

MoGLE : Mixture of Genre LoRA Expert

1조

소프트웨어학과 202020709 공종혁

소프트웨어학과 202020768 김지환

소프트웨어학과 202126864 노민혁

<https://github.com/minhyeok2001/MoGLE>

1.Motivation

1) TRPG(Tabletop Role-Playing Game)

- 대화를 통해 진행되는 아날로그 롤 플레이 게임
- 플레이어가 상상하는 모든 행동을 시도할 수 있는 자유도가 특징



< Fig 1. TRPG를 플레이하는 모습>



< Fig 2. TRPG D&D 플레이 세트>

Fig 1. <https://timharford.com/2025/11/whats-it-like-to-play-with-a-professional-dungeon-master/>

Fig 2. <https://naturebridge.kr/product/%EB%8D%98%EC%A0%84-%EC%95%A4-%EB%93%9C%EB%9E%98%EA%B3%A4-%EC%8A%A4%ED%83%80%ED%84%B0-%EC%84%B8%ED%8A%B8-5th-edition-dnd-%EC%8A%A4%ED%83%80%ED%84%B0-%ED%82%A4%ED%8A%B8-%EA%B2%80%EC%9D%80-%EA%B0%80%EB%B0%A9%EC%97%90-%EC%A3%BC%EC%82%AC%EC%9C%84-%EC%84%B1%EC%9D%B8%EC%9D%84%EC%9C%84%ED%95%9C-%EC%9E%AC%EB%AF%B8%EC%9E%88%EB%8A%94-dnd-%EB%A1%A4%EB%A7%81-%EB%B3%B4%EB%93%9C/73040>

1.Motivation

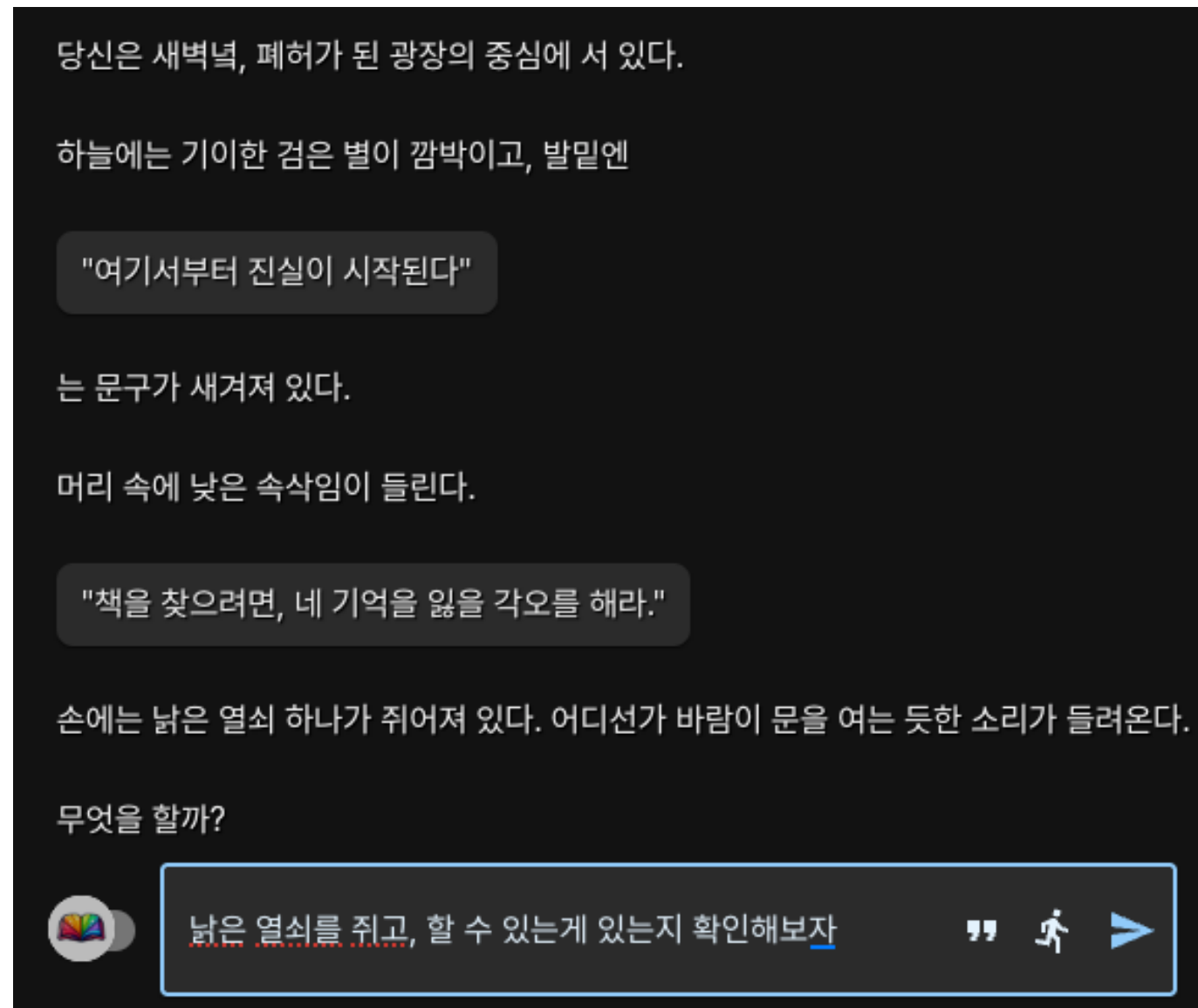
2) TRPG에서 GM(Game Master)의 역할

- 세계관 묘사, 규칙 판정, NPC 연기 등을 실시간으로 수행하는 TRPG의 운영자

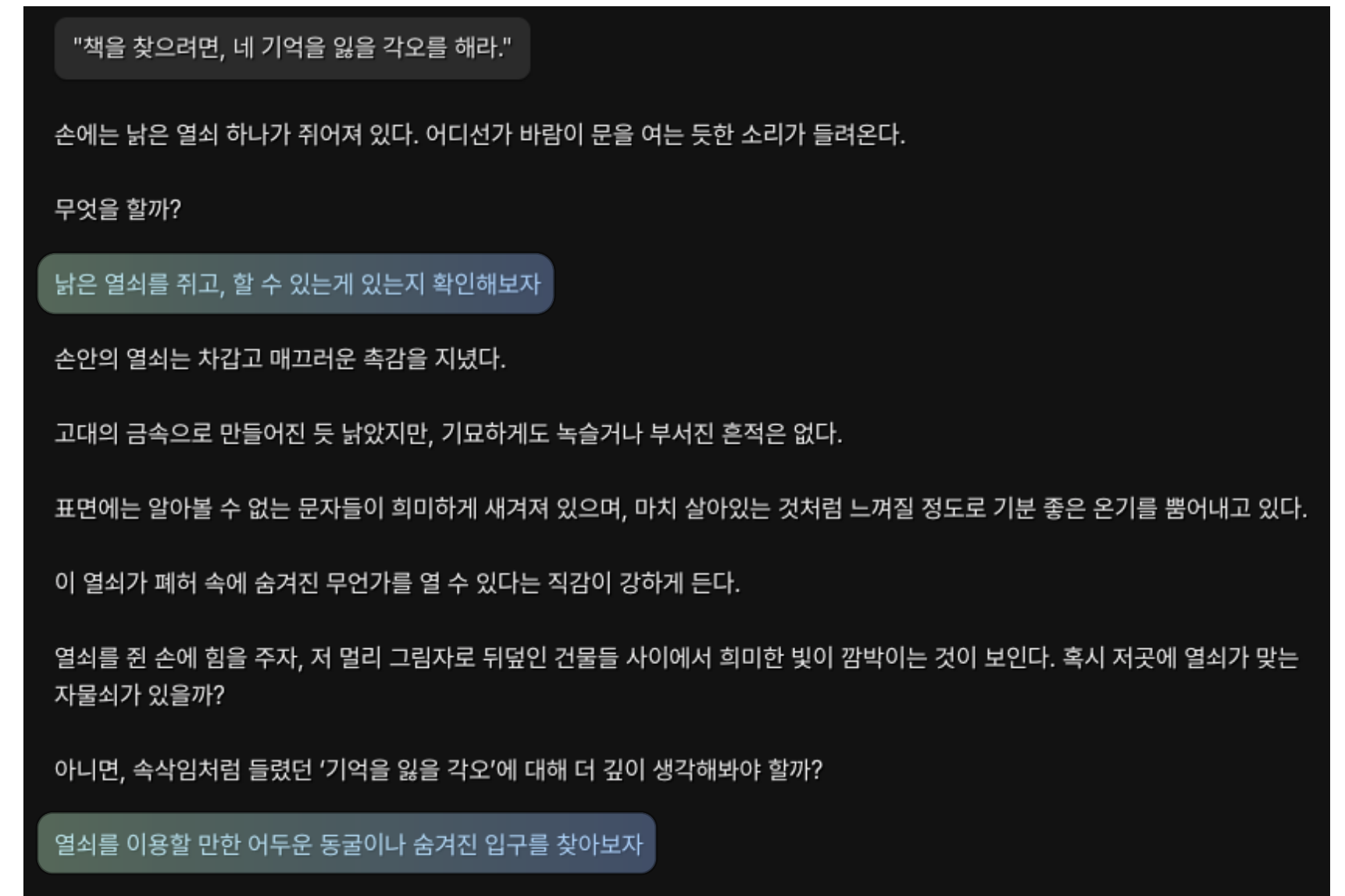
3) LLM based TRPG

- 인간 GM의 역할을 LLM이 대신 수행하는 형태
- 시간과 장소, 인원 수의 제약 없이 언제든지 ‘나만의 맞춤형 이야기’를 즐길 수 있음
- 장기 기억, 일관된 규칙 적용, 문체적 몰입감이 중요

1.Motivation



< Fig 1. GPTRPG를 플레이하는 모습>



< Fig 2. GPTRPG가 사람의 입력에 따라 장면 묘사, NPC연기, 규칙 판정을 하는 모습>

1.Motivation

4) LLM based TRPG의 문제점

- 장르에 맞지 않는 장면 묘사나 상황, 소재가 나오는 문제
- 사용자의 몰입감과 경험을 크게 해침

... For example, I tried my scenario and made a scream coming from the living room and when we investigate, the man was dead with a knife in his chest. It sounds more like a criminal set than zombie set.

→ ... 예를 들어, 내 시나리오를 한 번 돌려봤거든? 거실에서 비명 소리가 들리게 연출했어. 근데 막상 가서 보니까, 어떤 남자가 가슴에 칼이 꽂힌 채 죽어있는 거야. 이러면 좀비물이 아니라 완전 범죄 스릴러 같잖아.

< Fig 1. AIDungeon 플레이 중, 상황이 장르에 맞지 않는 문제를 토로하는 레딧 글 >

1.Motivation

5) 해결 방안 구상

(1) 장르 별로 Full Finetune하는 경우

- 장르 별로 모델을 필요할 때마다 개별 학습하는 것은 비용적, 현실적으로 어려움

→ LoRA 사용이 불가피

(2) 단일 LoRA로 모든 장르를 학습하여 사용하는 경우

- 여러 장르 데이터를 LoRA 하나에 모두 학습 시키면 스타일이 충돌하고 **성능이 떨어짐**

→ 장르 별 LoRA를 개별 학습하는 것이 최적

1.Motivation

5) 해결 방안 구상

(3) Base모델은 Freeze하고, 각 장르별로 각 LoRA를 학습하는 경우 (장르별로 각각 맞는 LoRA 사용)

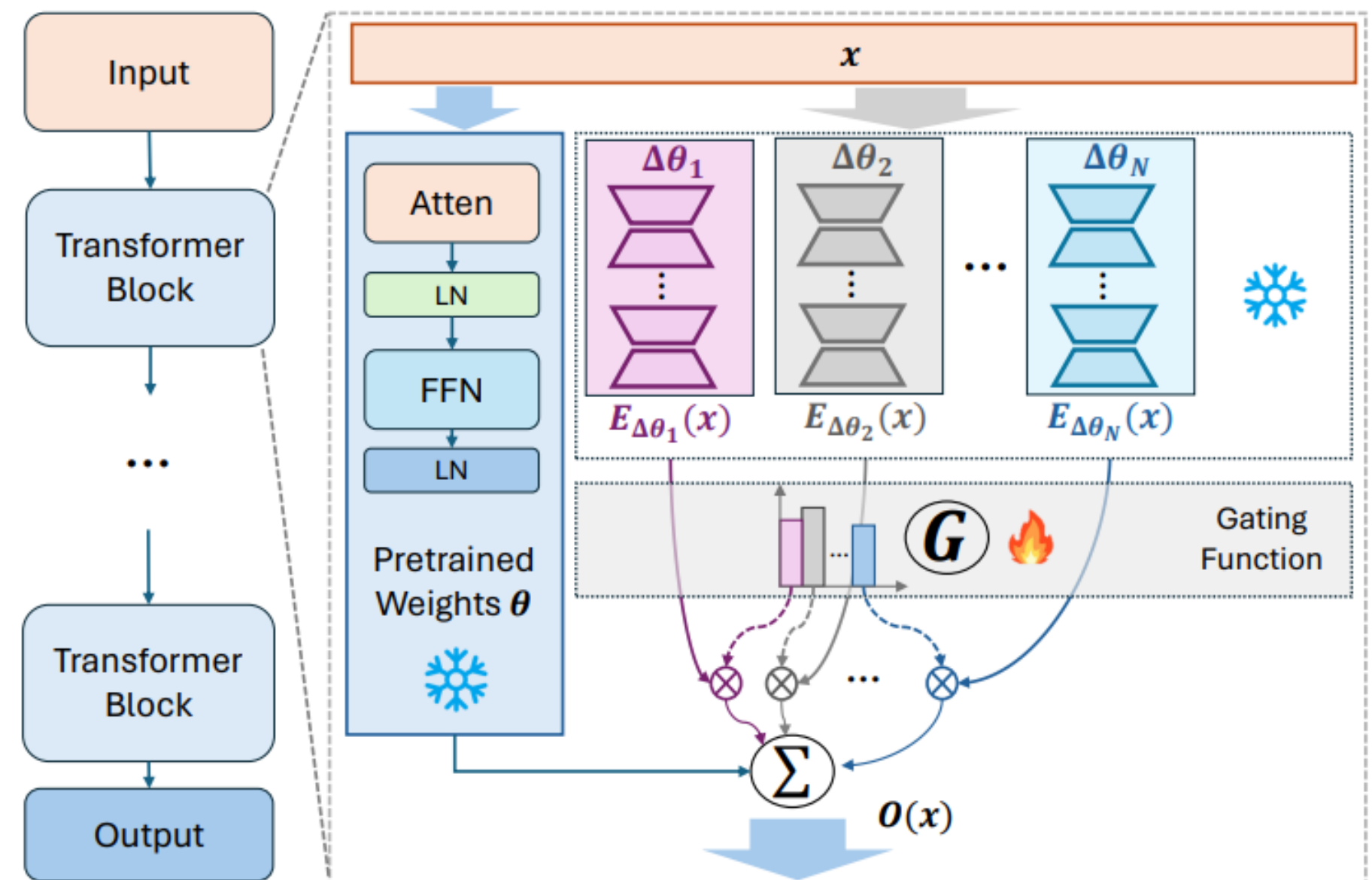
- 이용 측면에서 한 개의 모델만 사용하는 것이 더 편의성/효율성이 높음
- LoRA를 동적으로 로드/언로드 해서 사용하는 모델은 반응성이 떨어짐
- 사용자가 장르를 중간에 전환하거나, 혹은 장면 분위기에 따라 강도를 조절할 필요가 있음

→ 단일 장르 LoRA들을 한 모델 내에서 유연하고 동적으로 활용해야 함

1.Motivation

6) MoLE(Mixture of LoRA Experts)

- 여러 개의 LoRA들을 MoE 구조의 Gating을 이용해서 결합한 모델 구조
- 각 레이어마다 어떤 LoRA를 얼마나 사용할 지 Gating Function을 이용해서 자동 결정
- 이전까지 있었던 여러 LoRA를 단일 모델에서 사용하는 아키텍처에서 있던 비효율을 해결



< Fig 1. MoLE 구조 >

1.Motivation

7) MoLE사용의 타당성 검증

다른 기법들과의 비교

(1) Arithmetic Composition 기법의 경우

- LoRA의 결과 값들을 단순 합산 하는 방식
- LoRA 값이 중첩되어, 표현력이 붕괴함

(2) Weight Normalization 기법의 경우

- LoRA의 결과 값들을 정규화해서 합산하는 방식
- LoRA 간의 장르 스타일이 희석됨

→ MoLE 논문의 결과와 동일

```

,... -, -, ..., -...- , -... - , -, ...- . . . . -.-
-...-, , , ..., -...-, ... -,
... down down down down, down down
- further, further, further , , down , down, ,... further,
urther ... **_** ,
-, further , further down down ... deeper , ... , ... ,
, **
, ,
,
, down, further , more , ... , further ,, ,, ,,, deeper, ,
e, ,, deeper,, more, ,

```

<Fig 1. 여러 LoRA 사용으로 인한 표현력 붕괴 예시>

	Soft Label 비교	Accuracy
Baseline(no finetuned)	<u>0.5294</u>	<u>0.5334</u>
Baseline + LoRA(각 장르로)	0.5832	0.607
Weight Normalized	<u>0.5351</u>	0.5323

< Table 1.두 기법의 genre classifier 결과>

1.Motivation

7) MoLE사용의 타당성 검증

Layer별로 역할이 다른 지에 대한 관찰

- MoLE논문에서, LLM에 대한 주요 observation 중 하나로, 각 layer별로 gating function을 두어, LoRA output의 비율을 다르게 섞어주는 것이 성능 향상에 도움이 된다는 주장의 근거가 됨

	ANLI-R1	ANLI-R2	QNLI
Full LoRA	81.65	80.03	76.42
0%-20%	78.72	78.35	78.14
20%-40%	76.10	77.96	77.85
40%-60%	76.95	81.47	74.57
60%-80%	77.25	78.19	75.71
80%-100%	82.59	77.91	75.48

< Fig 1. MoLE 논문에서 실험한 레이어 별 task에 따른 성능 기여 지표>

ANLI : 기존 모델들이 풀기 어렵도록 적대적으로 만들어진 데이터셋
QNLI : 질문과 단락 쌍으로 구성된 자연어 추론 데이터셋

1.Motivation

7) MoLE사용의 타당성 검증

Layer별로 역할이 다른 지에 대한 관찰

- Task는 동일하지만 style이 다른 우리 경우에도
그러한지 실험 진행
 - Progression: 서사 진행력
 - Tone: 분위기/어조
 - Vividness: 묘사력

→ 우리 문제에도 레이어별 성능이 다른 것이 관찰 됨

Progression	Tone	Vividness	활성화된 LoRA 위치
0.476665	0.530835	0.549335	0-10 %
0.53667	0.622085	<u>0.59875</u>	10-20 %
0.575835	<u>0.597915</u>	0.575415	20-30 %
<u>0.565835</u>	0.5925	0.612085	30-40 %

< Fig 1. LoRA 활성화 레이어 별 task에 따른 성능 기여 지표>

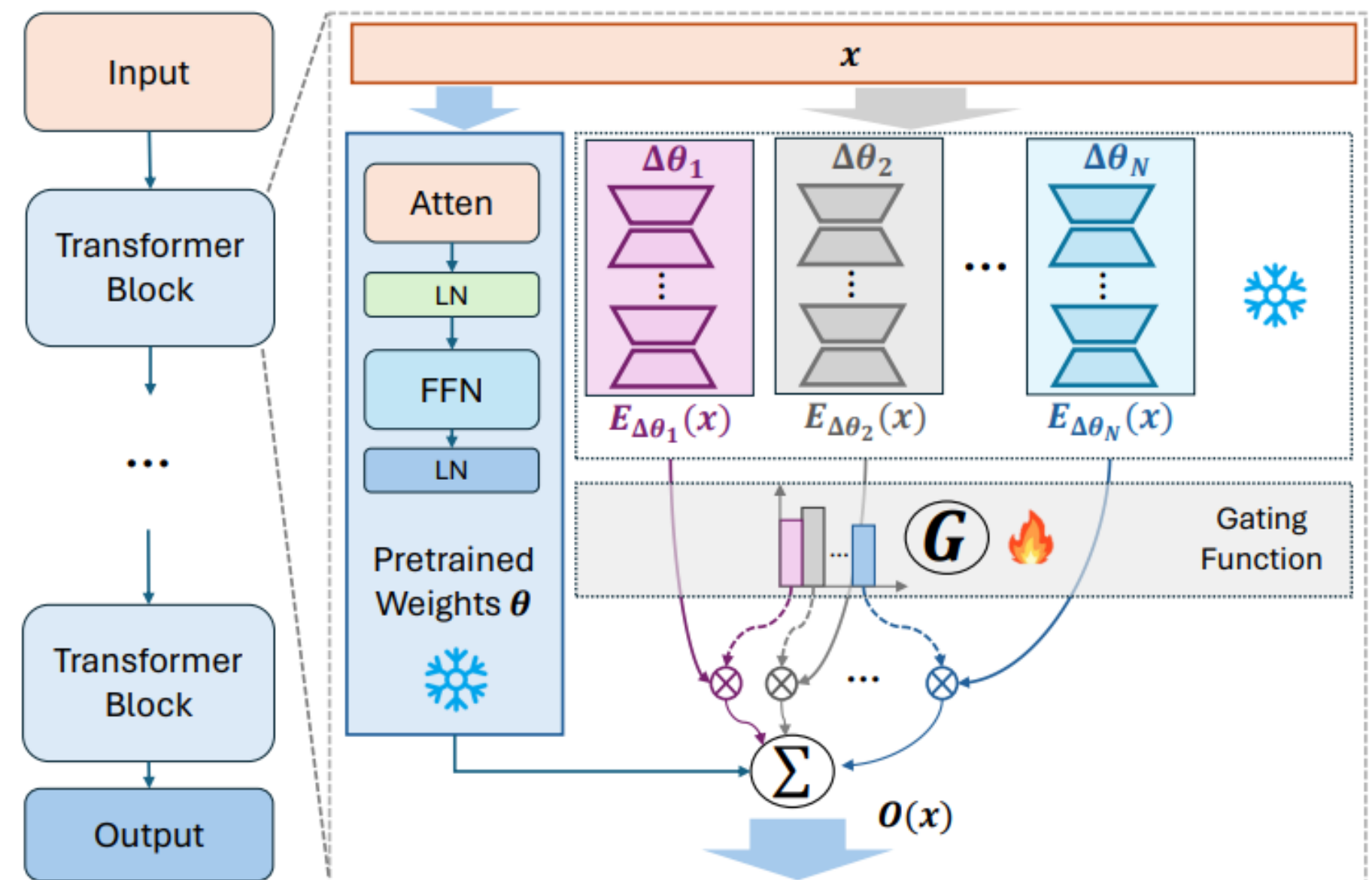
(40-100% Layer들만을 활성화 시켰을 때는, 동일 문장이 반복되는 현상이 지속적으로 발생하여, 실험에서 제외)

1.Motivation

7) MoLE사용의 타당성 검증

결론

- 즉 MoLE 논문에서 제안한 문제가, TRPG용 LLM 모델에서도 발견되었으므로, MoLE를 차용하는 것이 타당하다고 판단됨
- 레이어별로 gating function을 도입함으로써, 각 LoRA 레이어가 더 잘하는 부분을 전담하도록 함



< Fig 1. MoLE 구조 >

1.Motivation

8) 추가 가설

- 기존 TRPG LLM이 특정 장르의 특성을 충실히 반영하지 못하는 문제 해결을 위해 MoLE 도입
- MoLE의 Gating은 입력 맥락에 따라 최적의 LoRA 비중을 동적으로 결정함
- 따라서 기대한 효과 (장르 일관성 유지) 이외에도, "**플레이 도중 장르가 급변하는 시나리오**"에서도 별도의 조작 없이 유연한 대처가 가능할 것으로 가정함
- 위 가설을 검증하고자, 의도적으로 장르를 전환시키는 시나리오를 추가적인 평가 데이터셋으로 설정

1.Motivation

9) 최종 목표

- **MoGLE : Mixture of Genre LoRA Expert 모델 제안**
 - **가설 1. 장르 적합성**
 - MoGLE가 TRPG GM으로써, 특정 장르의 특성을 충실히 반영한 문장 생성할 것
 - **가설 2. 장르 전환 적응성**
 - MoGLE가 “플레이 도중 장르가 급변하는 시나리오”에도 별도의 조작 없이 높은 성능을 보일 것

2. Problem Definition

1) Problem to Solve

- TRPG의 GM역할을 수행하는 LLM의 생성을 다음과 같은 조건부 확률 분포 최적화 문제로 정의

$$\hat{y} = \arg \max_y P(y|x, g; \theta)$$

- 입력(x): 플레이어의 행동 및 현재 게임 문맥
- 조건(g): 목표로 하는 특정 장르
- 출력(y): 해당 장르의 문체적 특성과 서사 구조를 반영한 GM의 묘사
- 파라미터(θ): 모델의 파라미터

2. Problem Definition

2) Inference I/O

- MoGLE이 장르에 적절한 장면 묘사를 하는지 확인하기 위해, 두 가지 시나리오 가정
 - 장르 적합성을 위한 시나리오
 - 한 가지 장르로 진행되는 시나리오
 - 장르 전환 적응성을 위한 시나리오
 - 장르가 자연스럽게 전이되는 시나리오

2. Problem Definition

2) Inference I/O example

1. 한 가지 장르로 진행되는 시나리오

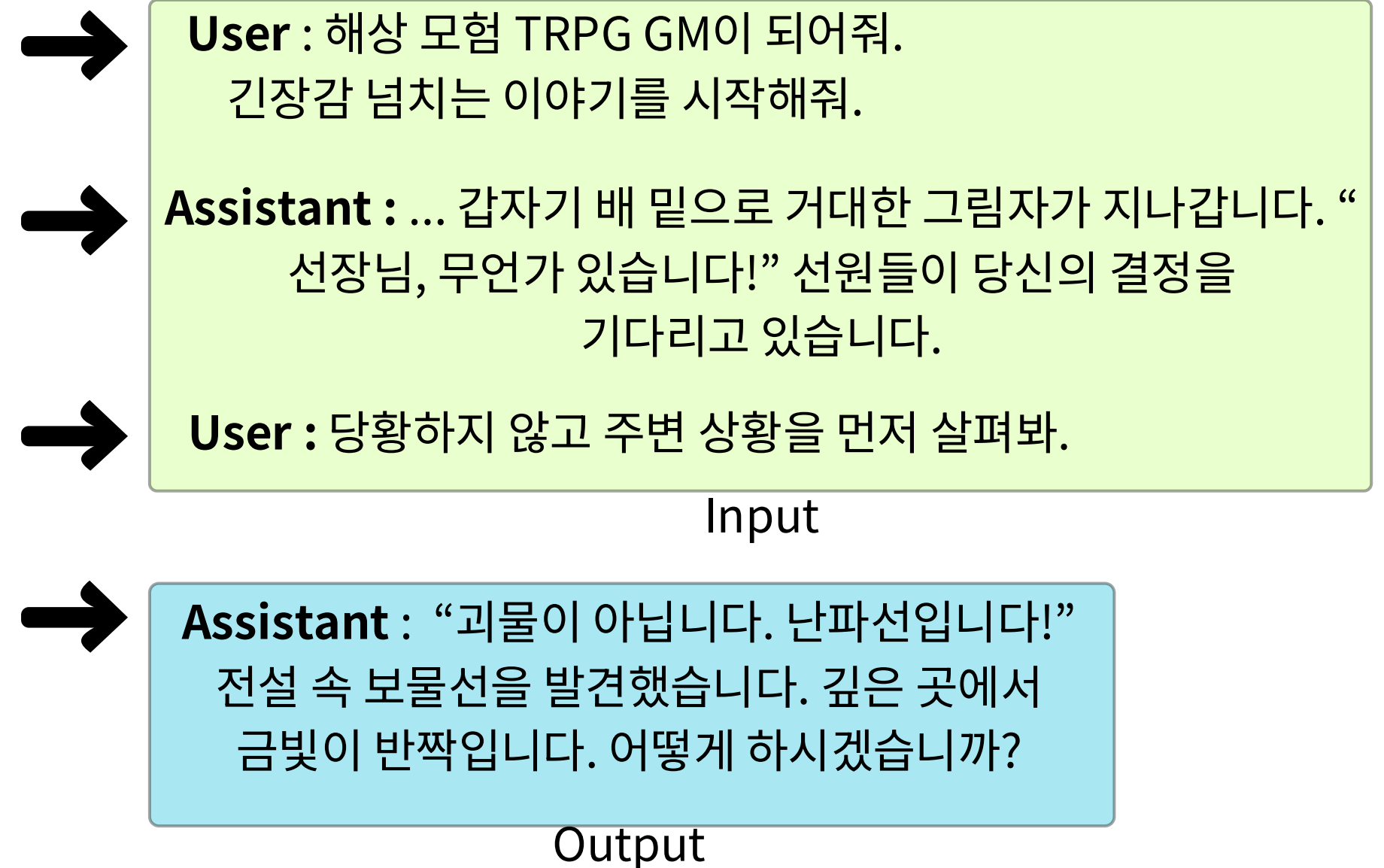
User: Function as a GM for naval adventure ...

Assistant: Welcome, Captains! ... The lookout shouts: "Captain! Something is moving under us!" ... Your crew looks to you for a decision.

User: I choose to survey the surroundings ...

Assistant: You scan the water ... "Captain... that's no creature. That's a shipwreck." ... It is the sunken Star Galleon. Gold glitters in the depths ... What will you do next?

< Fig 1. Inference time에서의 I/O 예시 >



2. Problem Definition

2) Inference I/O example

2. 장르가 자연스럽게 전이되는 시나리오

User: Function as a GM for naval adventure ...

Assistant: Welcome, Captains! ... The lookout shouts: "Something massive is moving under us!" ... Your crew looks to you for a decision.

User: Whatever is beneath us does not feel alive ... A cold spike crawls up my spine ... Then I see it: A ghost's face ... staring directly into mine.

Assistant: A sharp chill sweeps across the deck ... The lantern flame stretches sideways ... "C-Captain... the temperature-!" ... A wet, dragging sound echoes ... A drowned woman's ghost rises onto the deck. She only points toward you ... What will you do next?

< Fig 1. Inference time에서의 I/O 예시 >

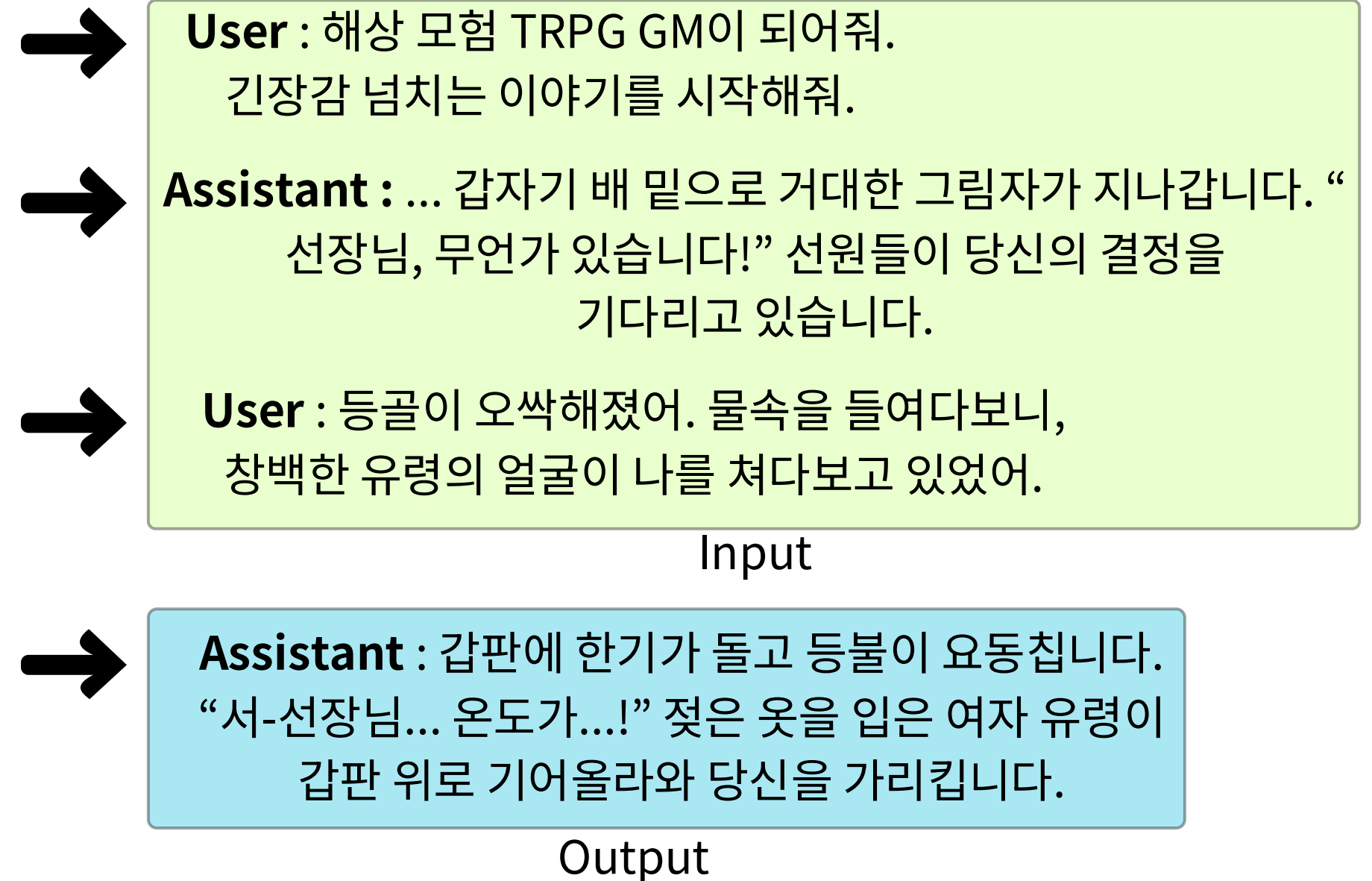


Fig 1. Playthrough on our model. User from real human, Assistant from LLM

3. Challenges

1) Data 확보

- 장르-스토리가 매칭된 데이터셋 중, 고품질의 데이터셋이 많지 않음
 - **1000 Stories 100 Genres**: 장르별로 절대적인 수가 모자람
 - **Book Genre Classification Dataset**: 시놉시스에 대한 장르 매칭 → TRPG 학습엔 부적합
 - **CMU Book Summary Dataset**: 스토리 요약에 대한 장르 매칭 → TRPG에 부적합
 - **Project Gutenberg**: Copyright Claim, Index, Author등, 노이즈가 너무 많고, 너무 오래 된 책들 뿐 → 고어적인 표현이 들어가는 특정 장르에만 부분적으로 활용 가능
 - **WritingPrompts**: 온라인 커뮤니티(Reddit) 기반의 특성 상 노이즈가 너무 많음
→ 특수한 데이터 전처리 파이프라인이 필요함

3. Challenges

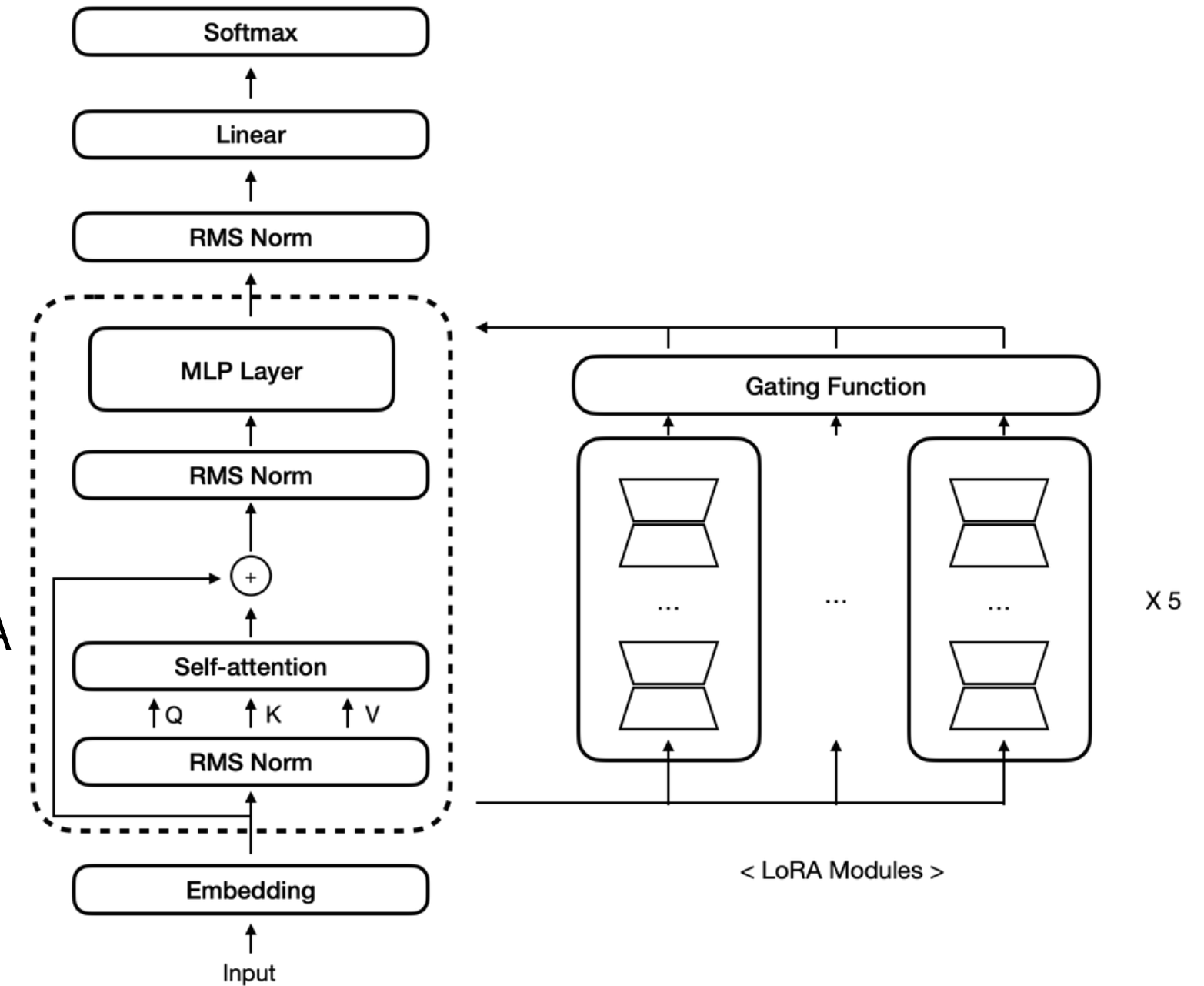
2) 구현 및 평가

- 라이브러리의 부재
 - MoLE 논문에서 제안한 모델은 라이브러리에 존재하지 않았으며, MoLE 논문의 깃허브에는 LLM이 아닌 Diffusion model에 대한 MoLE 구조만 구현되어있었음
- 평가의 모호성
 - 특정 장르의 특성을 얼마나 충실히 반영했는지를 정량적으로 평가할 수 있는 평가지표가 부재함.
 - 플레이 도중 장르가 급변하는 시나리오에서 모델이 얼마나 자연스럽게 적응하는지를 측정하는 평가 체계 또한 존재하지 않음

4. Methods

1) Architecture

- 기반 모델
 - Llama 3.1 8B - Instruct
- 구조
 - MLP에 3개, QKV에 3개, 총 6개의 부분에 LoRA 적용
 - 각 레이어 별로 Gating Function을 통해 레이어마다 LoRA 간 비율 동적으로 적용



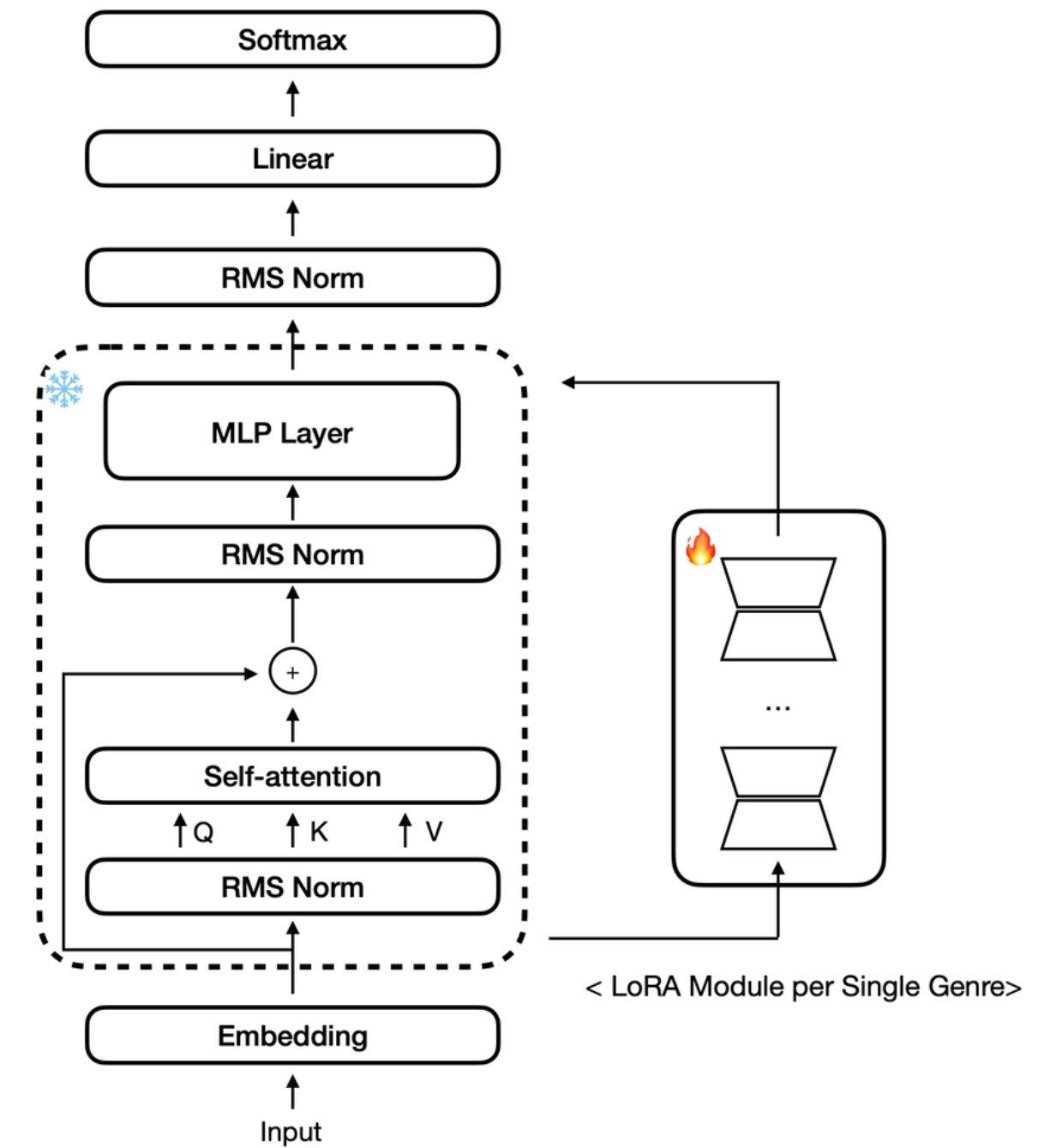
< Fig 1. Model Architecture >

4. Methods

2) Pipeline

1. Single LoRA Training Phase

- Input / Output : $Token_{1..k}$ $Token_{2...k+1}$
- model : LLM with **single** LoRA module
- 위 작업을 각 장르별로 수행 → 장르 개수만큼의 LoRA module 생성
- 총 장르 개수는 5가지
 - Adventure, Horror, Dystopian, Sci-Fi, Fantasy
 - 즉 5개의 LoRA 개별 학습



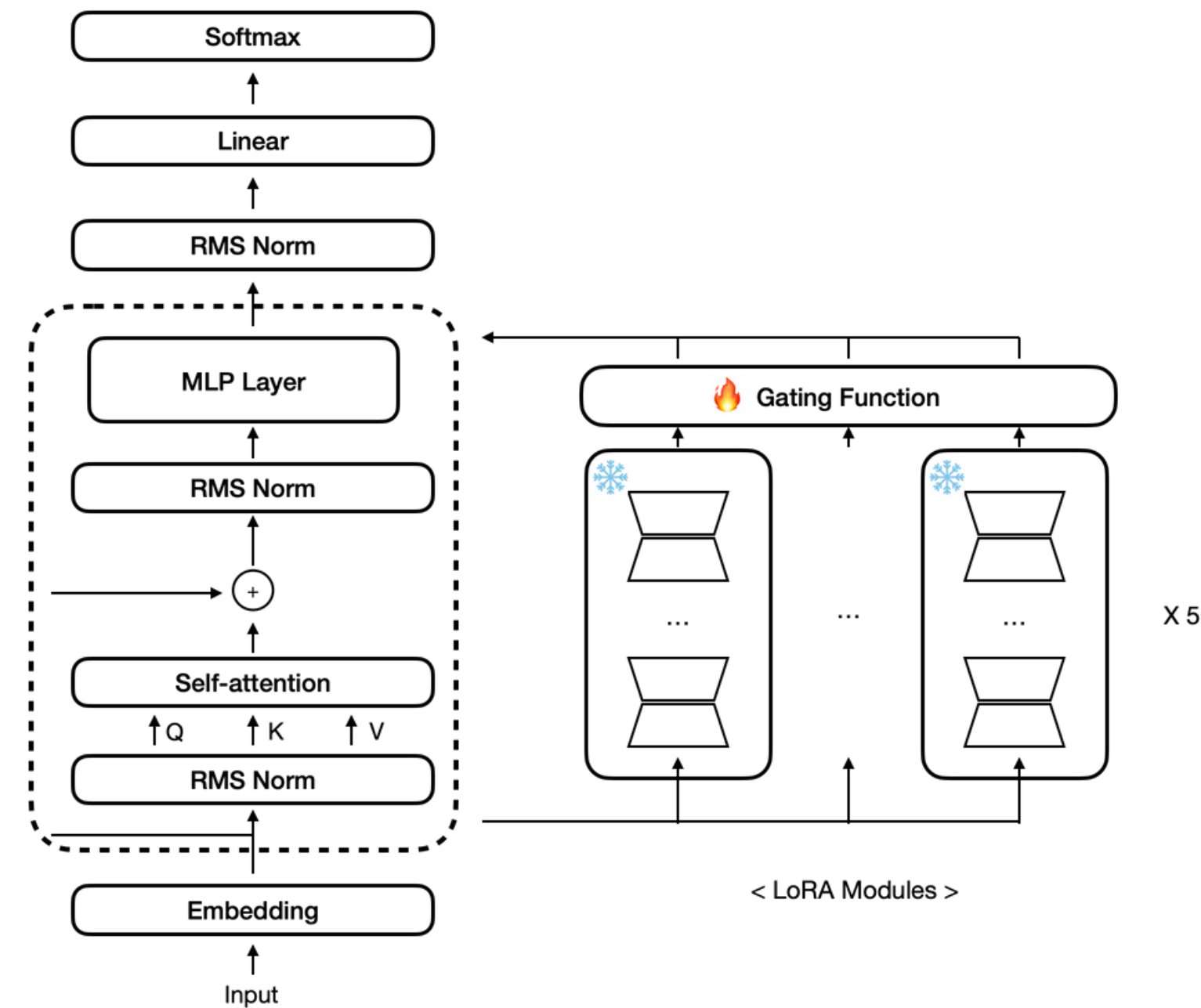
< Fig 1. Single LoRA Training Phase >

4. Methods

2) Pipeline

2. Gating Function Training Phase (MoLE Training Phase)

- Input / Output : $Token_{1..k}$ $Token_{2...k+1}$
- model : LLM with **Multi**(장르 개수) LoRA module & **Gating function**
- 전체 파라미터 개수
 - Llama 3.1 8B \rightarrow 8B
 - LoRA x 5 \rightarrow 16M x 5 = 80 M
 - Gate \rightarrow 23M
 - Total : 8.14B



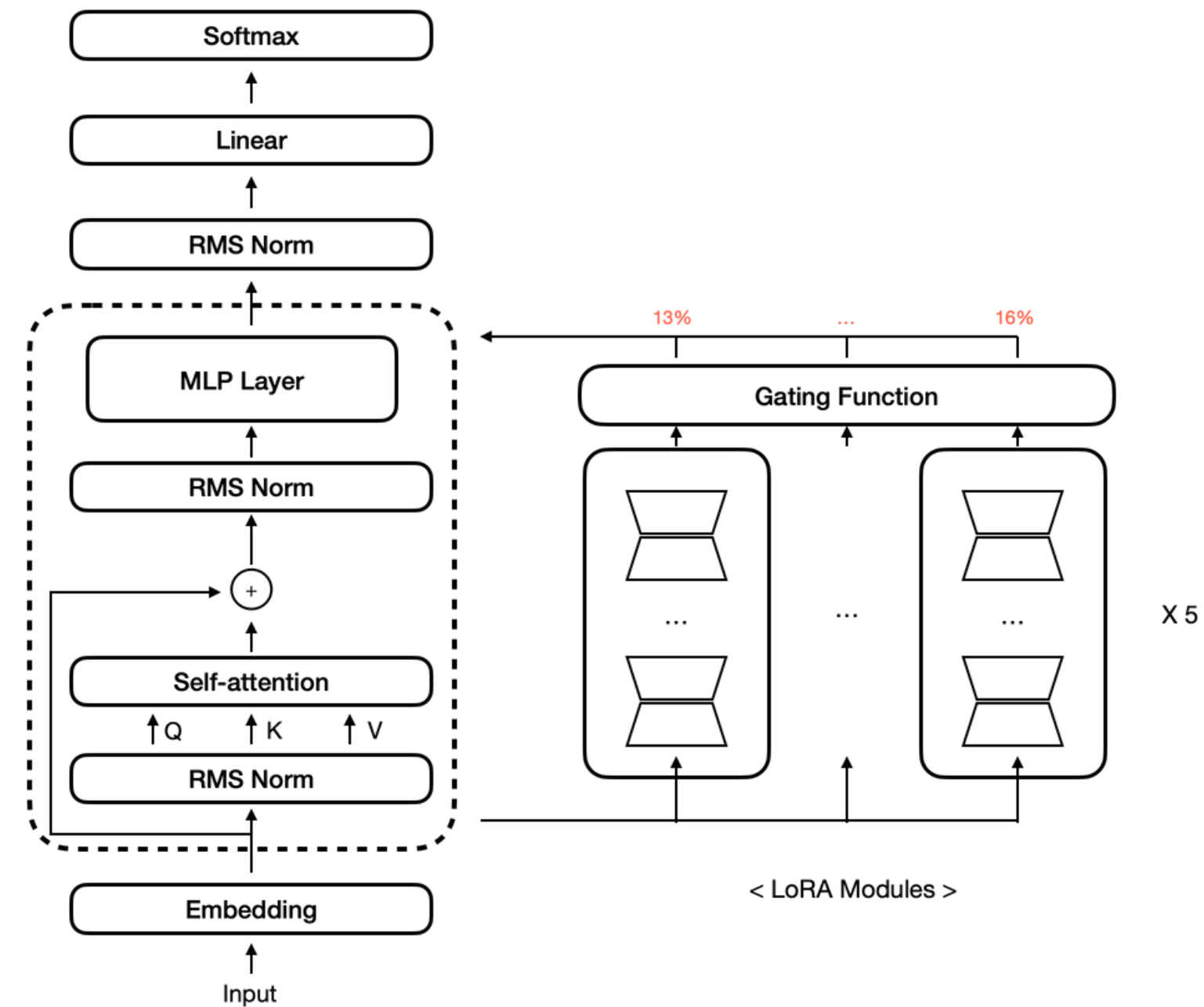
< Fig 1. Gating Function Training Phase >

4. Methods

2) Pipeline

3. Inference Phase

- Input / Output : $Token_{1..k}$ $Token_{2...k+1}$
- model : LLM with Multi(장르 개수) LoRA module & Gating function
- 그림처럼, 특정 LoRA의 비율을 input에 의거하여 gating function이 동적으로 조정



< Fig 1. Inference Phase >

4. Methods

3) Loss

- Single LoRA Training Phase Loss
 - $\mathcal{L} = \mathcal{L}_D$ (D는 domain. 즉 NLP domain loss인 CE)

- Gating Function Training Phase Loss

$$\mathcal{L} = \mathcal{L}_D + \alpha \mathcal{L}_{\text{balance}}$$

- Gating Balancing Loss

- Early stage에 수렴한 LoRA로만 정보가 흐르는 것을 막기 위한 loss

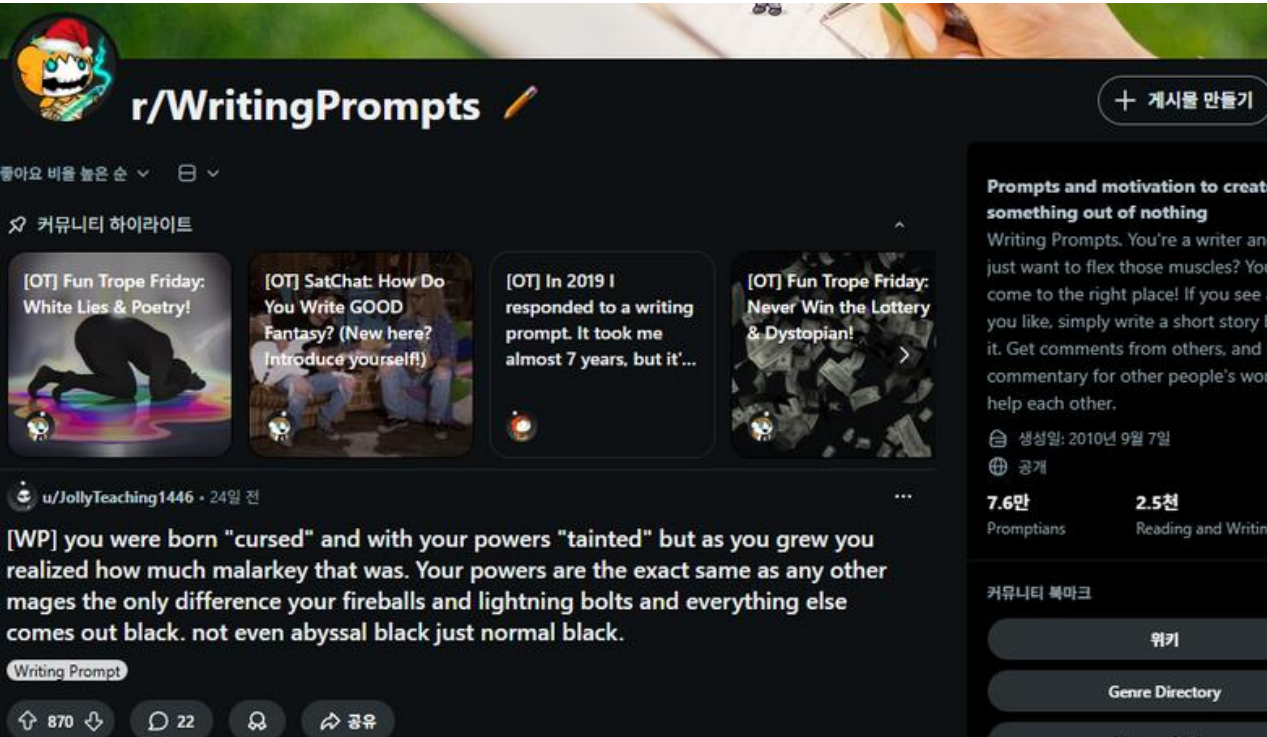
$$\mathcal{L}_{\text{balance}} = -\log \left(\prod_{i=0}^N \mathbf{q}^{(i)} \right) \quad \mathbf{q}^{(i)} = \frac{1}{M} \sum_{k=1}^M \frac{\exp(\varepsilon_i^k / \tau)}{\sum_{j=1}^N \exp(\varepsilon_j^k / \tau)}$$

M : layer 개수
N : layer 내의 LoRA 개수

5. Experiments

1) Training Dataset

- Data Source
 - WritingPrompts Dataset (euclaise)
 - 주제(prompts)에 맞춰 인간이 직접 작성한 Reddit 기반의 대규모 창작 소설 데이터셋
 - 장르별로 4000개의 데이터를 추출



< Fig 1. Reddit의 WritingPrompts 게시판>

Adventure	journey, map, explore, treasure, ...
Dystopian	government, regime, rebellion, ...
Fantasy	magic, wizard, dragon, spell, ...
Horror	blood, scream, ghost, fear, ...
Sci-Fi	space, robot, alien, ship, planet, ...

< Table 1.장르 추출에 사용된 장르별 키워드 예시>

Fig 1. Reddit, r/Writing Prompts (<https://www.reddit.com/r/WritingPrompts/>)

Table 1. From our data preprocessing pipeline

5. Experiments

1) Training Dataset

- Preprocessing Pipeline

1. 초기 필터링

- [removed], [deleted] 등 삭제된 포스트 제거

2. 키워드 기반 장르 라벨링

- 특정 장르 키워드가 4회 이상 등장 시 해당 장르로 분류
- Reddit의 엄격한 규칙(장르 이탈 시 ban)을 신뢰

3. LLM 기반 noise 제거

- Llama 3.1 8B 모델에 few-shot prompting으로
is_story, start_phrase, end_phrase를 생성

[System Prompt]

You are an expert editor. Identify the EXACT START and END of the fictional narrative.

Exclude meta-text (prompts, edits, author notes).

Example 1:

Input: "The dragon fell.\n\nThanks for reading!"

JSON Output: {"is_story": true, "start_phrase": "The dragon fell.", "end_phrase": "The dragon fell."}

INSTRUCTIONS:

1. start_phrase: First sentence of the ACTUAL story.
2. end_phrase: Last sentence of the ACTUAL story.
3. Ignore "Thanks for reading", "Edit:", "r/WritingPrompts".
4. If not a story, return "is_story": false.

Return JSON ONLY.

< Fig 1. LLM 기반 noise 제거를 위해 사용한 프롬프트 예시 >

5. Experiments

1) Training Dataset

1.Rule 기반 정리

- 토큰 수 100 이하, 2048 이상 데이터는 제거
- HTML Tags, Entities 제거
- Invisible Characters(Zero-width space, \u200b) 제거
- Markdown 문법 정규화(백틱, 강조 구문, 반복 부호 축소 등)

Adventure	2876
Dystopian	3493
Fantasy	3452
Horror	3484
Sci-Fi	3450
SUM	16755

< Table 1. 최종적으로 수집된 데이터 수>

The ship landed in the parking lot outside the office building ... Will reached down and searched the alien for some kind of weapon.

< Fig 1. 최종 수집된 데이터의 샘플(Sci-Fi장르)>

Table 1. From our final dataset

Fig1. From our final dataset

5. Experiments

2) Evaluation Dataset

1. Evaluation of Genre-Appropriate LLM Generation

2. Evaluation of Genre-Shift Generation

User : Act as a GM guiding a classic adventure campaign with treasure hunting.

Assistant : The oppressive humidity of the Xylos Jungle clings to you like a second skin, So you're in trouble. What would you do?

User : Take some wood and try to start a fire with flint.

<Fig.1 Eval 1에 사용되는 데이터셋중 일부>

→ 처음에 정한 장르인 Adventure를 그대로 따르는 User

User : Act as a GM guiding a classic adventure campaign with treasure hunting.

Assistant : The oppressive humidity of the Xylos Jungle clings to you like a second skin, So you're in trouble. What would you do?

User : While searching for flint to start a fire, you hear something wierd. Following the sound leads you to a small cave ... A quick glance through it reveals a sky filled with flying dragons and young wizards playing with magic among lush trees.

<Fig.2 Eval 2에 사용되는 데이터셋중 일부>

→ 처음에 정한 장르가 아닌, Fantasy로의 흐름을 원하는 User

5. Experiments

3) Evaluation Pipeline

1. SOTA comparison

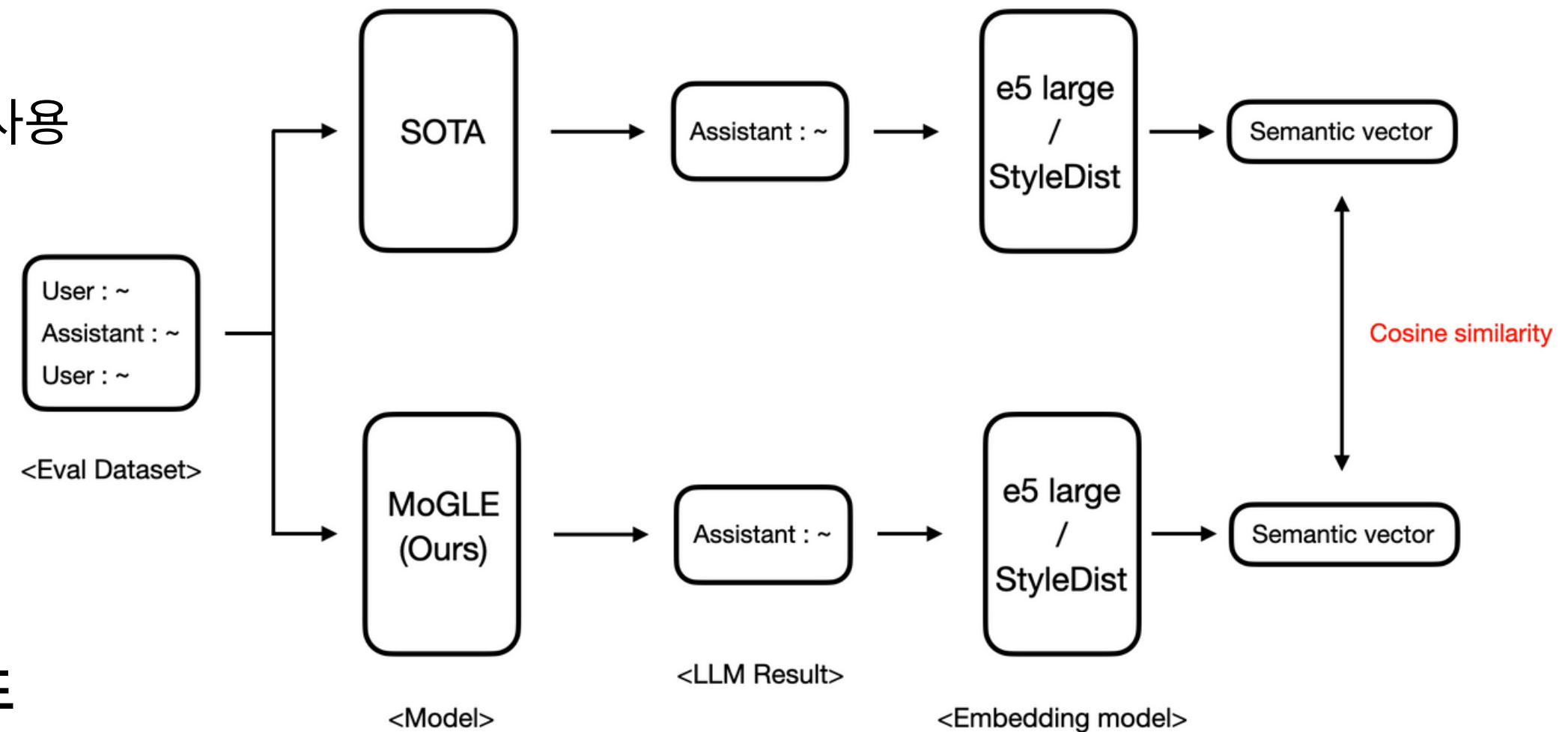
- SOTA model의 결과를 gold standard로 사용

- Llama4 Maverick 17B
- GPT OSS 120B
- MoonShotAI KIMI-k2 1T

- 두가지 비교사항

1. e5-large 임베딩 모델을 통한 **종합 유사도**

2. Style-Distance 임베딩 모델을 통한 **문체 유사도**

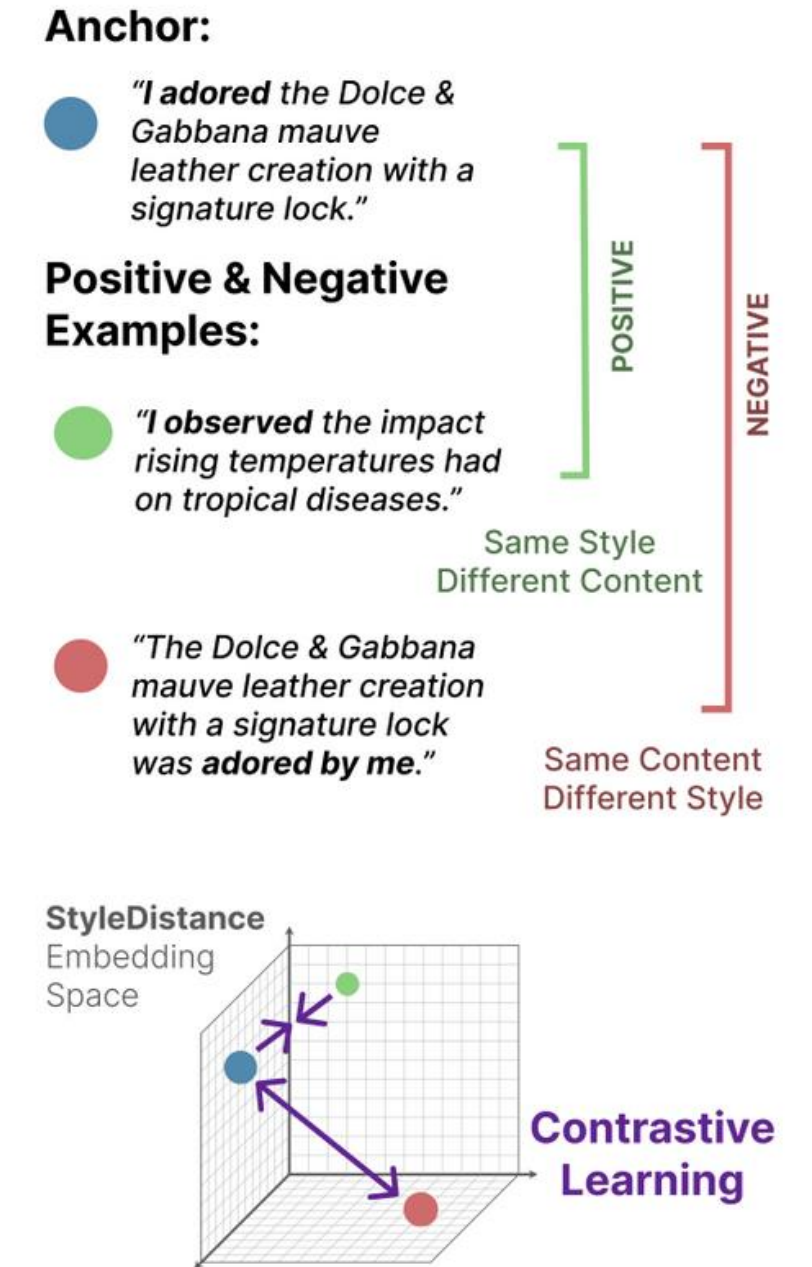


<Fig.1 SOTA를 활용한 평가 파이프라인>

5. Experiments

3) Evaluation Pipeline

- StyleDistance(NAAACL 2025)
 - 내용과 무관하게 스타일만을 잡아내기 위해 만든 임베딩 모델
 - 스타일이 같은 문장끼리 유사해지도록 Contrastive learning
 - StyleDistance 모델을 활용하여 서로 다른 문장 쌍의 스타일 유사도를 측정



< Fig 1. StyleDistance 모델 학습 방식 >

5. Experiments

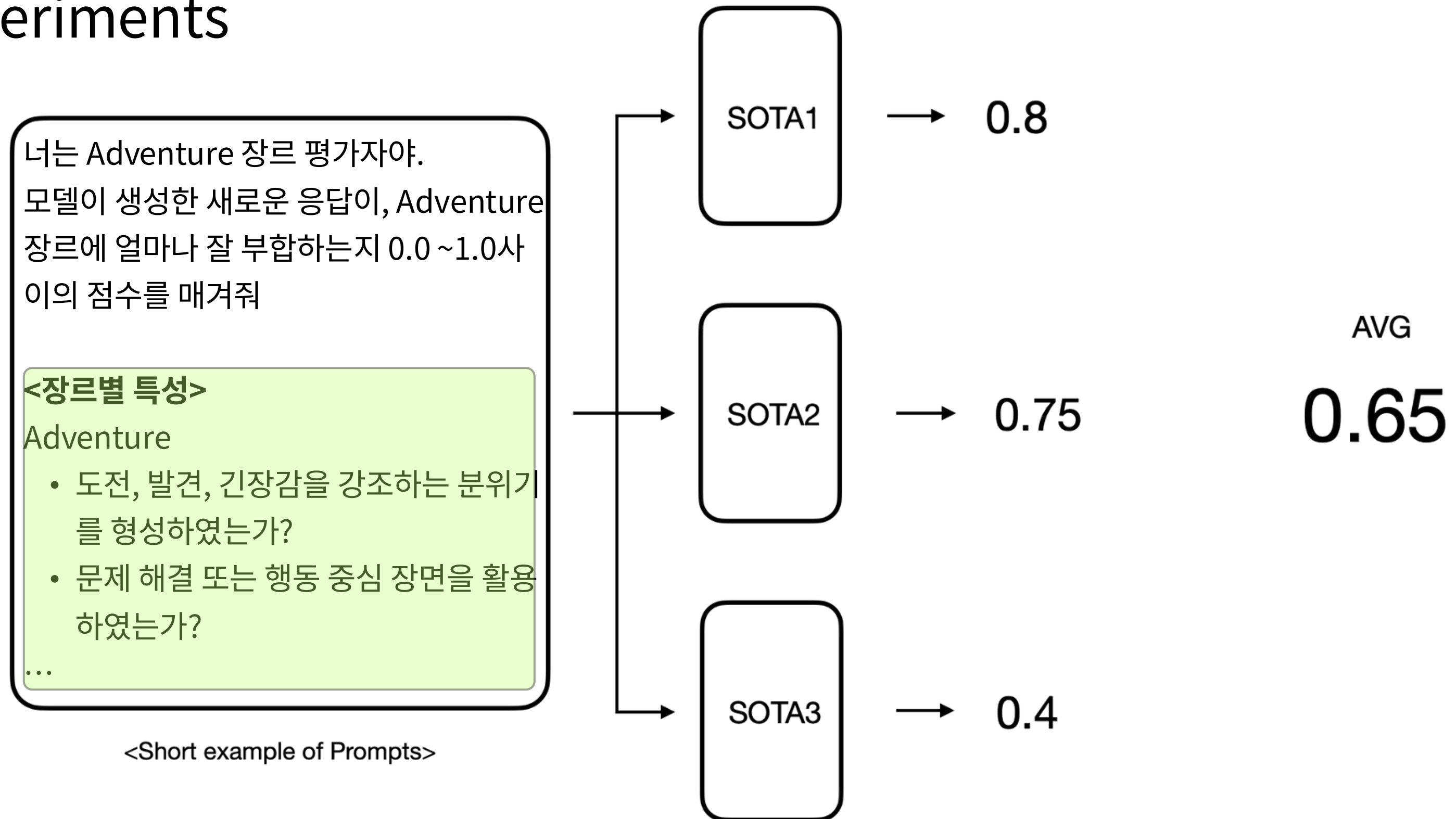
3) Evaluation Pipeline

2. LLM Judge

- 장르별 특성을 기반으로, 앞서 언급한 3개의 SOTA를 LLM Judge로 활용
- 장르별 특성은 위키피디아를 기반으로 팀원 대다수가 동의하는 4가지 기준 선정
 - 가상의 시간선상의 미래, 혹은 역사학적이거나 고고학적인 진실과 모순된 역사적 배경 등의 시간적 설정
 - 외우주(예시: 우주 여행), 다른 세계, 지저 세계 등의 공간적 배경, 장면^[9]
 - 외계인, 돌연변이, 안드로이드 혹은 휴머노이드 로봇, 그리고 인류의 진화 과정에서 나타난 다른 캐릭터들
 - 광선총, 순간이동 장치, 휴머노이드 컴퓨터 같은 미래적이거나 그럴듯한 기술^[10]
 - 시간 여행, 원형, 초광속 여행, 앤서블을 비롯한 새로운, 혹은 기존의 물리법칙과 모순되는 과학적 법칙

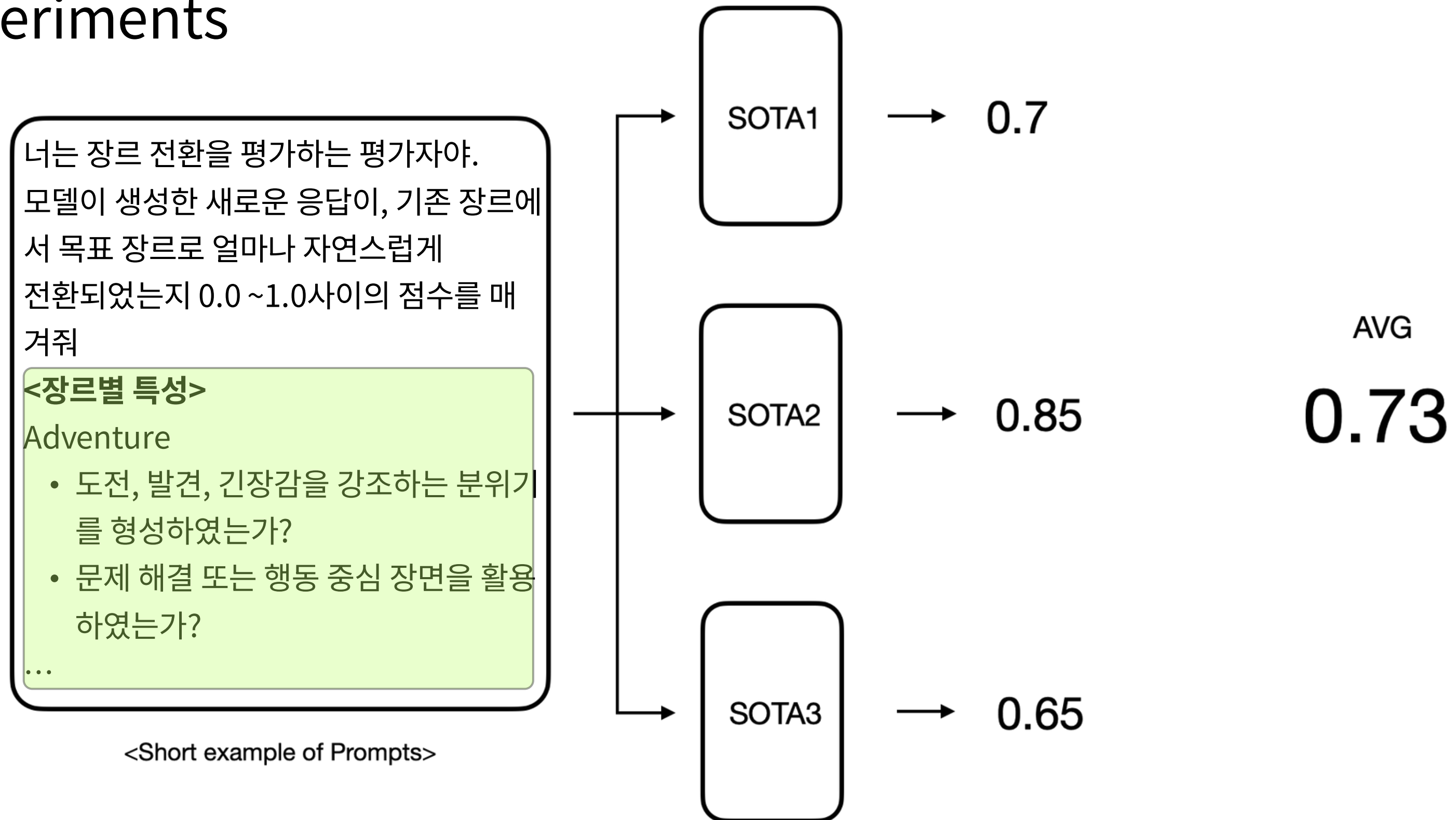
< Fig 1. 위키피디아 - SF장르의 특징 >

5. Experiments



<Fig.1 LLM Judge 파이프라인 for eval 1 (장르 적합성 평가)>

5. Experiments



<Fig.1 LLM Judge 파이프라인 for eval 2 (장르 전환 적응성 평가)>

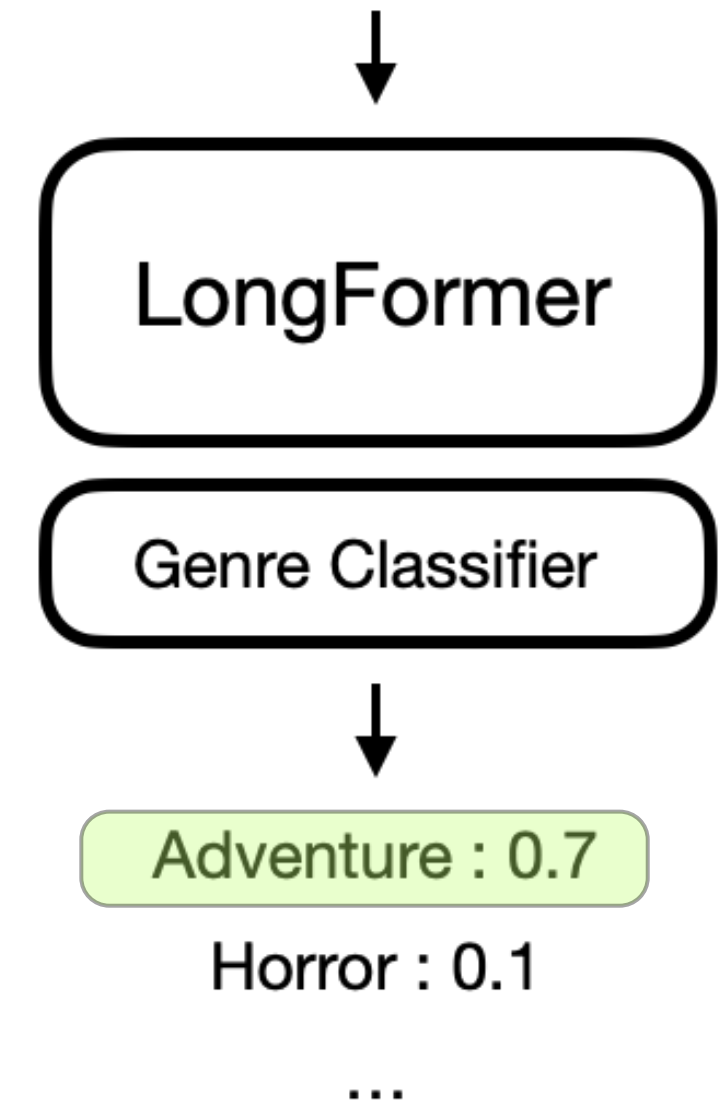
5. Experiments

3) Evaluation Pipeline

3. Genre Classifier

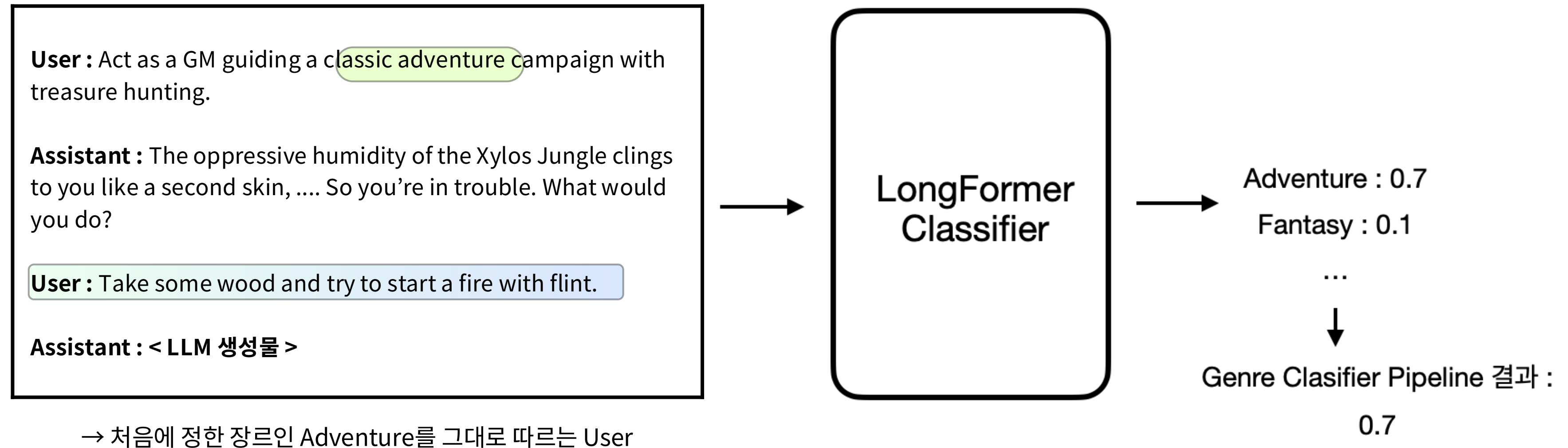
- LongFormer를 Finetune하여 Genre classifier로 사용
 - 자세한 Finetune 과정 및 성능은 Appendix A 에 기술
- LongFormer가 출력하는 확률값들중, 실제 정답 장르의 확률값을 활용
 - Evaluation1 (장르 적합성 평가) 의 경우
→ Genre classifier result = $P(\text{genre})$
 - Evaluation2 (장르 전환 적응성 평가) 의 경우
→ Genre classifier result = $P(\text{Source genre}) + P(\text{Target genre}) \times 1.5$

LLM Outputs (Adventure)



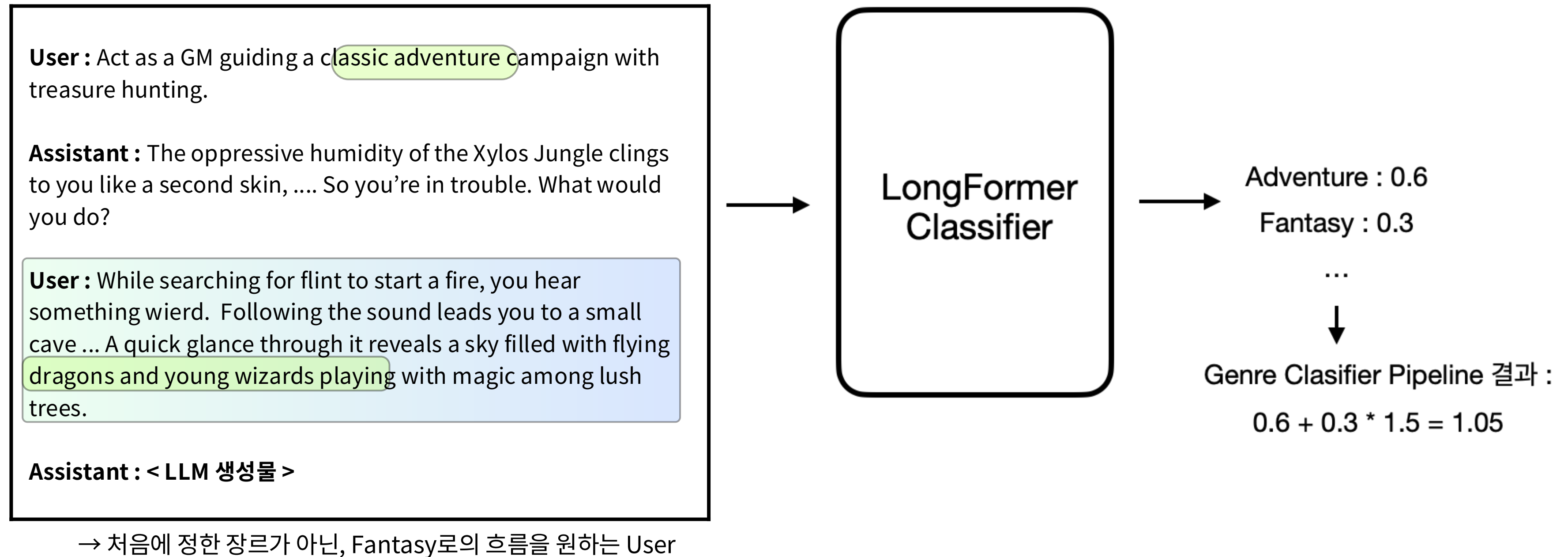
<Fig.1 Genre Classifier 파이프라인 >

5. Experiments



<Fig.1 Genre Classifier 파이프라인 for eval 1 (장르 적합성 평가)>

5. Experiments



<Fig.1 Genre Classifier 파이프라인 for eval 2 (장르 전환 적응성)>

5. Experiments

3) Evaluation Pipeline

1. SOTA comparison

- 동일 프롬프트를 SOTA와 MoGLE모델의 input으로 넣고, 그 결과를 임베딩 모델에 넣은 것을 비교
 - SOTA가 정답은 아니지만, 창의적 글쓰기는 “정답”이 없다는 점에서 최소한의 보증 지표로 판단

2. LLM Judge

- 장르별 문장 특징을 제공하고, 이를 적절히 수행하였는지 평가

3. Genre Classifier

- 생성된 문장이 특정 장르에 얼마 가까운지 평가

5. Experiments

4) Evaluation result

1. MoGLE의 Balance weight 별 결과

- a. Eval 1(장르 적합성 평가)
- b. Eval 2(장르 전환 적응성 평가)
- c. 결과 정리

2. 타 모델과 MoGLE 비교 결과

- a. Eval 1(장르 적합성 평가)
- b. Eval 2(장르 전환 적응성 평가)
- c. 결과 정리

5. Experiments

4) Evaluation result → Balance Loss weight 별 성능 비교 on Eval 1(장르 적합성 평가)

	SOTA comparison - e5	Genre classifier	SOTA comparison - sd	LLM judge
0.0 MoGLE	<u>0.93086</u>	0.652985	0.9836	0.62084
0.01 MoGLE	0.9313	0.60265	0.9879	<u>0.620025</u>
0.05 MoGLE	0.9306	<u>0.6054</u>	0.98665	0.56835
0.1 MoGLE	0.92525	0.5727	<u>0.98675</u>	0.5678
0.3 MoGLE	0.92655	0.60475	0.98405	0.570575
0.5 MoGLE	0.9261	0.58595	0.98575	0.5367

MoGLE 앞의 숫자는 loss함수에서 Balance loss의 계수를 의미

Ex) 0.01 MoGLE → $\mathcal{L} = \mathcal{L}_D + \alpha \mathcal{L}_{\text{balance}}$ 에서 α 는 0.01

5. Experiments

4) Evaluation result → Balance Loss weight 별 성능 비교 on Eval 2(장르 전환 적응성 평가)

	SOTA comparison - e5	Genre classifier	SOTA comparison - sd	LLM judge
0.0 MoGLE	0.880165	0.644945	0.856245	0.560555
0.01 MoGLE	0.878140	0.739185	0.861365	0.551110
0.05 MoGLE	0.879305	0.80856	0.85157	0.562225
0.1 MoGLE	<u>0.880365</u>	0.908030	<u>0.860535</u>	<u>0.588335</u>
0.3 MoGLE	0.877860	<u>0.784180</u>	0.857550	0.590000
0.5 MoGLE	0.880420	0.771445	0.857210	0.568890

MoGLE 앞의 숫자는 loss함수에서 Balance loss의 계수를 의미

Ex) 0.01 MoGLE → $\mathcal{L} = \mathcal{L}_D + \alpha \mathcal{L}_{\text{balance}}$ 에서 α 는 0.01

5. Experiments

4) Evaluation result

- Balance loss weight별 성능 비교 결과
 - Eval 1 (장르 적합성 평가)
 - Balance loss를 사용하지 않거나, weight를 0.01정도로 아주 작게 두는 것이 유리
 - 즉 한 가지 장르를 따를때에는, 낮은 Balance weight가 성능 향상 유도
 - Eval 2 (장르 전환 적응성)
 - Balance weight를 0.1정도로 크게 두는 것이 유리
 - 장르 변환이 있을때에는, 높은 Balance weight가 여러 장르를 사용하게 하여 성능 향상 유도

5. Experiments

4) Evaluation result

- 타 모델
 - Base (finetune x) : 파인튜닝 하지 않은 Llama 3.1 8B로 전체 eval 데이터셋 평가
 - Base x 5 (finetune per genre) : Llama 3.1 8B 모델을 5개 장르에 각각 별도로 파인튜닝하여, 장르별 성능을 개별적으로 평가한 뒤 결과를 평균
 - Base (Huge lora) : Llama 3.1 8B + LoRA 0.15B (LoRA r=52) 로 파인튜닝 후 전체 eval 데이터셋 평가

5. Experiments

4) Evaluation result

1. Genre-Appropriate LLM Generation

	SOTA comparison - e5	Genre classifier	SOTA comparison - sd	LLM judge
Base (finetune x)	0.930	0.5294	0.99145	0.5375
Base x 5 (finetune per genre)	0.93196	0.69874	0.98408	0.64418
Base (Huge lora)	0.92955	0.6221	<u>0.98415</u>	0.32165
MoGLE (ours, 0.0)	<u>0.93086</u>	<u>0.652985</u>	0.9836	<u>0.62084</u>

<Table 1. Eval 1 (장르 적합성 평가) >

5. Experiments

4) Evaluation result

2. Genre-Shift Generation

	SOTA comparison - e5	Genre classifier	SOTA comparison - sd	LLM judge
Base (finetune x)	0.864820	0.642815	0.848823	0.552842
Base (Huge lora)	<u>0.873102</u>	0.651128	0.850192	0.561242
MoGLE (ours, 0.1)	0.880365	0.908030	<u>0.860535</u>	0.588335

→ Eval 2의 경우, 두가지 장르가 함께 나오기 때문에, 각 장르를 학습한 5개의 Llama는 평가에서 제외

<Table 2. Eval 2 (장르 전환 적응성 평가)>

5. Experiments

4) Evaluation result

- 타 모델과의 성능 비교 결과
 - Eval 1 (장르 적합성 평가)
 1. Base x 5 (finetune per genre) vs MoGLE (ours)
 - MoGLE의 결과가 Llama 3.1 8B 모델을 5개 사용한 것과 결과가 유사함을 확인 (성능 하락 2.068%이내)
 - 파라미터 차이 : 40B (Base x 5) vs 8.142B (MoGLE)
 2. Base (Huge lora) vs MoGLE (ours)
 - 성능은 모든 지표에서 MoGLE이 더욱 뛰어남
 - 파라미터 차이 : 8.156B (Base Huge LoRA) vs 8.142B (MoGLE)
 - 즉 동일한 파라미터를 사용해도, LoRA보다 MoE에 LoRA를 결합시킨 MoLE 구조가 더욱 효율적임

5. Experiments

4) Evaluation result

- 타 모델과의 성능 비교 결과
 - Eval 2 (장르 전환 적응성 평가)
 1. Base (Base Huge LoRA) vs MoGLE (ours)
 - 파라미터 차이 : 8.156B (Base Huge LoRA) vs 8.142B (MoGLE)
 2. style distance를 제외한 모든 지표에서, MoGLE가 더 뛰어남

5. Experiments

5) Hypothesis Verification

- 장르 적합성 평가
 - 파라미터 차이가 극도로 많이 나는 모델과, 성능 차이가 미미한 결과를 얻음
 - 파라미터가 거의 동일한 모델에 비하여 더 나은 성능을 보임
 - MoGLE 가 장르에 충실한 문장들을 생성해냄을 증명
- 장르 전환 적응성 평가
 - 파라미터가 거의 동일한 모델에 비하여 더 나은 성능을 보임
 - 파라미터를 신경쓰지 않아도, 다른 타 모델에 비해 더 나은 성능을 보임
 - MoGLE 가 장르 전환에 유연한 대처가 가능함을 증명

6. Conclusion

- **Project Achievements**

- 단일 모델 내에서 다양한 장르를 소화하는 TRPG LLM 모델을 제작
- 레이어별 Gating function을 통해, LoRA의 개입 정도를 동적으로 조절하는 방법이 효과적임을 확인

- **Key Results**

- 장르 적합성 및 장르 전환 적응성 확보

- **Evaluation Pipeline**

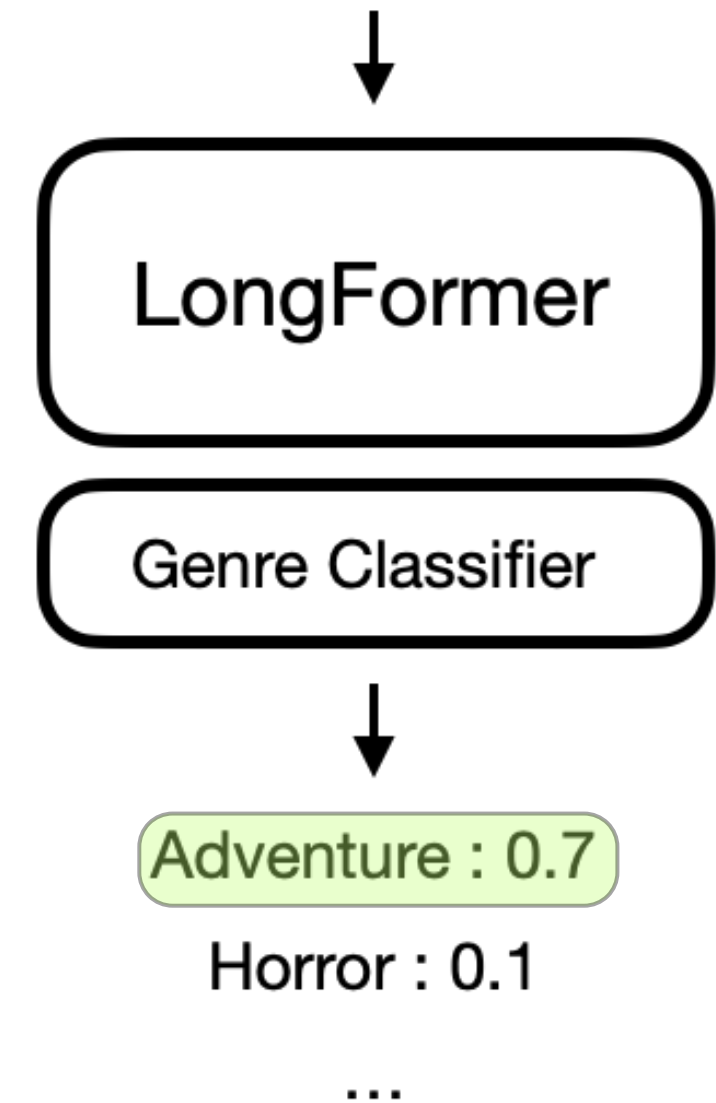
- 기존에 없던 장르 적합성 및 장르 전환 적응성을 평가하기 위한 Evaluation data 및 Pipeline 제작

APPENDIX. A

Genre Classifier

- 목적
 - Data Validation: 데이터가 실제로 특정한 문체적 패턴이 있는지를 확인하고자 함
 - Evaluation Matrix: 모델이 생성한 텍스트가 원본 데이터의 특성을 얼마나 반영하고 있는지 판단하고자 함
- Model 선정
 - 데이터셋의 구성 상(평균 1046 토큰), 긴 호흡의 문맥을 파악하는 것이 중요하므로, 최대 4096 토큰까지 처리 가능한 Longformer를 채택함

LLM Outputs (Adventure)



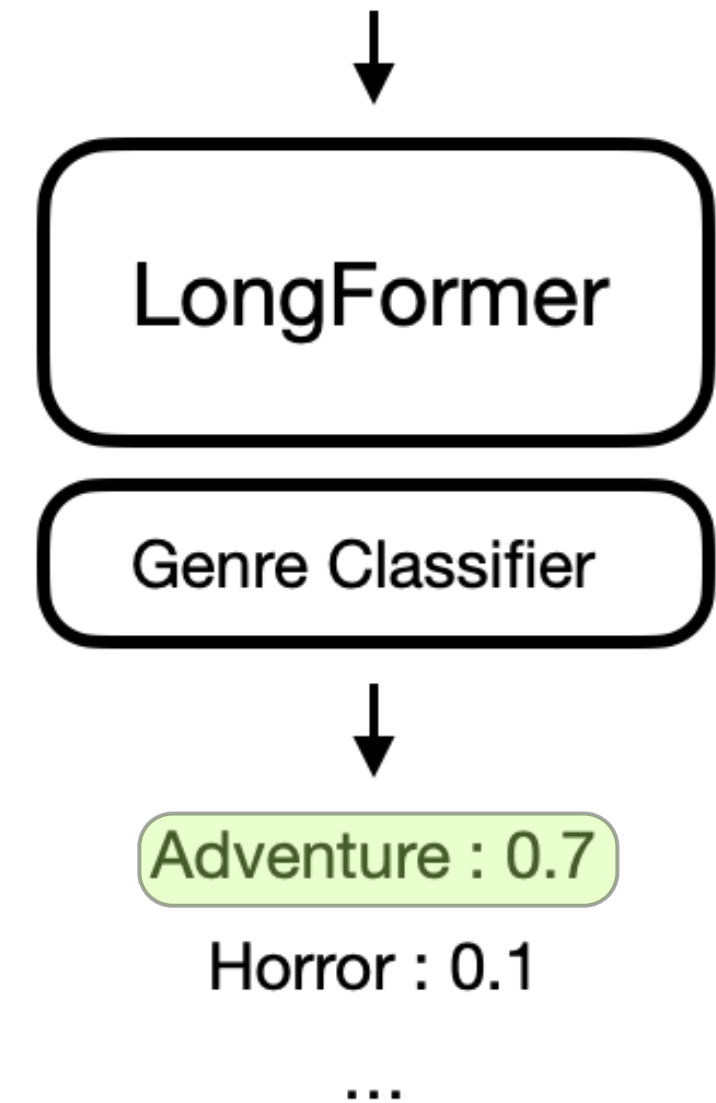
<Fig.1 Genre Classifier 파이프라인 >

APPENDIX. A

Genre Classifier

- 학습 방법
 - 전체 데이터셋 중, validation과 test를 위한 소수의 데이터만 남기고, 모두 이용해서 full-finetuning
- 학습 결과
 - Validation Accuracy 0.9308, F1 Score 0.9304 달성
 - 이를 통해, 데이터셋 내부에 뚜렷한 언어적 특성이 존재함을 증명
 - 이후 실제로 평가의 한 지표로 이용했음

LLM Outputs (Adventure)



<Fig.1 Genre Classifier 파이프라인 >

APPENDIX. B

Layerwise experiment criteria

```
evaluation_criteria = {  
    "vividness": ""  
    **1. Descriptive Vividness & Immersion**  
    - Does the model use rich sensory imagery (visual, tactile, auditory)?  
    - Does it use literary modifiers effectively to stimulate imagination?  
    - Does it provide a sense of presence beyond simple information delivery?  
    "",  
    "tone": ""  
    **2. Tone & Atmosphere Consistency**  
    - Does it maintain a consistent tone appropriate for the genre (e.g., Archaic, Cyberpunk, Urgent)?  
    - Does it effectively build suspense or a dramatic atmosphere?  
    - Is the sentence length and pacing effective for the mood?  
    "",  
    "progression": ""  
    **3. Narrative Progression & Coherence**  
    - Is there logical continuity with the previous context?  
    - Does it introduce new conflicts, mysteries, or characters to drive the story?  
    - Are NPC actions and dialogues consistent with their established characters?  
    ""}
```

APPENDIX. C

LLM-as-a-Judge evaluation-1 criteria

context_map = {

"Adventure": ""

- Fast or steady forward-moving pacing
 - Clear sense of journey, exploration, or mission
 - Action-oriented scenes or problem-solving moments
 - Atmosphere emphasizing challenge, thrill, or discovery
- "".strip(),

"Horror": ""

- Dark, tense, or unsettling atmosphere
 - Gradual build-up of fear, dread, or anxiety
 - Elements of threat, mystery, or the unknown
 - Emotional tone that evokes discomfort or suspense
- "".strip(),

"Fantasy": ""

- Fantasy setting with magic or supernatural elements
 - Consistent worldbuilding and internal logic
 - Emotional but not melodramatic tone
 - Characters, events, or visuals reflecting a mythical or otherworldly feel
- "".strip()

"Sci-Fi": ""

- Technology, science, or futuristic concepts integrated into the narrative
- Logical or speculative worldbuilding
- Analytical or reflective tone rather than purely emotional
- Themes involving innovation, artificial intelligence, space, or advanced society

"".strip(),

"Dystopian": ""

- Bleak, oppressive, or controlled societal structure
- Themes of surveillance, inequality, or loss of freedom
- Dark, reflective emotional tone
- Protagonist perspective highlighting resistance, suffering, or systemic issues

APPENDIX. D

LLM-as-a-Judge evaluation-2 criteria

The genre information is given as:

"{genre}"

The part before "->" is the previous genre, and the part after "->" is the new target genre.

Your job is to read:

1) the previous conversation context,

2) the model's new reply,

and then evaluate ****how well the new reply transitions into the new target genre**** while staying coherent with the prior context.

Use the following criteria:

1. Genre shift:

- How clearly does the new reply adopt the style, tone, atmosphere, and typical devices of the ****new target genre**** described by "{genre}"?

2. Contextual coherence:

- Does the new reply still make logical sense given the previous conversation context?

- Even if the mood changes, the situation, characters, and events should not become incoherent without reason.

3. Transition quality:

- Does the change in genre feel intentional, smooth, and motivated by the story (e.g., through setting, mood, vocabulary, or events)?

- Or does it feel random, abrupt, or out of place