챗 가능한 게임에서 AI 캐릭터와의 대화를 위한 LLM 활용

최명재[1], 신지호[1], 이세영[1], 정동주[2], 이병정[1]
[1]서울시립대학교 컴퓨터과학부
[2](주) 스마트잭
ssy07124@uos.ac.kr, sjm010529@gmail.com, ocdee39@gmail.com, lostcode7@gmail.com,
bjlee@uos.ac.kr

Utilizing LLM for Conversations with AI Character in Chat-enabled Games

Myoung-Jae Choi[1], Ji-Ho Shin[1], Se-Yeong Lee[1],
Dong-Ju Jung[3], Byung-Jeong Lee[2]
[1]Dept. of Computer Science and Engineering, University of Seoul
[2]Smart Jack Co., Ltd.

요 약

본 연구에서는 게임에서 자연스런 대화를 통해 스토리 몰입을 제공하는 AI 캐릭터를 위한 LLM 활용을 소개한다. 사용자는 게임 속 AI 캐릭터와 대화하며 스토리를 이어간다. 게임 속에서 사용자는 정해진 대사를 선택할 수도 있고, AI 캐릭터와 대화할 때 직접 대사를 입력할 수도 있다. 대사를 입력하면 그에 맞는 AI 캐릭터의 답변이 제공되고, 앞으로의 스토리에도 영향을 미친다. 결론적으로 LLM 기반 AI 캐릭터와의 자연스러운 대화를 통해 게임의 몰입도와 접근성을 높이는 것이 본 연구의 목표이다.

1. 서론

근래 LLM(Large Language Model)의 발전으로 챗봇의 성능이 비약적으로 향상되었다. 그에 따라 게임 캐릭터에 AI 챗봇을 도입하여 보다 생생한 반응을 이끌어내자는 움직임이 나타났고, 여러 게임 업체에서 AI 캐릭터의 개발을 발표하거나 프로토타입을 선보였다. 하지만 사용자에게 무한한 자유를 주었을 때 발생할 수 있는 다양한 변칙 상황의 고려와 이전 대화와의 일관성 유지 등에서 어려움을 겪고 있다. 또한 AI 캐릭터가 지난 대화를 기억하지 못하거나 게임의 세계관과 어울리지 않는 환각적인 정보를 제공하는 문제도 존재한다[1]. 이처럼 사용자의 다양한 대화 전개에 따라 맥락에 어울리지 않는 반응이나 질 낮은 대답이 돌아오게 된다면 오히려 게임의 몰입을 방해하는 악영향을 줄 수 있기에 상용화에 걸림돌이 된다.

따라서 본 연구에서는 사용자와 AI 캐릭터의 대화를 자연스럽게 유도하고, 거시적으로는 맥락을 일정하게 유지하는 시스템을 제안한다. 이야기의 큰 흐름은 고정되게 주어지므로 메인 스토리의 진행을 벗어나지 않도록 방향성을 설정할 수 있으며, 이를 통해 답변의 변칙성과 관계없이 기획 의도대로 사용자를 이끄는 스토리를 형성할 수 있다. 더 나아가 이전의 스토리와 현재 세계관을 참고하여 대답을 생성하도록 설계하여 일관성을 보장한다. 이렇듯 기획된 캐릭터를 AI로 생성하는 방법을 제안하여, 스토리를 갖는 게임에서 해당 시스템을 사용할 수 있도록 한다.

2. 관련 연구

2.1 QLoRA (Quantized Low Rank Adapters)

QLoRA는 LoRA와 양자화 기법을 적용하여 매우 효율적인 파인튜닝이 가능하도록 한 기법이다. LoRA는 이미 학습이 완료된 LLM 모델의 파라미터를 고정하고, 학습 가능한 순위분할행렬(Rank Decomposition Matrix)을 트랜스포머(transformer) 구조에 삽입한다[2]. 모든 LLM의 파라미터를 학습하지 않고 추가한 작은 차원의 행렬만을 학습시킨다. QLoRA는 LoRA 구조의 역전파 과정을 진행하며 추가적으로 기존 LLM의 파라미터를 4비트 양자화하여 사용한다[3]. 본 연구에서는 AI 캐릭터의 성격과 말투를 학습시키기 위해 캐릭터 성격에 맞는 수 천개의 대사 데이터를 사용하여 오픈소스 Llama2-70B 모델을 QLoRA로 파인튜닝한다. QLoRA를 사용하면 풀-파인튜닝보다 훨씬 적은 자원으로 파인튜닝이 가능하다.

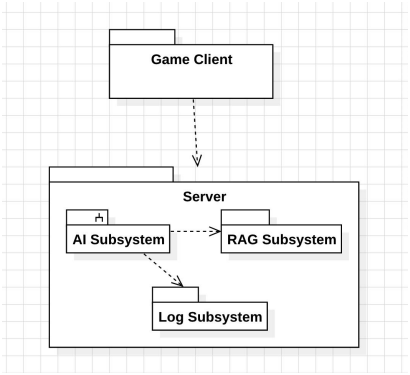
2.2 RAG (Retrieval-Augmented Generation)

RAG 모델은 검색 시스템과 생성 모델을 결합한 형태로, 정보 검색 시스템을 통해 획득한 지식을 활용하여 자연어 생성을 개선한다. RAG 모델은 검색기(Retriever)와 생성기(Generator)를 중심으로 구성되는데, 검색기는 Llama, Bert와 같은 언어 모델을 사용하여 질문(Query)와 문서(Document)를 벡터 공간에 임베딩 하여 유사도를 비교해 관련성이 높은 정보를 검색한다. 생성기는 이전의 검색된 정보를 입력으로 사용하여, 이 정보를 바탕으로 최종적인 답변을 생성한다. RAG 모델을 사용하게 되면 외부 데이터를 통합하여 지식 활용 능력을 향상시키고, 더 정확하고 섬세한 검색을 할 수 있게 된다[4]. 본 연구에서는 게임에 관한 다양한 정보들을 임베딩 하여 데이터베이스에 저장한 후, LLM 모델에게 AI 캐릭터의 답변 생성 요

구 프롬프트를 작성할 때 활용한다.

3. AI 캐릭터를 위한 LLM

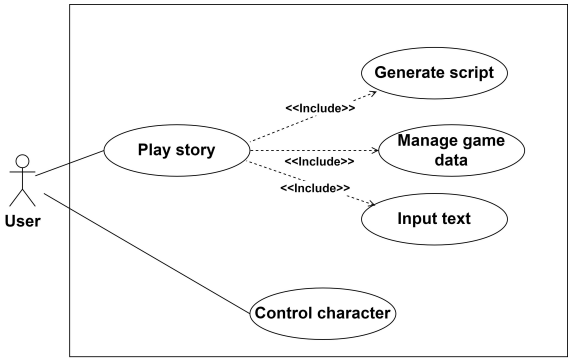
3.1 개요



(그림 1) 아키텍처도

그림 1은 본 연구의 아키텍처도이다. AI 서브시스템은 LLM을 추론하여 AI 캐릭터를 구현한다. 캐릭터의 성격과 말투를 학습하기 위해 수 천개의 대사 데이터를 QLoRA로 파인튜닝한 Llama2-70B 모델로 구축한다. RAG 서브시스템은 벡터 데이터베이스를 구동하는 서버로 AI 캐릭터가 문맥에 맞는 대답을 하기 위해 RAG 기법으로 추가 정보를 제공한다. AI 서브시스템은 RAG 서브시스템에서 받은 추가 정보를 프롬프트에 추가하여 답변을 추론한다. 로그 서브시스템은 RAG 서브시스템이 생성한 프롬프트와 AI 서브시스템이 생성한 답변을 로그로 저장한다.

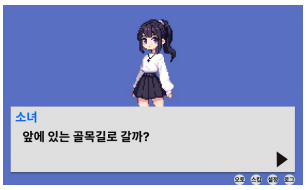
3.2 유스케이스도



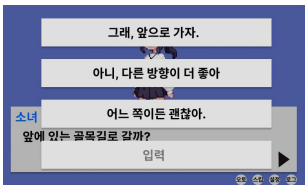
(그림 2) 유스케이스도

그림 2는 본 연구의 AI 캐릭터 기반 게임의 유즈케이스도이다. 본 게임에서 게임 사용자는 게임의 기본 기능으로서 게임 속 캐릭터를 조작할 수 있다(Control character). 또한 사용자는 게임 스토리를 진행할 수 있다(Play story). 스토리를 진행하면서, AI 캐릭터는 사용자에게 질문을 하는데, 이때 사용자는 3개의 정해진 대답을 고르거나, 직접 대답을 입력할 수 있다(Input text). 사용자의 대답과 데이터베이스에 저장된 게임 속 정보를 기반으로 LLM 모델은 사용자의 대답에 대한 AI 캐릭터의 반응을 생성하여 사용자에게 게임 속 화면을 통해 보여준다(Generate script). 또한 게임 품질의 높이기 위해 데이터베이스에 사용자가 입력한 대사 정보를 로그 데이터로 저장하고 관리하는 기능도 포함한다(Manage game data).

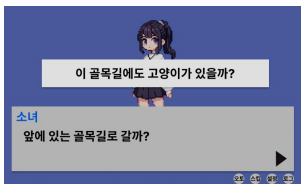
3.3 UI 설계



(그림 3) AI 캐릭터의 질문



(그림 4) 제공된 선택지



(그림 5) 선택지 입력

(그림 6) 입력에 대한 답변

이야기 진행 중 사용자는 그림 3과 같이 AI 캐릭터로부터 정해진 질문을 받게 되고, 그에 따른 답을 선택해야 한다. 그림 3에서 AI 캐릭터는 앞에 있는 골목길로의 진행 여부를 질문하고 있다. 이처럼 중요한 분기점에서는 그림 4과 같이 4개의 선택지가 제공된다. 세 개의 선택지는 긍정, 부정, 중립과 같이 특정한 방향성을 가지며, 마지막 선택지는 사용자가 직접 대사를 입력할 수 있게 한다. 사용자가 그림 5와 같이 “이 골목길에도 고양이가 있을까?”를 직접 입력했을 경우, 시스템은 RAG 서브시스템에서 ‘고양이’와 관련된 정보를 호출한다. RAG 서브시스템에는 스토리의 스크립트와 세계관 정보가 저장되어 있으므로, 해당 호출로 이전에 등장한 고양이 관련 에피소드 데이터가 검색된다. 사용자의 입력에 따라 이전 스토리의 내용이 선택적으로 프롬프트에 함께 전달되고, 결과적으로 AI 캐릭터가 그림 6과 같이 자연스러운 대사를 생성한다.

4. 결론

본 연구에서 제시한 AI 캐릭터를 도입할 경우 사용자가 능동적으로 캐릭터에게 대답하고 AI 캐릭터 또한 그에 맞춰서 사용자에게 다양한 답변을 하기 때문에 사용자는 더욱 몰입감있게 게임을 즐긴다. 또한, 게임 개발에서 시나리오 스크립트 작성에 소요되는 시간을 획기적으로 줄여 게임 개발 생산성을 향상시킬 것으로 기대한다.

참고문헌

[1] S. R. Cox and W. T. Ooi, "Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback," in Proc. of International Workshop on Chatbot Research and Design, pp. 167-184, 2023.
[2] E. Hu, et al. "Lora: Low-rank adaptation of large language models," in Proc. of International Conference on Learning Representations, 2021.
[3] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in Proc. of NeurIPS, 2023.
[4] A. Piktus, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. of NeurIPS, 2020.