Journal of the Korean Institute of Industrial Engineers
Vol. 51, No. 5, pp. 376-387, October 2025.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2025.51.5.376

거대 언어 모델과 인간 피드백 기반 강화학습을 활용한 절차적 게임 레벨 생성

이준범 · 김정인 · 허종국 · 이정민 · 김재훈 · 김성범†

고려대학교 산업경영공학과

Procedural Game Level Generation Using Large Language Models and Reinforcement Learning with Human Feedback

Junbeom Lee · Jung In Kim · Jongkook Heo · Jungmin Lee · Jaehoon Kim · Seoung Bum Kim

School of Industrial and Management Engineering, Korea University

Designing game levels during development is both time-consuming and resource-intensive. To address this challenge, recent studies have focused on automated level generation using deep neural networks trained through supervised learning. However, game level datasets suitable for training are often unavailable or contain structural errors that violate game mechanics, resulting in unplayable levels. To overcome these limitations, this study proposes Mario level generation using human preferences (MarioPref) that uses low-quality Super Mario Bros game level data and conditional prompts to generate playable levels that satisfy specified constraints. The framework uses a two-stage training approach: first, a large language model (LLM) is fine-tuned using supervised learning to reconstruct masked level elements; second, reinforcement learning with human feedback (RLHF) is applied to enhance playability and prompt adherence. Experimental results indicate that MarioPref effectively generates playable levels that satisfy input prompts, even when trained on low-quality data. These findings demonstrate the practical value of using human feedback to enhance content quality under data scarcity and suggests the potential for reliable content generation across diverse game genres.

Keywords: Deep Learning, Large Language Models, Procedural Content Generation, Reinforcement Learning For Human Feedback

1. 서 론

절차적 레벨 생성(procedural level generation)은 콘텐츠를 자동으로 생성하는 절차적 콘텐츠 생성(procedural content generation, PCG)의 대표적인 하위 분야로, 게임 개발 과정에서 중요한 역할을 담당한다. 게임 레벨은 하나의 플레이 단위로, 맵의 지형 구조를 기반으로 다양한 구성요소와 이벤트가 배치되어 설계된다. 이때 적의 위치, 아이템 분포, 난이도 조정 등은 플레이어의 도전과 흥미를 유도하도록 조절되며, 이러한 설계는 특정 목표를 향한 몰입감 있는 플레이 경험을 제공하는 데 핵심적인 역할을 한다. 따라서 게임 레벨 생성은 단순한 콘텐츠 배치를 넘어 명확한 게임 규칙과 설계 조건을 만족하는 고품질의 플레이 가능한 콘텐츠를 생성하는 것이 중요하다(Liu *et al.*, 2021).

최근 게임 산업의 지속적인 성장과 게임의 복잡성이 증가함에 따라, 성공적인 게임 레벨을 제작하는 데 필요한 인적, 물적 자원의 수요가 꾸준히 증가하고 있다. 그러나, 현재 게임 산업에서 대부분의 게임 레벨은 레벨 디자이너의 수작업에 의존하고 있으며, 이로 인해 많은 시간과 비용이 소요되는 문제를 가지고 있다(Shaker *et al.*, 2016). 이러한 문제를 해결하기 위한

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화체육관광 연구개발사업으로 수행되었음(과제명 : 중소 게임 기업의 게임 제작 검증 효율화를 위한 AI 기반의 대규모 게임 자동검증 기술 개발, 과제번호 : RS-2024-00393500, 기여율: 100%)
† 연락저자 : 김성범 교수, 02841 서울특별시 성북구 안암로 145 고려대학교 산업경영공학부, Tel: 02-3290-3397, Fax: 02-929-5888, E-mail : sbkim1@korea.ac.kr
2025년 4월 22일 접수; 2025년 5월 22일 수정본 접수; 2025년 5월 26일 게재 확정.

방안으로, 콘텐츠를 자동으로 생성하는 PCG가 주목받고 있으며, 고품질 콘텐츠를 효율적으로 생성함으로써 개발 비용과 시간을 절감하려는 연구의 필요성이 점점 더 커지고 있다.

초기PCG 연구는 주로 검색 기반 기법(search-based techniques)과 메타 휴리스틱 접근법(metaheuristic approaches)을 사용하여 게임 레벨을 생성하는데 초점을 맞췄다(Togelius *et al.*, 2011). Oranchak(2010)은 유전 알고리즘(genetic algorithm)과 담금질 기법(simulated annealing)을 사용하여, 레벨 생성을 탐색 공간에서 최적화 문제로 정의하고, 규칙 기반으로 생성된 콘텐츠의 품질을 평가 및 개선하는 데 적용하였다. 그러나 이러한 방법들은 사전에 정의된 규칙에 의존하기 때문에 새로운 디자인 요소에 유연하게 대응하지 못하며, 복잡한 레벨 구조를 충분히 표현하는 데 한계가 있다. 최근 딥러닝 기술의 발전으로, 명시적인 규칙 없이 데이터로부터 숨겨진 패턴과 규칙을 효과적으로 학습할 수 있는 딥러닝 기반 PCG 기술이 주목받고 있다(Hendrikx *et al.*, 2013). Volz *et al.*(2018)은 적대적 생성 신경망(generative adversarial networks, GAN)(Goodfellow *et al.*, 2014)을 사용하여 슈퍼 마리오 브로스(Super Mario Bros.) 게임의 레벨 잠재 공간을 학습하고, 진화 알고리즘을 통해 다양한 플레이 가능한 레벨을 생성하는 연구를 진행하였다. Awiszus *et al.*(2020)은 GAN을 활용하여 단일 예제를 학습하여 게임 레벨의 스타일과 구조를 유지하면서 새로운 레벨을 생성하였다. 그러나, GAN 기반 접근법은 조건에 부합하는 레벨을 생성하기 위해 잠재 공간을 탐색하는 과정에서 높은 연산 비용이 발생하는 한계를 가진다.

이러한 한계를 극복하기 위해, 최근에는 거대 언어 모델(large language models, LLM)을 활용한 게임 레벨 생성 연구가 진행되고 있다. Sudhakaran *et al.*(2023)은 LLM을 활용한 텍스트 프롬프트 기반 레벨 생성 모델인 MarioGPT를 제안하였다. MarioGPT는 자연어 프롬프트를 입력 받아 사용자의 요구에 부합하는 게임 레벨을 생성할 수 있음을 보여주었다. 또한, Nasir and Togelius(2023)는 데이터가 부족한 상황에서도 사람의 개입을 통해 레벨을 효과적으로 생성할 수 있는 human-in-the-loop 기반 미세 조정 학습을 제안하였다. 더 나아가 Hu *et al.*(2024)은 게임 규칙과 레벨을 동시에 생성할 수 있는 LLM 기반 프레임워크를 제안하였다. 하지만, LLM 학습에는 고품질 데이터가 필수적이며, 대부분의 게임에서는 고품질 데이터가 사전에 구축되어 있지 않아 수집과 정제 과정에서 상당한 시간과 비용이 소요된다(Summerville *et al.*, 2018). 또한 LLM은 기존 데이터의 패턴을 단순히 모방하는 경향이 있어, 복잡한 게임 규칙이나 공간적 관계를 효과적으로 파악하지 못하는 문제점이 존재한다(Mao *et al.*, 2024). 특히 저품질 데이터로 학습한 경우 플레이가 불가능한 지형이나 잘못된 오브젝트를 반복적으로 생성할 가능성이 높다. 따라서, 저품질 데이터에도 불구하고 정상적인 플레이가 가능한 고품질 레벨을 생성할 수 있는 알고리즘 개발이 요구된다.

본 논문에서는 LLM을 활용하여 저품질의 슈퍼 마리오 브로스 레벨 데이터와 조건 프롬프트를 입력 받아 오브젝트 및 지형을 재배치함으로써 조건 프롬프트에 부합하고 실제로 플레이 가능한 레벨 데이터를 생성하는 Mario level generation using human preferences(MarioPref)을 제안한다. 슈퍼 마리오 브로스의 레벨은 하늘, 몬스터, 블록 등 다양한 구성 요소를 나타내는 2차원 배열의 타일(tile) 배열로 구성된다. 본 연구에서는 이러한 2차원의 타일 데이터를 순차적인 1차원 시퀀스 데이터로 변환하여, LLM을 통해 효과적으로 레벨을 생성할 수 있도록 하였다. MarioPref는 길이 1,000 이상의 타일 시퀀스를 입력 및 출력하기 위해 efficient text-to-text transformer for long sequences (LongT5)(Guo *et al.*, 2022)을 활용한다. 또한, 오브젝트의 개수를 나타내는 조건 프롬프트는 robustly optimized bidirectional encoder representations from transformers pretraining approach (RoBERTa)(Y. Liu *et al.*, 2019)를 통해 임베딩하여, 프롬프트가 가지는 언어적 의미를 효과적으로 파악하도록 한다. 더불어, 학습 데이터의 다양성을 확보하기 위해 오브젝트 재구축 기반 데이터 마스킹 기법을 적용하여 모델의 일반화 성능을 높였다. 마지막으로, 플레이 불가능한 레벨을 생성하는 문제를 해결하기 위해 인간 피드백 기반 강화학습(reinforcement learning with human feedback, RLHF)을 활용한 미세 조정을 수행함으로써, 사용자 조건에 부합하고 실제 플레이가 가능한 고품질의 레벨을 생성하도록 하였다. 본 논문의 주요 기여점은 다음과 같다.

- 본 연구는 불완전한 구성 요소 또는 잘못된 지형 배치로 인해 플레이가 불가능한 저품질 슈퍼 마리오 브로스 레벨을 학습 데이터로 활용하여, 입력된 조건 프롬프트에 따라 지형 및 오브젝트를 효과적으로 재구축함으로써 고품질의 플레이 가능한 게임 레벨을 생성하는 MarioPref 모델을 제안한다. 이를 통해 게임 콘텐츠 제작 과정에서 레벨 디자인에 소모되는 시간과 비용을 현저히 줄일 수 있을 것으로 기대된다.
- 본 연구는 절차적 콘텐츠 생성 분야에서 RLHF을 도입한 최초의 시도로, 복잡한 도메인 지식이나 수학적 모델링 없이도 인간 피드백만을 활용하여 조건을 만족하고 플레이 가능성을 보장하는 게임 레벨을 효과적으로 생성할 수 있음을 실험적으로 입증하였다.

본 논문은 다음과 같은 구성으로 진행된다. 제2장에서는 제안 방법론에서 활용한 관련 연구에 대해 소개하고, 제3장에서는 제안 알고리즘에 대하여 설명한다. 제4장에서는 실험 방법 및 실험 결과를 설명한다. 마지막 제5장에서는 본 논문의 결론과 기대 효과를 다루도록 한다.

2. 배경 방법론

2.1 트랜스포머(transformer) 기반 언어 모델

Vaswani *et al.*(2017)은 기계 번역 분야에서 긴 시퀀스에 대한 장기 의존성(long-term dependency)을 해결하기 위해 transformer

모델을 제안하였다. Transformer는 self-attention 메커니즘을 통해 입력 및 출력 시퀀스 내 요소들 간 상관관계를 효과적으로 포착함으로써, 전체 시퀀스에 대한 정보를 반영하고 길이에 상관없이 복잡한 패턴을 학습할 수 있다. Transformer는 encoder-decoder 구조를 기반으로 하며, encoder는 입력 시퀀스 내 데이터 간의 상관관계를, decoder는 입력과 출력 시퀀스 간의 관계를 모델링한다. 특히, 다수의 독립적인 모듈로 이루어진 multi-head self-attention을 통해 다양한 표현 공간에서 정보를 추출할 수 있도록 하여 모델의 표현력을 강화한다. 이러한 구조적 특성을 바탕으로 transformer기반 언어 모델들이 발전되어 왔으며, 기계 번역뿐만 아니라 감성 분석, 자연어 추론 등 다양한 자연어 처리 작업에서 뛰어난 성능을 입증하고 있다(Devlin *et al.*, 2019; Radford *et al.*, 2019; Raffel *et al.*, 2020).

Raffel *et al.*(2020)은 번역, 요약, 질의 응답, 분류 등 다양한 자연어 처리 작업을 하나의 통합된 프레임워크로 수행하기 위해 text-to-text transfer transformer(T5)를 제안하였다. T5는 transformer의 encoder-decoder 구조와 대량의 텍스트 데이터를 활용한 사전 학습을 통해 모델이 언어의 문맥을 잘 이해하도록 한다. 이 모델은 특정 작업에 대한 프롬프트와 질문을 입력으로 받아, 이에 적절한 형태의 응답을 생성함으로써 다양한 자연어 처리 과제를 일관된 방식으로 해결한다. 하지만 T5는 dense self-attention을 기반으로 하므로, 비교적 긴 입력 시퀀스를 처리하는 데 한계가 존재한다. 이를 극복하기 위해 Guo *et al.*(2022)은 최대 입력 크기를 확장하고 긴 입력 시퀀스를 효과적으로 처리할 수 있도록 설계된 LongT5를 제안하였다. LongT5는 local attention을 통해 입력 길이 증가에 따라 급격히 증가하는 계산 복잡도를 완화함으로써 긴 텍스트 시퀀스를 보다 효과적으로 다룰 수 있다. 본 연구에서는 타일의 2차원 배열로 이루어진 길고 복잡한 시퀀스를 처리하기 위해 LongT5를 활용하였다.

Transformer의 encoder-decoder 구조를 모두 활용하는 모델뿐만 아니라 encoder만을 사용한 언어 모델들도 큰 발전을 이루어 왔다. 그 중 대표적인 연구인 bidirectional encoder representations from transformers(BERT)는 마스킹(masking)을 활용한 양방향 문맥 학습을 통해 일반적인 수준의 언어 이해가 가능하도록 사전 학습되었다(Devlin *et al.*, 2019). 하지만 BERT의 학습이 충분하지 않다는 점이 지적되었고, 이를 개선하기 위한 다양한 접근법들이 제안되었다. RoBERTa는 더 긴 문장 데이터를 활용하고 마스킹 패턴을 동적으로 변화시키는 방식으로 BERT의 학습 과정과 성능을 개선하였으며, 이를 통해 보다 강건한 언어 표현을 학습할 수 있도록 하였다(Liu *et al.*, 2019). 본 연구에서는 RoBERTa를 활용하여 생성하고자 하는 레벨에 대한 조건 프롬프트의 언어적 의미를 효과적으로 반영하고자 한다.

2.2 인간 피드백 기반 강화학습

RLHF는 언어 모델이 생성하는 결과물의 품질을 향상시키기 위해 인간 평가자의 피드백 데이터를 보상 신호로 활용하고, 이를 강화학습 목적 함수와 결합하여 모델을 고도화하는 방법론이다(Ouyang *et al.*, 2022; Rafailov *et al.*, 2023; Stiennon *et al.*, 2020). RLHF는 인간의 주관적 평가를 학습에 반영함으로써, 모델이 생성하는 출력이 실제 사용자 요구에 보다 부합하도록 미세 조정하는 데 중점을 둔다. RLHF는 특히 자연스러운 응답, 유용성, 안전성과 같은 정량화하기 어려운 품질 기준을 반영할 수 있다는 점에서 주목받고 있다.

일반적인 RLHF 프로세스인 proximal policy optimization (PPO) 방법론은 두 단계로 나뉘는데, 첫 번째 단계는 보상 함수 학습, 두 번째 단계는 학습된 보상 함수와 강화학습 목적 함수를 바탕으로 언어 모델을 미세 조정하는 것이다. 지도 학습으로 사전 학습된 언어 모델 π_θ과 입력 프롬프트 x에 대해, 동일 프롬프트에 대한 두 개의 출력$(y_1,\ y_2) \sim \pi_\theta(y|x)$을 생성한다. 이 때, 인간 평가자는 두 출력 중 선호하는 결과를 선택하고, 이를 바탕으로 선호된 출력 y_w와 비선호 출력 y_l로 레이블링 한다. 이렇게 구성된 크기 N의 선호 데이터셋 $D=\{x^i,\ y_w^i,\ y_l^i)\}_{i=1}^N$은 언어 모델 출력의 보상을 예측하는 보상 함수 r_ψ 학습에 활용된다. 이 때, 주어진 입력 프롬프트 x에 대해 두 출력 중 y_w가 선택될 확률 P_ψ는 보상 함수 r_ψ를 기반으로, Bradley-Terry 모델(Bradley and Terry, 1952)에 따라 다음과 같이 정의된다:

$$P_\psi(y_w > y_l \mid x) = \frac{\exp(r_\psi(x,y_w))}{\exp(r_\psi(x,y_w)) + \exp(r_\psi(x,y_l))} \quad (1)$$

이후 단계에서는 선호 출력 y_w의 선택 확률은 높이고, 비선호 출력 y_l의 선택 확률은 낮추도록 하기 위해, 이진 교차 엔트로피(binary cross entropy, BCE) 손실 함수를 통해 보상 함수 r_ψ를 학습한다:

$$L_\psi = -E_{(x,y_w,y_l)\sim D}\left[\log\sigma(r_\psi(x,y_w) - r_\psi(x,y_l))\right] \quad (2)$$

여기서 σ는 시그모이드(sigmoid) 함수를 의미한다. 학습된 보상 함수는 이후 다음의 강화학습 목적 함수를 통해 언어 모델 π_θ를 미세 조정하는 데 활용된다:

$$\max_\theta E_{x\sim D, y\sim\pi_\theta(y|x)}[r_\psi(x,y)] - \beta D_{KL}[\pi_\theta(y|x)\|\pi_{ref}(y \mid x)] \quad (3)$$

여기서 π_{ref}는 일반적으로 지도 학습을 통해 학습된 초기 언어 모델을 의미하며, β는 모델이 초기 모델에서 크게 벗어나지 않도록 규제하는 하이퍼파라미터(hyperparameter)이다. β의 값이 클수록 최종 모델이 초기 모델과 유사하게 유지된다.

인간 피드백을 통한 보상 함수 학습 방식은 복잡한 전문 지식 없이도 모델을 인간의 의도에 맞게 고도화할 수 있다는 장점이 있다(Christiano *et al.*, 2017). 하지만, 이러한 방식은 보상

함수 학습에 따른 높은 계산 비용과 인간 선호의 부정확한 반영으로 인해 잘못된 정책 업데이트가 이루어질 수 있다는 한계도 존재한다(An *et al.*, 2023; Hejna and Sadigh, 2023; Rafailov *et al.*, 2023). 이를 극복하기 위해, Rafailov *et al.*(2023)은 보상 함수 학습 단계를 생략하고, 인간 선호 정보를 직접 정책 최적화에 반영하는 direct preference optimization(DPO)를 제안하였다. 우선 식 (3)의 최적해 π_θ^*는 다음과 같이 표현될 수 있다(Peng *et al.*, 2019):

$$\pi_{\theta^*} = \frac{1}{Z(x)}\pi_{ref}(y \mid x)\exp\left(\frac{1}{\beta}r_\psi(x,y)\right), \quad \text{where } Z(x) = \Sigma_y \pi_{ref}(y \mid x)\exp\left(\frac{1}{\beta}r_\psi(x,y)\right) \tag{4}$$

따라서, 보상 함수는 다음과 같이 정리된다:

$$r_\psi(x,y) = \beta\log\frac{\pi_\theta^*(y|x)}{\pi_{ref}(y|x)} + \beta\log Z(x) \tag{5}$$

이때 식 (5)의 r_ψ를 식 (2)에 대입함으로써, DPO의 최종 손실 함수는 식 (6)과 같이 정의된다. 이를 통해 보상 함수를 별도로 학습하지 않고 선호 데이터 D만을 활용하여 언어 모델 π_θ을 직접 학습할 수 있다.

$$L_{DPO} = -E_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right] \tag{6}$$

식 (6)은 주어진 프롬프트 x에 대해, 선호 출력 y_w의 생성 확률은 π_{ref}보다 높이고, 반대로 비선호 출력 y_l의 생성 확률은 π_{ref}보다 낮추도록 학습시키며, 그 최적해는 식 (4)에서 정의한 π_{θ^*}와 일치한다.

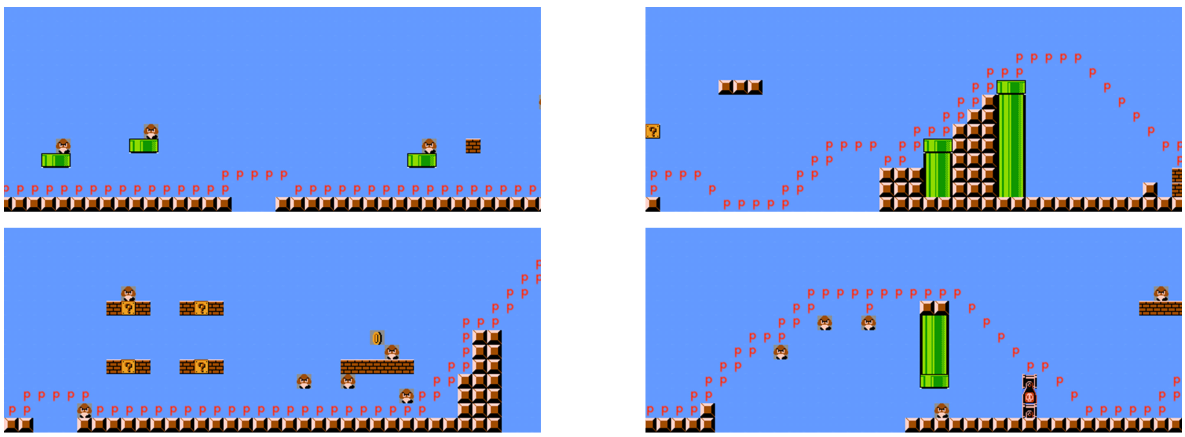
3. 제안 방법론

3.1 데이터 수집 및 전처리

슈퍼 마리오 브로스의 레벨은 <Table 1>과 같이 video game level corpus(VGLC)에서 제공하는 형식에 따라 각 타일에 대응하는 기호로 구성된 2차원 배열 형태로 표현된다(Summerville *et al.*, 2016). 이러한 2차원의 배열 형태의 레벨은 1차원 시퀀스로 변환된 후, 언어 모델에 입력 사용되어 레벨을 생성하는 데 활용된다. 이때 변환은 왼쪽에서 오른쪽, 즉 게임의 이동 방향을 기준으로 순차적으로 진행되며, 각 열의 타일 정보는 아래에서 위로 변환한다. 다시 말해, 맵의 가장 왼쪽 열부터 시작해 각 열을 아래에서 위로 순차적으로 읽고, 그 다음 오른쪽 열로 이동하는 방식으로 전체 타일 정보를 이어 붙여 하나의 1차원 시퀀스로 변환한다.

Table 1. Tile type and symbol in Super Mario Bros

Tile Type	Symbol	Tile
Empty	-	
Unbreakable Block	**X**	
Breakable Block	**S**	
Question Block	**? / Q**	
Coin	**o**	
Enemy	**E**	
Left pipe top	**<**	
Right pipe top	**>**	
Left pipe lower	**[**	
Right pipe lower	**]**	
Cannon	**B**	
Cannon Body	**b**	



(a) Floating Enemies and Incomplete Pipes (b) Non-playable Levels

Figure 1. Examples of errors in levels generated by MarioGPT: (a) Structural errors such as incompletely generated pipes and enemies floating without proper support; (b) Level configurations with excessive gaps or misaligned platforms that prevent the player from progressing, resulting in unplayable gameplay

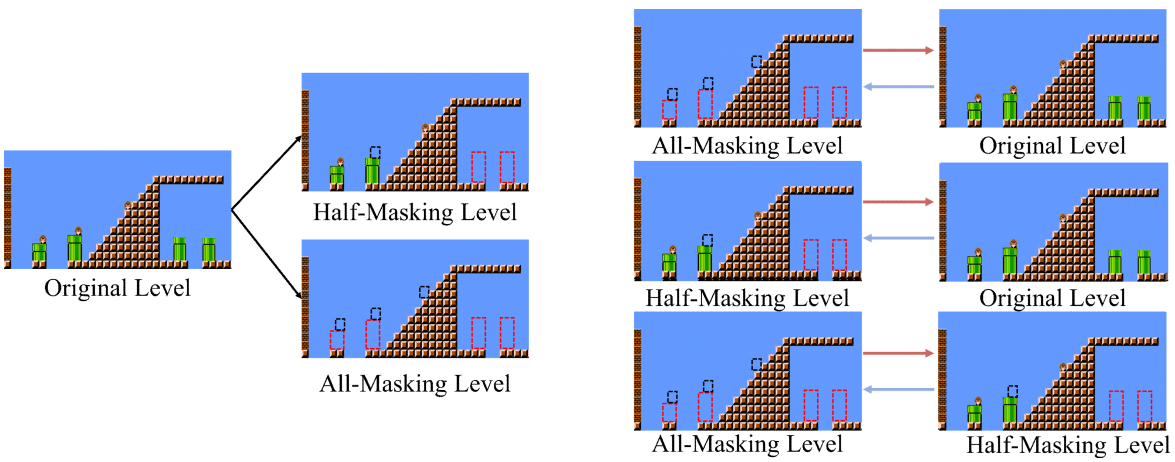
지도 학습 기반의 절차적 레벨 생성 연구에서는 고품질 데이터가 필수적이다. 하지만, 대부분의 게임 환경에서는 고품질의 데이터를 확보하는 데 상당한 비용과 시간이 소요된다. 따라서 본 연구에서는 이러한 제약을 고려하여, LLM기반 마리오 레벨 생성 모델인 MarioGPT를 활용해 학습에 사용할 저품질 데이터를 생성하였다. MarioGPT를 통해 생성한 데이터에는 두 가지 주요 문제가 존재한다. 첫 번째 문제는 <Figure 1(a)>와 같이 적이 공중에 떠 있거나 물체가 불완전하게 생성되는 구조적 오류로, 이는 게임의 물리적, 시각적 일관성을 해치고 플레이어의 몰입감을 저하시킬 수 있다. 두 번째는 <Figure 1(b)>와 같이 지형의 잘못된 배치로 인해 캐릭터가 레벨 끝까지 도달하지 못하는 플레이 불가능한 오류로, 이는 게임을 완료할 수 없게 만들어 플레이어의 성취감을 저해할 수 있다.

본 연구에서는 앞서 제시한 두 가지 문제를 각각 다른 접근법을 사용하여 해결하였다. 먼저, 구조적 오류(예: 공중에 떠 있는 몬스터, 불완전한 파이프)는 규칙 기반의 데이터 전처리 과정을 통해 보완하였다. 공중에 떠 있는 적은 해당 적의 위치 아래에 지형 블록이나 파이프가 존재하지 않으면 비정상적 위치로 간주하였다. 이러한 경우, 적을 동일한 열에서 가장 가까운 지형 혹은 파이프 타일 위에 정상적으로 위치하도록 조정하였다. 비정상적인 파이프는 다음 두 가지 경우로 정의하였다. 첫째, 하나의 파이프를 구성하는 타일들이 좌우에서 짝을 이루지 않는 경우이다. 둘째, 동일한 열에서 파이프 타일 사이에 하늘 타일이 존재하여 파이프가 연속되지 않는 경우이다. 이러한 이상 구조가 감지되면 해당 파이프를 구성하는 모든 타일들을 하늘 타일로 변환하여 제거하였다. 이러한 전처리 과정을 통해 비정상적인 요소가 학습 데이터에 포함되는 것을 방지하고, 모델이 보다 안정적이고 일관된 레벨을 생성할 수 있도록 유도했다. 두 번째 문제인 플레이 불가능한 레벨에 대한 학습은 인간 피드백 기반 미세 조정 방법론인 RLHF를 통해 개선했으며, 이에 대한 구체적인 접근 방식은 3.3절에서 자세히 설명하였다.

3.2 언어 모델 기반 레벨 생성 학습

본 연구에서는 언어 모델과 RLHF를 활용한 MarioPref 모델을 제안하였다. MarioPref는 자연어 조건 프롬프트와 저품질의 슈퍼 마리오 브로스 레벨을 동시에 입력 받아, 주어진 프롬프트의 조건에 맞는 레벨을 생성하는 역할을 수행한다. 이러한 과정을 학습하기 위해 MarioPref는 두 단계의 학습과정을 거친다. 첫 번째 단계에서는 마스킹된 레벨을 복원하도록 모델을 지도 학습 방식으로 학습하며, 두 번째 단계에서는 인간 피드백 기반 미세조정 방법론인 RLHF를 사용하여 프롬프트 조건을 만족하고 플레이 가능성이 높은 레벨을 생성하도록 미세조정한다.

MarioPref는 마스킹된 구성 요소를 복원하는 학습을 통해, 파이프와 적이 제거된 상황에서도 해당 레벨에 요구되는 구성 요소를 이해하고 이에 맞추어 올바르게 재구축하는 것을 목표로 한다. 이를 위해, 원본 레벨에서 파이프와 적을 하늘 타일로 대체하는 두 가지 마스킹 기법을 적용하였다. 첫 번째 방식은 <Figure 2(a)>와 같이 전체 구성 요소를 마스킹하는 방법이며, 두 번째 방식은 파이프와 적 중 절반만을 무작위로 마스킹하는 방법이다. 마스킹은 규칙 기반 알고리즘을 통해 진행했으며, 이를 통해 생성된 세 가지 데이터 조합으로 <Figure 2(b)>와 같은 총 여섯 가지 학습 과정을 구성하였다. 여섯 가지 학습 과정을 통해 모델은 단순히 마스킹된 구성 요소를 복원하는 것에 그치지 않고, 상황에 따라 필요한 구성 요소를 제거하는



(a) Masking Super Mario Level (b) Training Task

Figure 2. Masking Strategies for Training Generation Level Components. (a) Two masking approaches — complete masking of all pipes and enemies, and random masking of half — are applied to input levels; (b) Six distinct training scenarios are constructed by combining original and masked levels, allowing the model to learn both the generation and removal of specific level components

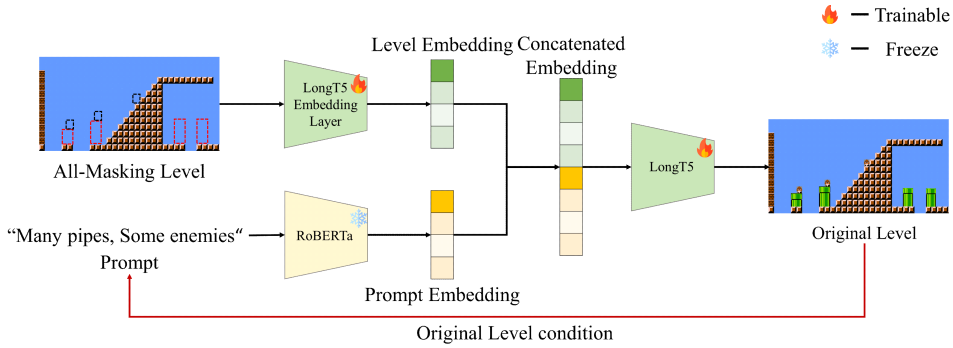


Figure 3. MarioPref Supervised Learning Architecture: LongT5 and RoBERTa based Model Integrating Structural Information from Mario Levels and Semantic Information from Natural Language Prompts

능력까지 습득하였다. 이를 통해 다양한 프롬프트 조건에 맞는 레벨을 유연하게 생성할 수 있도록 하였다.

본 연구에서는 LongT5의 기본 토큰 32,100개 전체를 사용하지 않고 레벨의 구성 요소를 나타내는 13가지 기호만을 제한적으로 토큰화하여 학습을 진행하였다. 이는 자연어 모델의 방대한 어휘를 그대로 사용할 경우, 불필요한 계산 비용이 발생하고 레벨 구성과 무관한 토큰이 포함되어 학습 효율이 저하될 수 있기 때문이다. 그러나, 토큰 수를 제한함으로써 사전 학습된 언어 모델이 보유한 풍부한 언어적 의미가 소실되는 문제가 발생할 수 있다. 이를 보완하기 위해 본 연구에서는 LongT5 모델과 RoBERTa 모델을 결합한 아키텍처를 제안하였다. 전체적인 구조는 <Figure 3>과 같다. 제안된 아키텍처에서 LongT5는 레벨의 구성 요소를 생성하는 역할을 수행하며, RoBERTa는 자연어 프롬프트를 입력 받아 조건 정보를 담은 임베딩 벡터를 산출한다. RoBERTa를 LongT5와 함께 학습할 경우, 이미 사전 학습된 RoBERTa에 내포된 일반적인 의미 구조가 데이터셋의 편향에 의해 변형되거나 왜곡될 가능성이 존재하여 학습의 안정성과 효율성을 저해할 수 있다(Radford *et al.*, 2021). 따라서 RoBERTa는 사전 학습된 언어적 의미를 유지하기 위해 학습하지 않고 고정된 상태로 사용한다. 입력된 마리오 레벨은 LongT5의 임베딩 층을 거쳐 레벨 임베딩 벡터로 변환되었고, 동시에 입력된 자연어 프롬프트는 RoBERTa 모델을 거쳐 프롬프트 임베딩 벡터를 생성한다. 이 두 임베딩 벡터를 결합하여 LongT5 모듈의 입력으로 전달함으로써, 모델이 레벨의 구조적 정보와 자연어 프롬프트가 가진 의미적 조건 정보를 동시에 고려할 수 있도록 하였다. 이때 학습은 지도학습 방식으로 진행되며, 6가지 학습 과정에 따라 정답 레벨 시퀀스를 생성하고, 생성된 시퀀스와 정답 시퀀스 간의 토큰 단위 차이를 최소화하는 교사 강제(teacher forcing) 기반의 최대 우도 추정 방식으로 모델을 학습시킨다.

3.3 인간 피드백 기반 미세 조정

지도 학습 과정에서는 두 가지 주요 문제가 발생한다. 첫째, 모델이 조건 프롬프트를 정확히 반영하지 못해 레벨을 원하는 형태로 생성하는 데 어려움을 겪는다. 둘째, 플레이 불가능한 레벨을 학습하는 문제가 존재하여, 실제 플레이 가능한 구조를 안정적으로 생성하지 못하는 한계를 가진다. 이는 지도 학습 방법만으로는 조건 프롬프트에 따라 다양한 레벨을 수정하는 과정을 충분히 학습하기 어려우며, 생성된 레벨의 플레이 가능성을 고려하여 학습하지 않기 때문이다. 따라서 본 연구는 이러한 한계를 극복하기 위해, RLHF방법론을 도입하여 모델의 성능을 개선하고자 하였다.

본 연구에서 제안한 RLHF방법론은 <Figure 4>와 같이 구성된다. 먼저 <Figure 4(a)>와 같이, 레벨의 구성 요소를 학습한 LLM 기반 생성 모델에 동일한 조건 프롬프트와 레벨을 입력으로 제공하고, temperature 값을 조절하여 서로 다른 세 가지 레벨을 생성하였다. 이때 temperature 값은 모델이 다음 토큰을 선택할 때의 확률 분포를 조정하는 역할을 한다. Temperature 값을 낮추면 가장 확률이 높은 토큰들이 선택되어 예측 결과가 결정적이고 일관된 특성을 갖게 되며, 반대로 temperature 값을 높이면 토큰 선택 과정에 무작위성이 더해져 다양하고 창의적인 결과가 생성된다. 생성된 세 가지 레벨 중에서 인간 평가자가 가장 적합한 레벨을 하나 선택한 뒤, 조건 프롬프트를 만족하면서 실제 플레이가 가능하도록 수정하였다. 이러한 인간 피드백 과정을 반복 수행하여 인간 피드백 데이터를 축적하고, 수정된 데이터를 선호 레벨(chosen level)로, 수정되지 않은 데이터를 비선호 레벨(rejected level)로 하여 인간 선호 데이터셋을 구축하였다.

구축된 인간 선호 데이터셋을 기반으로, 본 연구에서는 두 가지 RLHF 방법론인 DPO와 PPO를 사용하여 모델을 미세 조정하였다. 먼저, DPO 방법론을 사용한 MarioPref-DPO는 <Figure 4(b)>와 같이 별도의 보상 함수 없이 인간 선호 데이터를 직접 사용하여, 선호 레벨의 생성 확률은 높이고 비선호 레벨의 생성 확률을 낮추도록 모델을 미세 조정한다. 반면, PPO 방법론을 사용한 MarioPref-PPO는 <Figure 4(c)>와 같이 인간 선호 데이터를 기반으로 레벨 품질을 예측하는 보상 모델을 먼저 학습한다. 이후 해당 보상 모델이 산출한 신호를 활용해 정책을 반복적으로 업데이트하는 방식으로 미세 조정이 이루어진다. 이와 같은 인간 피드백 기반 미세 조정을 통해 모델이

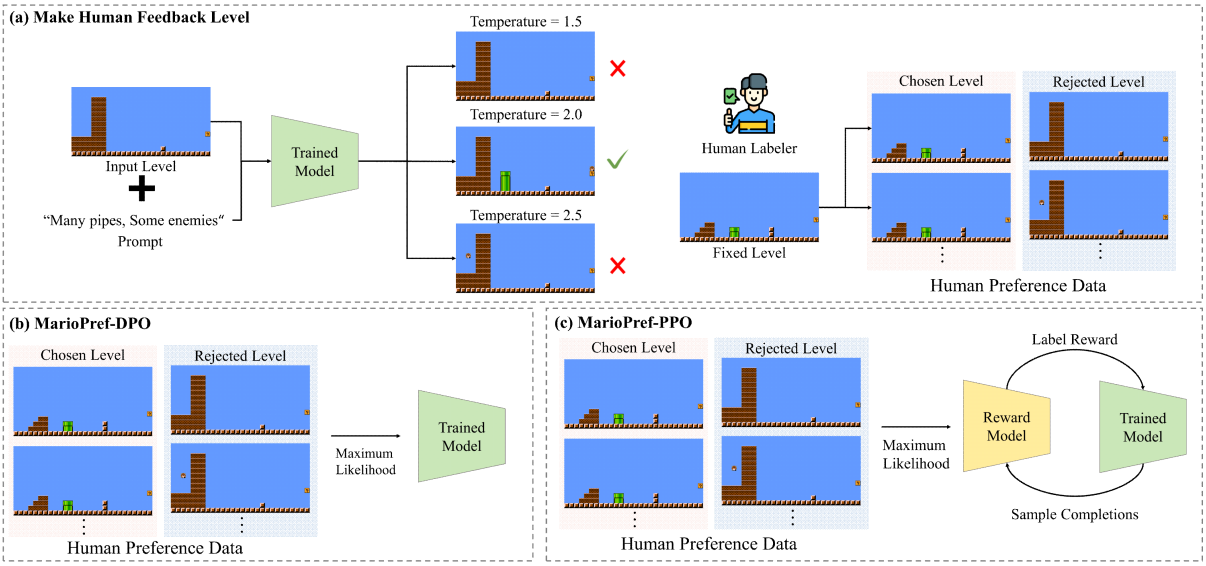


Figure 4. Overall Architecture of the Fine-tuning based on Human Feedback. (a) the process of constructing human preference data; (b) the model fine-tuning using DPO; and (c) the model fine-tuning through PPO based on a reward function

조건 프롬프트를 보다 정확히 반영하는 동시에, 실제 플레이 가능한 레벨을 생성할 수 있도록 유도하였다.

4. 실 험

4.1 실험 설계

본 연구에서는 MarioGPT를 활용하여 총 3,000개의 학습 데이터를 구축하였다. 이 과정에서는 temperature 값을 조정함으로써, 700~1,200 토큰 길이를 가진 다양한 저품질의 레벨을 생성하였다. MarioPref 모델은 생성된 학습 데이터를 바탕으로, 자연어 조건 프롬프트와 마스킹 된 레벨 데이터를 입력 받아 주어진 조건에 맞는 레벨을 복원하도록 지도 학습 방식으로 학습된다. 이후 RLHF 절차를 위해 지도학습 된 MarioPref 모델에 조건 프롬프트와 MarioGPT가 생성한 1,000개 입력 데이터를 입력하고, temperature 값을 1.5, 2.0, 2.5로 조정하여 각 조건별로 총 3,000개의 레벨을 추가로 생성하였다. 동일한 조건에서 생성된 세 가지 레벨 중, 인간 평가자가 한 개의 레벨을 선택하고 프롬프트 조건을 만족하고 플레이 가능하도록 수정하였다. 인간 피드백 데이터는 슈퍼 마리오 브로스 게임에 대한 도메인 지식을 보유한 네 명의 공저자에 의해 생성되었다. 각 공저자는 250개의 맵을 수정하였고, 이를 통해 총 1,000개의 인간 피드백 데이터를 구축하였다. 1,000개의 인간 피드백 데이터를 기존의 수정되지 않은 데이터와 결합하여 총 2,000개의 인간 선호도 데이터셋을 구축하였다. 이렇게 구축된 데이터셋을 기반으로 PPO와 DPO 방법론을 적용하여 미세 조정을 수행하였다.

본 연구에서는 앞서 구축한 인간 선호 데이터셋을 통해 미세 조정한 MarioPref-PPO와 MarioPref-DPO가 조건 프롬프트를 반영하여 실제 플레이 가능한 레벨을 생성할 수 있는지 검증하기 위한 실험을 수행하였다. 해당 실험에서는 학습에 사용되지 않은 MarioGPT가 생성한 별도의 40개 레벨을 평가 데이터로 사용하였으며, 파이프 조건 기준으로 No, Little, Some, Many의 4가지 조건을 각각 10개씩 균형 있게 구성하였다. 프롬프트 조건은 <Table 2>와 같이 설정하였다. 각 평가 레벨 대해 No, Little, Some, Many의 네 가지 파이프 조건과 네 가지 적 조건을 서로 조합하여 총 16가지 프롬프트 조건을 구성하였고, 이를 평가 레벨과 함께 모델의 입력으로 제공하였다. 결과적으로 각 레벨마다 16회씩 총 640회의 평가 실험을 수행함으로써, 제안된 방법론이 프롬프트 조건에 따라 레벨 구성 요소를 얼마나 효과적으로 생성하는지 정량적으로 분석하였다.

Table 2. Number of pipes and enemies with four prompt conditions (No, Little, Some, Many)

	No	Little	Some	Many
Number of pipes	0	1	2~3	4~5
Number of enemies	0	1	2~3	4~5

4.2 평가 지표

본 연구에서는 제안한 모델의 성능을 평가하기 위해 개선도(improvement score) 와 유효 생성률(playability)을 평가지표로 사용하였다. 개선도는 입력된 레벨이 조건 프롬프트에 따라 얼마나 효과적으로 생성되었는지를 정량적으로 평가하는 지표로, 식 (7)과 같이 정의된다. 여기서, I_v(initial value)는 입력 레벨의 파이프와 적의 조건 값을, T_v(target value)는 프롬프트

가 요구하는 목표 값을, G_v(generated value)는 모델이 생성한 레벨의 파이프와 적의 조건 값을 의미한다. 이 식을 통해 모델이 프롬프트 조건을 얼마나 정확하게 반영했는지를 입력 레벨과 비교하여 상대적인 개선 정도를 측정할 수 있다.

$$\text{Improvement score}(\%) = \left(1 - \frac{|I_v - G_v|}{|I_v - T_v|}\right) \times 100 \quad (7)$$

유효 생성률은 모델이 생성한 레벨 중 실제로 플레이 가능한 레벨의 비율을 나타내는 지표로써, 식 (8)과 같이 정의된다. 여기서 $N_{playable}$는 모델이 생성한 레벨 중 실제로 플레이 가능한 환경의 개수이며, N_{total}은 모델이 생성한 전체 환경의 수를 나타낸다. 플레이 가능 여부는 슈퍼마리오 브로스 플레이에 높은 성능을 보이는 A* 에이전트를 활용하여 측정하였다 (Togelius *et al.*, 2010). A* 에이전트가 끝까지 플레이할 수 있는 레벨의 비율을 유효 생성률로 산정함으로써, 실제 게임 내에서 생성된 레벨이 정상적으로 플레이할 수 있는지를 객관적으로 평가하였다.

$$Playability(\%) = \frac{N_{playable}}{N_{total}} \times 100 \quad (8)$$

본 연구는 위의 두 가지 평가지표를 활용하여 제안된 모델이 입력된 프롬프트 조건을 얼마나 효과적으로 반영하는지와 함께, 실제 게임 환경에서 플레이 가능성한 콘텐츠를 생성할 수 있는지를 종합적으로 검증하였다.

4.3 실험 결과

본 연구에서는 MarioPref가 효과적으로 레벨을 생성할 수 있는지를 검증하였다. 이를 위해, 지도 학습 기반 SFT, RoBERTa 프롬프트 임베딩과 PPO를 결합한 MarioPref-PPO, 그리고 RoBERTa 프롬프트 임베딩과 DPO를 결합한 MarioPref-DPO 세 가지 모델의 성능을 비교한 실험 결과를 제시하였다.

<Table 3>은 자연어 프롬프트 조건별로 생성된 레벨의 파이프와 적에 대한 개선도를 나타낸 결과이다. 실험 결과, SFT는 평균 64.23%의 개선도를 보였으나, 특히 구성 요소가 많이 요구되는 Some과 Many 조건에서는 파이프와 적을 충분히 생성하지 못하는 한계를 보였다. 이에 비해 MarioPref-DPO은 모든 조건에서 성능 향상을 보였으며, 특히 기존 모델이 어려움을 겪었던 Some과 Many 조건에서도 높은 개선도를 기록하였다. 이를 통해, 프롬프트 조건에 맞춰 레벨 구성 요소를 효과적으로 생성하는 데 인간 피드백 기반 방법론이 유효함을 확인할 수 있었다. 그러나 MarioPref-PPO은 MarioPref-DPO에 비해 상대적으로 낮은 성능 개선을 보였다. PPO 방식은 보상 함수를 별도로 학습하는 과정이 필요하나, 1,000개의 제한적인 선호 데이터만으로는 프롬프트 조건의 만족 여부를 정확히 반영한 보상 모델을 학습하는 데 한계가 있었고, 이러한 보상 모델의 불완전한 학습이 최종 성능 저하로 이어졌다.

<Table 4>는 생성된 레벨이 실제 플레이 가능한지를 평가하기 위해 유효 생성률을 비교한 실험 결과를 제시한다. 이를 통해 인간 피드백 기반 미세 조정 방법론이 생성된 레벨의 플레이 가능성을 향상시키는 데 얼마나 효과적인지를 분석하였다. 실험 결과, MarioPref-DPO는 66.25%의 가장 높은 유효 생성률을 기록하였으며, 이는 SFT대비 약 5% 향상된 수치였다. 이러한 결과는 인간 피드백 기반 방법론이 실제 플레이 가능한 레벨을 생성하는 데 효과적임을 입증한다. 반면, MarioPref-PPO는 60.00%의 유효 생성률을 기록하며 오히려 SFT모델보다도 낮은 성능을 보였다. 개선도 실험 결과와 마찬가지로, MarioPref-PPO는 보상 함수를 통해 간접적으로 플레이 가능성을 학습해야 하는 방식이기 때문에, 1,000개의 인간 피드백 데이터만으로는 보상 모델을 충분히 안정적으로 학습시키지 못하였고, 이로 인해 성능 향상도 제한적이었다. 반면 MarioPref-DPO은 별도의 보상 함수를 거치지 않고 인간의 선호 데이터를 직접 모델에 전달하여 학습하는 방식이기 때문에 상대적으로 적은 데이터만으로도 인간의 선호를 효과적으로 반영할 수 있었다. 또한, 인간이 직접 수정한 레벨을 명시적으로 선호하도록 미세 조정함으로써 플레이 가능한 레벨과 그렇지 않은 레벨 사이의 차이를 보다 정밀하게 학습할 수 있었으며, 결과적으로 DPO방법론은 유의미한 수준의 유효 생성률 향상을 달성하였다.

Table 3. Comparison of pipe and enemy improvement scores (%) by prompt conditions: supervised fine tuning (without prompt embedding), MarioPref-PPO, and MarioPref-DPO

Model		Prompt				
		No	Little	Some	Many	ALL
Pipe Improvement Score (%)	SFT	98.33	75.41	46.25	36.94	64.23
	MarioPref-PPO	87.08	83.33	73.75	61.11	76.31
	MarioPref-DPO	**99.72**	**95.00**	**85.00**	**72.22**	**87.98**
Enemy Improvement Score (%)	SFT	78.88	66.42	52.97	54.60	63.11
	MarioPref-PPO	83.07	70.71	56.54	50.65	65.24
	MarioPref-DPO	**100.00**	**96.42**	**77.97**	**94.53**	**92.23**

Table 4. Comparison playability (%): supervised fine tuning (without prompt embedding), MarioPref-PPO and MarioPref-DPO

Model	Playability
SFT	61.25
MarioPref-PPO	60.00
MarioPref-DPO	**66.25**

<Figure 5>는 본 연구에서 제안한 MarioPref-DPO가 주어진 프롬프트 조건에 따라 실제로 플레이 가능한 레벨을 효과적으로 생성하는지를 시각적으로 제시한다. <Figure 5(a)>는 플레이 가능한 입력 레벨과 조건 프롬프트 “Many pipes, Some enemies”를 입력으로 제공하였을 때의 결과를 보여준다. 기존 지도 학습 모델은 해당 조건에 따라 충분한 수의 파이프와 적을 생성하지 못하는 한계를 보인 반면, MarioPref-DPO는 프롬프트 조건을 효과적으로 반영한 레벨을 생성함을 확인할 수 있었다. 또한, <Figure 5(b)>는 플레이가 불가능한 구조의 입력 레벨과 프롬프트 “Some pipes, Many enemies”를 입력으로 제공하였을 때의 결과를 보여준다. SFT 은 프롬프트 조건을 만족하는 레벨을 생성하지 못 할 뿐만 아니라, 여전히 플레이 불가능한 구조가 유지되는 한계를 보였다. 그러나 MarioPref-DPO는 플레이어가 실제 게임 내에서 통과할 수 없는 구조를 효과적으로 수정하고, 동시에 주어진 조건에 부합하는 레벨을 생성하는 데 성공하였다. 이러한 결과를 통해 MarioPref-DPO 모델이 단순 지도 학습 기반 모델 대비 프롬프트 조건을 보다 정확하게 반영할 뿐만 아니라, 플레이 가능한 레벨을 생성하는 데 더욱 효과적임을 확인하였다.

높은 개선도와 높은 유효 생성률을 나타낸 MarioPref-DPO를 대상으로, RoBERTa를 활용하여 언어적 의미를 반영한 프롬프트 임베딩의 효과를 검증하기 위한 추가 실험을 수행하였다. <Table 5>와 <Table 6>은 프롬프트 임베딩 방법론이 성능에 미치는 효과를 개선도 및 유효 생성률을 통해 비교한 결과이다. <Table 5>의 실험 결과에서 프롬프트 임베딩 방법론만 단독으로 적용한 MarioPref(W/O DPO)는 파이프의 평균 개선도에서 유의미한 향상을 보였으나, 적 개선도는 오히려 감소하는 결과를 보였다. DPO방법론만 단독으로 적용한 MarioPref-DPO(W/O PE)는 SFT 모델 대비 성능이 일부 향상되었으나, 그 정도가 미미하였다. 이는 DPO방법론만으로는 프롬프트가 가진 언어적 의미를 충분히 반영하지 못해, 프롬프트 조건을 효과적으로 구현하는데 한계가 있음을 의미한다. 반면, DPO방법론과 RoBERTa 기반 프롬프트 임베딩을 함께 적용하였을 때는 성능이 현저히 향상되었으며, 이는 두 방법론이 상호 보완적인 역할을 수행하여 프롬프트 조건을 보다 정확하고 효과적으로 반영하는 데 기여함을 시사하였다.

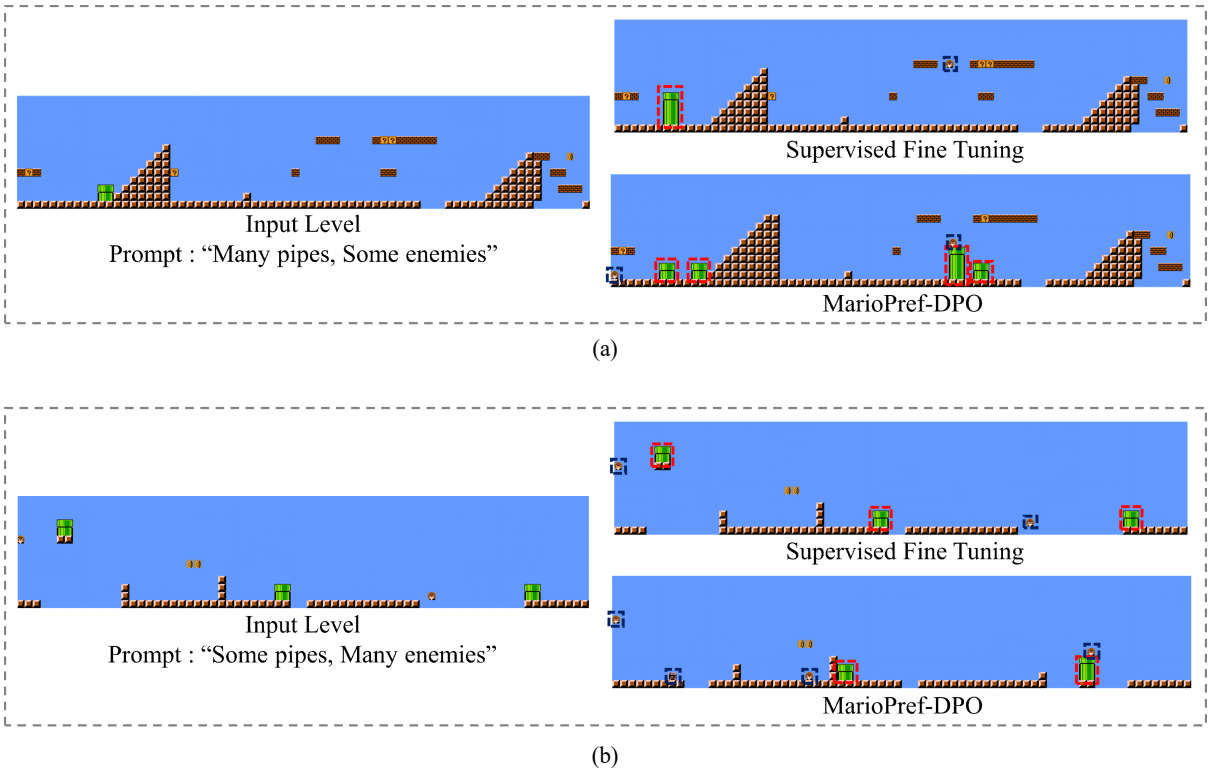


Figure 5. Comparison of Generated Mario Levels: Supervised Fine Tuning vs. MarioPref-DPO. The figure illustrates comparisons for two different prompts and input map. For each prompt, the left image shows the input level, the top-right image displays the level generated by the supervised fine-tuning model, and the bottom-right image presents the level generated by MarioPref-DPO

Table 5. Ablation Study on Prompt Embedding and DPO: Pipe And Enemy Improvement Scores (PE: prompt embedding)

Model		Prompt				
		No	Little	Some	Many	ALL
Pipe Improvement Score (%)	SFT	98.33	75.41	46.25	36.94	64.23
	MarioPref (W/O DPO)	84.30	77.91	68.33	56.38	71.73
	MarioPref-DPO (W/O PE)	99.58	81.66	79.58	34.58	65.72
	MarioPref-DPO	**99.72**	**95.00**	**85.00**	**72.22**	**87.98**
Enemy Improvement Score (%)	SFT	78.88	66.42	52.97	54.60	63.11
	MarioPref (W/O DPO)	72.52	73.57	63.09	37.36	61.64
	MarioPref-DPO (W/O PE)	85.67	73.21	70.23	70.83	74.99
	MarioPref-DPO	**100.00**	**96.42**	**77.97**	**94.53**	**92.23**

<Table 6>의 실험 결과에 따르면, LongT5모델에 프롬프트 임베딩 방법론만 단독으로 적용한 MarioPref(W/O DPO)는 오히려 SFT보다 낮은 58.90%의 유효 생성률을 보였다. 반면, DPO방법론을 추가로 적용한 MarioPref-DPO은 플레이 불가능한 레벨을 플레이 가능하도록 개선한 결과, 유효 생성률이 66.25%로 크게 향상되었다. 이는 DPO 로 미세 조정만을 수행한 MarioPref-DPO(W/O PE)의 유효 생성률 64.37%보다도 높은 수치였다. 이러한 결과는 프롬프트 임베딩과 인간 피드백 기반 미세조정이 함께 사용될 때, 프롬프트 조건을 보다 정확하게 반영할 뿐만 아니라 실제로 플레이 가능한 안정적인 레벨을 생성하는 데에도 효과적임을 입증하였다.

Table 6. Ablation Study on Prompt Embedding and DPO: Playability (PE: prompt embedding)

Model	Playability
SFT	61.25
MarioPref (W/O DPO)	58.90
MarioPref-DPO (W/O PE)	64.37
MarioPref-DPO	**66.25**

앞선 실험 결과를 통해 인간 피드백 데이터의 활용이 개선도와 유효 생성률 향상에 효과적으로 기여함을 확인하였다. 이에 따라 본 연구에서는 MarioPref-DPO 모델에 입력되는 인

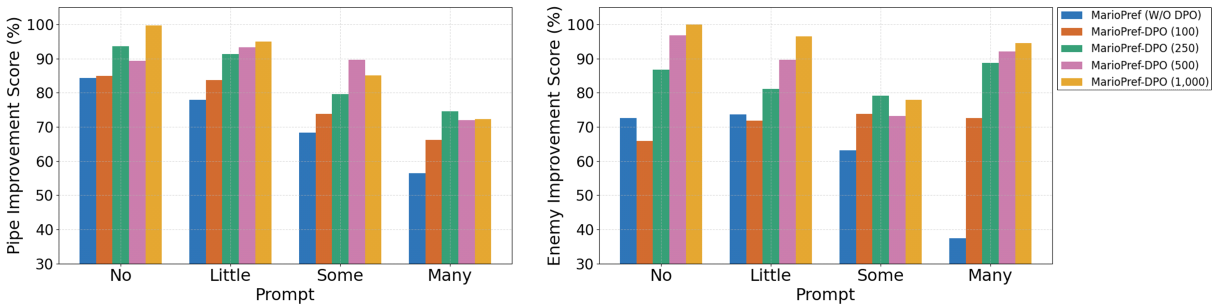


Figure 6. Comparison of Improvement Scores of MarioPref-DPO based on the Amount of Human Feedback Data

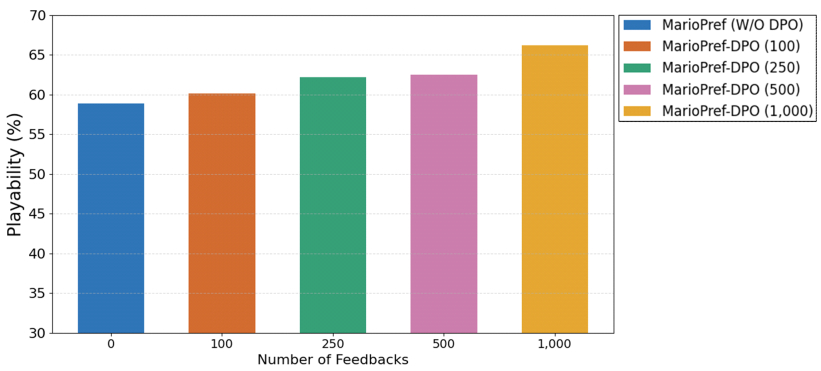


Figure 7. Impact of Human Feedback Data Size on the Playability of MarioPref-DPO

간 피드백 데이터의 양이 증가함에 따라 모델 성능에 어떤 변화가 발생하는지 확인하기 위한 추가 실험을 진행하였다. <Figure 6>은 인간 피드백 데이터의 개수 변화에 따른 개선도의 비교 결과를 나타낸 것이다. 실험 결과, 인간 피드백 데이터의 양이 증가할수록 전반적인 개선도가 점진적으로 향상되는 경향을 보였다. 특히 인간 피드백 데이터의 개수가 100개에서 1,000개로 증가함에 따라 평균 개선도가 지속적으로 증가했으며, 단 100개의 피드백 데이터만을 사용한 경우에도 DPO 방법론을 사용하지 않은 모델에 비해 대부분의 조건에서 개선된 성능을 보였다. 또한, <Figure 7>은 인간 피드백 데이터의 개수에 따른 유효 생성률의 비교 결과를 나타냈으며, 인간 피드백 데이터의 개수가 증가함에 따라 유효 생성률도 증가하는 경향을 확인할 수 있었다. 이러한 결과는 인간 피드백 데이터가 증가할수록 모델 성능 개선에 실질적인 영향을 미치며, 인간 피드백 데이터를 적극적으로 활용할수록 레벨 생성 모델의 품질을 향상시킬 수 있음을 시사하였다.

5. 결 론

본 연구에서는 슈퍼 마리오 브로스 레벨과 자연어 프롬프트를 입력 받아, 프롬프트 조건을 반영한 레벨을 생성하는 언어 모델 기반의 MarioPref을 제안하였다. 특히, 저품질의 학습 데이터 셋을 사용함에도 불구하고 프롬프트 정보를 효과적으로 반영할 수 있는 RoBERTa와 RLHF방법론인 DPO와 PPO를 활용하여, 플레이 가능하고 프롬프트 조건을 만족하는 레벨을 성공적으로 생성하였다. 실험 결과, 지도 학습과 DPO 기반 미세조정을 거친 언어 모델이 생성한 레벨은 개선도와 실제 플레이 가능성 측면에서 모두 유의미한 성능 개선을 보였으며, 이를 통해 제안된 프레임워크의 효과를 입증하였다.

본 연구는 고품질의 데이터 확보가 현실적으로 어려운 게임 개발 환경에서 인간 피드백의 적극적인 활용이 레벨의 품질을 효과적으로 향상할 수 있음을 보여주었다는 점에서 의의가 크다. 나아가, 인간과 인공지능의 협력적 접근 방식이 데이터 제약을 극복하는 데 실질적인 해법이 될 수 있음을 제시하였다. 다만, 생성된 레벨을 기반으로 인간 피드백 데이터를 구축하는 과정에는 직접적인 맵 수정이 요구되며, 이로 인해 시간과 비용이 소요되는 한계가 존재한다. 또한, 본 연구는 구조적 요소가 비교적 명확한 슈퍼 마리오 브로스 환경을 대상으로 진행되었기 때문에, 제안된 프레임워크의 다양한 게임 장르에 대한 일반화 가능성은 추가적인 검증이 필요하다. 따라서 향후 연구에서는 보다 적은 양의 피드백으로도 효과적인 학습이 가능한 알고리즘 개발과 함께, 플레이 로그와 같은 간접적인 피드백 신호를 활용한 학습 전략에 대한 탐구가 필요하다. 나아가 다양한 게임 환경과 장르에 본 프레임워크를 적용함으로써, 범용성과 안정성에 대한 체계적인 평가가 필요하다. 이러한 방향은 인간 피드백의 효율성을 극대화함과 동시에, 현실적인 개발 환경에서도 높은 품질의 게임 콘텐츠를 생성할 수 있는 기반 기술로 이어질 것으로 기대된다.

참고문헌

An, G., Lee, J., Zuo, X., Kosaka, N., Kim, K.-M., and Song, H. O. (2023), Direct preference-based policy optimization without reward modeling, *Advances in Neural Information Processing Systems*, **36**, 70247-70266.

Awiszus, M., Schubert, F., and Rosenhahn, B. (2020), TOAD-GAN: Coherent style level generation from a single example, *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, **16**(1), 10-16.

Bradley, R. A. and Terry, M. E. (1952), Rank analysis of incomplete block designs: I. The method of paired comparisons, *Biometrika*, **39**(3/4), 324-345.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017), Deep reinforcement learning from human preferences, *Advances in Neural Information Processing Systems*, **30**.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019), Bert: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171-4186.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014), Generative adversarial nets, *Advances in Neural Information Processing Systems*, **27**.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y. H., and Yang, Y. (2022, July), LongT5: Efficient Text-to-Text Transformer for Long Sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 724-736.

Hejna, J. and Sadigh, D. (2023), Inverse preference learning: Preference-based rl without a reward function, *Advances in Neural Information Processing Systems*, **36**, 18806-18827.

Hendrikx, M., Meijer, S., Van Der Velden, J., and Iosup, A. (2013), Procedural content generation for games: A survey, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **9**(1), 1-22.

Hu, C., Zhao, Y., and Liu, J. (2024), Game generation via large language models, In *2024 IEEE Conference on Games (CoG)*, IEEE, 1-4.

Liu, J., Snodgrass, S., Khalifa, A., Risi, S., Yannakakis, G. N., and Togelius, J. (2021), Deep learning for procedural content generation, *Neural Computing and Applications*, **33**(1), 19-37.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019), Roberta: A robustly optimized bert pretraining approach, ArXiv Preprint ArXiv:1907.11692.

Nasir, M. U. and Togelius, J. (2023), Practical PCG through large language models, *2023 IEEE Conference on Games (CoG)*, 1-4.

Oranchak, D. (2010), Evolutionary algorithm for generation of entertaining shinro logic puzzles, Applications of Evolutionary Computation: EvoApplicatons 2010: EvoCOMPLEX, EvoGAMES, EvoIASP, EvoINTELLIGENCE, EvoNUM, and EvoSTOC, Istanbul, Turkey, April 7-9, 2010, *Proceedings, Part I*, 181-190.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano,

P. F., Leike, J., and Lowe, R. (2022), Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems*, 35, 27730-27744.

Peng, X. Bin, Kumar, A., Zhang, G., and Levine, S. (2019), Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, ArXiv Preprint ArXiv:1910.00177.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021), Learning transferable visual models from natural language supervision, In *International Conference on Machine Learning*, PmLR, 8748-8763.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019), Language models are unsupervised multitask learners, *OpenAI Blog*, **1**(8), 9.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023), Direct preference optimization: Your language model is secretly a reward model, *Advances in Neural Information Processing Systems*, **36**, 53728-53741.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020), Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research*, **21**(140), 1-67.

Shaker, N., Togelius, J., and Nelson, M. J. (2016), Procedural content generation in games, *Springer International Publishing*.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020), Learning to summarize with human feedback, *Advances in Neural Information Processing Systems*, **33**, 3008-3021.

Sudhakaran, S., González-Duque, M., Freiberger, M., Glanois, C., Najarro, E., and Risi, S. (2023), Mariogpt: Open-ended text2level generation through large language models, *Advances in Neural Information Processing Systems*, **36**, 54213-54227.

Summerville, A. J., Snodgrass, S., Mateas, M., and Ontanón, S. (2016). The vglc: The video game level corpus, ArXiv Preprint ArXiv:1606.07487.

Summerville, A., Snodgrass, S., Guzdial, M., Holmgård, C., Hoover, A. K., Isaksen, A., Nealen, A., and Togelius, J. (2018), Procedural content generation via machine learning (PCGML), *IEEE Transactions on Games*, **10**(3), 257-270.

Togelius, J., Karakovskiy, S., and Baumgarten, R. (2010), The 2009 mario ai competition, *IEEE Congress on Evolutionary Computation*, 1-8.

Togelius, J., Yannakakis, G. N., Stanley, K. O., and Browne, C. (2011), Search-based procedural content generation: A taxonomy and survey, *IEEE Transactions on Computational Intelligence and AI in Games*, **3**(3), 172-186.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, **30**.

Volz, V., Schrum, J., Liu, J., Lucas, S. M., Smith, A., and Risi, S. (2018), Evolving mario levels in the latent space of a deep convolutional generative adversarial network, *Proceedings of the Genetic and Evolutionary Computation Conference*, 221-228.

저자소개

이준범 : 한양대학교 산업경영공학과에서 2024년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석사과정에 재학 중이다. 연구 분야는 Reinforcement Learning이다.

김정인 : 가톨릭대학교 정보통신전자공학부에서 2020년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석박통합과정에 재학 중이다. 연구 분야는 Reinforcement Learning, Self-Supervised Learning이다.

허종국 : 고려대학교 산업경영공학부에서 2021년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석박통합과정에 재학 중이다. 연구분야는 Self-Supervised Learning, Reinforcement Learning이다.

이정민 : 한양대학교 산업공학과에서 2022년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석박통합과정에 재학 중이다. 연구 분야는 Applications of Uncertainty Quantification, Large Language Models이다.

김재훈 : 동국대학교 경영학과에서 2019년 학사 학위를 취득하고, 고려대학교 산업경영공학과에서 석박통합과정에 재학 중이다. 연구 분야는 Reinforcement Learning이다.

김성범 : 고려대학교 산업경영공학부 교수로 2009년부터 재직하고 있으며, 인공지능공학연구소 소장, 기업산학연협력센터 센터장, 한국데이터마이닝학회 회장을 역임했다. 미국 University of Texas at Arlington 산업공학과에서 교수를 역임하였으며, 한양대학교 산업공학과에서 학사학위를 미국 Georgia Institute of Technology에서 산업시스템공학 석사 및 박사학위를 취득하였다. 인공지능, 머신러닝, 최적화 방법론을 개발하고 이를 다양한 공학, 자연과학, 사회과학 분야에 응용하는 연구를 수행하고 있다.