으로써, 생성된 스토리의 품질을 다각도로 측정할 수 있었다. 이 방법론은 AI 모델의 스토리텔링 능력을 객관적으로 평가하고, 향후 개선이 필요한 영역을 식별하는 데 유용한 틀을 제공한다.

IV. 결 과

본 연구에서는 Llama-3-8B, EEVE-Korean-10.8B-v1.0, EXAONE-3.0-7.8B-Instruct 모델을 대상으로 한국어 대중문화 콘텐츠 스토리 플롯 생성을 위한 fine-tuning을 수행하고 그 결과를 비교 분석하였다. 평가는 BLEU, Distinct, Coverage, 그리고 Order 점수를 통한 정량 평가의 결과는 표 2와 같다. 생성된 플롯의 질적 특성 또한 심층적으로 분석하였다.

실험 결과, fine-tuning을 거친 모든 모델들이 그럴듯한 플롯을 생성해내는 것을 확인할 수 있었다. 특히, 생성된 플롯은 학습 데이터인 스토리데이터와 유사하게 "스토리 유닛의 서사 단계: 해당 유닛의 주요 스토리라인" 형태를 유지하였으

표 2. 스토리 생성 모델의 성능 비교 평가
Table 2. Comparative performance evaluation of story generation models

Model	LLaMa-3-8B (baseline)	Fine-tuned -LLaMa-3-8B	Fine-tuned -EEVE-Korean-10.8B-v1.0	Fine-tuned -EXAONE-3.0-7.8B-Instruct
BLUE-1	.0675	.3521	.3723	.4247
BLUE-2	.0405	.2662	.2837	.3574
BLUE-3	.0128	.2083	.2232	.3093
BLUE-4	.0021	.1655	.1787	.2687
Distinct-1	.4171	.2835	.2867	.2125
Distinct-2	.7453	.4628	.4526	.3586
Distinct-3	.8548	.5857	.5687	.4508
Distinct-4	.8977	.6762	.6561	.5297
Coverage	.0649	.2757	.2729	.2682
Order	.4000	.9885	.9865	.9889

며, 스토리헬퍼 15단계의 순서를 대체로 준수하였다.

주목할 만한 특징으로, 모든 모델이 학습 데이터의 특성을 반영하여 인물을 'C000'과 'E000' 형태로 비식별화하여 생성하였다. 이러한 비식별화된 인물 표현은 생성된 플롯 내에서 일관성 있게 유지되었으나, 동일 인물의 특성이나 행위가 플롯 전후에서 불일치하는 경우도 관찰되었다. 그러나 학습 데이터가 최종적으로 완성되는 서사 텍스트가 아니라 아이디어와 구성의 단계에 있는 단순한 문장 형태의 플롯이라는 점을 고려하였을 때 복합적으로 판단해야 할 요인이 될 수 있다. 하나의 콘텐츠 내에서 인물의 전후관계가 자연스럽게 이어지지 않는다는 것은 스토리가 일관성 없이 이랬다저랬다 하는 것으로 볼 수도 있으나 대중서사 콘텐츠의 스토리를 풍성하게 만드는 반전 요소로 발전시킬 수도 있기 때문이다.

정량적 평가 결과, fine-tuning을 거친 모델들은 baseline 모델인 Llama-3-8B 대비 BLEU, Coverage, Order 점수에서 현저한 향상을 보였다. 특히 EXAONE-3.0-7.8B-Instruct 튜닝 모델이 BLEU 점수와 Order 점수에서 최고 성능을 기록했으며, EEVE-Korean-10.8B-v1.0 튜닝 모델은 Distinct 점수에서 우수한 성능을 나타냈다.

생성된 플롯과 원본 스토리데이터 간의 텍스트 유사도는 전반적으로 낮게 나타났으며, 로그라인에서 제시된 결말을 충실히 반영하지 못하는 경향이 관찰되었다. 이는 모델이 학습 데이터를 단순 복제하는 것이 아니라 새로운 내용을 생성하고 있음을 시사한다.

모델 별 특징을 살펴보면, 추가적인 한국어 사전 학습을 거치지 않은 Llama-3-8B 모델의 경우 학습 데이터가 한국어로 이루어졌음에도 불구하고 거의 한국어로 응답하지 않고 영어로 응답하였다. 그러나 fine-tuning을 거친 Llama-3-8B 모델은 한국어로 플롯을 생성하였다.

학습 데이터에서 서사 단계를 제외하고 플롯을 생성하도록 학습시킨 경우, 생성된 플롯은 서사 단계를 구분하여 학습한

모델에 비해 입력 조건과의 관련성, 플롯 내 일관성, 창의성 측면에서 상대적으로 낮은 성능을 보였다. 이는 서사 단계가 플롯의 구조화와 일관성 유지에 중요한 역할을 함을 시사한다.

종합적으로, 본 연구 결과는 서사단계로 구조화된 한국어 스토리데이터에 대한 fine-tuning을 통해 LLM의 한국어 대중문화 콘텐츠 스토리 플롯 생성 능력이 크게 향상될 수 있음을 보여준다. 특히 EXAONE 튜닝 모델은 일관성과 완성도 측면에서, EEVE-Korean-10.8B-v1.0 튜닝 모델은 다양성과 창의성 측면에서 강점을 보였다. 그러나 여전히 인물 특성의 일관성 유지, 결말 반영 등의 측면에서 개선의 여지가 있으며, 이는 향후 연구에서 중점적으로 다뤄야 할 과제로 판단된다. 결론적으로, 본 연구의 분석 결과는 각 모델의 강점과 약점을 명확히 보여준다. EXAONE 튜닝 모델은 일관성 있고 완성도 높은 스토리 생성에 탁월했으며, EEVE 튜닝 모델은 창의적이고 다양한 표현을 사용하는 데 강점을 보였다. LLaMA 계열 모델들은 전반적으로 균형 잡힌 성능을 나타냈으나, 한국어 표현의 자연스러움에서는 다소 부족함을 보였다.

4-1 모델 별 Fine-Tuning 결과 비교

본 연구에서 수행한 fine-tuning 결과, 모든 모델에서 baseline 모델인 LLaMA 3 8B 대비 전반적인 성능 향상이 관찰되었다. Fine-tuning 모델 간의 성능 평가 결과는 표 2와 같다.

모든 fine-tuned 모델들이 baseline 대비 현저히 향상된 BLEU 점수를 기록했다. EXAONE 튜닝 모델이 가장 높은 BLEU를 달성하여 참조 텍스트와의 유사성이 가장 높음을 나타냈다. 그러나 전반적으로 생성된 플롯과 원본 스토리데이터 간의 텍스트 유사도는 낮게 나타났는데, 이는 모델이 학습 데이터를 단순 복제하지 않고 새로운 내용을 생성하고 있음을 시사한다.

흥미롭게도 전혀 추가 학습을 하지 않은 baseline Llama-3-8B 모델이 가장 높은 Distinct를 기록했다. Fine-tuned 모델들 중에서는 Distinct-1을 제외하고는 Llama-3-8B 튜닝 모델이 가장 높은 Distinct 점수를 보이며, 다양한 한국어 표현 생성 능력은 가장 좋은 것으로 나타났다. EEVE 모델이 Llama-2를 기반으로 추가 학습을 한 모델이라는 점을 고려할 때, 이는 추가 학습 이전의 원본 모델의 성능에 Distinct 점수가 영향을 받는 것이라고 볼 수 있다.

Baseline 모델에 비해 모든 fine-tuned 모델들이 현저히 높은 Coverage 점수를 보였다. 이는 fine-tuning 과정이 모델의 주제 반영 능력을 크게 향상시켰음을 나타낸다. 그러나 로그라인에서 제시된 결말을 충실히 반영하지 못하는 경향이 관찰되어, 이 부분에 대한 개선이 필요할 것으로 보인다.

모든 fine-tuned 모델들이 매우 높은 Order 점수를 기록했다. EXAONE 튜닝 모델이 가장 높은 Order 점수를 보였다. 이는 fine-tuning 과정이 모델의 서사 구조 이해 및 생성 능력을 크게 향상시켰음을 시사한다. 생성된 플롯이 스토리헬

퍼 15단계의 순서를 대체로 준수한 것도 이를 뒷받침한다.

EXAONE 튜닝 모델의 경우 BLEU 점수와 Order 점수에서 최고 성능을 기록했다. 이는 모델이 일관성 있고 구조화된 플롯 생성에 탁월함을 보여준다.

결론적으로, fine-tuning 과정은 모든 모델의 한국어 스토리 플롯 생성 능력을 현저히 향상시켰다. 특히 주어진 조건에 부합하는 일관성 있는 스토리 구조 생성에 효과적이었음을 Order 점수의 큰 폭 상승을 통해 확인할 수 있었다. 그러나 Distinct 점수의 감소는 fine-tuning 과정에서 모델의 창의성과 다양성이 일부 제한되었을 가능성을 시사하며, 이는 향후 연구에서 개선이 필요한 부분으로 판단된다. 또한, 로그라인에서 제시된 결말을 충실히 반영하지 못하는 문제와 인물 특성의 일관성 유지 문제도 추가적인 개선이 필요한 영역으로 확인되었다.

4-2 생성된 스토리 플롯 분석

생성된 플롯은 스토리헵퍼 15단계의 순서를 대체로 준수하였으며, 서사 단계와 스토리라인이 적절히 대응되었다. Fine-tuned 모델들은 baseline에 비해 더 복잡하고 입체적인 캐릭터를 설정하는 경향을 보였다. EEVE 튜닝 모델이 캐릭터의 내적 갈등과 성장을 가장 잘 표현했다. 모든 fine-tuned 모델에서 한국적 요소와 대중문화 콘텐츠의 특성이 더 잘 반영되었다. EXAONE 튜닝 모델이 가장 자연스럽게 한국적 맥락을 스토리에 통합했다. EEVE 튜닝 모델이 가장 창의적이고 예측하기 어려운 전개를 보여주었다. 반면, EXAONE 튜닝 모델은 상대적으로 안정적이고 예측 가능한 스토리를 생성했다. EXAONE 튜닝 모델이 가장 높은 일관성과 완성도를 보여주었다. 그러나 모든 모델에서 동일 인물의 특성이나 행위가 플롯 전후에서 불일치하는 경우가 관찰되었다. Fine-tuned 모델들은 대체로 입력된 장르, 주제, 모티프를 잘 반영하였다. 그러나 로그라인에서 제시된 결말을 충실히 반영하지 못하는 경향이 관찰되기도 하였다. EEVE와 EXAONE 튜닝 모델은 LLaMA 계열 모델들에 비해 더 자연스러운 한국어 표현을 생성하였다. 존재하지 않는 단어나 비문, 어색한 한국어 문장의 빈도가 현저히 낮았다.

V. 결론 및 토의

본 연구는 LLM fine-tuning을 활용한 한국어 대중문화 콘텐츠 스토리 플롯 생성 방법을 제안하고, 그 효과성을 실증적으로 검증하였다. 연구 결과, fine-tuning을 거친 모델들이 baseline 모델 대비 현저한 성능 향상을 보였으며, 특히 서사 구조의 이해와 일관성 있는 플롯 생성 능력이 크게 개선되었음을 확인하였다.

본 연구를 통해 서사 단계의 순서와 구성이 줄거리의 일관성과 완결성에 미치는 영향을 실증적으로 확인하였다. 각 유닛에 해당하는 서사 단계와 줄거리를 동시에 제시하고 학습

했을 때 가장 자연스럽게 인과관계가 명확한 스토리가 생성되었다. 이는 모델이 학습 과정에서 내러티브 구조에 관한 암묵적 지식을 습득하고, 이를 바탕으로 줄거리를 전개해 나가는 것으로 해석할 수 있다. 또한, 한국어 대중문화 콘텐츠 데이터를 활용한 LLM의 fine-tuning 효과를 검증하였다. 특히 EXAONE 튜닝 모델의 경우 일관성과 완성도 측면에서, EEVE 튜닝 모델은 다양성과 창의성 측면에서 강점을 보여, 각 모델의 특성에 따른 차별화된 성능 향상을 확인하였다.

본 연구에서는 Coverage와 Order 등 스토리 구조 평가에 특화된 지표를 확장하여 한국어 콘텐츠 평가를 위한 체계적인 방법론을 제시하였다. 본 연구에서 제안된 방법론은 대중문화 콘텐츠 제작 과정에서 창의적인 스토리텔링을 지원하는 도구로 활용될 수 있다. 특히 초기 아이디어 구상 단계에서 다양한 플롯 옵션을 빠르게 생성하고 검토하는 데 유용할 것이다. 또한, 줄거리 생성 외에도 장르 분류, 인물 관계 예측, 서사 단계 분석 등 다양한 하위 태스크에 instruction tuning 방법론을 적용할 수 있어, 콘텐츠 분석 및 기획 과정의 효율성을 높일 수 있을 것으로 기대된다.

그러나 본 연구에서 제안된 학습 결과에서는 생성된 플롯에서 동일 인물의 특성이나 행위가 전후에서 불일치하는 경우가 관찰되었다. 이는 모델의 장기 의존성(long-term dependency) 학습 능력의 한계를 시사한다. 또한 로그라인에서 제시된 결말을 충실히 반영하지 못하는 경향이 있었다. 이는 모델이 전체적인 스토리 아크를 유지하면서도 세부적인 요구사항을 반영하는 능력이 아직 부족함을 나타낸다. 더해, 등장인물을 “C001”, “E001”과 같은 특정 기호로 지칭하는 방식은 인물의 고유명사와 각 캐릭터의 고유한 특성보다는 스토리텔링에서 드러나는 기호적인 기능성에 집중하도록 하게 해준다, 실제 독자에게 혼란을 줄 수 있어 비식별화된 각 인물이 나타내는 기호적 기능을 보다 명료하게 이해할 수 있는 인물 표현 방식의 필요로 한다. 이러한 한계는 인공지능을 활용한 한국어 대중문화 콘텐츠 스토리 생성의 후속연구의 필요성을 드러낸다.

본 연구는 이러한 결과를 통해 한국어 대중문화 콘텐츠 제작 과정에서 인공지능을 활용한 창의적 스토리텔링의 가능성을 보여주었으며, 향후 한국어 콘텐츠 생성 AI 연구의 기초 자료로 활용될 수 있을 것으로 기대된다. 나아가 향후 제시된 한계점들을 극복하고 보다 창의적이고 일관성 있는 스토리 생성이 가능한 모델을 개발함으로써, 인공지능 기술과 창의 산업의 융합에 기여할 수 있을 것으로 기대된다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2023년도 문화기술 연구개발 사업으로 수행되었음(과제명: 인공지능 기반 방송콘텐츠 스토리 분석 및 파급효과 예측 기술개발, 과제번호: RS-2023-00226907, 기여율: 100%).

부 록

한국어 대중문화 콘텐츠 스토리 플롯 생성을 위한 fine-tuning 학습 데이터의 한국어 입력 프롬프트 및 참조 플롯, 학습된 각 모델의 생성 플롯 샘플

본 부록에서는 모델의 원본 한국어 입력 프롬프트와 생성된 출력을 그대로 제시함으로써, 각 fine-tuning 모델의 한국어 콘텐츠 처리 및 생성 능력을 객관적으로 보여주고자 한다. 이는 모델의 언어적 성능을 평가하는 데 있어 중요한 실증적 데이터를 제공하며, 특히 한국어 특유의 언어적 뉘앙스와 문화적 맥락이 모델 출력에 어떻게 반영되는지를 직접적으로 관찰할 수 있게 한다. 이러한 접근은 모델의 한국어 처리 능력에 대한 깊이 있는 분석을 가능케 하며, 향후 한국어 자연어 처리 연구에 유용한 기초 자료로 활용될 수 있을 것으로 기대된다.

표 3. 한국어 대중문화 콘텐츠 스토리 플롯 생성을 위한 fine-tuning 학습 데이터의 한국어 입력 프롬프트 및 참조 플롯 샘플

Table 3. A sample of Korean input prompts and reference plots for fine-tuning training data in pop culture content story plot generation

Sample of prompt of StoryData-plotGen
You are an AI assistant. Develop a plot from the given conditions. Genre: 드라마, Theme: 도전 Motif: 파산 Log Line: 보건국에서 가게1의 위생을 점검하려고 온다. 가게1이 위생 점검에서 안 좋은 등급을 받아 파산 위기에 몰린다.
Sample of reference plot of StoryData-plotGen
Opening Salvo: 과거 C001은 고급 식당의 최고참 요리사에게 모욕적인 대우를 받으며 일했다. Main Character: C001이 주방의 위생 상태가 좋지 않음을 인지한다. C001이 주방을 청소한다. Setting-up: C001은 요리하는 꿈을 꾸며 실제로도 요리를 해 집에 불을 낼 뻔한다. Setting-up: C002가 C001에게 가게1의 문제점을 분석한 자료를 준다. 1st Accident: C009가 원래 담당자인 C008을 대신해 가게1에 위생 점검을 하러 온다. C006가 C009를 마음에 들어 하지 않는다. 1st Accident: C002가 C001에게 계속해서 가게1에 대한 문제점을 늘어놓는다. 1st Accident: 가게1이 위생 점검에서 안 좋은 등급을 받는다. C006와 C001이 서로를 탓한다. Making a Choice: C002와 C006가 실리콘을 사러 철물점에 도착한다. Making a Choice: C016가 가게1을 인수하고 싶다고 말한다. C001이 거절한다. 2nd Accident: C006가 전화로 딸을 위로한다. C002가 딸을 대하는 C006에게서 의외의 모습을 발견한다. Choice to Fight: C001이 동생 C015에게 자신의 이상 증세에 대해 털어놓는다. Resurrection: C006가 보조 주방 구역에서 C007의 편지 봉투를 발견한다. Resurrection: C001이 C002를 정식 채용한다. Final Salvo: C001은 위생 점검에서 문제가 되었던 담배가 사실은 자신의 것이었음을 깨닫는다.

*The texts in Korean is to present the actual input prompts and generated outputs.

표 4. Llama-3-8B(baseline) 모델의 한국어 대중문화 콘텐츠 스토리 플롯 생성 샘플

Table 4. Sample story plot generation for Korean pop culture content by Llama-3-8B (baseline model)

Sample of generated plot of Llama 3-8B (baseline)
Im developing a plot based on the given conditions.
Title: The Crisis of Health Genre: Drama, Theme: Challenge, Motif: BankruptcyLog Line: As an inspector from the Health Department comes to inspect the hygiene of Store 1, Store 1 is on the brink of bankruptcy after receiving a bad grade in the inspection.
Plot:
Act I: * Store 1, a small but popular store in the city, is struggling to stay afloat due to financial difficulties.* The owner, Joon-ho, is determined to turn the store around, but his efforts are met with resistance from the employees, who are frustrated with their low wages and poor working conditions.* Meanwhile, an inspector from the Health Department, Ms. Kim, is sent to inspect the stores hygiene. She is strict and demanding, and Store 1s employees are worried about receiving a bad grade.
Act II: * The inspection reveals several violations of health and safety regulations, including expired food products and dirty equipment.* Ms. Kim gives Store 1 a low grade, which is reported to the public, causing a significant drop in sales.* Joon-ho is devastated, as the stores reputation is ruined and he is unable to pay his employees wages.* The employees, feeling betrayed by Joon-hos lack of transparency and accountability, begin to question their loyalty to the store.
Act III: * Joon-ho realizes that he must take drastic measures to save the store. He decides to take on a new partner, a wealthy investor who promises to help him turn the store around.* However, the new partner has an ulterior motive: to use Store 1 as a front to launder illegal money.* Joon-ho is torn between his desire to save the store and his moral principles. He must decide whether to accept the new partners offer or risk everything to maintain his integrity.
Act IV: * Joon-ho decides to reject the new partners offer and instead comes up with a plan to rebrand the store and attract new customers.* With the help of his employees, Joon-ho works tirelessly to improve the stores hygiene and service, and the store begins to attract a new wave of customers.* However, the new partner is not willing to give up and tries to sabotage the stores efforts. Joon-ho must find a way to outsmart the new partner and save the store.
Epilogue: * Store 1 is saved, but not without significant financial losses and a bruised reputation. Joon-ho learns a valuable lesson about the importance of integrity and the consequences of compromising his values.* The employees, who had been on the verge of quitting, are grateful to Joon-ho for his bravery and commitment to the stores success. They vow to stand by him and help him rebuild the stores reputation.This plot explores the theme of challenge and the motif of bankruptcy, as Joon-ho faces a crisis that threatens to destroy his store. The drama unfolds as he struggles to maintain his integrity and find a way to save the store, despite the odds against him.

표 5. Fine-tuned Llama-3-8B 모델의 한국어 대중문화 콘텐츠 스토리 생성 샘플

Table 5. Sample story plot generation for Korean pop culture content by Fine-tuned LLaMA-3-8B model

Sample of generated plot of Fine-tuned Llama-3-8B Model
<p>Main Character: 가게1의 위생 점검 결과로 폐업을 결심하고 있다. Main Character: 가게1의 C001과 C002은 파산에 관한 대화를 한다. Setting-up: C001은 자신이 C005에게 빚을 받았다는 것을 알아차린다. Setting-up: C001은 C005에게 빚을 갚기 위해 일을 시작하려 한다. Setting-up: C001은 C005에게 빚을 갚기 위해 가게2에 일을 하러 간다. 1st Accident: C005는 가게2에서 나온 C001에게 차갑게 대한다. 1st Accident: C005는 가게2에서 나와 술을 마신다. Doubts & Debate: C001은 C002에게 식당 일을 하겠다고 한다. Making a Choice: C001은 가게2에서 식사하는 사람들을 본다. Making a Choice: 가게2에 C001이 돌아와서 식사하는 사람들이 불편해한다. 2nd Accident: C001은 가게2에 손님이 없어졌다고 C002에게 전한다. 2nd Accident: 가게2에 사람들이 다시 모이고 C001은 가게2에 새로운 요리들을 만든다. Defeat: C005가 다시 가게2에 돌아온다. Defeat: C005는 가게2에서 떠나려고 한다. Innermost Cave: C001은 가게2의 사람들을 돌려놓는다. Innermost Cave: C001은 가게2에서 일하는 사람들에게 인사를 나눈다. Innermost Cave: C005는 가게2에서 떠난다. Choice to Fight: C001은 가게2에서 새로운 요리를 만들고 C005를 떠올린다.</p>

*The texts in Korean is to present the actual input prompts and generated outputs..

표 6. Fine-tuned EEEV-Korean-10.8B-v1.0 모델의 한국어 대중문화 콘텐츠 스토리 플롯 생성 샘플

Table 6. Sample story plot generation for Korean pop culture content by Fine-tuned EEEV-Korean-10.8B-v1.0 model

Sample of generated plot of Fine-tuned EEEV-Korean-10.8B-v1.0 Model
<p>Main Character: C001은 식당 일을 그만두겠다고 말한다. Setting-up: C001은 C003에게 자신이 일하는 식당에서 요리사로 일하고 있다고 말한다. Setting-up: C002는 C001에게 자신이 만든 음식을 칭찬해달라고 말한다. 1st Accident: C001은 자신의 요리 실력에 대해 걱정한다. Doubts & Debate: C001은 C002의 음식을 맛보고, 맛있는 음식을 누가 만들어 버렸는지 알아내는 일에 재능이 있다고 말한다. Doubts & Debate: C001은 C002의 음식을 평론해준다. C002는 C001에게 감사의 표시를 한다. Doubts & Debate: C001은 C003가 자신의 식당에서 일하는 것을 허락해달라고 한다. Doubts & Debate: C003는 셰프로 일하기로 하고, 첫 출근길에 가게1을 찾아간다. Making a Choice: C003는 자신의 어머니인 C006에게 가게1에 취업했다고 말하지만, C006은 계속되는 말실수로 C003를 화나게 한다. Making a Choice: C002는 C001에게 식당1을 잘 운영하고 있는지 묻는다. C001은 식당1을 잘 운영하고 있다고 말한다. 2nd Accident: C002는 식당의 평론가로서 비평할 준비를 한다. 2nd Accident: C002는 C001에게 식당의 평론가가 되었다고 말한다. C001은 C002가 요리사로서 성장했다고 말하며 칭찬해준다. 2nd Accident: C002는 C001에게 셰프로써 자신을 인정해줄 것을 요구한다. Villains Move: C002는 C001가 평론가가 된 것을 의심한다. C001은 C002를 안심시킨다. Villains Move: C002는 C001의 평론에 화가 나지만, 평론으로 성공할 수 있다고 말한다. Villains Move: C001은 식당의 평론가가 되어 평론하는 것을 힘들어한다. C001은 평론가의 임무를 다하기로 결심한다. Defeat: C001은 식당의 평론가로 가게1을 방문한다.</p>

Innermost Cave: C003는 보건국 직원에게 위생 검사를 받는다.
 Choice to Fight: C001은 C003가 보건국 직원에게 위생 검사를 받고 있다는 이야기를 듣는다.
 Choice to Fight: 보건국 직원은 C003에게 가게1의 냉장고를 청소하라고 한다.
 Resurrection: 보건국 직원은 C003에게 냉장고를 청소하라고 한다. C003는 C001에게 도움을 요청하지만, 거절당한다.

*The texts in Korean is to present the actual input prompts and generated outputs.

표 7. Fine-tuned EXAONE-3.0-7.8B-Instruct 모델의 한국어 대중문화 콘텐츠 스토리 플롯 생성 샘플

Table 7. Sample story plot generation for Korean pop culture content by Fine-tuned EXAONE-3.0-7.8B-Instruct model

Sample of generated plot of Fine-tuned EXAONE-3.0-7.8B-Instruct Model
<p>Opening Salvo: E003이 가게1을 방문하여 가게1을 검사한다. Setting-up: C001은 E003에게 높은 위생등급을 받는다. 1st Accident: E004이 가게1을 방문하여 위생 점검을 한다. Making a Choice: C001은 위생등급이 하향 조정되었다는 것을 알게 된다. Trailer Moments: C001은 C003를 위해 가게1을 영업하는 것을 지속하기로 결심한다. 2nd Accident: C001은 E004의 말을 무시하고 고기를 진열하려다 큰 혼난다. Trailer Moments: C001이 손님들에게 음식을 제공한다. Villains Move: C002는 C001의 가게가 위생 점검에 실패했다는 사실을 알게 된다. Defeat: C001은 C002에게 가게1을 그만두겠다고 말한다. Choice to Fight: C001은 가게1 영업을 포기하지 않는다. Final Salvo: C001이 가게1을 영업하기 위해 보건당국에게 부탁한다.</p>

*The texts in Korean is to present the actual input prompts and generated outputs..

참고문헌

[1] M. Kreminski, M. Dickinson, M. Mateas, and N. Wardrip-Fruin, "Why Are We Like This?: The AI Architecture of a Co-Creative Storytelling Game," in *Proceedings of the 15th International Conference on the Foundations of Digital Games (FDG '20)*, Bugibba, Malta, 13, September 2020. <https://doi.org/10.1145/3402942.3402953>

[2] N. Akoury, S. Wang, J. Whiting, S. Hood, N. Peng, and M. Iyyer, "STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 6470-6484, November 2020. <https://doi.org/10.18653/v1/2020.emnlp-main.525>

[3] B. Branch, P. Mirowski, and K. W. Mathewson, "Collaborative Storytelling with Human Actors and AI Narrators," arXiv:2109.14728, September 2021. <https://doi.org/10.48550/arXiv.2109.14728>

- [4] H. Oson, J.-L. Lu, and Y. Ochiai, "BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese," in *Proceedings of 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, Yokohama, Japan, 19, May 2021. <https://doi.org/10.1145/3411763.3450391>
- [5] A. Tikhonov, I. Samenko, and I. P. Yamshchikov, "StoryDB: Broad Multi-Language Narrative Dataset," in *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, Punta Cana, Dominican Republic, pp. 32-39, November 2021. <https://doi.org/10.18653/v1/2021.eval4nlp-1.4>
- [6] M. Lee, P. Liang, and Q. Yang, "CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*, New Orleans: LA, 388, April-May 2022. <https://doi.org/10.1145/3491102.3502030>
- [7] 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>
- [8] H. Im, S. Jeon, H. Cho, S. Shin, D. Choi, and H. Hong, "AI Story Relay: A Collaborative Writing of Design Fiction to Investigate Artificial Intelligence Design Considerations," in *Proceedings of the 2023 ACM International Conference on Supporting Group Work (GROUP '23)*, Hilton Head: SC, pp. 6-8, January 2023. <https://doi.org/10.1145/3565967.3570973>
- [9] E.-S. You, O.-J. Lee, and J.-T. Kim, "Computational Narrative Generation: A Survey," *Journal of Information Technology and Applied Engineering*, Vol. 10, No. 1, pp. 9-18, June 2020. <http://doi.org/10.22733/JITAE.2020.10.01.002>
- [10] S. J. Park, "Thesis on the Feasibility of an 'AI Yi Yuksa'(1) - Suggestions for Building AI and Literary Discourse Functions," *The Journal of Literary Creative Writing*, Vol. 22, No. 2, pp. 87-119, August 2023. <https://doi.org/10.47057/jklcw.2023.58.04>
- [11] T. H. Kim and S. J. Park, "Implementation of AI Poet Using ChatGPT (2) - Focusing on the Creation of 'AI Yi Yuk-Sa' and the Possibility of Poetic Conversation," *International Language and Literature*, No. 55, pp. 59-92, August 2023.
- [12] S. J. Park, "A Study on the Poetic Rhetoric Implementation of 'AI Yi Yuk-sa' (3) - Reimagining 'Peak' and Centered around 'Iron Rainbow'," *Literary Criticism*, Vol. 89, pp. 207-242, September 2023. <https://doi.org/10.31313/LC.2023.09.89.207>
- [13] J. An, S. Leeyouk, Y. Jang, and D. Jeong, "Construction of Debaised Korean AI Fairytale Generator," *Journal of Digital Contents Society*, Vol. 24, No. 7, pp. 1513-1528, July 2023. <https://doi.org/10.9728/dcs.2023.24.7.1513>
- [14] Y.-Y. Kim and Y.-H. Nam, "Fairy Tale Reconstruction Service Using Conversational AI," *Journal of Digital Contents Society*, Vol. 24, No. 9, pp. 1955-1964, September 2023. <http://dx.doi.org/10.9728/dcs.2023.24.9.1955>
- [15] S. Kwon, D. Kim, and D. Jeong, "Play with AI: Play Writing with GPT-3 and Collaborative Acting with Virtual Human - A Case Study of a XR-Based Artwork 'Blue Space'," in *Proceedings of HCI Korea 2023*, Jeongseon, pp. 1428-1431, February 2023.
- [16] J. Y. Park, J. M. Shin, G. Y. Kim, J. Nam, and B.-C. Bae, "A Prototype Design of an Interactive Story Authoring Tool Based on a Generative Language Model," in *Proceedings of the Korean Computer Congress 2023 (KCC 2023)*, Jeju, pp. 1817-1819, June 2023.
- [17] J.-S. Kim and M.-J. Kim, "Analysis of the Distribution of Creative Elements in StoryData," *Journal of Digital Contents Society*, Vol. 24, No. 12, pp. 3005-3012, December 2023. <https://doi.org/10.9728/dcs.2023.24.12.3005>
- [18] B. Li, S. Lee-Urban, G. Johnston, and M. Riedl, "Story Generation with Crowdsourced Plot Graphs," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue: WA, pp. 598-604, July 2013. <https://doi.org/10.1609/aaai.v27i1.8649>
- [19] 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>
- [20] L. Yao, N. Peng, R. Weischedel, K. Knight, D. Zhao, and R. Yan, "Plan-and-Write: Towards Better Automatic Storytelling," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, Honolulu: HI, pp. 7378-7385, July 2019. <https://doi.org/10.1609/aaai.v33i01.7378>
- [21] 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>

[22] P. Tambwekar, M. Dhuliawala, L. J. Martin, A. Mehta, B. Harrison, and M. O. Riedl, “Controllable Neural Story Plot Generation via Reward Shaping,” in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, Macao, China, pp. 5982-5988, August 2019. <https://doi.org/10.24963/ijcai.2019/829>

[23] 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, January 2020. https://doi.org/10.1162/tacl_a_00302

[24] 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>

[25] P. Xu, M. Patwary, M. Shoeybi, R. Puri, P. Fung, A. Anandkumar, and B. Catanzaro, “MEGATRON-CNTRL: Controllable Story Generation with External Knowledge Using Large-Scale Language Models,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 2831-2845, November 2020. <https://doi.org/10.18653/v1/2020.emnlp-main.226>

[26] 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>

[27] J. Guan, Z. Feng, Y. Chen, R. He, X. Mao, C. Fan, and M. Huang, “LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation,” *Transactions of the Association for Computational Linguistics*, Vol. 10, pp. 434-451, April 2022. https://doi.org/10.1162/tacl_a_00469

[28] S. J. Paech, “EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models,” arXiv:2312.06281v1, December 2023. <https://doi.org/10.48550/arXiv.2312.06281>

[29] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, ... and Z. Papakipos, “The Llama 3 Herd of Models,” arXiv:2407.21783v1, July 2024. <https://doi.org/10.48550/arXiv.2407.21783>

[30] S. Kim, S. Choi, and M. Jeong, “Efficient and Effective

Vocabulary Expansion towards Multilingual Large Language Models,” arXiv:2402.14714, February 2024. <https://doi.org/10.48550/arXiv.2402.14714>

[31] LG AI Research, “EXAONE 3.0 7.8B Instruction Tuned Language Model,” arXiv:2408.03541, August 2024. <https://doi.org/10.48550/arXiv.2408.03541>



김지선 (Ji-Seon Kim)

2021년 : 이화여자대학교
융합콘텐츠학과 (학사)

2022년~현재 : 이화여자대학교 융합콘텐츠학과 석박사 통합과정

※ 관심분야 : 디지털 스토리텔링(Digital Storytelling), 스토리 데이터(Story Data), 인공지능(Artificial Intelligence) 등



김명준 (Myoung-Jun Kim)

1991년 : 한국과학기술원
(전산학 석사)
1996년 : 한국과학기술원
(전산학 박사)

1996년~1997년: University of Washington Visiting Scholar

1997년~2000년: 한국전자통신연구원 선임연구원

2001년~현재 : 이화여자대학교 융합콘텐츠학과 교수

※ 관심분야 : 컴퓨터그래픽스, 영상처리, 스토리 데이터, 인공지능 등