# AMAN: Agent for Mentoring and Assisting Newbies in MMORPG

**Jeehyun Lee**[*†]
Sogang University
jhlee22@sogang.ac.kr

**Seung-Moo Yang**[*†]
Seoul National University of Science & Technology
ydaniel0826@ds.seoultech.ac.kr

**Won Ik Cho**[**†]
Seoul National University
tsatsuki@snu.ac.kr

## Abstract

In online games with diverse contents and frequent patch updates, newcomers first learn gameplay mechanics by community intelligence but soon face challenges as they grow up. This calls for their collaboration with other senior gamers, which leads to real-time and practical guidance in the gaming. To let the users easily access to such supportive experience, we introduce AMAN, Agent for Mentoring and Assisting Newbies in MMORPG – a companion chatbot designed to engage novice gamers through multi-turn dialogues. Unlike typical tutorial bots, our model functions as a human-like chat buddy that interacts with users in a friendly manner while providing substantive informational depth. In this light, we propose a multi-stage learning approach that incorporates continual pre-training using a sequence of online resources and depth up-scaling to enhance the model's capabilities with curated dialogues, thereby mimicking the learning process a newcomer might experience through diverse online sources and community feedbacks. To align with gamers' specific needs, we first analyze user-oriented topics from online communities regarding a widely played MMORPG and construct a domain-specific dataset. Furthermore, we develop a multi-turn dialogue data to foster dynamic conversations with users. The evaluation result with the model trained upon publicly available language model shows our practical applicability on how conversational assistant in online games can help novice gamers.

## 1 Introduction

MMORPG refers to massive multiplayer online role-playing games and their social communities (Jon, 2010). Players develop their avatars by completing quests, earning experience points, and enhancing abilities and items (Sourmelis et al., 2017). These elements help the player progress to higher stages or levels of the game, even as the process becomes increasingly repetitive and challenging (Achterbosch et al., 2008). However, MMORPG often presents high barriers to entry due to the complexity of in-game elements and specialized terms. New players, or "newbies," frequently struggle with understanding the gameplay mechanics, i.e. the events within the game employed by agents intended to interact with the game state (Sicart, 2008).

MMORPG players create communities centered around knowledge sharing and social networking (Hsiao and Chiou, 2012; Junghoon Moon and Jo, 2013). Interactions within games are significant motivating factors (Ducheneaut et al., 2006; Alsén et al., 2016), driving engagement and retention through the impact of player connections on the overall gaming experience (El-Nasr et al., 2016). In the communities, senior gamers offer expert advice and learning opportunities to help novices (Gandolfi et al., 2023). However, the vast amount of information and the use of gamer slang words can make it difficult for newcomers to fully participate. Gamer slang encompasses varying levels of linguistic knowledge and expertise (Ensslin, 2011). New gamers require a real-time, readily accessible dictionary to understand the vocabulary and varied meanings within the gaming world.

In response to these challenges and to support novices, we propose a user-friendly chatbot that can simplify complex gameplay mechanics and terminology. Chatbots, interactive agents that offer immediate responses to users (Smutny and Schreiberova, 2020), have demonstrated efficacy in breaking down complex concepts and enhancing user engagement. Users favor the human-like and friendly chatbot over the mechanical, task-focused one (Islind et al., 2023). Applying this concept to MMORPG, our chatbot provides clear, fun explanations to help new players, reducing the knowledge

[*]Equal Contribution.
[**]Corresponding Author.
[†]Work done after graduation.

gap and preparing them to fully enjoy and participate in the community.

As far as the state-of-the-art model benchmarks prove, large language models (LLMs) have been effective in various tasks. Besides, when further trained on domain-specific datasets, they significantly enhance their performance within that particular domain (Wu et al., 2023c,a). This capability makes custom LLMs ideal for applications like assistant chatbots in gaming industry. LLMs have been applied in the gaming in various roles (Gallotta et al., 2024). They can act as in-game players (Toshniwal et al., 2022; Ciolino et al., 2020), serve as non-player characters (NPCs), and function as game masters (GM) directing the game's flow (Triyason, 2023a; Zhu et al., 2023b). While we recognize player assistance as a crucial role for LLMs, existing research has mostly focused on other aspects (Gallotta et al., 2024).

Our paper presents a novel approach to developing a game-specific multi-turn chatbot that supports novice gamers, reflecting traits of real-life conversation. The conversational capabilities emphasized in LLMs are well-suited for providing hints and walkthroughs in game strategies. Unlike typical tutorial bots, our chatbot engages in friendly conversations and provides gameplay tips. Using a custom game-specific dataset built on a well-known MMORPG 'Lost Ark Online' (Figure 1) and a multi-stage learning approach, our chatbot aims to enhance the gaming experience for newcomers by offering relevant support in a specific persona style.

## 2 Related Works

### 2.1 LLMs in Game Domain

The emergence of LLMs, which demonstrate near-human intelligence based on extensive prior knowledge, has opened up new possibilities for the integration across diverse domains. In gaming, LLMs have primarily emerged in two application areas: as game player agents and as supportive tools in as in-game supportive tools.

In the first scenario, LLMs serve as in-game player agents, displaying their abilities in games such as Chess and StarCraft (Toshniwal et al., 2022; Ciolino et al., 2020; Ma et al., 2023). Additionally, LLMs have been used to evaluate each other in text-based games like 20 Questions or Wordle, providing a controlled environment for benchmarking their capabilities (Chalamalasetti et al., 2023).

Figure 1: Screenshots of an actual game scene, for a newbie first encountering an MMORPG, the sheer volume of information can be overwhelming.

In the second role, LLMs are adopted to generate dialogue for NPCs, exploring dimensions like context-aware conversations (Paduraru et al., 2023) and story-focused dialogues (Taveekitworachai et al., 2023). Studies showed that LLM-generated dialogue and quests can enhance player experience (Paduraru et al., 2023), indicating potential to improve user engagement and narrative quality (Paduraru et al., 2023). LLMs can also serve as Game Masters to craft in-game plots and enhance gameplay in games such as Dungeons & Dragons (Zhu et al., 2023a; Triyason, 2023b).

### 2.2 Domain Adaptation

In language model research, there have been continuous findings that domain-adaptive pre-training significantly enhances the capabilities of natural language understanding models (Gururangan et al., 2020; Cheng et al., 2022).

Two principal approaches are used for pre-training domain-specific language models: build-

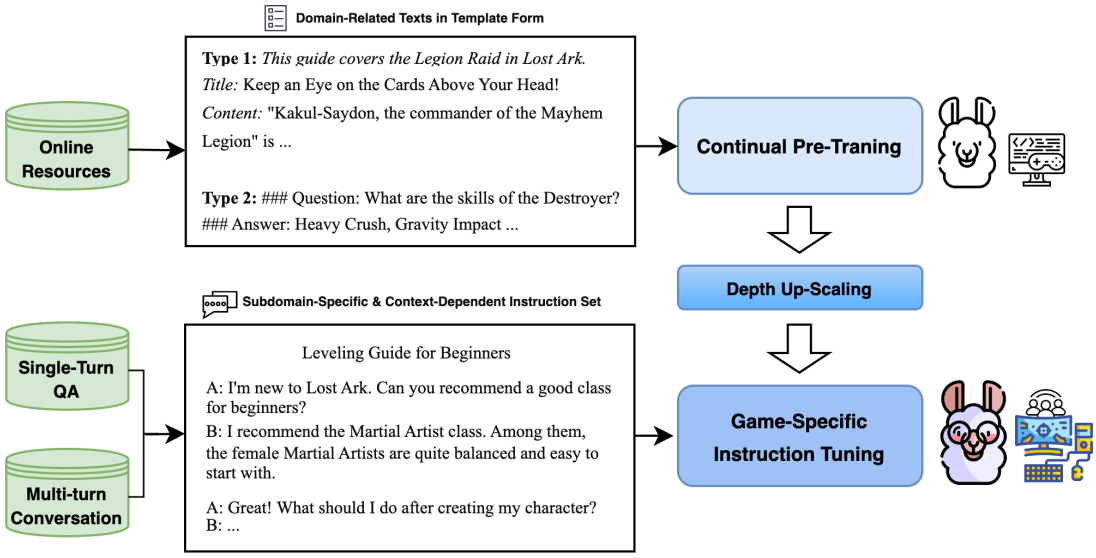



Figure 2: The proposed multi-stage learning approach includes: Stage 0: Continual Pre-Training, Stage 1: Depth Up-Scaling, Stage 2: Game-Specific Instruction Tuning. This process develops a "game buddy bot" capable of engaging in game-related conversation like a friend. It provides both in-depth knowledge and interactive support to newbies, enhancing their overall gaming experience.

ing from scratch and employing continual pre-training. SciBERT (Beltagy et al., 2019), an encoder-only model tailored for scientific literature, exemplifies the from-scratch approach. Models such as BloombergGPT (Wu et al., 2023b), designed for finance, have adopted decoder-only architectures, building on this approach. On the other hand, models like BioBERT (Lee et al., 2020) and Pmc-llama (Wu et al., 2023a), tailored for the medical domain, showcase the advantages of continual training. Through this process, the models undergo gradual refinement to improve their performance in domain-specific tasks. Based on a prior work demonstrating that continual pre-training can achieve competitive results with less data and fewer resources compared to training from scratch (Xie et al., 2023), we have adopted the continual pre-training approach in our work.

### 2.3 Role-Playing Language Models

Recently, role-playing language models (LMs) have been making significant advancements (Shanahan et al., 2023), enabling a variety of innovative applications. These include the creation of digital replicas of individuals (Tik Ng et al., 2024), AI-driven characters in chatbots (Wang et al., 2023b), and role-playing video games (Wang et al., 2023a). As role-playing LMs become more integrated into our daily lives, it is crucial to promote a society that thrives through the synergistic coexistence of humans and these intelligent agents.

In our research, we aim to create a sense of familiarity for users, as if they were conversing with someone with extensive gaming experience. To achieve this, we develop the system by integrating not only knowledge and information that assist in gaming but also multi-turn conversation data.

## 3 Methodology

We propose a multi-stage learning approach to develop an effective conversational assistant for playing MMORPG, focusing on both knowledge acquisition and interactive capabilities. Our approach is inspired by usual journey of the novice gamers of MMORPG. Newcomers that just completed the tutorial would primarily be informed by community intelligence or in-game NPCs (Stage 0). However, as they become senior and their expertise grows (Stage 1), the challenges they face become more complex and tricky. In this circumstances, real-time conversation with other senior gamers would greatly help the player achieve the desired goal (Stage 2). Our multi-stage learning process is outlined as follows:

### 3.1 Stage 0: Continual Pre-Training

The initial stage is the warm-up phase, during which the model undergoes continual learning us-

ing game domain corpora to incorporate game-specific knowledge that the pre-trained LM has not encountered before. Before instruction-tuning, domain-adaptive pre-training is conducted to enable the model to learn complementary representations of the game domain. This process reflects how newcomers initially acquire game knowledge through online resources, encompassing an understanding of the game's lore, mechanics, character backgrounds, and strategic elements.

We train an LM on domain-related texts using two kinds of templates, as outlined in Figure 2. The first type guides the model to understand the relationships between game entities and concepts within specific game subdomains by providing structured content, including instruction, document title, and content. The second type, designed for the single-turn question answering (QA) format used in the next instruction tuning stage, equips the model with the ability to answer game-related questions accurately.

### 3.2 Stage 1: Depth Up-Scaling

In the intermediate stage, we scale up the model parameter size following the approach applied in Kim et al. (2023). Similar to how gamers progressively expand their foundational game knowledge, we replicate the model of Stage 0 and extend it by duplicating selected layers.

### 3.3 Stage 2: Game-Specific Instruction Tuning

The final training stage reflects the shift in how players learn the game as their proficiency grows. To mirror this, we employ instruction tuning on a combined dataset of single-turn and multi-turn dialogues.

As players gain experience, their questions become more specific. Single-turn QA data enables the model to answer specific queries about the game. This includes topics like patch notes, class skills, equipment details, and dungeon guides.

Also, experienced players often engage in deeper conversations with others. Multi-turn conversations allows the model to engage in context-aware interactions. This data, consisting of transcripts from real player conversations, helps the model understand the flow of conversation, and be exposed to how players might ask questions in different ways.

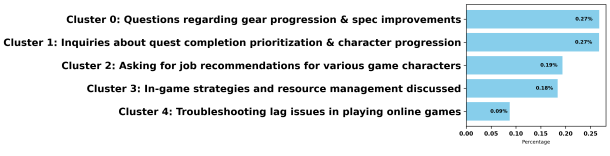
## 4 Dataset

For concrete implementation of our method, we chose a widely played MMORPG 'Lost Ark Online' serviced by Smilegate (Figure 1) which holds 1.2M global users. Though the game is serviced all around the world, here we target Korean gamer communities where the discussion on the gaming strategies and patch updates is active.

### 4.1 User Analysis

| Topic |
|---|
| **Topic 0: Raids & Gear Progression**<br>• Keywords: Abyss dungeon, Gear, Raids, Valton gate, TriPod, Commanders, Quest, Argos, Chaos dungeon |
| **Topic 1: In-Game System**<br>• Keywords: Engravings, (Ancient) Accessories, Setting, Jewelries, Combat engravings, Avatar |
| **Topic 2: Character Selection & Development**<br>• Keywords: Class, vs., Main / Alternate character, Recommend, Barracks, Preferable, Growing |
| **Topic 3: Beginner's Tips & Strategies**<br>• Keywords: Jumping Ticket, Newbie, Mokoko, Card, Event, Story, Super mokoko express |

(a) LDA Topic Modeling Result




(b) K-means Clustering Result

Figure 3: Topic Distributions of sampled questions from gaming community.

We conducted user analysis to explore the topics that gamers mostly inquire about while playing. We collected 40,299 posts from QA section of the game community website[1] between January 20, 2022, and April 9, 2024. These posts were processed by combining titles and questions into single sentences, removing stopwords, and then vectorizing them based on frequency using the top 1,000 most frequent words. We applied topic modeling (Jelodar et al., 2019) to these vectorized representations. The result is presented in Figure 3a. Users show interest in topics related to Raids and Gear Progression (e.g., Abyss dungeon, Quest), In-Game Systems (e.g., Engravings), Character Selection and Development (e.g., Main/Alternate characters), and Beginner's Tips and Strategies (e.g., Jumping Ticket, Newbie).

[1] https://www.inven.co.kr/board/lostark/4822

Next, as shown in Figure 3b, we performed K-means clustering to analyze topic distribution. To compute sentence embeddings, we used Korean version[2] of SentenceTransformers (Reimers and Gurevych, 2019). The number of clusters was determined using the elbow method. For each cluster, we selected the 100 sentences closest to the centroid and use ChatGPT (OpenAI, 2023) to summarize the central topics of each cluster. Most questions are about leveling guide and game walkthrough. People seek advice on upgrading equipment settings and making spec improvements. Prioritizing quests for character development and requesting class recommendations are also popular questions.

### 4.2 Dataset Collection

| Stage | Data Type | Domain | Statistics | Number |
|---|---|---|---|---|
| Stage 0 | Raw Document | - | Total # Sentences | 52,395 |
| | | | # Tokens | 8.7M |
| Stage 2 | Single-Turn QA | Story | Total # | 237 |
| | | Class (subclass, skill, engraving) | Total # | 445 |
| | | Balance Patch Update | Total # | 1,561 |
| | | Dungeons & Raids guide | Total # | 766 |
| | | Set Effect of Equipment | Total # | 284 |
| | Multi-Turn Conversation | Game-Specific Knowledge | Total # Dialogues | 61 |
| | | | Avg. # Turns per Dialogue | 21.67 |
| | | | Total # Turns | 1322 |
| | | Chit-chat | Total # Dialogues | 950 |
| | | | Avg. # Turns per Dialogue | 20 |
| | | | Total # Turns | 19,003 |

Table 1: Data Statistics

The detailed statistics of the dataset for each stage are summarized in Table 1. For Stage 0, we collected 52,395 sentences in total, containing 8.7M tokens, using Korean wiki articles about Lost Ark and posts from the Q&A and Story sections in the aforementioned gaming community. We filtered out the noisy posts and remove any profanity. This stage emphasizes the width of strategic content and discussions within the gaming community.

In Stage 2, we put further effort into collecting the data for specific knowledge and context-aware interactions. Firstly, we gathered 3,293 QA sets under various categories. Two gamers who are well-versed in Lost Ark, having played for over three years, organized documents summarizing essential in-game information covering categories such as story, class, gear, etc. Based on these documents and the dungeon guide from Stage 0, we developed a set of questions and answers. These questions were informational and objective, designed with no open-ended responses. We enhanced the QA set with query augmentation to ensure robust model performance across various question phrasings.

Additionally, we used the multi-turn dialogues collected in-house to reflect the conversations of players. First, we asked for the same human participants to create dialogues on topics such as basic game information, character classes, user culture, dungeons & raids, and leveling methods, in a self-play manner, all while maintaining the friendly persona. Detailed guidelines are outlined in Appendix A. Nextly, we constructed chit-chat data to enhance the chatbot's persona consistency and improve its daily conversation capabilities.

For the conversation on game-specific knowledge, the total dialogue count is relatively small, reaching 61 instances, which is smaller than previous multi-turn dialogue datasets (Feng et al., 2023; Cui et al., 2020). However, this does not necessarily mean the quality degradation of the dataset (Kryscinski et al., 2019). Our dataset facilitates in-depth conversations across multiple topics. Also, the average number of turns per dialogue is quite longer than in other benchmark datasets (Li et al., 2017; Feng et al., 2023; Bai et al., 2024) since most dialogues cannot be terminated in a short turn due to the complex mechanics of the game. The last stage highlights the intricate depth of knowledge sharing about Lost Ark, ensuring that chatbot can understand questions posed in various forms about the same topic.

### 4.3 Dataset Composition

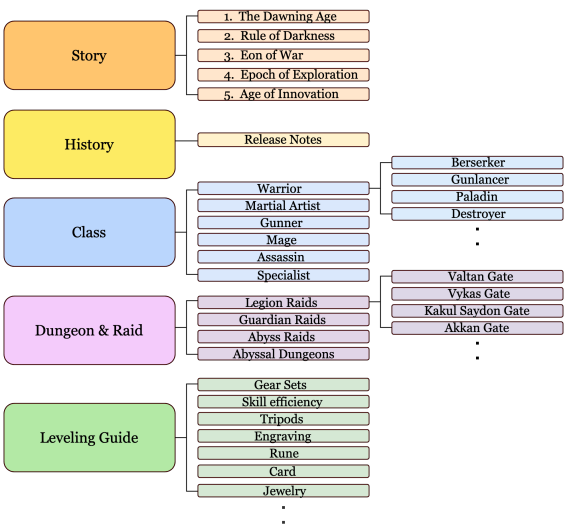



Figure 4: Hierarchical Taxonomy of Lost Ark

[2] https://huggingface.co/jhgan/ko-sroberta-multitask

| Single-Turn QA | |
|---|---|
| **Category** | **Example** |
| Story | **Q:** I am curious about the story of The Dawning Age, the Birth of the World in Lost Ark.<br>**A:** In the beginning, a world existed where only chaos prevailed, and from it, order was born ... |
| Class | **Q:** Tell me about the skills of the Warlord class.<br>**A:** Hook Chain, Rising Spear, Dash Upper Fire ... |
| Leveling Guide | **Q:** What is the Tripod system?<br>**A:** The Tripod system is a skill enhancement system. You can raise the skill level by acquiring amulets ... |
| Dungeon & Raid | **Q:** What is the reward of the Abyss Dungeon 'Ivory Tower of Chaos'?<br>**A:** The primary reward is the 'Elixir of Wisdom and Energy of Wisdom' ... |
| Balance Patch Update | **Q:** Can you tell me how the Sorceress class was updated in the May 8, 2024 update?<br>**A:** The effects of the 'Igniter and Reflux' engraving have been changed as follows ... |
| **Multi-Turn Conversation** | |
| **Category** | **Example** |
| Leveling Guide | **Speaker A**: Thinking about going to Argos today!! Is Argos a manageable raid∼?<br>**Speaker B:** It depends on your level! Are you around level 1460?<br>**Speaker A:** Yeah, about that.<br>**Speaker B:** Then no worries! If you know the essential mechanics, you can skip almost all of the damage ... |
| Dungeon & Raid | **Speaker A:** I'm thinking of trying Valtan now, what kind of raid is Valtan?<br>**Speaker B:** Finally trying Valtan, huh?! Valtan is, um... a raid to see if the player and Lost Ark really click!! Haha!<br>**Speaker A:** It's a test to see if the player and Lost Ark click?<br>**Speaker B:** Yep! Usually, if someone enjoys Valtan, they tend to enjoy the raids that come after ...<br>⋮ |

Table 2: Example of our dataset.

The hierarchical taxonomy of the game is illustrated in the Figure 4. Lost Ark features a highly complex data structure where characters are developed based on the epic story, each possessing unique skills and engravings. Players participate in raids to level up items and use runes and jewelries to enhance their skills. We categorize the single-turn QA set based on these game's features.

**Story** Lost Ark encompasses the story and background of Arkrasia, the in-game world, exploring past events in-depth and elucidating the unique traits and culture of its characters. We utilize a worldview composed of five main parts.

**Leveling Guide** In Lost Ark, both gear level and growth level are crucial for character development and progression. Gear level impacts a character's strength and enables access to higher-level content such as advanced dungeons and raids. Growth level enhances skills and stats, improving combat efficiency and unlocking new game features. To facilitate understanding of these concepts for model training, we have created a QA dataset.

**Class** Lost Ark offers 6 main class archetypes that divide into many advanced sub-classes, each with its own unique skills. Class engravings, which provide various effects and bonuses, enhance the gameplay experience for each class. We craft QA sets that cover the main classes, their sub-classes, and their skills and engravings.

**Dungeon & Raid** Lost Ark centers on legion raids, which are endgame content, where players enhance their items through refinement to increase their item levels and advance their progress. We transform dungeon guides gathered from the gaming community into a QA format.

**Balance Patch Update** Game updates with new content are crucial in online gaming, enhancing player engagement and retention (Hyeong et al., 2020). Staying updated with patch notes helps players adapt to game evolution. When constructing a single-turn QA set, we compile balance patch updates by date and class from the official website.

## 5 Experimental Setup

### 5.1 Model and Training

For training and evaluation, we adopted the model derived from LLaMA2 (Touvron et al., 2023). We trained the Korean version of LLaMA2 (L. Junbum, 2023) for Stage 0 and Stage 2, with the original model depth up-scaled from parameter 7B to 10B in Stage 1, with the duplication of 8 layers. The model was trained for 3 epochs with a batch size of 64. We utilized AdamW (Loshchilov and Hut-

ter, 2017) as our optimizer, incorporating cosine learning rate scheduling and weight decay.

### 5.2 Evaluation

Accuracy in QA systems is often more critical in specific domains compared to general QA (Siriwardhana et al., 2023), and the same holds for the gaming domain.

To evaluate our models, we have created a test dataset specifically curated for open-domain QA within the MMORPG domain. This test set was build by aforementioned gamers of Lost Ark. They have crafted questions that a real user might be curious about, covering topics such as leveling, class, story, and raid. For single-turn questions, the set contains 20 instances. For multi-turn questions designed to assess context understanding, two additional sentences were added to each single-turn question, resulting in a total of 60 instances. Utilizing this set, we systematically conduct comparative analyses of our models using diverse methods. The prompt used can be found in Appendix B.

**Human Evaluation** We adopted three human evaluation criteria to measure the quality of the generated responses: (1) fidelity in reflecting **knowledge**, assessed on a binary scale (0 or 1) (2) adherence to a friendly conversational **style** (0 to 2 scale), and (3) **fluency** and appropriateness of the response in relation to the context (0 to 2 scale)

**Automatic Evaluation** To compare the model performance in replicable manner and relate them with human evaluation, we additionally compute the BERTScore (Zhang* et al., 2020), which measures F1 scores by matching token embeddings between the human reference and chatbot response. Besides, to further evaluate the conversational style, we trained a style classifier and measured its average probability of predicting a target style (StyleProb). Detailed training method of style classifier is in Appendix C.

## 6 Results & Analysis

The Cohen Kappa scores (Cohen, 1960) between the two raters are as follows: 0.318 for Knowledge, 0.4622 for Style, and 0.1586 for Fluency. Knowledge scores remained stable through the stages after continued pre-training, with some categories (Class, Story) even showing improvement. In the Class category, a significant improvement was shown in Knowledge scores across stages, likely due to the category's nature being closer to closed QA than open-ended QA. Indeed, the Knowledge Cohen Kappa score for the Class category is 0.4696, indicating a high level of agreement between evaluators. As progression advances from Stage 0 to Stage 2, Fluency scores increase. This conveys that training the model with multi-turn conversations can help it better reflect context and improve dialogue fluency. Besides, Style scores significantly rise from Stage 0 to Stage 2-1 and then to Stage 2-2, which means that adding chit-chat dataset results in more friendly and conversational responses. However, Style scores were generally lower in the story category due to its tendency for longer, narrative responses. The experimental results show that both contextual understanding and stylistic representation improve in the latter stages.

| Methods | Category | Human Evaluation | | | Automatic Evaluation | |
|---|---|---|---|---|---|---|
| | | Knowledge | Style | Fluency | BERTScore | StyleProb |
| Stage 0 | Class | 0.417 | 0.028 | 0.815 | 0.668 | 0.234 |
| | Leveling | 0.614 | 0.162 | 0.654 | 0.661 | 0.375 |
| | Raid | 0.533 | 0.222 | 0.744 | 0.651 | 0.324 |
| | Story | 0.237 | 0.921 | - | 0.689 | 0.129 |
| | Mean | 0.450 | 0.333 | 0.738 | 0.667 | 0.266 |
| Stage 2-1 | Class | 0.722 | 1.204 | 1.713 | 0.726 | 0.470 |
| | Leveling | 0.727 | 1.055 | 1.578 | 0.699 | 0.508 |
| | Raid | 0.533 | 1.156 | 1.8 | 0.720 | 0.288 |
| | Story | 0.526 | 0.237 | - | 0.714 | 0.072 |
| | Mean | 0.627 | 0.913 | 1.697 | 0.715 | 0.335 |
| Stage 2-2 | Class | 0.725 | 1.298 | 1.442 | 0.685 | 0.772 |
| | Leveling | 0.523 | 1.432 | 1.515 | 0.688 | 0.863 |
| | Raid | 0.6 | 1.422 | 1.6 | 0.7 | 0.781 |
| | Story | 0.553 | 1.605 | - | 0.714 | 0.183 |
| | Mean | 0.6 | 1.439 | 1.519 | 0.697 | 0.64 |

Table 3: Test set results: Reflection of Knowledge, Contextual Fluency, and Friendly Style. Scores are based on: (1) fidelity in reflecting knowledge (0 or 1 scale), (2) adherence to a friendly conversational style (0 to 2 scale), and (3) fluency and appropriateness in context (0 to 2 scale). In the table, Stage 2-2 is our final proposed methodology, AMAN.

Figure 5 displays the real conversation history of the test set across various stages. In response to the user's first question, Stage 0 model provides a detailed answer but it is excessively long and includes irrelevant content. In contrast, Stage 2-1 and Stage 2-2 models address the core aspect of the query and answer it accurately, and in particular, the latter responds as if it were an actual mate that the user encounters while playing Lost Ark. In the following question, Stage 0 model focuses on the term 'item level' and fails to understand the context. However, Stage 2 models provide accurate answer while maintaining a friendly tone, which guarantees the experience that the user might get from human mates in his/her onboarding and growth.

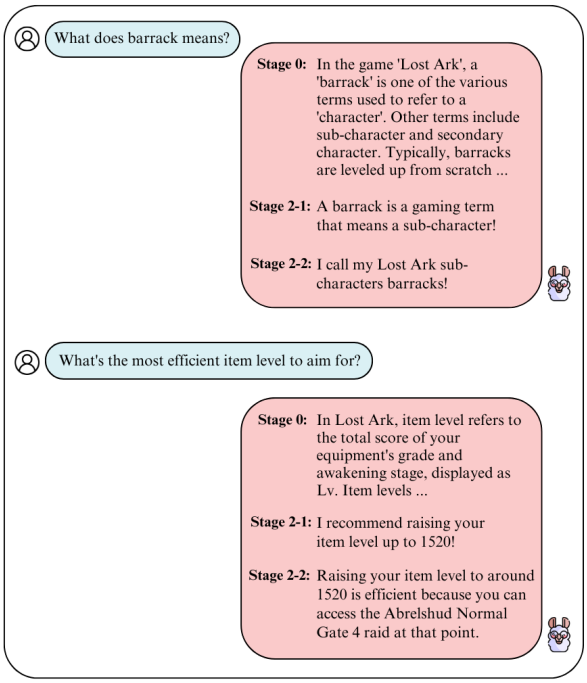



Figure 5: Example of a multi-turn conversation from the test set, showing the responses generated by each stage for given context.

## 7 Conclusion

We introduce a chatbot named 'AMAN' for newbie gamers. Our experimental results show that the model developed with this method successfully balances friendliness and accuracy. Our study demonstrates the potential for LLM to effectively serve the gaming community by covering multiple in-game topics, particularly in reducing the challenges faced by newer players. We hope our approach encourages further research in this field and leads to the development of more accessible and friendly game companion chatbots.

## Limitations

Though this study was conducted specifically for the MMORPG genre, it is likely to be suitable for most game genres that require question and answer interactions. However, we note that our experiment was conducted only with a single MMORPG 'Lost Ark' which holds significant amount of users worldwide, that our methodology may not be necessarily effective for independent or small-scaled games, and games in other genres as well.

Besides, despite our efforts to capture comprehensive game knowledge, our current approach faces limitations in terms of data storage capacity. Incorporating retrieval-augmented generation (RAG) into our framework would significantly enhance the promptness and accuracy of our model's responses, owing to its ability to retrieve and integrate relevant information from external sources.

## References

Leigh Achterbosch, Robyn Pierce, and Gregory Simmons. 2008. Massively multiplayer online role-playing games: the past, present, and future. *Comput. Entertain.*, 5(4).

Adam Alsén, Julian Runge, Anders Drachen, and Daniel Klapper. 2016. Play with me? understanding and measuring the social aspect of casual gaming. *CoRR*, abs/1612.02172.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *CoRR*, abs/2402.14762.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11174–11219. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, Jianfeng Liu, Yuefeng Zhan, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2022. Snapshot-guided domain adaptation for electra. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2226–2232.

Matthew Ciolino, Josh Kalin, and David Noever. 2020. The go transformer: Natural language modeling for game play. In *Third International Conference on Artificial Intelligence for Industries, AI4I 2020, Irvine, CA, USA, September 21-23, 2020*, pages 23–26. IEEE.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1406–1416. Association for Computational Linguistics.

Nicolas Ducheneaut, Nicholas Yee, Eric Nickell, and Robert J. Moore. 2006. "alone together?": exploring the social dynamics of massively multiplayer online games. In *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, pages 407–416. ACM.

Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. 2016. *Game analytics*. Springer.

Astrid Ensslin. 2011. *The Language of Gaming*.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 7348–7363. Association for Computational Linguistics.

Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis. 2024. Large language models and games: A survey and roadmap. *CoRR*, abs/2402.18659.

Enrico Gandolfi, Richard E Ferdig, and Ilker Soyturk. 2023. Exploring the learning potential of online gaming communities: An application of the game communities of inquiry scale. *New Media & Society*, 25(6):1374–1393.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Cheng-Chieh Hsiao and Jyh-Shen Chiou. 2012. The effect of social capital on community loyalty in a virtual community: Test of a tripartite-process model. *Decision Support Systems*, 54(1):750–757.

Ji Hyeon Hyeong, Kang Jun Choi, Jae Young Lee, and Tae-Hyung Pyo. 2020. For whom does a game update? players' status-contingent gameplay on online games before and after an update. *Decision Support Systems*, 139:113423.

Anna Islind, María Óskarsdóttir, Svanhvít Smith, and Erna Arnardottir. 2023. The friendly chatbot: Revealing why people use chatbots through a study of user experience of conversational agents.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211.

Allan Jon. 2010. The development of mmorpg culture and the guild. *Australian Folklore: A Yearly Journal of Folklore Studies*, 25:97–112.

G. Lawrence Sanders Edward J. Garrity Junghoon Moon, Md. Dulal Hossain and Sooran Jo. 2013. Player commitment to massively multiplayer online role-playing games (mmorpgs): An integrated model. *International Journal of Electronic Commerce*, 17(4):7–38.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 540–551. Association for Computational Linguistics.

L. Junbum. 2023. llama-2-ko-7b (revision 4a9993e).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Junbum Lee. 2021. Kcelectra: Korean comments electra. `https://github.com/Beomi/KcELECTRA`.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Weiyu Ma, Qirui Mi, Xue Yan, Yuqiao Wu, Runji Lin, Haifeng Zhang, and Jun Wang. 2023. Large language models play starcraft II: benchmarks and A chain of summarization approach. *CoRR*, abs/2312.11865.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ciprian Paduraru, Marina Cernat, and Alin Stefanescu. 2023. Conversational agents for simulation applications and video games. In *Proceedings of the 18th International Conference on Software Technologies, ICSOFT 2023, Rome, Italy, July 10-12, 2023*, pages 27–36. SCITEPRESS.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Miguel Sicart. 2008. Defining game mechanics. *Game Stud.*, 8(2).

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Pavel Smutny and Petra Schreiberova. 2020. Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, 151:103862.

Theodoros Sourmelis, Andri Ioannou, and Panayiotis Zaphiris. 2017. Massively multiplayer online role playing games (mmorpgs) and the 21st century skills: A comprehensive research review from 2010 to 2016. *Computers in Human Behavior*, 67:41–48.

Pittawat Taveekitworachai, Febri Abdullah, Mustafa Can Gursesli, Mury F. Dewantoro, Siyuan Chen, Antonio Lanatà, Andrea Guazzini, and Ruck Thawonmas. 2023. What is waiting for us at the end? inherent biases of game story endings in large language models. In *Interactive Storytelling - 16th International Conference on Interactive Digital Storytelling, ICIDS 2023, Kobe, Japan, November 11-15, 2023, Proceedings, Part II*, volume 14384 of *Lecture Notes in Computer Science*, pages 274–284. Springer.

Man Tik Ng, Hui Tung Tse, Jen-tse Huang, Jingjing Li, Wenxuan Wang, and Michael R Lyu. 2024. How well can llms echo us? evaluating ai chatbots' role-play ability with echo. *arXiv e-prints*, pages arXiv–2404.

Shubham Toshniwal, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2022. Chess as a testbed for language model state tracking. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11385–11393. AAAI Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tuul Triyason. 2023a. Exploring the potential of chatgpt as a dungeon master in dungeons & dragons tabletop game. In *Proceedings of the 13th International Conference on Advances in Information Technology, IAIT 2023, Bangkok, Thailand, December 6-9, 2023*, pages 3:1–3:6. ACM.

Tuul Triyason. 2023b. Exploring the potential of chatgpt as a dungeon master in dungeons & dragons tabletop game. In *Proceedings of the 13th International Conference on Advances in Information Technology*, IAIT '23, New York, NY, USA. Association for Computing Machinery.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023c. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *arXiv preprint arXiv:2311.08545*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. 2023a. Calypso: Llms as dungeon masters' assistants. In *Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE '23. AAAI Press.

Andrew Zhu, Lara J. Martin, Andrew Head, and Chris Callison-Burch. 2023b. CALYPSO: llms as dungeon masters' assistants. *CoRR*, abs/2308.07540.

# Appendices

## A Guidelines for participants

The dataset for Stage 2 was compiled with the participation of gamers who have over three years of experience and had logged in at least once within the two weeks preceding the data collection period. Participants created a multi-turn dataset based on key gaming categories, such as basic game information, story, leveling methods, classes, user culture, and raids. The dataset was constructed according to the following guidelines.

- Choose a subdomain within a larger theme to compose the dialogue.
- Each conversation includes an average of more than 20 turns.
- Maintain natural and everyday conversation, but ensure it contains meaningful game knowledge.
- The character should be friendly and approachable, like a friend.
- Do not include violent or explicit content.
- Do not use profanity.
- Allow the use of slang and abbreviations that appear in games.
- At the end of each dialogue session on a specific topic, an administrator reviews it.

## B Prompt Templates

System prompt: *"A user discusses various aspects of Lost Ark with an expert, covering topics such as raids, the story, character development, and classes. When explaining the story, the expert elaborates on the Lost Ark universe, highlighting key eras and events including the Dawning Age, Rule of Darkness, Eon of War, Epoch of Exploration, and Age of Innovation."*

## C Detailed Training Methodology of the Style Classifier

We trained the style classifier as a binary model: 0 represents data without distinctive character style, specifically using the formal bot responses from the OIG-small-chip2-ko dataset[3] and 1 corresponds to the chit-chat dataset used in stage 2, as in the Table 1. Additionally, we incorporated 3,880 sentences from the AI-hub Korean text style conversion dataset[4], categorizing them formal (0) or colloquial (1) based on style, with each category having 1,940 sentences. We employed the Korean comment ELECTRA (Clark et al., 2020; Lee, 2021). This model is pre-trained on NAVER news comments, which often contain typos and informal expressions not typical in formal datasets.

[3] https://huggingface.co/datasets/heegyu/OIG-small-chip2-ko

[4] https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=287