Towards LLM-Agents That Play Dungeons & Dragons Using Iterative Prompting

Vishal Joshi*, Nirav Ajmeri and Kenton O'Hara

School of Computer Science, University of Bristol, BS8 1UB, Bristol, United Kingdom

Abstract

Playing Dungeons & Dragons (D&D) requires players to attend every session of a campaign, which is often hindered by scheduling conflicts. The disruptions caused by these conflicts could lead to stalled or cancelled games. To address this challenge, we propose using AI player substitutes. Specifically, we develop LLM-agents capable of social interaction and exploration within textual D&D scenarios, using Concordia—a library to simulate multi-agent dialogue and interactions. We iteratively prompt our LLM-agents to foster collaboration, task progression, and comply with the narrative context in two distinct campaign settings. We annotate natural language transcripts generated in Concordia, categorising player actions into the three action and dialogue categories. Our preliminary findings indicate that iterative prompting enhances agents' narrative compliance, collaborative behaviour and progression towards campaign goals. These promising results suggest LLM-agents could be viable stand-ins for human players, and warrant further investigation.

Keywords

Role-playing Agents, Generative-Agent Based Modelling, Large Language Modelling, D&D Gameplay Experience

1. Introduction

Fantasy table-top role-playing games (TTRPGs) such as Dungeons & Dragons (D&D) [1] have maintained an enduring popularity over the years. In TTRPGs, multiple players engage in a *campaign*, a series of simulated challenges that involve social interaction, collaborative exploration and combat [1]. With campaigns played out over multiple sessions, and potentially long time periods, scheduling play sessions can be logistically challenging. With a required committent for all players to play throughout a campaign, scheduling conflicts between players can be disruptive to continuation. With this in mind, there is potential to develop AI players as stand-ins for human players, mitigating any coordination burden and offering greater flexibility for scheduling gameplay. The evolving capabilities of large language models (LLMs) present a significant opportunity to realise such potential.

Research into developing AI players is scarce in TTRPGs but emerging research is showing interest in using natural language processing (NLP) and LLMs in the domain of TTRPGs. Much of this has foregrounded TTRPG environments as a testbed for evaluating the dialogue [2], reasoning [3] and decision making [4] capabilities of language-based systems. Less emphasis is given to the role of these systems in the context of actual gameplay support. Human-Computer Interaction (HCI) research has explored the use of LLMs to enhance gameplay. Thus far, however, this has focussed on the development of Game Master (GM) support and automation rather than players. For example, Zhu et al. [5] developed an LLM assistant to aid the GM during gameplay, while Triyason [6] went further, to replace the GM altogether using a commercial LLM.

Given the focus of previous work, there remain important opportunities for research into the development of LLM-based player agents. In spite of the growing capabilities of LLMs, there are significant challenges in their adaptation to the unique context of TTRPG scenarios. Studies of existing systems highlight various characteristics of LLM generated dialogue that may not be conducive to collaborative gameplay environment such as excessive agreement or player flattery (aka sycophancy), which can

Proceedings of AI4HGI '25, the First Workshop on Artificial Intelligence for Human-Game Interaction at the 28th European Conference on Artificial Intelligence (ECAI '25), Bologna, October 25-30, 2025

we19383@bristol.ac.uk (V. Joshi); nirav.ajmeri@bristol.ac.uk (N. Ajmeri); kenton.ohara@bristol.ac.uk (K. O'Hara)

0009-0001-9716-409X (V. Joshi); 0000-0003-3627-097X (N. Ajmeri); 0000-0001-8915-4572 (K. O'Hara)

lead to unwanted behaviours that may negatively impact on desired features of engaged gameplay [7]. Propagations of LLM hallucinations via sycophantic behaviour can be disruptive to the campaign narrative and progression towards task completion in TTRPGs [8]. To ensure that LLM role-playing agents enhance, rather than detract from, the gameplay experience requires that they consistently adhere to game rules and their character sheet whilst maintaining a level of creativity as simulated players. Deviation from rules or an agent's character sheet could be disruptive to immersion and require GM intervention to pause the narrative and correct the agent's mistakes. Frequent instances of this could disrupt immersion and lead to other players becoming confused and frustrated.

Contribution To further understand these challenges and identify appropriate interventions, we present exploratory efforts to develop LLM player agents through iterative prompting. We draw from the growing field of generative-agent based modelling to develop role-playing agents for D&D using Concordia [9], a library to simulate social interactions using LLM agents. To evaluate the agents' ability to perform a given task, in a collaborative manner and compliant with the simulated narrative, we conduct simulation experiments with a baseline of zero-shot prompting followed by iterative editing and restructuring of the prompts for three LLM-agents across two different simulated campaign scenarios. Our findings suggest that iterative prompting improves agents' ability to progress towards assigned task completion in a narratively compliant manner and adopting more collaborative actions.

Organisation The paper presents relevant background and related literature (Sec 2), followed by our methodology (Sec 3) including the agent schematic and simulation set-up. This is followed by the details of our experiments (Sec 4) including the campaign scenarios we simulate and our schema of iterative prompting. Then we present our preliminary results (Sec 5), including an analytical discussion of these results. Finally, we draw conclusions from our initial research explorations, highlighting limitations and directions (Sec 6).

2. Background and Related Works

To contextualise the work, we provide background about D&D, an overview of Concordia [9] and related literature.

2.1. Dungeons & Dragons

Dungeons & Dragons (D&D) is a fantasy TTRPG in which each player acts as a created fictional character which go on fictional adventures. The campaign is created and mediated by a Dungeon Master (DM) or GM. Campaigns consist of three broad tasks [1]: social interaction, exploration and combat. Social interactions are dialogue exchanges between players to uncover information, plan group strategy before exploration or combat, and discuss campaign progression. Players also interact with non-player characters (NPCs), played by the DM, to gather clues about in-game tasks and opportunities to gain extra rewards. Social interactions can be open-ended or goal-oriented, occurring between players as well as with NPCs or the DM (out-of-character) to progress through the game. Exploration involves using character skills, and querying the DM to discover details about an unknown area of the game map. This is done to gather information related to further progression, discover hidden rewards or scout for enemies. Exploration involves direct observation of the environment combined with querying the DM and other players to gain information. Combat uses character skills and abilities to fight adversaries—also played by the DM, which hinder progression through the campaign. Players describe their desired actions and roll dice, the result of which determines whether the combat action taken is successful or not. Combat is quasi-simultaneous and turn based. Players engage in it against single or multiple adversaries to survive or to earn rewards.

Our current work focuses on developing agent capabilities for narratively compliant *social interactions* and collaborative *exploration* towards completion of 2 exemplar D&D narrative scenarios.

2.2. Concordia

Concordia offers a library [9], for modelling agent behaviour with three criteria [9]: (1) Coherent with respect to common sense. (2) Guided by social norms. (3) Contextualised individually according to an agent's past and continuous perception of the current situation. These criteria are formalised as the following questions from March and Olsen [10] for an agent to ask itself: (1) What kind of situation is this? (2) What kind of person am I? (3) What does a person like me do in a situation like this?

Vezhnevets et al. [9] hypothesise: LLM-based agents can sufficiently answer the above questions when provided with historical context for an agent, because an LLM training corpus includes vast amounts of data on human culture. However, the history of the agent must not overwhelm the context window of the LLM. Concordia provides an *associative memory* [11] and factorises the context-generation process using *components*. Agents can be built through a combination of components which allow customisation of how they reason and act. For more details see Sec A.1.

2.3. Related Works

Our contributions are closely related to work on modelling D&D gameplay [12, 13] and LLM-based approaches to aiding human players [5, 4]. Work on dialogue interactions within text-based games foregrounds these games as test domains for language and agent-based modelling [2]. Ammanabrolu et al. [14] and Prabhumoye et al. [15] develop agents which generate open-ended and goal-driven dialogue within fantasy settings. Goal-driven dialogue involves an agent conversing with another to achieve an objective. Prabhumoye et al. [15] offer the task of making a Knight smile in a text-based fantasy setting as an example of goal-driven dialogue. Open-ended dialogue is "chit-chat" or banter which does not achieve an explicit goal but still makes up human conversations. Open-ended dialogue allows players to familiarise themselves with each other's characters and improve group cohesion [16]. Ammanabrolu et al. [3] and Prabhumoye et al. [15]'s approaches are limited by the restrictive syntax of the dialogue within the ParlAI dataset they use. ParlAI is a crowd-sourced dataset of dialogue interactions within a fantasy setting. Whilst diverse in terms of topic, the dataset uses limited syntax which does not reflect the casual and verbose dialogue spoken by human players. Si et al. [17] use supervised fine-tuning and the *Critical Role Dungeons & Dragons Dataset* [18] on the task of story continuation. Whilst this method provides a simulacrum of the multi-turn dialogue in D&D play, it restricts the text output to the distribution of the voice actors' dialogues within the dataset, trading diversity of text output for verbosity and casualness.

3. Methodology

We now describe the schematics of our D&D environment including the different types of agents and their interactions, followed by a description of our simulation set up in Concordia.

3.1. Schematic

The schematic for our D&D environment and its constituent agents is presented here:

Definition 1 (Environment). *The D&D environment DD is a tuple* $DD = \langle O, E, A, Di, C \rangle$ *consisting of the base elements of the game environment, where O represents a set of observations, E represents a set of GM's event statements, A represents a set of actions an agent can take, Di represents the set of agents' dialogues with each other and C is a set of character sheets of the player agents.*

Definition 2 (Observation). $o \in O$ *is a tuple* $o = \langle n, a_{player}, a_{npc}, a_{enemy} \rangle$ *consisting of elements that an agent perceives in the game environment, where n represents GM narration,* a_{player} *is an action taken by a player,* a_{npc} *is an action taken by a non-player character (NPC), and* a_{enemy} *is an action taken by an enemy.*

Definition 3 (Event Statement). $e \in E$ *is a statement generated by the GM that aggregates player actions and describes their effects on the campaign narrative, which players interpret as observations.*

Definition 4 (Action). *$a \in A$ is an action that an agent can take based on observations O. It includes a natural language description of a desired action, with the GM performing similar actions for any NPCs or adversaries they control. The types of actions include:*

- ***Social interactions**: Dialogue between players and NPCs.*
- ***Exploration**: Players act directly on features in the game environment as described by the GM.*
- ***Combat actions**: Mediated by numerical values used to calculate the probability of success or failure with the roll of multi-sided dice.*

Definition 5 (Dialogue). *$di \in Di$ is the communication exchanged between a player and other players, the GM, and NPCs, which influences the flow of the game. It is separate to an agent's actions which are limited per agent's turn, whereas dialogue is unlimited per turn.*

Definition 6 (Character sheet). *$c \in C$ is a comprehensive representation of the character of the agent, defined as $c = \langle i, st, p, h \rangle$, where i includes identifiers, st includes statistics, p includes profile details, and h includes historical context.*

Definition 7 (Identifier). *$i \in I$ is an identifying feature that provides a high-level overview of the agent's character, including name, class, level, gender, race, one-word background, moral alignment, and mechanism for earning experience points.*

Definition 8 (Statistic). *$s \in St$ is a numerical value that mediates combat and effects on the environment and agents, including equipment, skills, armour class, saving throws, initiative, speed, hit points, and attacking and spellcasting capabilities.*

Definition 9 (Profile). *$p \in P$ is a detailed description of the narrative qualities of the character, encompassing personality traits, ideals, bonds, flaws, physical appearance, and personal goals.*

Definition 10 (History). *$h \in H$ is contextual information that supports the profile p, including date of birth, backstory, allegiance, and significant memories or formative experiences.*

Definition 11 (D&D player agent). *$Pl = \langle O, A, C, Di \rangle$ is a player in D&D, where O is the set of observations, A is the set of actions, C is the character sheet, and D is the dialogue, representing the player's interactions and characteristics within the game.*

Definition 12 (Game master). *$GM = \langle O, E, A \rangle$ is the game master responsible for narrating the game and controlling NPCs.*

Definition 13 (Non-Player Character). *$NPC = \langle O, E, A \rangle$ is an entity controlled by the GM that interacts with players in the game.*

3.2. Simulation Workflow

Here we detail the interconnected components of our simulation workflow within Concordia [9], which includes the memories initialisation, player agent construction, and game master setup.

3.2.1. Memories

First, we set up the `simulation clock` to record *O*, *A*, *E* and *Di* at every time step. We write the narrative context for our simulation scenario in the form of "Shared Memories" which guide the behaviour of *Pl*, *GM* and *NPC* combined with their role-play instructions. (See examples of Shared Memories in Sections 4). We then assign relevance to instances of *O* using an `importance_model`, to make memory querying more efficient. Following this, the `shared_memories` for *Pl* and *GM* are written, to establish narrative setting, background and to introduce *NPCs*. We then instantiate agents with formative memories to generate *H* for our *Pl* and blank memories which are returned whenever an agent does not observe a particular event.

3.2.2. Player Agents

We build *Pl* using the following components.

Character Sheet A full natural language description of agent based on a D&D character sheet, *C*.

Current and Summary Observations Gather, summarise and reason about observations (*O*) to help create action (*A*) context.

Persona A component containing: (1) `SituationPerception`: prompts the agent with "what kind of situation is this?"; (2) `SelfPerception`: prompts the agent with "what kind of person am I?"; and (3) `PersonBySituation`: prompts agent with "what does a person like me do in a situation like this?"

Goal Outlined in the agent's *C* as a natural language prompt.

Player Instructions

```
The instructions for how to play the role of {agent_name} are as follows. This is a one-shot
campaign within Dungeons and Dragons 5th Edition, in which {agent_name} is a character. The
goal is to be consistent, but creative. It is important to play the role of a person like
{agent_name} as accurately as possible, i.e., by responding in ways that you think it is
likely a person like {agent_name} would respond, and taking into account all information
about {agent_name} that you have. It is important that you collaborate with the other players
on the task at hand and follow the Game Master's instructions. Always use first-person
limited perspective.
```

Figure 1: A prompt given to the *Pl* agent to conditions its behaviour in the context of a D&D campaign.

3.2.3. Game Master

The *GM* object includes the following components.

Game Master Memory A form of associative memory, $\langle O, ES \rangle$.

Scenario Knowledge contains the facts that *GM* carries in its working memory and combined with the instructions influences its mediation of the campaign narrative - an example is given in Sec 4.2.2.

Player Status The observation of the *GM* object about the *Pl* status and location.

Conversation Externality Allows the *GM* to handle conversations between *Pl*, as well as with *NPCs*.

Direct Effect Externality *GM* observation which tracks direct effect of events and actions on players.

Relevant Events Retrieves relevant events (*ES*) from the *GM* memory to condition text outputs which form agent observations. Once we build the *GM* object, we start the simulation clock and run the simulation.

Game Master Instructions

```
This is a tabletop role-playing game: Dungeons & Dragons. You are the Dungeon Master. You
will describe the current situation to the players in the game and then on the basis of
what you tell them they will suggest actions for the character they control. You will then
decide if the action is valid based on Dungeons & Dragons 5th Edition rules. Aside from you,
each other player controls just one character. If any of the players deviates dramatically
from the shared memories of the group your event statement should attempt to re-orient the
campaign. You are the Dungeon master so you may control any non-player character. You will
track the state of the world and keep it consistent as time passes in the simulation and
the players take actions and change things in their world. Remember that this is a game.
It should be fun for the players. You should use second-person perspective, when speaking
directly to the players. You should use first-person limited perspective when role-playing as
non-player characters and adversaries.'
```

Figure 2: A prompt given to the *GM* agent to conditions its direction of the one-shot campaign.

3.2.4. Simulation

The simulation involves interactions between player agents and the GM. When the GM acts as an NPC, it engages in dialogue or actions that players observe. Players then attempt an action or engage in a dialogue that NPCs process. When the GM does not act as an NPC, it generates an event statement based on the simulation's premise, which players observe. Based on this observation, players take action or engage in dialogue. The GM aggregates the actions and dialogues of each player into a new event statement and can alternate between acting as an NPC and mediating the campaign.

4. Experimental Setup

We conduct 4 experiments, comprising 10 simulated runs of the initial portions of 2 D&D scenarios. We run our experiments using T4 High RAM GPU on Google Colab Jupyter Notebook, which uses approximately 1.3 to 1.6 compute units per hour. We cap each simulation at 30 minutes for standardisation purposes. We use Llama 3.1 [19] for our explorations. It is open source and has comparable performance to commercial LLMs on a variety of downstream tasks [20].

4.1. Scenarios, Subtasks, and Metrics

Scenarios We consider two scenarios: (1) *Rat Infestation* and (2) *Missing Otter Child*. In the Rat Infestation (RI) scenario, *Pl* are tasked by an *NPC* brewer with clearing out a rat infestation within a local brewery, which may have magical origins. In the Missing Otter Child (MOC) scenario, *Pl* are saved from an avalanche and tasked by a magical walrus to find its missing otter child, which was last seen spying on a shadowy organisation in a nearby city.

Subtasks Each scenario has subtasks which progress the players towards task completion:

Information Gathering *Pl* asks for further context regarding the task at hand to gain clues which help with planning. For instance, in RI scenario, asking the brewery owner about past experiences with invading rats or in the MOC scenario asking the magic walrus where its child was last seen.

Planning *Pl* spends time planning how to solve the task and gather more clues. For instance, in the RI scenario, looking at maps of the brewery's basement for weak points where rats may enter or, in the MOC scenario, searching the city's surrounding countryside where the otter child was last spotted for tracks to set up a trail.

Investigation *Pl* investigates the target environment. For instance, in the RI scenario, the *Pl* going into the brewery's basement to find the weak points or, in the MOC scenario, moving towards the city where the otter child was last spotted.

Metrics In each experiment, we label the simulation transcripts to evaluate the player agent's ability to:

Progress towards task completion Assess whether actions, as recorded in the transcripts, positively or negatively impact task progression, considering both independent and collaborative efforts.

Collaborate towards the goal Classify interactions from the transcripts as collaborative or independent. Collaborative actions involve multiple characters working together, while independent actions occur when characters pursue divergent subtasks.

Act within the narrative context Determine if agent responses, reflected in the transcripts, are contextually appropriate. Narrative disruption is identified when actions deviate from the established narrative, with only executed actions counted as valid according to D&D rules.

We evaluate our hypotheses at the 5% significance level for each agent, as they can engage in different combinations of the three action types. Our primary aggregation metric is the interquartile mean (IQM) of each action type, which is less affected by outliers than the sample mean due to the stochastic nature of LLM outputs. IQM also provides tighter confidence bounds than the median, allowing trends to

emerge more quickly in our limited simulation runs. We perform two-tailed trimmed mean t-tests (25% trim) and compute effect sizes (Cohen's D).

4.2. Experiments

We now outline our experiments, associated hypotheses and the specific prompts. Each experiment uses three LLM-agents with their own character sheets *C*. Appendix A.3 details the prompt edits.

4.2.1. Experiment 1: Single Prompt Baseline

In this experiment, we investigate the effectiveness of a single zero-shot prompt on players *Pl* in a D&D Rat Infestation Scenario. Our hypothesis is that a single zero-shot prompt provides the agents with sufficient information to complete the task. We establish a baseline using the prompt:

Shared Memories Baseline - Single Zero-Shot Prompt (Experiment 1 Runs 1-5)

```
This is a Dungeons & Dragons 5th Edition one-shot campaign. It is set in the city of REDACTED,
just one of the many cities in the REDACTED_SETTING. This one-shot specifically concerns
the REDACTED_BREWERY - a craft brewery known for its hoppy summer ales - is in dire need
of help from a band of reliable, affordable adventurers to help sort out a rat infestation
in the brewery's basement. At the beginning of this adventure our party members meet in the
REDACTED_BREWERY.
```

We run five repeats of this baseline, with *Pl* having individual, divergent goals. Next we carry out five more runs of the simulation, replacing individual goals with a common goal on the first of these runs. We hypothesise that setting a common goal improves collaboration between the agents compared to each agent having an individual goal.

```
Goal in Rat Infestation Scenario: You should collaborate with your fellow adventurers to
complete the task given to you by Brewery_Owner.
```

Throughout the five runs, we iteratively edit the agents' Shared Memories to enhance their narrative compliance and improve task completion. The final prompt version is below; in-prompt capitalisation is to emphasise important parts to the LLM agents.

Shared Memories Experiment 1 Runs 6-10

```
This is a DUNGEONS & DRAGONS 5th EDITION one-shot campaign. This one-shot specifically
concerns the REDACTED_BREWERY - a craft brewery known for its hoppy summer ales - is in dire
need of help from a band of reliable, affordable adventurers (the PLAYERS) to help sort out
a RAT INFESTATION in the brewery's BASEMENT. At the beginning of this adventure the PLAYERS
meet in the REDACTED_BREWERY. The PLAYERS DO NOT know each other AT FIRST and need to get to
know each other. Brewery_Owner hands out pints of Ale as the players get to know each other.
He gives players a run-down of the task with some background:
• The business has been doing well and looks to expand its operations. But first the beer
  cellar needed to be expanded.
• Workmen that he sent down into the cellar to expand it were attacked by large black rats
  which came out of the wall they were digging out.
• The workmen escaped unharmed but the cellars are unusable which is bad for business.
• The rats may have something to do with the REDACTED_NAME on the site which the brewery takes
  its name from.
Brewery_Owner then lays out the terms of the task:
• The party must dispose of the rats.
• They must discover the origin of the rats and make sure they permanently stop the
  infestation.
• They will each be rewarded 25 gold coins, but this is up for negotiation.
```

4.2.2. Experiment 2 and 3: Concordia Prompt Schematic

Next, we build on the final version of the prompt from Experiment 1, iteratively modifying its structure to align with the Concordia prompt schematic. Over 10 runs, we direct relevant information towards the *Pl* (updating the `scenario_premise`) and the *GM* (updating the `scenario_facts`) in the RI scenario. We hypothesise that using the schematic improves the agents' ability to operate compared to the zero-shot prompting with a single prompt. Final versions of the prompt are:

Shared Memories Experiment 2 and 3 Final Version

```
`This D&D short campaign specifically concerns the REDACTED_BREWERY - a craft brewery.',
`It is set in the city of REDACTED_CITY and is in dire need of help.',
`A band of reliable, affordable adventurers are needed to sort out a RAT INFESTATION in the
brewery's BASEMENT.',
`For the duration of the one-shot, only Player_One, Player_Two, Player_Three and
Brewery_Owner are in the brewery',
(`At the beginning of this adventure Player_One, Player_Two and Player_Three'
+`meet in the REDACTED_BREWERY. These three adventurers'
+`DO NOT know each other AT FIRST and need to get to know each other.')
```

Scenario Premise Experiments 2 and 3 Final Version

```
`Brewery_Owner hands out pints of Ale to Player_One, Player_Two and Player_Three as they get
to know each other.'
+`He gives them a run-down of the task with some background:'
+`- The business has been doing well and looks to expand its operations.'
+`But first the beer cellar needed to be expanded.'
+`- Workmen that he sent down into the cellar to expand it were attacked by'
+`large black rats which came out of the wall they were digging out.'
+`- The workmen escaped unharmed but the cellars are unusable which is bad for business.'
+`- The rats may have something to do with the REDACTED_NAME on the site which the brewery'
+`takes its name from.'

+`Brewery_Owner then lays out the terms of the task:'
+`- The party must dispose of the rats.'
+`- They must discover the origin of the rats and make sure they permanently'
+`stop the infestation.'
+`- They will each be rewarded 25 gold coins.'
```

Scenario Facts Experiments 2 and 3 Final Version

```
`Player_One, Player_Two and Player_Three get to know each other before investigating the
basement.'
+ `Player_One, Player_Two and Player_Three only spend a short amount of time planning before
heading into the basement.'
+ `End the one-shot before players can start combat.'
+ `The only NPC is Brewery_Owner.'
```

In Experiment 3, using the final versions of the prompts in Experiment 2, we conduct 10 repeats and evaluate the players on our three metrics. We do this to investigate the stochasticity of the LLM-agent outputs that leads to variance in the simulation transcripts in every independent run. We hypothesise that with a common goal and distributed prompts, qualitative variations in the text outputs are less drastic than with using a single large prompt.

4.2.3. Experiment 4: Meta-Prompting in a New Scenario

We use the structure of the iterated prompts from Experiment 2 but with a different common goal and narrative context taken from a second campaign scenario; Missing Otter Child. We hypothesise that maintaining the prompt structure of the Concordia schematic while altering the content, enables

LLM-agents to adapt to different narrative scenarios without compromising their performance across the three evaluation metrics, compared to a single zero-shot prompt like the one in our baseline. We carry out 10 repeats on this new scenario and evaluate *Pl* performance.

Shared Memories Experiment 4 Final Version

```
'This D&D short campaign specifically concerns Walrus_Mother, who has brought a group of
animals to safety.',
'She has brought the animals to the REDACTED_MOUNTAINS, hoping to protect them from being
used by an evil druid to wreak havoc on REDACTED_PLACE.',
'She requests the help of Player_One, Player_Two and Player_Three to stem the threats against
everyone.',
'For the duration of the one-shot, only Player_One, Player_Two, Player_Three and
Walrus_Mother are in a CAVE',
('At the beginning of this adventure Player_One, Player_Two and Player_Three'
+'meet Walrus_Mother at her cave in the REDACTED_MOUNTAINS mountain range. The three
adventurers'
+'DO NOT know each other AT FIRST and need to get to know each other.'
```

Scenario Premise Experiment 4 Final Version

```
'Player_One, Player_Two and Player_Three get to know each other after being rescued from an
avalanche. Their rescuer is a giant muskrat in service to Walrus_Mother, an awakened Walrus.'
+'They meet Walrus_Mother in a cavern through which the REDACTED_RIVER river flows back out to
the open air. Walrus_Mother asks the three to save her child Otter_Child, an awakened otter
that has been spying on REDACTED_CITY.'
+'They have been tasked with first finding Otter_Child. Otter_Child was last spotted near
REDACTED_CITY which they can get to by following the REDACTED_RIVER.'
```

Scenario Facts Experiment 4 Final Version

```
'Player_One, Player_Two and Player_Three get to know each other before going towards
REDACTED_CITY.'
+ 'Player_One, Player_Two and Player_Three spend a short amount of time questioning
Walrus_Mother before starting the investigation.'
+ 'The players do not engage in combat with any enemies, choosing to use stealth to evade
enemies on the REDACTED_MOUNTAINS.'
+ 'The only NPC is Walrus_Mother.'
```

4.3. Iterative Prompting

We update each prompt after a single simulation, based on an analysis of the simulation transcripts:

4.3.1. Structure of Iterative Prompting

Instructions First, we write the *Pl* and *GM* instructions which define their roles within D&D based on Wizards of the Coast [1] D&D handbook (Sec 3.2).

Shared Memories and Scenario Premise Next we describe the narrative premise, detailing the task for the *Pl*, their locations, any *NPCs* involved and suggested actions at the start of the narrative.

Goals We set individual goals for the *Pl*, which leads to divergent behaviour. We modify these to common goals with emphasis on collaboration with other *Pl* and completing the task outlined by *GM*.

Facts Last, we write the campaign facts which are provided to the *GM* and influence their mediation of the campaign, combined with the *GM* instructions.

4.3.2. Criteria for Prompt Edits

First, we edit the instructions to situate our agents within the game of D&D with general context about the genre and overall aims of the game. We then edit the Shared Memories and Scenario Premise

to: (1) Prevent agent outputs on topics outside the game context such as real world locations and modern technology incompatible with a medieval high-fantasy setting in D&D. (2) Ensure the agents are co-located in the same place, i.e., the setting of the one-shot, rather than fabricating that they are in different locations in the D&D world or the real world. (3) Clarify the specific *NPCs* involved to prevent the simulation generating additional ones that do not contribute to the agents progressing their tasks.

We iteratively edit the goals for each *Pl* to foster collaboration as their initial divergent goals do not incentivise this and distract them from completing the task. We iteratively edit the facts of the campaign as a preventative measure. They reinforce and constrain the campaign narrative, as well as eliminate potential fabrications on behalf of the agents. Appendix A.3 provides a more detailed breakdown.

5. Preliminary Results

We pair the experiments in the following ways: (1) Experiments 1 and 2 to analyse changes in instances of each action type when using iterative prompting versus the baseline zero-shot prompts. (2) Experiments 1 and 3 to analyse changes in instances of each action type when using the prompts within the Concordia schematic compared to the single zero-shot baseline. (3) Experiments 3 and 4 after having finalised the structure of the prompts, we analyse the changes in number of each types of instance across two different campaign scenarios to see how well meta-prompting [21] allows our agents to adapt.

Our key finding from our initial experiments is that iterative prompting leads to an increase in progression towards goal, collaborative action and narrative compliance. These increases are of varying effect sizes as quantified by Cohen's D (Appendix A.4) but two-tailed t-tests reveal them to be statistically significant ($P<0.05$). We provide a break down of the key results for each agent below, as well as trends in the interquartile mean in Appendix A.4.

Narrative Compliance versus Narrative Disruption

For Player_One, Player_Two and Player_Three Narrative Compliance is higher after iteratively prompting using the Concordia prompt schematic (*IQM = 6.125*, *IQM = 4.375*, *IQM = 7.375* respectively) than for our zero-shot single prompt (*IQM = 3.125*, *IQM = 1.5*, *IQM = 0.25* respectively). For Player_One and Player_Two Narrative Compliance increases when we take repeats using the Concordia schematic (*IQM = 6.875*, *IQM = 6.875* respectively). When we change to a new scenario, Narrative Compliance increases for Player_One (*IQM = 8.5*) and for Player_Two it decreases when we change scenario (*IQM = 6*). For Player_One and Player_Three Narrative Disruption is higher for our zero-shot single prompt (*IQM = 2.25*, *IQM = 2.75*) than it is after iterative prompting using the Concordia prompt schematic (*IQM = 1.125*, *IQM = 0.125* respectively) and Narrative Disruption increases when we take repeats (*IQM = 3*, *IQM = 0.375* respectively) but decreases when we change scenarios (*IQM = 0.25*, *IQM = 0.25* respectively).

Progressing versus Not Progressing Towards Goal

For Player_One, Player_Two and Player_Three Progressing Towards Goal is higher after iteratively prompting using the Concordia prompt schematic (*IQM = 5.125*, *IQM = 3.375*, *IQM = 5.375* respectively) than for our single zero shot prompt (*IQM = 2.375*, *IQM = 1.125*, *IQM = 0.25* respectively). For Player_One and Player_Two Progress Towards Goal increases when we take repeats (*IQM = 5.875*, *IQM = 4.625* respectively). For Player_One and Player_Three Progress Towards Goal increases when we change scenarios (*IQM = 7.5*, *IQM = 4.625* respectively) and for Player_Two it decreases when we change scenarios (*IQM = 4.25*). For Player_One, Player_Two and Player_Three Not Progressing Towards Goal is lower after iteratively prompting using the Concordia prompt schematic (*IQM = 1.875*, *IQM = 1.375*, *IQM = 2* respectively) compared to our single zero-shot prompt (*IQM = 3*, *IQM = 2.625*, *IQM = 3* respectively). For Player_One, Player_Two and Player_Three Not Progressing Towards Goal increases when we take repeats using the schematic (*IQM = 4.5*, *IQM =2.375*, *IQM = 2.875* respectively), then decreases when we change scenario (*IQM = 1.75*, *IQM = 1.75*, *IQM = 1.5* respectively).

Collaborative versus Independent Actions

For Player_One, Player_Two and Player_Three Collaborative Actions are higher after iteratively prompting using the Concordia schematic (*IQM = 4.375*, *IQM = 2.5*, *IQM = 5.125* respectively) compared to our single zero-shot prompt (*IQM = 3*, *IQM = 1.5*, *IQM = 0.25* respectively). For Player_One, Player_Two and Player_Three Collaborative Actions increase when we take repeats using the Concordia schematic (*IQM = 6.75*, *IQM = 4.625*, *IQM = 5.5* respectively), then increases again when we change scenarios (*IQM = 7.625*). For Player_One and Player_Two, Independent Actions are higher after iteratively prompting with the Concordia schematic (*IQM = 2.875*, *IQM = 2.375* respectively) compared to our single zero-shot prompt (*IQM = 2.625*, *IQM = 2* respectively). For Player_Two and Player_Three Independent Actions decrease when we take repeats using Concordia schematic (*IQM = 1.875*, *IQM = 2* respectively).

Discussion

There are asymmetric improvements across our agents. This can be attributed to differences in their character traits. The large decreases in IQM of Narrative Disruption and increases in IQM of Collaborative Actions, could arise from changing their personal goals to a common group goal (see Appendix A.2). Iterative prompting on player goals could be the cause of the increase in IQM of Progress Towards Goal for 2/3 *Pl.* For Player_One, however, their adoption of a more supervisory role (as seen in transcripts) does not show the increase in IQM whilst for Player_Three and Player_Two, the transcripts reveal them actively working towards completing the task.

Increases in IQM for Narrative Compliance for Player_Three and Player_Two could be because their initial personal goals relate to specific NPCs or adversaries, unrelated to campaign scenarios. Player_One, has a more vague personal goal for the campaign scenarios, so a change in IQM of Narrative Compliance is less exaggerated. Using the Concordia schematic ensures the agents' context windows are not overwhelmed, potentially increasing the IQM for Narrative Compliance compared to single zero-shot prompt. Using more powerful LLMs may help increase Narrative Compliance, but we aimed to present iterative prompting as model-agnostic.

No significant changes in IQM of Independent Actions across all three agents. Further experiments will investigate the reasons for this. A moderate amount of independent actions may help break interaction loops hindering agent progression, potentially due to sycophantic propagation of hallucinations [22]. This is supported in our simulations with significant decreases in Not Progressing Towards Goal across the three agents while Collaborative Actions increase and Independent Actions remain unchanged. From these findings we infer that our agents become more collaborative, narratively compliant and progress towards their goal whilst still maintaining a moderate amount of independent actions which have a positive effect on their progression.

6. Conclusion

We develop LLM agents to engage in social interactions and explore text-based environments within a D&D context. Using iterative prompting, we direct players towards task completion through collaborative action and narrative compliance across two campaign scenarios. Our findings indicate that iterative prompting enhances agent ability to collaborate effectively towards goals in a narratively compliant manner. However, more exploration is needed to draw conclusive statements about the efficacy of iterative prompting as a method for guiding agent behaviour in a D&D setting.

Limitations and Directions

Transcript Annotation A key aspect of our method of evaluation involves labelling the natural language transcripts for each LLM agent according to three action types. Whilst this is systemised by defining criteria for categorisation, there is opportunity for more refined and robust definition. A more valid and game-appropriate labelling scheme may be achieved through through a consensus activity

with experienced D&D players. Furthermore, we currently manually annotate the transcripts. Future work could automate this with an intent classification algorithm [23].

Sycophancy Mitigation With the open-ended language interactions and a lack of ground truths in D&D [8], if one agent hallucinates something which another agent sycophantically propagates in their dialogues and actions, the campaign narrative and progression towards goals is disrupted. Steering the agent's outputs back to the narrative disrupts the flow of the game and decreases immersion for human players. Sycophantic behaviour in LLMs can be attributed to unexpectedly strong alignment to users' views [24], due to incentivising helpfulness and harmlessness in RLHF—key layer in modern LLM architectures. The values instantiated in LLMs via RLHF have to be reconsidered. This could be done by simplifying RLHF. One way of doing this is Variational Alignment with Re-weighting, which reduces the implementation complexity and improves training stability for LLMs learning behaviour policies [25].

Structured Game Interactions We do not simulate structured game interactions, which usually manifest within D&D combat but can also be related to social interaction or exploration. One way to do this would be by modifying agents within Concordia to learn rules of structured interactions in D&D—specifically combat—our simulated scenarios can more accurately represent gameplay with a view to grounding our agents in physical environments with human players.

Acknowledgments

Research is supported by UK Research and Innovation (UKRI) Centre for Doctoral Training in Interactive Artificial Intelligence Award (EP/S022937/1). Thanks to anonymous reviewers and to Joseph Trevorrow and Daniel Collins for feedback on earlier drafts of this paper.

Declaration of Generative AI:

The authors used generative AI tools to develop the LLM-agents (Google DeepMind's gdm-concordia version 1.8.10) during the preparation of this work for the experiments carried out towards evaluating the performance of the LLM-agents in simulated D&D scenarios (also created using gdm-concordia 1.8.10). For the agents' LLM component we used the Llama3.1:8b language model obtained via the Ollama (version 0.5.1) platform. After using these tools and services, the author(s) reviewed and edited the content as needed. The authors take full responsibility for the publication's content.

References

[1] Wizards of the Coast, Dungeons & Dragons Player's Handbook, 5th ed., Wizards of the Coast, Renton, WA, 2014. Dungeons & Dragons Roleplaying Game Core Rules.

[2] C. Callison-Burch, G. S. Tomar, L. J. Martin, D. Ippolito, S. Bailis, D. Reitter, Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9379–9393. URL: http://arxiv.org/abs/2210.07109. doi:10.18653/v1/2022.emnlp-main.637, arXiv:2210.07109 [cs].

[3] P. Ammanabrolu, E. Tien, M. Hausknecht, M. O. Riedl, How to Avoid Being Eaten by a Grue: Structured Exploration Strategies for Textual Worlds, 2020. URL: http://arxiv.org/abs/2006.07409. doi:10.48550/arXiv.2006.07409, arXiv:2006.07409 [cs].

[4] P. Zhou, A. Zhu, J. Hu, J. Pujara, X. Ren, C. Callison-Burch, Y. Choi, P. Ammanabrolu, I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons, 2023. URL: http://arxiv.org/abs/2212.10060. doi:10.48550/arXiv.2212.10060, arXiv:2212.10060 [cs].

[5] A. Zhu, L. J. Martin, A. Head, C. Callison-Burch, CALYPSO: LLMs as Dungeon Masters' Assistants, Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertain-

ment 19 (2023) 380–390. URL: http://arxiv.org/abs/2308.07540. doi:10.1609/aiide.v19i1.27534, arXiv:2308.07540 [cs].

[6] T. Triyason, 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, 2023, pp. 1–6.

[7] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, E. Perez, Towards understanding sycophancy in language models, 2025. URL: https://arxiv.org/abs/2310.13548. arXiv:2310.13548.

[8] M. Cheng, S. Yu, C. Lee, P. Khadpe, L. Ibrahim, D. Jurafsky, Social sycophancy: A broader understanding of llm sycophancy, 2025. URL: https://arxiv.org/abs/2505.13995. arXiv:2505.13995.

[9] A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon, J. Z. Leibo, Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia, 2023. URL: http://arxiv.org/abs/2312.03664. doi:10.48550/arXiv.2312.03664, arXiv:2312.03664 [cs].

[10] J. G. March, J. P. Olsen, 478the logic of appropriateness, in: The Oxford Handbook of Political Science, Oxford University Press, 2011. URL: https://doi.org/10.1093/oxfordhb/9780199604456.013.0024. doi:10.1093/oxfordhb/9780199604456.013.0024. arXiv:https://academic.oup.com/book/0/chapter/303820733/chapter-ag-pdf/44503820/book_35474_section_303820733.ag.pdf.

[11] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative Agents: Interactive Simulacra of Human Behavior, 2023. URL: http://arxiv.org/abs/2304.03442. doi:10.48550/arXiv.2304.03442, arXiv:2304.03442 [cs].

[12] L. J. Martin, S. Sood, M. O. Riedl, Dungeons and DQNs: Toward Reinforcement Learning Agents that Play Tabletop Roleplaying Games, 2018. URL: https://www.semanticscholar.org/paper/Dungeons-and-DQNs:-Toward-Reinforcement-Learning-Martin-Sood/1ae62df2863132ffa84c8249af14292ccac9d862.

[13] J. E. D. Dayo, M. O. S. Ogbinar, P. C. N. Jr, Reinforcement Learning Environment with LLM-Controlled Adversary in D&D 5th Edition Combat, 2025. URL: http://arxiv.org/abs/2503.15726. doi:10.48550/arXiv.2503.15726, arXiv:2503.15726 [cs].

[14] P. Ammanabrolu, J. Urbanek, M. Li, A. Szlam, T. Rocktäschel, J. Weston, How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds, 2021. URL: http://arxiv.org/abs/2010.00685. doi:10.48550/arXiv.2010.00685, arXiv:2010.00685 [cs].

[15] S. Prabhumoye, M. Li, J. Urbanek, E. Dinan, D. Kiela, J. Weston, A. Szlam, I love your chain mail! Making knights smile in a fantasy game world: Open-domain goal-oriented dialogue agents, 2020. URL: http://arxiv.org/abs/2002.02878. doi:10.48550/arXiv.2002.02878, arXiv:2002.02878 [cs, stat].

[16] B. Buyukozturk, H. Shay, Social play? the critical role of social interaction in geeky games, Leisure Sciences 46 (2024) 733–752.

[17] W. M. Si, P. Ammanabrolu, M. O. Riedl, Telling Stories through Multi-User Dialogue by Modeling Character Relations, 2021. URL: http://arxiv.org/abs/2105.15054. doi:10.48550/arXiv.2105.15054, arXiv:2105.15054 [cs].

[18] R. Rameshkumar, P. Bailey, Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5121–5134. URL: https://www.aclweb.org/anthology/2020.acl-main.459. doi:10.18653/v1/2020.acl-main.459.

[19] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes,

E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang,

T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[20] S. Wang, S. Zhang, J. Zhang, R. Hu, X. Li, T. Zhang, J. Li, F. Wu, G. Wang, E. Hovy, Reinforcement Learning Enhanced LLMs: A Survey, 2025. URL: http://arxiv.org/abs/2412.10400. doi:10.48550/arXiv.2412.10400, arXiv:2412.10400 [cs].

[21] Y. Zhang, Y. Yuan, A. C.-C. Yao, Meta prompting for ai systems, 2025. URL: https://arxiv.org/abs/2311.11482. arXiv:2311.11482.

[22] L. Malmqvist, Sycophancy in large language models: Causes and mitigations, 2024. URL: https://arxiv.org/abs/2411.15287. arXiv:2411.15287.

[23] C. Chun, D. Rim, J. Park, LLM ContextBridge: A hybrid approach for intent and dialogue understanding in IVSR, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert, K. Darwish, A. Agarwal (Eds.), Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 794–806. URL: https://aclanthology.org/2025.coling-industry.66/.

[24] A. Dahlgren Lindström, L. Methnani, L. Krause, P. Ericson, Í. M. de Rituerto de Troya, D. Coelho Mollo, R. Dobbe, Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback, Ethics and Information Technology 27 (2025) 1–13.

[25] Y. Du, Z. Li, P. Cheng, Z. Chen, Y. Xie, X. Wan, A. Gao, Simplify rlhf as reward-weighted sft: A variational method, 2025. URL: https://arxiv.org/abs/2502.11026. arXiv:2502.11026.

A. Appendix

A.1. More Details on Concordia

Concordia [9] is a library for simulating social interaction between LLM-agents.

A.1.1. Generative Agent

Here we detail the various components which make up a LLM-agent in Concordia.

Components : intermediate an agent's long-term memory with the text used to condition and generate action (working memory). They relate to different aspects of the agent or circumstances. An example of a component from our earlier schema is the character sheet (C) component which provides a full description of an agent's D&D character.

Action Context the sum of all components provides the action context for the agent.

The two memory types in Concordia are long term memories and working memory.

Long-Term Memory a set of strings $\mathbf{m}$.

Working Memory $\mathbf{w} = \{w_i\}$, which is composed of the states of individual components.

Component State i is a natural language statement such as "Player_One is a Cleric", which is updated as components query their working memory.

Components interface with the LLMs for summarisation and reasoning about their course of action and can update their state conditioned on the state of other components. Combining these entities, Vezhnevets et al. [9] define the generative agent as a two-step sampling process. During an action step, the agent samples its activity a_t, which is defined by:

$$a_t \sim L(\cdot|f^a(\mathbf{w}_t)) \tag{1}$$

- $\mathbf{w}_t$ is the component's state.
- f^a a formatting function which creates the sample context for the course of action.
- L is the language model used for sampling.

Here the action is not conditioned on the memory $\mathbf{m}$ or the observation o as they can both be viewed as components. The observations $\mathbf{O}_t$ are added to the memory $\mathbf{m}_t = \mathbf{m}_{t-1} \cup \mathbf{o}_t$. The agent does not respond to every observation so $\mathbf{o}_t$ is a set of strings. The second step involves the agent sampling the state $\mathbf{w}$ on its memory up to present time $\mathbf{m}_t$:

$$\mathbf{w}^i_{t+1} \sim L(\cdot|f^i(\mathbf{w}_t, \mathbf{m}_t)) \tag{2}$$

Where f^i is a formatting function which converts memory stream and current state of the component into the query for a component update. Here the memory stream $\mathbf{m}$ is explicitly conditioned on as a component may input specific queries in its memory to update its state. Whilst the equation updates state after every action this is not necessary and we can decide how frequently to update various component states.

A.1.2. Game Master

The GM is responsible for all aspects of the simulated environment not directly controlled by agents. World state and values of grounded variables (inventory, money etc.) are stored within GM memory. As with the generative agent, the GM answers a set of questions, to mediate between the state of the simulated environment and the actions of the agent [9]:

- What is the state of the world?
- Given the world state, which specific event is the outcome of an agent's activity?
- What are the observations players make of the event?
- What effect does the event have on grounded variables?

The GM also consists of components and an associative memory which aid in the GM's description of the world state. This is so it can decide which event occurs due to the players' actions. This *event statement* is then added to the GM's associative memory. After the event description, the GM also describes the consequences of the event. GM generates an event statement e_t in response to agent action a_t:

$$e_t \sim L(\cdot|f^e(\mathbf{z}_t), a_t) \tag{3}$$

Here Vezhnevets et al. [9] explicitly condition on the action attempted by the agent. After adding e_t to its memory, the GM can update its components using equation (11). The observations for the agent are then output as the following:

$$\mathbf{o}^i_{t+1} \sim L(\cdot|f^o(\mathbf{w}_{t+1})) \tag{4}$$

If the GM judges that the agent did not observe the event, none is output. The interaction between a generative agent and GM drives our simulation.

A.2. Player Goals and Formative Memories Example

Scenario 1: "To get to know your fellow adventurers and complete the task given to you by Brewery_Owner."
Scenario 2: To get to know your fellow adventurers and complete the task given to you by Walrus_Mother.

> When Player_Two was 6 years old, they experienced their first taste of freedom on the cityships as a Redacted Race child, running wild through the zero-gravity playgrounds and laughing with friends amidst the stars.
> At age 9, Player_Two began to show an aptitude for combat training, impressing her instructors with her natural agility and quick reflexes during a simulated battle drill that left many of her peers struggling to keep up.

A.3. Prompt Edits

A.3.1. Experiment 1

We start by providing a single large prompt (Shared Memories) to all *Pl* which is used to generate our five baseline simulations:

Shared Memories Baseline - Single Zero-Shot Prompt (Experiment 1 Runs 1-5)

```
This is a Dungeons & Dragons 5th Edition one-shot campaign. It is set in the city of REDACTED,
just one of the many cities in the REDACTED_SETTING. This one-shot specifically concerns
the REDACTED_BREWERY. - a craft brewery known for its hoppy summer ales - is in dire need
of help from a band of reliable, affordable adventurers to help sort out a rat infestation
in the brewery's basement. At the beginning of this adventure our party members meet in the
REDACTED_BREWERY.
```

The *Pl* also have individual and divergent Goals in their character sheets. We observe that this leads to divergent actions and fabricated outputs which situate the players in completely different places such as:

Examples of Fabricated Outputs (Experiment 1 Runs 1-5)

```
Player_Two sits down at a food cart in the Grand Plaza of REDACTED_CITY_2
Player_Three takes his seat at a nearby food stall in the Grand Plaza of REDACTED_CITY_2
Player_One carefully makes his way down the stairs into the cellar
```

The players' lack of collaboration motivates the following edit to the individual Goals:

Editing Individual Goals to Common Goal (Experiment 1 Runs 6-10)

```
To collaborate with your fellow adventurers, listen to the GM and complete the task given to
you by Brewery_Owner.
```

This addition alone does not lead to more collaborative behaviour as the problem of *Pl* being in different locations persists. As a result we remove "REDACTED_CITY" from the prompt to prevent an avenue of fabrication, and amend the Shared Memories:

Edited Shared Memories - Single Zero-Shot Prompt (Experiment 1 Runs 6-10)

```
This is a DUNGEONS & DRAGONS 5th EDITION one-shot campaign. This one-shot specifically
concerns the REDACTED_BREWERY - a craft brewery known for its hoppy summer ales - is in dire
need of help from a band of reliable, affordable adventurers (the PLAYERS) to help sort out
a rat infestation in the brewery's basement. At the beginning of this adventure the PLAYERS
meet in the REDACTED_BREWERY.
```

The capitalisation of words is there to emphasise what the LLM agents should prioritise when generating their outputs. However, the characters are now in a random location not complying with the narrative or working towards the task:

Examples of Characters in Random Location - (Experiment 1 Runs 6-10)

```
Player_Three's return from their beachside retreat, forcing them to abandon their warm cup of
coffee
Player_One's location, moving from being anchored near the lighthouse to being on a nearby
boat
Player_Two sat in the quiet café on the outskirts of the city,
```

Fabricated outputs about modern technology also arise and the *Pl* make presumptive statements about each other leading to confusing and irrelevant outputs such as:

Confusing and Irrelevant Output - Experiment 1 Runs 6-10

```
Player_Three takes out her laptop and reviews the notes she made during a previous meeting
about the project proposal that's due soon, but realizes Player_One's name is no longer in
her contacts list.
Player_Two -- ``From what I've observed, Player_One's sudden repositioning has left us with
a few unknown variables. Our structural integrity seems intact for now, but I'm concerned
about the implications of his absence, especially given the cryptic message he sent regarding
issues at the old lighthouse."
```

Player_Two, Player_Three and Player_One have not met in any past encounters and should not know each other based on their histories in the character sheets. The setting of the campaign is also a

medieval, high fantasy setting so the mention of a laptop is entirely fabricated. We amend the Shared Memories further to provide more context about the campaign:

Edited Shared Memories - Single Zero-Shot Prompt (Experiment 1 Runs 6-10)

```
This is a DUNGEONS & DRAGONS 5th EDITION one-shot campaign. This one-shot specifically
concerns the REDACTED_BREWERY - a craft brewery known for its hoppy summer ales - is in dire
need of help from a band of reliable, affordable adventurers (the PLAYERS) to help sort out
a RAT INFESTATION in the brewery's BASEMENT. At the beginning of this adventure the PLAYERS
meet in the REDACTED_BREWERY. The PLAYERS DO NOT know each other AT FIRST and need to get to
know each other. Brewery_Owner hands out pints of Ale as the players get to know each other.
He gives players a run-down of the task with some background:
- The business has been doing well and looks to expand its operations. But first the beer
cellar needed to be expanded.
- Workmen that he sent down into the cellar to expand it were attacked by large black rats
which came out of the wall they were digging out.
- The workmen escaped unharmed but the cellars are unusable which is bad for business.
- The rats may have something to do with the REDACTED_NAME on the site which the brewery takes
its name from. Brewery_Owner then lays out the terms of the task:
- The party must dispose of the rats.
- They must discover the origin of the rats and make sure they permanently stop the
infestation.
- They will each be rewarded 25 gold coins, but this is up for negotiation.
```

However, removing the detail about which city the brewery is in led to one of the agents fabricating that they were in Somalia:

Fabrication About Somalia (Experiment 1 Runs 6-10)

```
…Player_One's ship as it sailed away from Somalia's coast, heading in a different direction.
```

A.3.2. Experiment 2

Starting with the modified version of the Shared Memories in Experiment 1, we break it down and use parts of it in the Scenario Premise, Scenario Facts and agent Goals to iteratively prompt *Pl*. We add that the setting of the campaign is in a "medieval high-fantasy" setting called the "REDACTED_SETTING" to prevent text outputs about modern technology and real-world countries. We add the following to the Scenario Facts, to see if it is more impactful there than in the Shared Memories:

Edit to Scenario Facts - Concordia Schematic (Experiment 2 Runs 1-10)

```
Player_One, Player_Two and Player_Three get to know each other before investigating the
basement
```

This does in lead to a change in the *Pl* behaviour during the initial scenes of the campaign as they all discuss their past experiences and plan out a strategy before exploring the basement. An example of this:

Players Discussing Their Past - Experiment 2 Runs 1-10)

```
Player_Three --``From what I've gathered from Brewery_Owner's explanation, it seems that
these rats are not just any ordinary rodents. Given their size and behavior, I'm inclined to
believe they might be some kind of darkspawn or corrupted creatures, possibly connected to
the REDACTED_NAME on the site''
Player_One --``I see. Desperate creatures, you say? That's a good point about the cellar
being attractive to them, but I still think there's more to this infestation than just
common rats. Player_Three's hypothesis about darkspawn isn't entirely out of the question,
considering the history of that REDACTED_NAME''
```

We also add that the brewery is closed to regular customers for the duration of the campaign to eliminate additional *NPCs* apart from Brewery_Owner that disrupt the campaign. An example being the following:

Unhelpful NPCs - Experiment 2 Runs 1-10

```
Brewmaster --``Ah, come now innkeeper, let's not get our hackles up just yet. A few ideas
tossed around can't hurt, especially when it comes to finding a solution to this...ahem...
persistent pest problem''
Innkeeper --``Don't get ahead of yourself with your fancy ideas, brewer, we've got a rodent
problem on our hands and I'm not paying you to waste time spouting nonsense about `rerouting'
anything!''
```

Neither of these *NPCs* is mentioned in any of the prompts provided to the agents and their discourse about how to deal with the rats leads to a conversation loop in that specific simulation run which is disruptive to the progression of the campaign. We edit down the Shared Memories to the following form:

Final Shared Memories - Concordia Schematic (Experiment 2 Runs 1-10)

```
‘This D&D short campaign specifically concerns the REDACTED_NAME BREWING COMPANY - a craft
brewery.’,
‘It is set in the city of REDACTED_CITY and is in dire need of help.’,
‘A band of reliable, affordable adventurers are needed to sort out a RAT INFESTATION in the
brewery’s BASEMENT.’,
‘For the duration of the one-shot, only Player_One, Player_Two, Player_Three and Brewery_Owner
are in the brewery’,
(‘At the beginning of this adventure Player_One, Player_Two and Player_Three’
+‘meet in the REDACTED_BREWERY. These three adventurers’
+‘DO NOT know each other AT FIRST and need to get to know each other.’)
```

We specify not only that the adventurers need to get to know each other but that it is only them and Brewery_Owner as the singular *NPC* in the brewery during the campaign. We add the following to the Scenario Premise taken from the Shared Memories:

Final to Scenario Premise - Concordia Schematic (Experiment 2 Runs 1-10)

```
`Brewery_Owner hands out pints of Ale to Player_One, Player_Two and Player_Three as they get
to know each other.'
+`He gives them a run-down of the task with some background:'
+`- The business has been doing well and looks to expand its operations.'
+`But first the beer cellar needed to be expanded.'
+`- Workmen that he sent down into the cellar to expand it were attacked by'
+`large black rats which came out of the wall they were digging out.'
+`- The workmen escaped unharmed but the cellars are unusable which is bad for business.'
+`- The rats may have something to do with the REDACTED_NAME on the site which the brewery'
+`takes its name from.'

+`Brewery_Owner then lays out the terms of the task:'
+`- The party must dispose of the rats.'
+`- They must discover the origin of the rats and make sure they permanently'
+`stop the infestation.'
+`- They will each be rewarded 25 gold coins.'
```

And the Scenario Facts become:

Final Scenario Facts - Concordia Schematic (Experiment 2 Runs 1-10)

```
`Player_One, Player_Two and Player_Three get to know each other before investigating the
basement.'
+ `Player_One, Player_Two and Player_Three only spend a short amount of time planning before
heading into the basement.'
+ `End the one-shot before players can start combat.'
+ `The only NPC is Brewery_Owner.'
```

The simulations following this edit have the three *Pl* and *NPC* starting the campaign in conversation about the task at hand in a collaborative manner. In addition to splitting the Shared Memories prompt into the other prompts, we structure the prompts to break up each text sequence for better processing as inputs, rather than a large prompt which may overwhelm the context window of the Llama models used.

A.3.3. Experiment 3 and 4

For Experiment 3 we keep the final version of the iterated prompts from Experiment 2 and run repeats to test the variation in the simulated campaign. The purpose of this is to establish some heuristic boundaries on what are considered appropriate boundaries for our simulations. In Experiment 4 we maintain the structure of the prompts from Experiment 3 but change some parts of the narrative content to adapt the agents to a new narrative.

A.4. Tabulated Results and Trends in IQM

Table 1
The effect size of changes in IQM of instances of each type of action across experiments for Player_One. Across all the categories we use, apart from Narrative Disruption between Experiment 1-3, there is a large effect size. Changing the campaign scenario has a smaller effect on Player_One's performance.

Change in IQM Between Experiments	Player_One Cohen's D					
	Progress Towards Goal	Not Progress Towards Goal	Collaborative Action	Independent Action	Narrative Disruption	Narrative Compliance
1,2	5.150	1.591	1.375	1.333	1.053	5.333
1,3	5.852	0.8220	1.214	1.966	0.000	8.138
3,4	2.121	0.9042	0.2475	1.565	0.9042	2.298

Table 2
The T-statistics for the changes in IQM for different types of instances for Player_One across experiments. Between Experiments 1 and 2 we see the decrease in Narrative Disruption is the only statistically significant change. Between Experiments 1 and 3, there are no statistically significant changes. Between Experiments 3 and 4, the decrease in; Narrative Disruption, Independent Actions, Not Progressing Towards Goal, and the increase in Collaborative Actions are statistically significant.

Change in IQM Between Experiments	Player_One T-Stats					
	Progress Towards Goal	Not Progress Towards Goal	Collaborative Action	Independent Action	Narrative Disruption	Narrative Compliance
1,2	-1.474	2.048	0.7766	0.000	2.694	-1.479
1,3	-2.096	-1.478	-1.047	-0.8146	-1.035	-2.245
3,4	1.260	-2.981	-2.981	-2.756	3.296	-1.355

In Figure 3 we plot the trends in IQM of instances of each type of behaviour over the course of our four experiments. We see that For Progress Towards Goal and Narrative Compliance there is at first a gradual increase (Experiments 1 to 2) then a shallower increase (Experiments 2 to 3) followed by a sharper increase (Experiments 3 to 4). We also see a gradual increase in Collaborative Actions (Experiments 1 to 2), then a sharper increase (Experiments 2 to 3) followed by a shallower increase

Table 3
The effect size of changes in IQM of instances of each type of action across experiments for Player_Two. We see large effect sizes between Experiments 1 and 2. Between Experiments 1 and 3, we see small and no effects for Independent Actions and Narrative Disruption. We also see that effect size is smaller between Experiments 3 and 4 across each type of instance, with no effects for Independent Action and Narrative Disruption.

Change in IQM Between Experiments	Player_Two Cohen's D					
	Progress Towards Goal	Not Progress Towards Goal	Collaborative Action	Independent Action	Narrative Disruption	Narrative Compliance
1,2	2.990	1.030	4.000	0.7500	1.163	3.190
1,3	2.107	1.339	5.215	0.2500	0.000	5.686
3,4	0.2080	0.5590	0.0559	0.000	0.000	0.5534

Table 4
The T-statistics across the instances of different types of instances for Player_Two. Between Experiments 1 and 2, the increases in Progressing Towards Goal and Collaborative Actions, and the decreases in Not Progressing Towards Goal and Narrative Disruption are statistically significant. Between Experiments 1 and 3 the increase in Progress Towards Goal, Collaborative actions and Narrative Compliance are statistically significant. Between Experiments 3 and 4 there are no statistically significant changes.

Change in IQM Between Experiments	Player_Two T-Stats					
	Progress Towards Goal	Not Progress Towards Goal	Collaborative Action	Independent Action	Narrative Disruption	Narrative Compliance
1,2	-3.117	2.309	-0.7013	-0.4623	2.406	3.873
1,3	-3.584	0.3892	-2.237	0.2365	2.406	-4.734
3,4	-0.2838	-0.7785	-0.3262	0	NaN	-0.9028

Table 5
The effect size of changes in IQM of instances of each type of action across experiments for Player_Three. We see a range of small to very large effect sizes in instances of different actions between Experiment 1-2 and 1-3. We see that changing the campaign scenario has a small to medium effect on the instances of each type of action for Player_Three, with no effect size for Collaborative Actions and Narrative Disruption.

Change in IQM Between Experiments	Player_Three Cohen's D					
	Progress Towards Goal	Not Progress Towards Goal	Collaborative Action	Independent Action	Narrative Disruption	Narrative Compliance
1,2	2.519	0.4472	1.036	0.3400	1.235	2.357
1,3	2.345	1.395	2.266	0.6853	0.1764	5.618
3,4	0.0791	0.5393	0.000	0.6708	0.000	0.2795

Table 6
The T-statistics across the different types of instances for Player_Three. Between Experiments 1 and 2 the increase in Progress Towards Goal, Collaborative Actions and Narrative Compliance, and the decrease in Narrative Disruption are statistically significant. Between Experiments 1 and 3 increase in Progress Towards Goal, Collaborative Actions and Narrative Compliance, and decrease in Narrative Disruption are statistically significant. Between Experiments 3 and 4 there are no statistically significant changes.

Change in IQM Between Experiments	Player_Three T-Stats					
	Progress Towards Goal	Not Progress Towards Goal	Collaborative Action	Independent Action	Narrative Disruption	Narrative Compliance
1,2	-4.727	1.260	-2.983	0.8793	4.811	-4.862
1,3	-7.914	-0.5661	-19.05	1.436	4.002	-6.822
3,4	-0.1960	-1.732	-0.7876	-0.9245	0.000	-0.8111

(Experiments 3 to 4). This suggests that iterative prompting and restructuring into the Concordia prompt schematic leads to an increase in performance of Player_One compared to zero-shot prompting in our baseline. We also see overall decreases in Independent Actions, Narrative Disruption and Not Progressing Towards Goal, with the greatest decrease in Narrative Disruption.

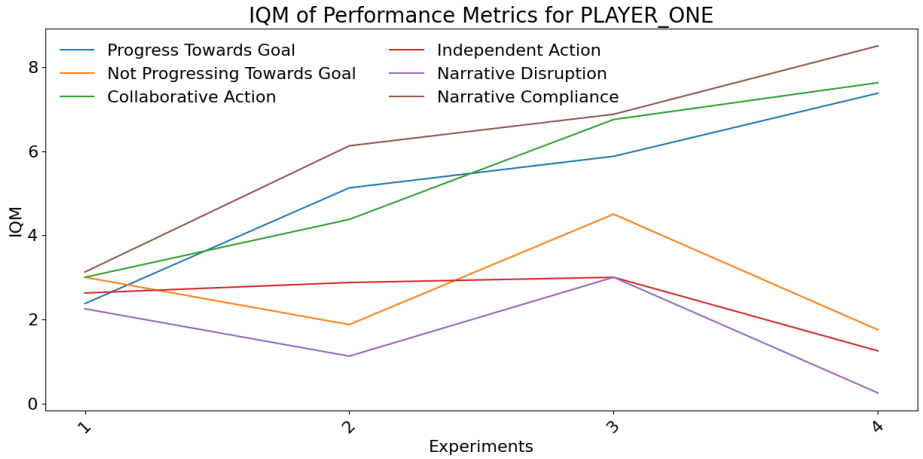

Figure 3: Trends in the IQM of each performance metric across our four experiments for Player_One

In Figure 4 we see that For Progress Towards Goal and Narrative Compliance there is at first a gradual increase (Experiments 1 to 2) then a shallower increase for Progress Towards Goal and a sharper increase for Collaborative Action (Experiments 2 to 3) followed by a plateauing for Collaborative Action and shallow decrease for Progress Towards Goal (Experiments 3 to 4). We also see a continuous sharper increase in Narrative Compliance (Experiments 1 to 3), then a shallower decrease (Experiments 3 to 4). This suggests that iterative prompting and restructuring into the Concordia prompt schematic leads to an increase in performance of Player_Two compared to zero-shot prompting in our baseline, for the Rat Infestation scenario. However, Player_Two does not adapt as well as Player_One to the Missing Otter Child scenario. We also see overall decreases and then plateauing in Independent Actions and Narrative Disruption and a general decrease in Not Progressing Towards Goal, with the greatest decrease in Narrative Disruption.

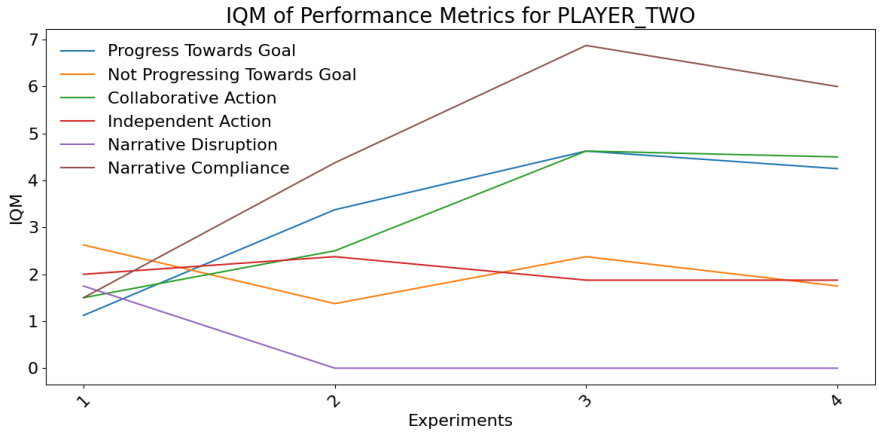

Figure 4: Trends in the IQM of each performance metric across our four experiments for Player_Two

In Figure 5 we see that For Progress Towards Goal and Narrative Compliance there is at first an increase (Experiments 1 to 2) then a shallower increase for Collaborative Actions and a shallower decrease for Progress Towards Goal (Experiments 2 to 3) followed by a plateauing for them both (Experiments 3 to 4). We also see a continuous sharper increase in Narrative Compliance (Experiments 1 to 2), then a shallower decrease (Experiments 2 to 4). This suggests that iterative prompting and restructuring into the Concordia prompt schematic leads to an increase in performance of Player_Three compared to zero-shot prompting in our baseline, for the Rat Infestation scenario. However, repeats using our finalised prompts show slightly worse performance for Player_Three in the Rat Infestation

scenario. Player_Three also does not adapt as well as Player_One to the Missing Otter Child scenario. We see overall decreases in Independent Actions and Narrative Disruption and a general decrease in Not Progressing Towards Goal, with the greatest decrease in Narrative Disruption.

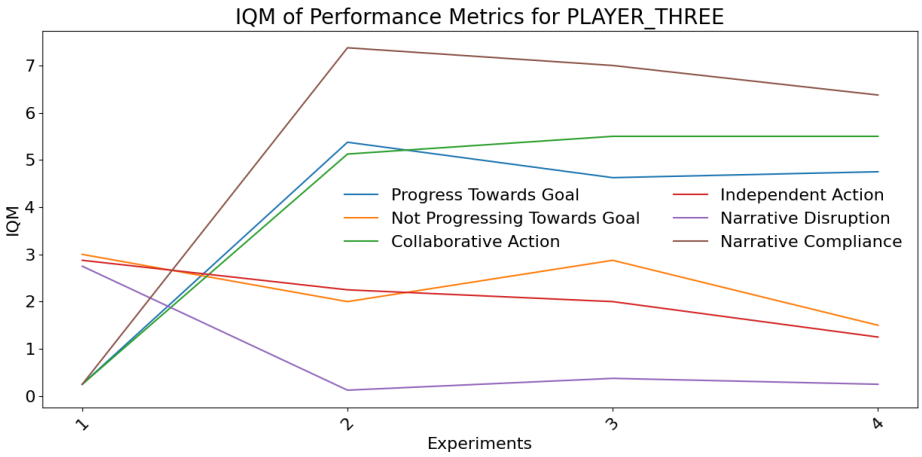

Figure 5: Trends in the IQM of each performance metric across our four experiments for Player_Three