Large Language Models and Video Games: A Preliminary Scoping Review

PENNY SWEETSER, The Australian National University, Australia

Large language models (LLMs) hold interesting potential for the design, development, and research of video games. Building on the decades of prior research on generative AI in games, many researchers have sped to investigate the power and potential of LLMs for games. Given the recent spike in LLM-related research in games, there is already a wealth of relevant research to survey. In order to capture a snapshot of the state of LLM research in games, and to help lay the foundation for future work, we carried out an initial scoping review of relevant papers published so far. In this paper, we review 76 papers published between 2022 to early 2024 on LLMs and video games, with key focus areas in game AI, game development, narrative, and game research and reviews. Our paper provides an early state of the field and lays the groundwork for future research and reviews on this topic.

CCS Concepts: • **Software and its engineering** → **Interactive games**; • **Computing methodologies** → **Natural language generation**; • **Applied computing** → **Computer games**.

Additional Key Words and Phrases: large language models, LLMs, games, videogames, GPT

1 INTRODUCTION

Since OpenAI's release of ChatGPT in late 2022, awareness and usage of Large Language Models (LLMs) has surged in research, development, and the general population across a broad spectrum of domains. LLMs are powerful tools for language processing and prediction, pre-trained on vast collections of natural language, and capable of performing diverse language analysis and generation tasks [23]. The release of ChatGPT, along with the many other available LLMs (e.g., GPT-4, LLaMa, Codex, BERT) has opened new doors to research and development potential, which has seen a recent increase in related research. Like many fields, LLMs hold interesting possibilities for video games, which has prompted many researchers to hasten to investigate the potential for applying LLMs to various aspects of video game research and development. Although the concept of generative AI is not new to video games, with decades of prior work in AI-powered generation of game content [26, 46], LLMs have the potential to revolutionise generation and co-creation of video game content, along with game development tools and processes, and games research approaches. As research and development of LLMs and games is occurring and evolving quickly, it is difficult to capture a full picture of how LLMs are being used in games research. The aim of this paper is to provide a preliminary scoping review of LLMs and video games, surveying the related research conducted between 2020 and 2023. We aim to identify the ways in which researchers have been exploring the use of LLMs for game development and research to date. To identify the relevant papers, we conducted a Google Scholar search for papers published between 2020-2023 (and very early 2024). We identified 76 relevant papers from 2260 results returned in the search. Our review of the papers revealed

Author's address: Penny Sweetser, penny.kyburz@anu.edu.au, The Australian National University, Canberra, Australia.

Table 1. Relevant papers on video games and LLMs by theme.

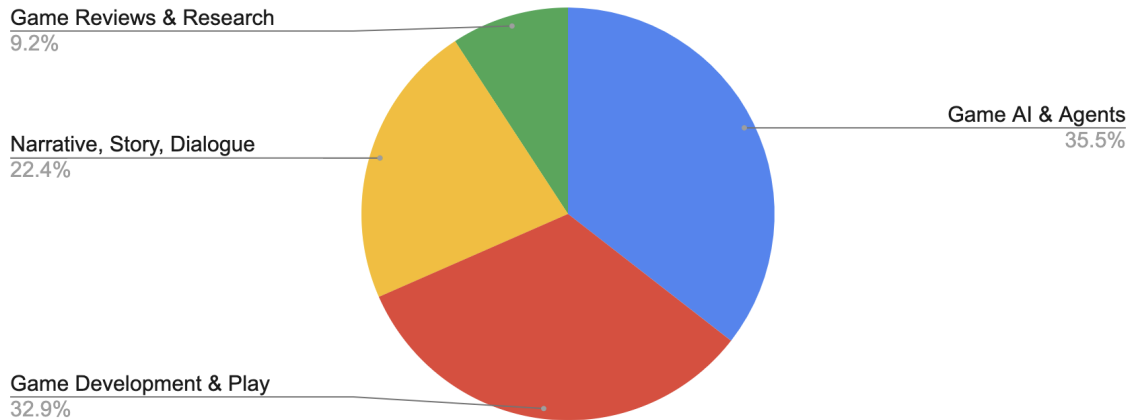| Theme	Papers
Game AI and Agents	[1, 9, 12, 14, 15, 18, 19, 21, 29, 30, 36, 37, 39–42, 47, 48, 50, 51, 73, 76–80, 88]
Game Development and Play	[4, 6, 8, 11, 13, 20, 27, 28, 31, 34, 44, 53–57, 59, 60, 65–67, 70, 71, 83, 87]
Narrative, Story, and Dialogue	[2, 3, 5, 16, 22, 32, 35, 43, 45, 49, 61, 62, 68, 69, 72, 75, 82]
Game Research and Reviews	[17, 24, 25, 33, 52, 64, 74]
Recommendation	[7, 10, 38, 58, 63, 81, 84–86]

Fig. 1. Percentage of relevant papers identified in search by theme.

that 27/76 (35.5%) of the papers related to game AI, 25/76 (32.9%) to game development, 17/76 (22.4%) to narrative, and 7/76 (9.2%) to game research and reviews. We additionally found another 9 papers that used game data sets as part of broader recommendation research. In this paper, we analyse the work conducted using LLMs for games research to date, providing a snapshot of the state of the field and a grounding for future research in this area.

2 SCOPING REVIEW

In order to carry out our scoping review, we conducted a Google Scholar search for papers related to LLMs and video games published between 2020 to 2023 (and very early 2024). We used Google Scholar as it includes a wide variety of sources and publication types and allows full text search. We conducted our search on 15 January 2024, using the search string "(LLM OR LLMs OR "large language model" OR "large language models") AND ("video game" OR "video games" OR videogame OR videogames)" for the period 2020 to 2024. We used a full text search (rather than title or abstract only). Our search returned 2260 results, which we sorted by relevance. We reviewed the first 500 returned records to identify relevant papers. We found that the majority (80%) of relevant papers were in the first 250 records, with relevance rapidly decreasing after this point.

To identify relevant papers, we reviewed the title and abstract of each paper, and searched the full text for "game" and "large language model" to find the usage. Most papers mentioned games in previous work or as an example, but did

not include games in their own research. Some papers used the acronym "LLM" for another meaning. We considered papers relevant if they reported on original research related to video games and LLMs, either with a focus on games or using games as a testbed. We excluded general reviews that mentioned games and broader work that could be applied to games but that had not been applied to games in the paper. Through our search, we identified 71 relevant papers from the 2260 results. We found an additional five relevant papers while reviewing the 71 identified papers, for a full set of 76 papers. The majority (68/76, 90.7%) of papers were published (including preprints) in 2023, with 2 (2.7%) in 2024 and 6 (8.0%) in 2022. This suggests that any future searches with the same focus could start from 2022. A search with the same search string from 2022 onwards returned 1860 records (400 fewer).

During our search, we coded our identified relevant papers into broad categories, so that we could further review the similar papers together and identify the key themes. Following the initial coding, we finalised 4 key themes that captured all papers and recoded all papers into these themes. The most common theme was "Game AI and Agents" with 27/76 (35.5%) of the papers. The second theme was "Game Development and Play" with 25/76 (32.9%) the papers. The third theme was "Narrative, Story, and Dialogue" with 17/76 (22.4%). The fourth and final theme, "Game Research and Reviews", captured the remaining 7/76 (9.2%) of the papers. We additionally found another 9 papers that used game data sets as part of broader recommendation research. Although we did not consider this to fit the focus of our search, we believe this related topic is of interest and relevant to future work on games and LLMs, so we also captured these papers for analysis.

The final step in our scoping review was to read and analyse each of the relevant papers. The following sections present our analysis of the papers, grouped by each of our key themes. We summarise the focus of the research conducted under each of these themes, extract sub-themes, and discuss relevant findings. Across our key themes, we found that by far the most commonly used LLM was GPT (65/76, 85.5%), followed by LLaMA (7/76, 9.2%), and Codex and BERT (each 5/76, 6.6%), with some papers using multiple LLMs.

2.1 Game AI and Agents

Over a third (27/76, 35.5%) of the relevant papers identified in our search related to the theme of Game AI and Agents. The papers within this theme related to game agent design and behaviours (11/27) [14, 15, 29, 36, 39, 40, 42, 50, 51, 76, 88], LLMs and reinforcement learning (9/27) [18, 30, 37, 47, 48, 77–80], and collaboration and coordination (7/27) [1, 9, 12, 19, 21, 41, 73].

The papers on game agent design and behaviours experimented with applying LLMs to different aspects of agents, including reasoning [29, 39], playing games [40, 42], embodied agents [51, 76], goals and tasks [14, 36], emotion simulation [15], believable agents [50], and general AI [88]. The papers on reasoning tested LLM capabilities at logical reasoning via Minesweeper [39] and social reasoning in a custom multi-agent environment [29]. Li et al. [39] found that LLMs possessed foundational abilities, but that they could not integrate these into a multi-step logical reasoning process to solve Minesweeper. Kaiya et al. [29] argued that their Lyfe Agents exhibited human-like self-motivated social reasoning, which could enrich human social experiences in virtual worlds. The papers on playing games applied GPT-4V(ision) to analyse video to predict next steps when playing a Mario video game [40] and to test the capabilities of LLMs to play StarCraft II (SC2) [42]. Lin et al. [40] claimed that their agent displayed an understanding of the Mario game and generated reasonable action controls to play the game effectively. Ma et al. [42] concluded that LLMs possess the relevant knowledge and complex planning abilities to play SC2 (in their textual SC2 environment). The papers on embodied agents investigated the use of diff history for environment observations in NetHack [51] and an LLM-powered embodied lifelong learning agent (Voyager) in Minecraft [76]. Piterbarg et al. [51] found that using diff history allowed

observational compression and abstraction that yielded a 7x improvement in game score and outperformed agents that use visual observations by over 40%. Wang et al. [76] argued their their Voyager agent shows strong in-context lifelong learning capability and exceptional proficiency at playing Minecraft. The paper on tasks [36] presented an interactive task learning system (VAL) that uses LLMs only for specific tasks to support interactive learning of hierarchical task knowledge from natural language. They found that the acquired knowledge was human interpretable and generalised to support execution of novel tasks without additional training and that users could successfully teach VAL using natural language in the Overcooked-AI video game setting. The paper on goals [14] introduced an LLM-based agent to support the representation, generation and learning of diverse, abstract, human-relevant goals in a text-based environment, Cooking World. They found that their agent learned to master a large diversity of skills without reward functions, curriculum, or hand-coded goal representations. The remaining papers investigated using LLMs for agents to solve emotional intelligence tasks and to simulate emotions in a conversational RPG game [15], to synthesise memories over time into high-lever reflections used to plan behaviour in a Sims-like environment [50], and to create generally capable agents in Minecraft [88]. Croissant et al. [15] concluded that their work provided a first step towards better affective agents represented in LLMs. Park et al. [50] argued that their generative agents produced believable individual and emergent social behaviors. Finally, Zhu et al. [88] concluded that their agent demonstrated the potential of LLMs in developing capable agents for handling long-horizon, complex tasks, and adapting to uncertainties in open-world environments.

The papers on reinforcement learning (RL) investigated combining LLMs with RL in game environments, focusing on state descriptions [47, 48], exploration and intrinsic motivation [18, 30], chain-of-thought reasoning [77, 79], rewards [37], embodied agents [80], and adversarial play [78]. The work focusing on state descriptions [47, 48] proposed a method for automatically selecting concise state descriptions for LLM actors in NetHack. They found they could reduce the length of state descriptions by 87% and improve task success rates by 158% in NetHack. The research on exploration and intrinsic motivation used background knowledge from LLMs to shape exploration in the Crafter game environment (2D Minecraft) [18] and an LLM to construct intrinsic rewards in NetHack [30]. Du et al. [18] found that their agents had better coverage of common-sense behaviours, while Klissarov et al. [30] found that their method significantly outperformed existing approaches. For rewards, Lee and Lee [37] investigated whether an LLM can postprocess reward signals for RL in MineDojo. They reported that LLMs can suggest a new combination of existing heuristic functions, but did not observe improved learning results. On adversarial play, Xu et al. [78] proposed a new framework for LLM-based agents with strategic thinking ability to play the language game Werewolf. They found that combining LLMs with the RL policy produced a variety of emergent strategies, achieved a higher win rate, and was robust against human players. The chain-of-thought (CoT) papers found that LLMs prompted with CoT outperformed state-of-the-art RL baselines in Crafter [77] and benchmarks such as Overcooked and FourRoom [79]. Finally, Yang et al. [80] introduced Octopus, a novel large vision-language model for embodied agents, tested in a custom experimental environment (OctoVerse).

The papers on collaboration and coordination focused on human-AI collaboration [9, 19, 21, 41, 73] and multi-agent coordination [1, 12, 19]. Gong et al. [19] proposed a new infrastructure, MindAgent, to evaluate planning and coordination capabilities in games for multi-agent systems and human-AI collaboration, and introduced a new test environment, Cuisineworld. Guan et al. [21] proposed using an LLM to develop an action plan to guide both humans and AIs by facilitating a clear understanding of tasks and responsibilities. Their experiments in the Overcooked-AI environment with a human proxy model and real humans found that their method outperformed existing learning-based approaches. Brawer et al. [9] proposed a preliminary design for a natural language interface for a task assignment system using an LLM, which they plan to evaluate in the Overcooked environment. Verma et al. [73] investigated using

LLMs to serve as effective human proxies by capturing human preferences for collaborating with AI agents. They explored the ability of LLMs to model mental states and understand human reasoning processes in Overcooked. Liu et al. [41] proposed a hierarchical language agent for human-AI coordination in Overcooked to provide both strong reasoning abilities and real-time execution. They reported that their agent outperformed baseline agents, with stronger cooperation abilities, faster responses, and more consistent language communications. Agashe et al. [1] introduced the LLM-Coordination (LLM-Co) Framework, designed to enable LLMs to play coordination games, and assessed the effectiveness of LLM-based agents in different coordination scenarios in the Overcooked-AI benchmark. They reported that LLM-based agents can infer a partner's intention and reason actions, coordinate with an unknown partner in complex long-horizon tasks, and outperform RL baselines. Chen et al. [12] proposed a multi-agent framework that can collaboratively and dynamically adjust its composition as a greater-than-the-sum-of-its-parts system in Minecraft. They found that their framework can effectively deploy multi-agent groups that outperform a single agent and identified the emergence of social behaviours.

2.2 Game Development and Play

Almost a third (25/76, 32.9%) of the papers related to the theme of Game Development and Play. The papers within this theme related to game and content generation (13/25) [6, 11, 13, 31, 44, 53–55, 57, 59, 60, 67, 71], serious games and game-based learning (8/25) [4, 8, 13, 28, 31, 56, 70, 83], game design (6/25) [4, 27, 28, 34, 56, 70], role-playing games (3/25) [20, 56, 87], and automated bug detection tools (2/25) [65, 66].

The papers on content generation related to game level generation [6, 31, 44, 54, 59, 60, 67, 71], game generation [11, 13, 55, 57], and art generation [53]. Papers on level generation investigated using LLMs to generate levels for the games Sokoban (2D puzzle game) [71], Super Marios Bros (tile-based game levels) [59, 60], Metavoidal (2D-game rooms for an under-development game) [44], Future Worlds (strategy game focused on environmental sustainability education) [31], Minecraft (3D sandbox game) [6], VR Pong [54], and Science Birds (based on Angry Birds) [67]. These papers focused on experimenting with controlling LLM level generators [71], human-in-the-loop fine-tuning [44], developing an end-to-end procedural level generation framework [31], using bootstrapping to generate enough data for training [44], and establishing a competition environment for participants to compete at creating effective prompts for ChatGPT to generate game levels [67]. Researchers found that the performance of LLMs in generating game levels scales dramatically with dataset size, which is consistent with other PCGML (Procedural Content Generation via Machine Learning) techniques, creating the same tension between the cost/benefits of data versus level creation [71]. They also found that LLMs can be text-prompted for controllable level generation, with varying performance across metrics, addressing a key challenge in current PCG techniques [59, 71]. Papers on game generation went beyond level generation to look at games holistically or larger game structures/systems. These papers introduced approaches to generate content and mechanics in a VR Pong game during development and play (Codex VR Pong) [54, 55], a multi-agent collaborative framework (GameGPT) to automate game development [11], and a simulation game system (Infinitia) that uses generative image and language models at play time to reshape the game world (terrain, vegetation, buildings, inhabitants) analogous to the fictional Holodeck [57]. Researchers found that LLMs could be used to generate non-deterministic and emergent gameplay [54, 55] and their proposed/developed systems had potential in generating games [11, 13, 55, 57], but that further work was needed to realise this potential and that human game developers still surpass AI models [55]. The paper on art generation [53] aimed to create lifelike 3D avatars from text descriptions using a combination of large language and vision models. The researchers found that their system offered an intuitive

approach for users to craft controllable, realistic, and fully-realised 3D characters within 2 minutes, showing potential for integration into a game art pipeline.

The papers on serious games and game-based learning related to using LLMs to support educators in developing games [4, 28, 70, 83], as well as more specifically to role-playing game design [56], board game design [28, 70], and level [31] and game generation [13] for serious games. A common goal of these papers was to support time-poor teachers in developing educational games for their classrooms in an expedited manner and without the need for expertise in game development [28, 70, 83]. LLMs were used to support educators with brainstorming [28], choosing a game [70], personalising the game for constructive alignment and inclusion [70], suggesting game themes and mechanisms aligned with curriculum and learning goals [28], providing templates or exemplars of game components [28], tailoring strategies to address specific learning challenges [83], supporting teachers in crafting personalised learning blueprints [83], and offering feedback on game prototypes and identifying areas for improvement [70]. One paper also explored the creation of games as a learning mechanisms in combination with playing the game, which was found to have "double impact" (first through design, then through play) [56]. Overall, researchers found that generative AI offers great potential for supporting and enhancing (rather than replacing) designers and educators in the creation of serious games and game-based learning education [13, 28, 70]. They also found that generative AI can be an effective design tool for diverse populations (e.g., educators, students, domain experts) [31].

Apart from the previously discussed papers on design of serious games and game-based learning, the game design papers focused on use of LLMs for game idea generation [34] and game mechanic design [27]. The idea generation paper [34] presented a collaborative design framework that aimed to simulate the typical human design process. As part of this framework, LLMs were used for the recombination and variation of ideas. The game mechanic paper [27] used an LLM for an element synthesis game, where players can combine elements to create new content based on physical and chemical properties of the combined elements, conceptual association, combination principles, and other logical reasoning rules. The researchers found that their LLM-based framework was effective in simulating element synthesis in video games with high freedom, high logic, repeatable playability, and more content. However, they noted that the inherent uncertainty in text generated by LLMs can have adverse effects on game mechanics.

The role-playing game (RPG) papers [20, 56, 87] involved using LLMs to design RPGs [56], to aid in playing RPGs online [87], and act as game masters in RPGs [20]. As previously discussed in relation to serious games, using LLMs to design RPGs was found to be an effective learning tool for children, particularly paired with the children also playing the games they designed [56]. For aiding people to play RPGs (in this case Dungeons and Dragons, or DnD) online, researchers [87] collected a dataset with 25,000 unique DnD sessions from real games on Discord that used the Avrae bot, including language, game commands, and underlying game state information. They found that their dataset (FIREBALL) improved natural language generation according to both automated metrics and human judgements of quality. For modelling RPG game masters (GMs), researchers [20] evaluated three LLMs (ChatGPT, Bard, OpenAssistant) as out-of-the-box GMs. Considering the skills needed by a GM (creating and managing a fictional world, tracking the game state, understanding the players' actions), they found that ChatGPT and Bard provided satisfying game experiences, although they struggled with commonsense reasoning. They found that OpenAssistant was unable to maintain the GM role during most tests.

2.3 Narrative, Story, and Dialogue

Most of the remaining papers (17/76, 22.4%) related to the theme of Narrative, Story, and Dialogue. The papers within this theme related to dialogue or conversation generation (11/17) [2, 3, 5, 16, 32, 35, 43, 45, 49, 62, 75], story generation

and interactive story (6/17) [2, 22, 32, 61, 68, 69, 82], and quest generation (2/17) [5, 72]. Two papers related to both dialogue and story generation [2, 32] and one paper to both dialogue and quest generation [5]. Two papers considered both NPC dialogue and NPC actions/tasks [49, 75].

The papers on dialogue generation mostly focused on non-player characters (NPCs) in games. However, there was one case where the dialogue was for an online streamer commentating matches in a fighting game (DareFightingICE) [45] and another where NPCs from a game were interacting with players in an online Discord community external to the game [62]. Papers on dialogue generation focused on different dimensions of dialogue and NPCs, including context-awareness and in-context dialogue [5, 16, 43], style and personality [35, 43], and story-focused dialogue [2, 32]. Most papers performed an evaluation of their systems with human players or game developers. The reported results were mixed, depending on the focus of the papers and the criteria for success. Some researchers reported on the difficulty of using LLMs to generate NPC dialogue [2] and evaluator preferences for game designer generated text compared to LLM-generated text [3]. Others reported that the LLM-generated text enhanced the player experience generally [5, 49, 62] or within specific conditions (e.g., with context-aware NPCs [16]). Some papers reported on the success of the approach via other performance metrics [32, 35, 43, 75].

The papers on story and quest generation investigated different aspects of game story endings [22, 68, 69], the use of LLMs for generating interactive story games [61, 82], and quest description generation [5, 35]. Papers focusing on story endings explored the valence (positive/neutral/negative) [68] and bias of the generated endings [69] and the impact of generating stories with/without a prompted ending (along with inspiration keywords) [22]. The interactive story generation papers studied players interacting with the commercial game AI Dungeon [82] and a custom game "1001 Nights" [61]. Research related to story endings found that generated stories were biased towards positive endings [69], that models classified stories into uninstructed endings [68], and that more detail in prompts led to higher coherence but less inspiration [22]. The interactive story researchers observed that player's mental models shifted over the play session from assuming a linear-branching narrative to one with open possibilities, impacting their motivations for repeat engagement [82]. One paper proposed the term "AI-Native games" for games where generative AI is fundamental to the game's mechanics [61], also noting that inconsistency, incoherence, and AI-transparency are key challenges in these games. One paper focused on quest generation found that LLM-generated quests approached hand-crafted quests in terms of fluency, coherence, novelty, and creativity according to human evaluation [5]. Conversely, the other quest generation paper found that only 1 in 5 generated quests were deemed acceptable by a human critic, with a large variation in quality of the generated quests [35].

2.4 Game Research and Reviews

The remaining papers (7/76, 9.2%) related to LLMs for use in analysing or generating data about games, including research data and game reviews. The papers in this theme used LLMs to analyse game reviews [33, 74] and interviews with video game players [17] and to generate synthetic data for hate speech in games [64], player responses to interview questions [24, 25], and utterances about video games [52]. Researchers found the LLMs to be effective for their research purposes in analysis and generation, including dealing with the complex and highly abbreviated lexicons related to games [64] and offering improvements in performance [74] and cost [24, 25]. However, researchers also noted some drawbacks, including variation in quality [24], lack of depth of content and flexibility [33], and potential for crowd-sourced participant data to become unreliable [25].

2.5 Recommendation

During our search process, we found an additional 9 papers [7, 10, 38, 58, 63, 81, 84–86] that used game data sets as part of broader LLM for recommendation (LLM4Rec) research. These papers did not fit the focus of our search, or of this preliminary scoping review. However, we felt it was important to capture them in this paper, given their proximity to the topic and their likely relevance to the games research community in future. Each of these papers made use of a games dataset (Amazon 7/9, Steam 3/9), alongside other datasets (e.g., movies, books). The LLM4Rec papers focused on performance of ranking-based recommendation [7, 84], sequential recommendation [81], instruction following and tuning [85], interpretation and explanation [38], collaborative semantics [86], user alignment [58], intent-aware session recommendation [63], and game features [10]. The LLM4Rec papers most frequently utilised LLaMA (4/9) and GPT (4/9), whereas the games-focused papers overwhelmingly made use of GPT.

3 CONCLUSIONS

We reviewed 76 papers on LLMs and video games, published between 2022 to 2024, in order to provide a first snapshot of the state of research on this topic and to lay the foundations for the rapidly moving ongoing and future work. We designed our review to be narrowly focused on the application of LLMs to video games. As noted earlier, we included papers that reported original research related to video games and LLMs. We excluded work that had not yet been applied to games (but that could be applied to games in future). We also did not explicitly search for areas related to games (e.g., gamification) or other types of generative models (e.g., image generation, multi-modal generation). We found that research clustered around topics related to game AI (agents, RL, collaboration), game development (content/game generation, serious games and game-based learning, game design, RPGs, bug detection), narrative (conversation/story/quest generation), and game research and reviews (analysing/generating game data). Much of the work reviewed constituted initial attempts at applying LLMs to different aspects of games. Researchers generally reported positive results or indicated promising future work related to their application of LLMs to games. Others reported success in combining LLMs with other approaches (e.g., RL) to outperform baselines. Some of the commonly reported highlights included human interpretability, social behaviours and experiences, foundational skills (e.g., can play games), and empowerment of non-developers to create games. Negatives included lack of logical reasoning ability and unpredictability (e.g., in the context of generating content in live games). Given the explosion of interest around LLMs, research will continue to move quickly on this topic. Our review aimed to provide a preliminary snapshot of the state of the work on LLMs and games, to help researchers to understand the key directions and findings so far, and to lay the foundation for future work. We expect that capturing the state of the field and keeping pace with new research will be an ongoing challenge, but it is exciting to see where LLMs will take the field of games research and development.

REFERENCES

[1] Saaket Agashe, Yue Fan, and Xin Eric Wang. 2023. Evaluating Multi-Agent Coordination Abilities in Large Language Models. *arXiv preprint arXiv:2310.03903* (2023).

[2] Nader Akoury, Ronan Salz, and Mohit Iyyer. 2023. Towards Grounded Dialogue Generation in Video Game Environments. (2023).

[3] Nader Akoury, Qian Yang, and Mohit Iyyer. 2023. A Framework for Exploring Player Perceptions of LLM-Generated Dialogue in Commercial Video Games. In *Findings of the Association for Computational Linguistics: EMNLP 2023.* 2295–2311.

[4] Ashish Amresh. 2023. Integrating Reinforcement AI Into the Design of Educational Games. In *Proceedings of the 17th European Conference on Game-Based Learning: ECGBL 2023.* Academic Conferences and publishing limited.

[5] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-Based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23).* Association for Computing Machinery, New

York, NY, USA, Article 290, 20 pages. https://doi.org/10.1145/3544548.3581441

[6] Maren Awiszus, Frederik Schubert, and Bodo Rosenhahn. 2022. Wor (l) d-GAN: Towards Natural Language Based PCG in Minecraft. *IEEE Transactions on Games* (2022).

[7] Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Fuli Feng, Xiangnaan He, and Qi Tian. 2023. A bi-step grounding paradigm for large language models in recommendation systems. *arXiv preprint arXiv:2308.08434* (2023).

[8] Jair Augusto Bottega, Victor Augusto Kich, Junior Costa de Jesus, Raul Steinmetz, Alisson Henrique Kolling, Ricardo Bedin Grando, Rodrigo da Silva Guerra, and Daniel Fernando Tello Gamarra. 2023. Jubileo: An Immersive Simulation Framework for Social Robot Design. *Journal of Intelligent & Robotic Systems* 109, 4 (2023), 91.

[9] Jake Brawer, Kayleigh Bishop, Bradley Hayes, and Alessandro Roncone. 2023. Towards A Natural Language Interface for Flexible Multi-Agent Task Assignment. *arXiv preprint arXiv:2311.00153* (2023).

[10] M Charity, Yash Bhartia, Daniel Zhang, Ahmed Khalifa, and Julian Togelius. 2023. A Preliminary Study on a Conceptual Game Feature Generation and Recommendation System. *arXiv preprint arXiv:2308.13538* (2023).

[11] Dake Chen, Hanbin Wang, Yunhao Huo, Yuzhao Li, and Haoyang Zhang. 2023. Gamegpt: Multi-agent collaborative framework for game development. *arXiv preprint arXiv:2310.08067* (2023).

[12] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848* (2023).

[13] Iván J Pérez Colado, Víctor M Pérez Colado, Antonio Calvo Morata, Rubén Santa Cruz Píriz, and Baltasar Fernández Manjón. 2023. Using new AI-driven techniques to ease serious games authoring. In *2023 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–9.

[14] Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. 2023. Augmenting Autotelic Agents with Large Language Models. *arXiv preprint arXiv:2305.12487* (2023).

[15] Maximilian Croissant, Madeleine Frister, Guy Schofield, and Cade McCall. 2023. An Appraisal-Based Chain-Of-Emotion Architecture for Affective Language Model Game Agents. *arXiv preprint arXiv:2309.05076* (2023).

[16] Lajos Matyas Csepregi. 2021. The Effect of Context-aware LLM-based NPC Conversations on Player Engagement in Role-playing Video Games. *Unpublished manuscript* (2021).

[17] Stefano De Paoli. 2023. Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach. *Social Science Computer Review* (2023), 08944393231220483.

[18] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023. Guiding pretraining in reinforcement learning with large language models. *arXiv preprint arXiv:2302.06692* (2023).

[19] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, et al. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971* (2023).

[20] Santiago Góngora, Luis Chiruzzo, Gonzalo Méndez, and Pablo Gervás. 2023. Skill Check: Some Considerations on the Evaluation of Gamemastering Models for Role-Playing Games. In *International Conference on Games and Learning Alliance*. Springer, 277–288.

[21] Cong Guan, Lichao Zhang, Chunpeng Fan, Yichen Li, Feng Chen, Lihe Li, Yunjia Tian, Lei Yuan, and Yang Yu. 2023. Efficient Human-AI Coordination via Preparatory Language-based Convention. *arXiv preprint arXiv:2311.00416* (2023).

[22] Mustafa Can Gursesli, Pittawat Taveekitworachai, Febri Abdullah, Mury F Dewantoro, Antonio Lanata, Andrea Guazzini, Van Khôi Lê, Adrien Villars, and Ruck Thawonmas. 2023. The Chronicles of ChatGPT: Generating and Evaluating Visual Novel Narratives on Climate Change Through ChatGPT. In *International Conference on Interactive Digital Storytelling*. Springer, 181–194.

[23] Muhammad Usman Hadi, qasem al tashi, Rizwan Qureshi, Abbas Shah, amgad muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, and Seyedali Mirjalili. 2023. Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. (Nov. 2023). https://doi.org/10.36227/techrxiv.23589741.v4

[24] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2022. Neural Language Models as What If?-Engines for HCI Research. In *27th International Conference on Intelligent User Interfaces*. 77–80.

[25] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[26] Mark Hendrikx, Sebastiaan Meijer, Joeri Van Der Velden, and Alexandru Iosup. 2013. Procedural content generation for games: A survey. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 1, Article 1 (feb 2013), 22 pages. https://doi.org/10.1145/2422956.2422957

[27] Lei Huang and Xing Sun. 2023. Create Ice Cream: Real-time Creative Element Synthesis Framework Based on GPT3. 0. In *2023 IEEE Conference on Games (CoG)*. IEEE, 1–4.

[28] Wilian Gatti Junior, Emily Marasco, Beaumie Kim, Laleh Behjat, and Marjan Eggermont. 2023. How ChatGPT can inspire and improve serious board game design. *International Journal of Serious Games* 10, 4 (2023), 33–54.

[29] Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe Agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172* (2023).

[30] Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. 2023. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166* (2023).

[31] Vikram Kumaran, Dan Carpenter, Jonathan Rowe, Bradley Mott, and James Lester. 2023. End-to-End Procedural Level Generation in Educational Games with Natural Language Instruction. In *2023 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[32] Vikram Kumaran, Jonathan Rowe, Bradford Mott, and James Lester. 2023. SCENECRAFT: automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 19. 86–96.

[33] Michael Lankes and Andreas Stöckl. 2023. Game Reviews Reviewed: A Game Designer's Perspective on AI-generated Game Review Analyses. In *2023 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[34] Pier Luca Lanzi and Daniele Loiacono. 2023. Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design. *arXiv preprint arXiv:2303.02155* (2023).

[35] Gaetan Lopez Latouche, Laurence Marcotte, and Ben Swanson. 2023. Generating video game scripts with style. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*. 129–139.

[36] Lane Lawley and Christopher J MacLellan. 2023. VAL: Interactive Task Learning with GPT Dialog Parsing. *arXiv preprint arXiv:2310.01627* (2023).

[37] Juyong Lee and Doohyun Lee. [n. d.]. Can Language Models Postprocess Rewards for Reinforcement Learning? ([n. d.]).

[38] Yuxuan Lei, Jianxun Lian, Jing Yao, Xu Huang, Defu Lian, and Xing Xie. 2023. RecExplainer: Aligning Large Language Models for Recommendation Model Interpretability. *arXiv preprint arXiv:2311.10947* (2023).

[39] Yinghao Li, Haorui Wang, and Chao Zhang. 2023. Assessing Logical Puzzle Solving in Large Language Models: Insights from a Minesweeper Case Study. *arXiv preprint arXiv:2311.07387* (2023).

[40] Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773* (2023).

[41] Jijia Liu, Chao Yu, Jiaxuan Gao, Yuqing Xie, Qingmin Liao, Yi Wu, and Yu Wang. 2023. LLM-Powered Hierarchical Language Agent for Real-time Human-AI Coordination. *arXiv preprint arXiv:2312.15224* (2023).

[42] 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. *arXiv preprint arXiv:2312.11865* (2023).

[43] Matthias Müller-Brockhausen, Giulio Barbero, and Mike Preuss. 2023. Chatter Generation through Language Models. In *2023 IEEE Conference on Games (CoG)*. IEEE, 1–6.

[44] Muhammad U Nasir and Julian Togelius. 2023. Practical PCG Through Large Language Models. *arXiv preprint arXiv:2305.18243* (2023).

[45] Chollakorn Nimpattanavong, Pittawat Taveekitworachai, Ibrahim Khan, Thai Van Nguyen, Ruck Thawonmas, Worawat Choensawat, and Kingkarn Sookhanaphibarn. 2023. Am I Fighting Well? Fighting Game Commentary Generation With ChatGPT. In *Proceedings of the 13th International Conference on Advances in Information Technology*. 1–7.

[46] Mark J. Nelson Noor Shaker, Julian Togelius. 2016. *Procedural Content Generation in Games*. Springer Cham. https://doi.org/10.1007/978-3-319-42716-4

[47] Kolby Nottingham, Yasaman Razeghi, Kyungmin Kim, JB Lanier, Pierre Baldi, Roy Fox, and Sameer Singh. 2023. Selective Perception: Learning Concise State Descriptions for Language Model Actors. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

[48] Kolby Nottingham, Yasaman Razeghi, Kyungmin Kim, JB Lanier, Pierre Baldi, Roy Fox, and Sameer Singh. 2023. Selective perception: Optimizing state descriptions with reinforcement learning for language model actors. *arXiv preprint arXiv:2307.11922* (2023).

[49] 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*. INSTICC, SciTePress, 27–36. https://doi.org/10.5220/0012060500003538

[50] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[51] Ulyana Piterbarg, Lerrel Pinto, and Rob Fergus. 2023. diff History for Long-Context Language Agents. *arXiv preprint arXiv:2312.07540* (2023).

[52] Angela Ramirez, Mamon Alsalihy, Kartik Aggarwal, Cecilia Li, Liren Wu, and Marilyn Walker. 2023. Controlling Personality Style in Dialogue with Zero-Shot Prompt-Based Learning. *arXiv preprint arXiv:2302.03848* (2023).

[53] Jianqiang Ren, Chao He, Lin Liu, Jiahao Chen, Yutong Wang, Yafei Song, Jianfang Li, Tangli Xue, Siqi Hu, Tao Chen, et al. 2023. Make-A-Character: High Quality Text-to-3D Character Generation within Minutes. *arXiv preprint arXiv:2312.15430* (2023).

[54] Jasmine Roberts, Andrzej Banburski-Fahey, and Jaron Lanier. 2022. Steps towards prompt-based creation of virtual worlds. *arXiv preprint arXiv:2211.05875* (2022).

[55] Jasmine Roberts, Andrzej Banburski-Fahey, and Jaron Lanier. 2022. Surreal vr pong: Llm approach to game design. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022). https://www. microsoft. com/en-us/research/publication/surreal-vr-pong-llm-approach-to-gamedesign*.

[56] Kenji Saito, Kayo Kobayashi, Waki Takekoshi, Atsuki Hashimoto, Nobukazu Hirai, Akifumi Kimura, Asuka Takahashi, Naoki Yoshioka, and Asuto Mano. 2023. Double Impact: Children's Serious RPG Generation/Play with a Large Language Model for Their Deeper Engagement in Social Issues. In *Joint International Conference on Serious Games*. Springer, 274–289.

[57] Ahad Shams, Douglas Summers-Stay, Vsevolod Metelsky, Arpan Tripathi, and Karan Malhotra. 2023. Towards a Holodeck-style Simulation Game. *arXiv preprint arXiv:2308.13548* (2023).

[58] Yubo Shu, Hansu Gu, Peng Zhang, Haonan Zhang, Tun Lu, Dongsheng Li, and Ning Gu. 2023. RAH! RecSys-Assistant-Human: A Human-Central Recommendation Framework with Large Language Models. *arXiv preprint arXiv:2308.09904* (2023).

[59] Shyam Sudhakaran, Miguel González-Duque, Claire Glanois, Matthias Freiberger, Elias Najarro, and Sebastian Risi. 2023. Prompt-Guided Level Generation. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*. 179–182.

[60] Shyam Sudhakaran, Miguel González-Duque, Claire Glanois, Matthias Freiberger, Elias Najarro, and Sebastian Risi. 2023. MarioGPT: Open-Ended Text2Level Generation through Large Language Models. arXiv:2302.05981 [cs.AI]

[61] Yuqian Sun, Zhouyi Li, Ke Fang, Chang Hee Lee, and Ali Asadipour. 2023. Language as reality: a co-creative storytelling game experience in 1001 nights using generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 19. 425–434.

[62] Yuqian Sun, Hanyi Wang, Pok Man Chan, Morteza Tabibi, Yan Zhang, Huan Lu, Yuheng Chen, Chang Hee Lee, and Ali Asadipour. 2023. Fictional Worlds, Real Connections: Developing Community Storytelling Social Chatbots through LLMs. *arXiv preprint arXiv:2309.11478* (2023).

[63] Zhu Sun, Hongyang Liu, Xinghua Qu, Kaidong Feng, Yan Wang, and Yew-Soon Ong. 2023. Large Language Models for Intent-Driven Session Recommendations. *arXiv preprint arXiv:2312.07552* (2023).

[64] Arina Svetasheva and Keeheon Lee. 2024. Harnessing Large Language Models for Effective and Efficient Hate Speech Detection. (2024).

[65] Mohammad Reza Taesiri, Tianjun Feng, Cor-Paul Bezemer, and Anh Nguyen. 2023. GlitchBench: Can large multimodal models detect video game glitches? *arXiv preprint arXiv:2312.05291* (2023).

[66] Mohammad Reza Taesiri, Finlay Macklon, Yihe Wang, Hengshuo Shen, and Cor-Paul Bezemer. 2022. Large Language Models are Pretty Good Zero-Shot Video Game Bug Detectors. *arXiv preprint arXiv:2210.02506* (2022).

[67] Pittawat Taveekitworachai, Febri Abdullah, Mury F. Dewantoro, Ruck Thawonmas, Julian Togelius, and Jochen Renz. 2023. ChatGPT4PCG Competition: Character-like Level Generation for Science Birds. arXiv:2303.15662 [cs.AI]

[68] Pittawat Taveekitworachai, Febri Abdullah, Mustafa Can Gursesli, Mury F Dewantoro, Siyuan Chen, Antonio Lanata, 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 *International Conference on Interactive Digital Storytelling*. Springer, 274–284.

[69] Pittawat Taveekitworachai, Mustafa Can Gursesli, Febri Abdullah, Siyuan Chen, Federico Cala, Andrea Guazzini, Antonio Lanata, and Ruck Thawonmas. 2023. Journey of ChatGPT from Prompts to Stories in Games: the Positive, the Negative, and the Neutral. In *2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 202–203.

[70] Andrea Tinterri, Marilena di Padova, Francesco Palladino, Giordano Vignoli, Anna Dipace, et al. 2024. AI in board Game-Based Learning. In *CEUR WORKSHOP PROCEEDINGS*.

[71] Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level Generation Through Large Language Models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*. 1–8.

[72] Susanna Värtinen, Perttu Hämäläinen, and Christian Guckelsberger. 2022. Generating role-playing game quests with gpt language models. *IEEE Transactions on Games* (2022).

[73] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2023. Preference Proxies: Evaluating Large Language Models in capturing Human Preferences in Human-AI Tasks. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*.

[74] Markos Viggiato and Cor-Paul Bezemer. 2023. Leveraging the OPT Large Language Model for Sentiment Analysis of Game Reviews. *IEEE Transactions on Games* (2023).

[75] Ryan Volum, Sudha Rao, Michael Xu, Gabriel DesGarennes, Chris Brockett, Benjamin Van Durme, Olivia Deng, Akanksha Malhotra, and William B Dolan. 2022. Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code. In *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)*. 25–43.

[76] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291 [cs.AI]

[77] Yue Wu, So Yeon Min, Shrimai Prabhumoye, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Tom Mitchell, and Yuanzhi Li. 2023. SPRING: GPT-4 Out-performs RL Algorithms by Studying Papers and Reasoning. *arXiv preprint arXiv:2305.15486* (2023).

[78] Zelai Xu, Chao Yu, Fei Fang, Yu Wang, and Yi Wu. 2023. Language agents with reinforcement learning for strategic play in the werewolf game. *arXiv preprint arXiv:2310.18940* (2023).

[79] Xue Yan, Yan Song, Xinyu Cui, Filippos Christianos, Haifeng Zhang, David Henry Mguni, and Jun Wang. 2023. Ask more, know better: Reinforce-Learned Prompt Questions for Decision Making with Large Language Models. *arXiv preprint arXiv:2310.18127* (2023).

[80] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. 2023. Octopus: Embodied vision-language programmer from environmental feedback. *arXiv preprint arXiv:2310.08588* (2023).

[81] Zhengyi Yang, Jiancan Wu, Yanchen Luo, Jizhi Zhang, Yancheng Yuan, An Zhang, Xiang Wang, and Xiangnan He. 2023. Large language model can interpret latent space of sequential recommender. *arXiv preprint arXiv:2310.20487* (2023).

[82] Qing Ru Yong and Alex Mitchell. 2023. From Playing the Story to Gaming the System: Repeat Experiences of a Large Language Model-Based Interactive Story. In *International Conference on Interactive Digital Storytelling*. Springer, 395–409.

[83] Lei Yue and Liang Guo. 2023. Combine DGBL With AI System: A Technical Guidance to Reduce Teacher's Burden in Digital Game-Based Learning. In *European Conference on Games Based Learning*. Academic Conferences International Limited, 826–XXV.

[84] Zhenrui Yue, Sara Rabhi, Gabriel de Souza Pereira Moreira, Dong Wang, and Even Oldridge. 2023. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. *arXiv preprint arXiv:2311.02089* (2023).

[85] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001* (2023).

[86] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Adapting large language models by integrating collaborative semantics for recommendation. *arXiv preprint arXiv:2311.09049* (2023).

[87] Andrew Zhu, Karmanya Aggarwal, Alexander Feng, Lara J Martin, and Chris Callison-Burch. 2023. Fireball: A dataset of dungeons and dragons actual-play with structured game state information. *arXiv preprint arXiv:2305.01528* (2023).

[88] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory. arXiv:2305.17144 [cs.AI]