Skill Check: Some Considerations on the Evaluation of Gamemastering Models for Role-playing Games

Santiago Góngora[1] [0000−0002−6256−7577], Luis Chiruzzo[1] [0000−0002−1697−4614], Gonzalo Méndez[2] [0000−0001−7659−1482], and Pablo Gervás[2] [0000−0003−4906−9837]

[1] Universidad de la República, Uruguay {`sgongora,luischir`}`@fing.edu.uy`
[2] Universidad Complutense de Madrid, Spain {`gmendez,pgervas`}`@ucm.es`

Abstract. In tabletop role-playing games a Game Master (GM) is the player in charge of the game, who must design the challenges the players face and narrate the outcomes of their actions. In this work we discuss some challenges to model GMs from an Interactive Narrative and Natural Language Processing perspective. Following those challenges we propose an initial set of categories for unit tests to evaluate these models, and we also use them to test ChatGPT, Bard and OpenAssistant as out-of-the-box GMs.

Keywords: Role-playing Games · Interactive Narrative · Natural Language Processing.

1 Introduction

Probably no one wants to hear somebody say "*Watch out! Behind that door there's a giant monster!*"; except if they are playing a role-playing game (RPG), using their imagination to visit endless worlds and having lots of fun.

Tabletop role-playing games (TTRPGs) consist of two or more players that collaborate in order to create a story. One of these players is the Game Master (GM), who is the one in charge of creating the world where the narrated events take place, describing the non-playable characters the human players meet and the situations they face [35]. Having a player acting as the GM is one of the characteristics that most TTRPGs share (e.g. *Dungeons & Dragons* or *Call of Cthulhu*), since a GM is not only a storyteller, but also a judge and a guide for the rest of the human players.

Capturing the essence of role-playing games has always been one of the goals of Interactive Storytelling research [49]. However, through the years only limited solutions have been found, typically by having tons of modular scenes that can be mixed to generate other narrative structures [3], but pushing the player's freedom aside[4] [46]. To automate a GM is a big challenge for Natural Language Processing (NLP) and Artificial Intelligence (AI), due its complexity on dialogue and

---

[3] For example "Call of Cthulhu: The Official Video Game", an adaptation of the RPG.

[4] An interesting example of this is "The Stanley Parable", a novel videogame that makes the players think about free will and the impact of their actions.

creativity [11]. Also role-playing games are themselves an interesting approach for narrative generation, since a general story can be built from the story of its characters [48].

Our long-term goal is to model the diverse set of skills that a GM needs to play RPGs. Although our main efforts will be focused on developing a solution for Spanish-speaking GMs, we will also work for English in order to contextualise the results. This long path must lead to an explainable, grounded, *thoughtful and critic* (as possible) model, so human-in-the-loop features should be taken into consideration to meet the needs reported by [1] and [24].

In this paper we will take a first step in this long path by proposing a set of unit test categories to evaluate such GM models. Since these categories are inspired in core aspects of role-playing games, they can be used to evaluate any system that models a GM, independently of its features, architecture or technical aspects. We also use these brand new tests to evaluate ChatGPT[5], Bard[6] and OpenAssistant[7] [28] as out-of-the-box automated GMs, both for Spanish and English.

2 Previous work

The study of the role of GMs in narrative is nothing new. [4] discusses some narrative problems when dealing with virtual environments and the role of the GM in role-playing games. [40] use Robin's Laws [29] as a framework for discussing a direction towards an interactive storytelling system. [50] present an in-depth analysis of role-playing games and the concepts involved.

Some efforts have been made to explicitly model the capabilities of a GM. [41] propose an initial model of a GM in order to tackle the *Interactive Dilemma* in Interactive Storytelling, the conflict and balance between the player's will and the designer's choices. Closely related to this concept is the GM's skill to improvise some aspect of a scene due to unexpected players' actions. [32] discuss an architecture for storytelling in open worlds. As it presents an insightful discussion about the improvisational problem, it is one of the most clarifying works for us to date.

Most of the latest works pursue the modelling of GMs for *Dungeons & Dragons* (D&D), called *Dungeon Masters* (DMs), since it is the most popular RPG game and finding data is easier than for other games. For example [11] argues for the complexity of modelling the (D&D) game, performing experiments with neural models and using *control features* to guide their outputs. They also describe a *gameplay dataset* in English used for training. [52] try to create a DM model with the ability to predict player's actions, modelling the *Dungeon Master-Player* interactions using Theory-of-Mind and Reinforcement Learning. [33] details an approach using Reinforcement Learning to model a D&D player.

---

[5] https://openai.com/blog/chatgpt

[6] https://bard.google.com

[7] https://huggingface.co/chat/

The recent published datasets are also centered on modelling the D&D game. [44] present one of the most complete datasets to study D&D interactions, which was possible due to a great effort in transcribing and curating dialogues from the popular *Critical role* web show. [31] also created a dataset of D&D player's interactions from online forums and [53] present a 8M utterances corpus from play-by-posts servers on Discord.

It is important to note that all of these recent works rely on English resources and all of them are for D&D, while our main objective is to work on the general aspects of a Spanish-speaking GM, regardless of the specific game or theme.

3 A list of gamemastering challenges

Most of the works mentioned in the previous section discuss difficulties faced while modelling some aspects of RPGs. However, as a way of introducing some details that guide our long-term goal and justify the test categories we propose, we would like to convey our thoughts on some challenges that GMs must face while running an RPG session. This list is not exhaustive and there may be other challenges that are not described here.

I. World and story design. As storytellers, GMs must manage a fictional but rich world where the players' characters will live and act, generating a rich coherent world, populated with diverse forms of life (e.g. plants, animals, creatures) and characters. They also need to create some interesting places (e.g. an old library) and challenges for the players, which can be logic puzzles, tactic battles, complex dialogues with characters, or other challenges (e.g. the library has hidden rooms). Usually these situations are intended to be solved by teaming up with other characters, collaborating and using the different skills that they may master [35]. It is useful if a GM can also measure how interesting these challenges are for the players, and how meaningful they are for the development of their characters or other characters that live in the fictional world. In the videogames community this is known as *Player Modelling* [12, 49]. *Narrative planning* — the process of creating a story while carefully designing its sequence of events [45, 19] — and what to expect from it is something that human GMs are usually aware of: what they want to make the players (or their characters) feel or do. In Narratology, the procedure used by narrators to choose what to tell and when is known as *focalization* [20], and it is crucial to GMs since this can help to deliver a suspenseful experience to the players [16]. It is also important that such a model can take *creative responsibility* [14, 10] while being able to explain what the plan, reasons and objectives of each utterance are. Finally, if the narrated story takes place in a real setting (e.g. Ancient Greece or Egypt) or uses specific concepts (e.g. Medicine or Mathematics), GM usually have to study to ensure narrative consistency to the players, so such a GM model should widely show *domain knowledge* of the chosen setting [51].

II. Extract player's actions from input. Since TTRPGs are played through a discussion between the players, these games have an inherent *conversational nature.* Therefore classic research problems related to dialogue systems

[13] are fundamental to model GMs, like Natural Language Understanding. Semantic parsing and representation of text is a long established area of research in NLP, which could allow to solve harder language problems [2]. An automated GM could be one of those cases. In order to *understand* the actions taken by the players, decide if they are possible in the fictional world and then determine the outcomes, the GM model should have the ability to semantically analyze their inputs. Also we should consider the context where those player actions are taken, so pragmatics play an important role when understanding what the players want or do not want to do, what they mean and what they do not [21].

III. Commonsense reasoning. Commonsense reasoning is an important research area within NLP [47], and despite the great advances made in the area, it remains as one of the hardest tasks [15, 22]. A recent work by [43] shows that this is one of the tasks in which Large Language Models (LLMs) like ChatGPT cannot achieve a good performance. The relation between this classic task and the challenges for a GM is direct: since commonsense is an inherent part of our human identity, it naturally arises when playing RPGs. This challenge is related but different to the previous Natural Language Understanding challenge; a model can semantically represent what a human is saying, but maybe the action does not make sense in some context. For instance, sometimes players may want to do actions that are possible in the real world but not in the fictional world (e.g. a character wants to play basketball but there is no gravity in her world).

IV. Track the game state. One of the core aspects of RPGs is to let the players act as they wish, what in Interactive Storytelling is usually called *user agency* [46]. Making the players feel this way while thoroughly tracking the state of items and characters is one of the greatest problems for interactive storytelling [8, 9, 32]. To track some component of the game is to know where it is, how hurt (in case of a creature or a character) or damaged it is (in case of an object), and other properties that it may have (e.g. intensity of the magic property of a sacred object). This game state must be constantly updated as the world changes and the story moves forward.

Finally we would like to mention other relevant aspects for this long path. In first place, we think it is crucial that the narrative structure and the game state may be represented using a human-readable format, in order to ensure that people can analyze and customize the game. Since RPG games are used in educational environments [42, 18], different gamemastering models could be used to create exercises or whole games with a wide range of learning objectives. Evidence shows that RPGs are also used in experimental therapy methodologies [37, 3]. Having the possibility of customizing the boundaries of an RPG session is extremely crucial for those of applications. In second place, it is fundamental that these models *act* ethically and respectfully. In modern RPGs like “Alice is Missing” there is a special tool called “the X card” which acts as a safety device for the players; whenever the GM or the players say that make some other player feel hurt or uncomfortable, then the X card can be touched and every one must leave that topic behind. Having this kind of mechanics is extremely important for a safe, responsible and ethic model for a GM. This is also crucial when working with neural or LLMs, which are known to *hallucinate* [25] and generate offensive outputs [6]. Last but not least, we have to keep in mind that GMs are constantly adapting the game to fit the player's choices, so they also have the additional requirement of facing every challenge described here on the fly, as [32] argue with their definition of *open-world improvisational storytelling*.

4 How to evaluate such models?

The procedure to evaluate creative systems (i.e. appropriate experiments and metrics) has long been a subject of debate, and remains one of the main problems of the field [27, 26, 14, 39]. Since TTRPGs can be modelled as a series of utterances in a complex dialogue [36, 34, 17], we will assume that a GM model will always have a conversational nature, as we mentioned in *challenge II*. This gives us a general guideline: there is always a player who is asking or trying to do something, and another player answering or reacting to it.

The first idea that comes to mind could be to ask humans to play and evaluate the models based on their reaction. Although we consider important to measure how fun it is to play with the models, the humans' judgments can be very subjective, not very specific, and also biased by the fluency of the generated text [5]. This is related to the fact that nowadays LLMs are particularly good at producing strings that sound very natural to the human reader since they are trained to exploit the patterns behind the form of massive amounts of texts [7], what can lead to distract the evaluators from their goal of judging the gamemastering capabilities of a model.

Hence we would like to take an approach on evaluating basic, almost essential, skills that a GM should master. We propose three different test categories related to the previously described challenges: *commonsense reasoning*, the capacity to track *items* in the world and the ability to coherently design *maps*. These categories were designed reflecting on core characteristics of RPGs, so we think they can be used to evaluate any system trying to model a GM, independently of the theme and features of the modelled game, and if they are modal or multi modal [30]. We also hope these categories work as a guide for human evaluators, helping them to judge models with logical decisions and therefore reducing the subjectivity, the mentioned biases and the evaluation noise as possible. We will describe each of them next.

4.1 GM-P-GM pattern

In *challenges II* and *III* we discussed the importance of pragmatics and commonsense reasoning for a GM model. In order to evaluate the performance on this challenges we propose the GM-P-GM pattern, a formalization of the most elemental interaction between a GM and a player, inspired in [32] and RPG rulebooks [36, 34, 17].

According to [32] formalization, an *exceptional branch* ocurrs when "the user's action causes the world to enter a state in which no successor in the plot graph can be executed because one or more preconditions of each successor plot point is contradicted". As a particular case of this exception, we propose to evaluate the model's ability to judge the feasibility of a player's action:

- *GM_1*: Narrates a **situation** to solve in some **context**.
- *Player*: Describes the **actions** to overcome that **situation**.
- *GM_2*: Validates if those **actions** are feasible for that **context**, and next narrates the **outcomes**.

To run this test we give the model the GM_1 and *Player* utterances and ask it to generate the GM_2 utterance. If the narrator prevents the action and explains why it is an inconsistency, the test is passed. A failure case is shown in table 1.

GM	You enter the building to look for the lost suitcase. However, the door is covered in flames. To your right you see a fire extinguisher, which seems to be ready to use. You can't use anything from your backpack because you can hardly move your arms. What will you do?
Player	I take a bucket of water from my backpack and put out the fire
GM	As the player grabs the bucket of water from their backpack, they realize it's not nearly enough to put out the flames completely. However, they quickly pour the water onto the fire, creating a small opening in the flames (...)
Result	Failed

Table 1. An example of a *GM-P-GM* unit test.

4.2 Item tracking

As we described in *challenge IV*, item tracking is one of the fundamental problems for gamemastering. In fact there is a popular concept for the items that each character has, and it is called **inventory**. To avoid the use of specific RPG words, we use *bag* or *backpack* as more neutral words for a similar concept.

To test if the model can track objects inside or outside the bag, we follow this schema:

1. *Player*: "Make a list of the items I have in my bag"
2. *GM*: $Object_1, Object_2, Object_3, ...$
3. *Player*: "I take $Item_K$ and leave it on *another location*"
4. *GM*: *Narrates outcome*
5. *Player*: "Make a list of the items I have in my bag"
6. *GM*: $Object_1, Object_2, Object_3, ...$
7. *Player*: "What's on *another location* ?"
8. *GM*: $Object_1, Object_2, Object_3, ...$

If the lists given in steps #2 and #6 are the same or contain other unrelated objects, then the test is failed. In case they are not the same, the step #8 tries to check if the model tracks the object to its new location. A failure case is shown in table 2.

GM	(...) Aside from the mysterious key, your backpack also contains the usual adventurer's equipment: a sturdy rope, a grappling hook, a water flask, some dried rations, a small first-aid kit, a compass, a map of the surrounding area, and a trusty dagger at your side.
Player	I take the rope, the map and the key and I leave them on the floor
GM	You carefully set the rope, map, and key down on the floor, making sure they are within reach in case you need them later. (...)
Player	What's in my backpack?
GM	As you take a closer look at your backpack, you notice that it contains a few essential items that any adventurer would find useful: A sturdy rope: (...) this rope is perfect for climbing or rappelling down steep surfaces. (...) A map of the surrounding area: this map shows the nearby landmarks, towns, and topography of the region, allowing you to plan your route. (...)
Result	Failed

Table 2. An example of an *Item tracking* unit test.

GM	(...) Current place: You find yourself standing at the entrance of a dark and damp cave. The entrance is surrounded by tall trees, and the sound of running water can be heard in the distance. Accessible places: Inside the cave, Back to the forest, Along the river
Player	I want to explore the cave
GM	(...) Current place: Inside the cave Accessible places: The cave entrance, Deeper into the cave
Player	I want to go outside
GM	(...) Current place: Outside the cave Accessible places: Inside the cave, Along the river, Deeper into the forest
Result	Passed

Table 3. An example of a *Map design* unit test.

4.3 Map design

When GMs (or game designers generally) create a fictional world they also have the challenge to design an interesting and coherent concrete map for the players to explore [23], as we described in *challenge I*. Although we also argued the importance of having interesting situations for the players, for now we do not focus on how attractive, large or complex the maps generated are.

Inspired in the schema of the *Item tracking* category, we propose to ask the model to describe the accessible locations from the current location. Then the player must choose to move to one of those locations and go back again to the original location, in order to test the map's coherence. If during this process the available locations are coherent, the test is passed. A successful case is shown in table 3.

5 Large Language Models as out-of-the-box GMs

In order to perform some experiments with our three proposed categories, we wrote five unit tests in Spanish and five in English for each of them. We then used those unit tests to evaluate the out-of-the-box gamemastering skills of ChatGPT[8], Bard[9] and also of the LLaMA-based OpenAssistant chatbot[10] for means

---

[8] "ChatGPT Mar 14 Version. Free Research Preview"

[9] Bard Experimental. Accessed on the 17th of July, 2023.

[10] "Model: OpenAssistant/oasst-sft-6-llama-30b". Accesed on the 14th of July, 2023.

of comparison. We consider these three models appropriate since they are dialogue systems i.e. they have a *conversational nature*, a fundamental requirement for a GM model.

Since these are only preliminary experiments we consider really important to make the experimental logs open, because it can help the reader to critically examine the results reported here and reflect on the real flaws and strengths of both our proposed test categories and the evaluated models. Therefore the detailed logs of the experiments and comments about why we consider that they were successful or not, are accessible on GitHub[11].

We detail the analysis in two subsections, the quantitative results and the qualitative observations.

5.1 Quantitative results

After we ran the tests using the aforementioned models, we carefully examined the outputs and determined the results for each test. They are shown in table 4.

Category	**OA** [ES]	**BARD** [ES]	**CGPT** [ES]	**OA** [EN]	**BARD** [EN]	**CGPT** [EN]
GM-P-GM	0 out of 5	2 out of 5	1 out of 5	1 out of 5	1 out of 5	0 out of 5
Items	0 out of 5	0 out of 5	2 out of 5	0 out of 5	3 out of 5	1 out of 5
Map	0 out of 5	3 out of 5	2 out of 5	0 out of 5	3 out of 5	3 out of 5
Total	0 out of 15	5 out of 15	5 out of 15	1 out of 15	7 out of 15	4 out of 15

Table 4. Number of passed tests for each of the categories described in section 4, testing OpenAssistant (OA), Google's Bard and ChatGPT (CGPT), both for English and Spanish. The last row shows the sum of the passed tests for each model.

As can be seen, the performance on the *GM-P-GM* category is really low, regardless of the language or model. This result is aligned with those in [43], where commonsense reasoning was one of the remarkable flaws of ChatGPT. However, *Item tracking* and *Map design* tests were quite good both for ChatGPT and Bard.

Although these preliminary experiments do not report a big gap in the results for ChatGPT or Bard between languages, they do unveil their strength over OpenAssistant. In most cases OpenAssistant just could not finish the test, generating nonsensical outputs that had nothing to do with the narrative situation. That problem was even deeper when performing the tests in Spanish.

5.2 Qualitative observations

The first and most important observation is that, at first glance, ChatGPT and Bard are really good at making the user feel that is playing with a real GM.

---

[11] Only some representative examples are available during review: https://github.com/sgongora27/skill-check-GM-tests

There is a world to interact with, characters to meet and items to use. Everything seems perfect if the player chooses an action from those suggested by the model, although it is not that perfect when having to improvise new scenes and keep it coherent. OpenAssistant, however, struggles to deliver a minimum interactive experience and the tests had to be repeated several times to obtain a reasonable output. Our evaluation schema does not distinguish that kind of errors, hence this aspect cannot be inferred by just comparing the quantitative results for each category between models (e.g. ChatGPT failed the Spanish *Item tracking* tests due to wrongly list the available items, while OpenAssistant failed them because could not even give a proper output). However we think that the quantitative results do represent the strengths and weaknesses of each model (e.g. ChatGPT is better with world coherence than with commonsense reasoning) but also the "Total" scores provide an accurate comparison of the experience provided by the different models.

The second observation is about the contents generated by the models. Unlike what happens to the GM-P-GM test, in the *Item tracking* and *Map design* tests the original prompt did not include any topic, scene or context, so it was the model itself who had to take the *creative responsibility.* Almost every scene *created* by the models took place in a medieval-fantasy setting, and the starting locations lacked diversity. The relation between RPGs and a medieval setting is aligned to what we mentioned in section 2 that most of the available data about RPGs is in fact about *Dungeons & Dragons.* More generally, this also reinforces the ideas of [6] about the relation between the performance of LLMs and the data used to train them. As evidenced by these results, more work on other role-playing games with different settings is needed. Also there is a related flaw regarding the absence of diversity of plots, since after playing a bunch of hours the narrative situations start to repeat. Although this is related to the previous comment about the settings and the biased data, it is important to have in mind that a great diversity of plots can be created using a medieval-fantasy setting[12], so they are independent flaws and might be studied separately.

Our third observation is about these models' tendency to constantly adjust the output to the prompt. If the players asks a question about some thing of the world, then the output will try to describe it without letting the player to perceive any mystery about the plot. The same happens if the player asks to do something that was not fully described by the model (e.g. leave an item on a table that was not mentioned in the description of the room); the model probably will immediately adapt its output to fit that object, room or character. This is not a good sign for the information management in narrative, one of the skills that a GM should have as we described in *challenge I.*

5.3 Limitations

Although we propose the test categories to assist the evaluation of GM models, the human subjectivity is still there. Even in the *Item tracking* and the *Map*

---

[12] This is evidenced by the massive amount of adventures published for RPGs with this theme, such as *Dungeons & Dragons* or *Pathfinder.*

design tests, that seem to be free from ambiguity since they only need the evaluator to check a list of items or places, there can be some degrees of correctness. In addition to the success/failure decision this subjectivity can also be present in the prompts design as well, as in the case of the *GM-P-GM* tests which need a specific human-designed case in order to run (i.e. a situation to solve and a player's solution to it).

It is important to highlight that the difficulties faced when evaluating a creative system added to the nearly-infinite input space that an RPG offers make the evaluation even harder. Not having a symbolic representation to guide the model also makes the execution of these tests more unpredictable, making the evaluator to take unexpected decisions on the fly. It would be positive for the *Map design* tests to have this kind of symbolic representation available in order to perform an in-depth analysis of the different reachable places in a given scene but without moving the story forward. Also these dialogue models compute the utterances each time a new input is sent, what makes the replication experiments harder. Additionally we share the same limitations found by [43] regarding the needed time to run a small set of tests.

Finally, to perform a deeper evaluation and extract stronger conclusions we would need a diverse team of human evaluators and a bigger number of tests. We invite the reader to access the experimental logs in order to evaluate by themselves the results reported here, and to try other tests for these categories and compare the results.

6 Conclusions and future work

In this paper we detailed some aspects to model automated Game Masters. We discussed some challenges to face, regarding gameplay requirements and other safety features, like the option to set boundaries for the game.

We proposed three basic categories for unit tests in order to evaluate any kind of game mastering model while reducing the impact of the evaluator's subjectivity. In the future we would like to improve them and design more categories (e.g. test the emotional variation of a character during an interaction with other character [38]).

We also performed preliminary experiments with ChatGPT, Bard and OpenAssistant using those three categories. We found that although ChatGPT and Bard can provide a satisfying gaming experience, they struggle when dealing with common-sense reasoning. OppenAssistant was unable to mantain the gamemastering role during the experiment. The difficulties faced to control their outputs while running the tests make us think that in the future more hybrid (e.g. *neuro-symbolic*) approaches should be explored. We think that having symbolic representations would help to keep the test phase more controllable, allow the players to examine the narrative details, avoid some scenes that they do not want to play and add another elements that they do.

References

1. Acharya, D., Mateas, M., Wardrip-Fruin, N.: Story improvisation in tabletop roleplaying games: Towards a computational assistant for game masters. In: 2021 IEEE Conference on Games (CoG). pp. 01–08 (2021). https://doi.org/10.1109/CoG52621.2021.9619006
2. Allen, J.: Natural language understanding. Benjamin-Cummings Publishing Co., Inc. (1995)
3. Arenas, D.L., Viduani, A., Araujo, R.B.: Therapeutic use of role-playing game (rpg) in mental health: A scoping review. Simulation & Gaming **53**(3), 285–311 (2022)
4. Aylett, R.: Narrative in virtual environments -towards emergent narrative. In: Proceedings of the AAAI fall symposium on narrative intelligence. pp. 83–86 (01 1999)
5. Behnke, H., Fomicheva, M., Specia, L.: Bias mitigation in machine translation quality estimation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1475–1487. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.104, https://aclanthology.org/2022.acl-long.104
6. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 610–623. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445922, https://doi.org/10.1145/3442188.3445922
7. Bender, E.M., Koller, A.: Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5185–5198. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.463, https://aclanthology.org/2020.acl-main.463
8. Benotti, L.: Clarification potential of instructions. In: SIGDIAL Conference (2009)
9. Benotti, L.: Implicature as an Interactive Process. (L'implicature comme un Processus Interactif). Ph.D. thesis, Henri Poincaré University, Nancy, France (2010), https://tel.archives-ouvertes.fr/tel-00541571
10. Botelho, L.M.: A guided journey through non-interactive automatic story generation. CoRR **abs/2110.11167** (2021), https://arxiv.org/abs/2110.11167
11. Callison-Burch, C., Tomar, G.S., Martin, L., Ippolito, D., Bailis, S., Reitter, D.: Dungeons and dragons as a dialog challenge for artificial intelligence. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 9379–9393. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.emnlp-main.637
12. Charles, D., McNeill, M., McAlister, M., Black, M., Moore, A., Stringer, K., Kücklich, J., Kerr, A.: Player-centred game design: Player modelling and adaptive digital games. In: Proceedings of DiGRA 2005 Conference: Changing Views ? Worlds in Play, pp. 285–298. Digital Games Research Association: DiGRA (2005), https://mural.maynoothuniversity.ie/12737/
13. Chen, H., Liu, X., Yin, D., Tang, J.: A survey on dialogue systems: Recent advances and new frontiers. SIGKDD Explor. Newsl. **19**(2), 25–35 (nov 2017). https://doi.org/10.1145/3166054.3166058, https://doi.org/10.1145/3166054.3166058

14. Colton, S., Wiggins, G.A.: Computational creativity: The final frontier? In: European Conference on Artificial Intelligence (2012)
15. Davis, E., Marcus, G.: Commonsense reasoning and commonsense knowledge in artificial intelligence. Communications of the ACM **58**(9), 92–103 (2015)
16. Delatorre, P., León, C., Salguero, A., Palomo-Duarte, M., Gervás, P.: Information management in interactive and non-interactive suspenseful storytelling. Connection Science **31**(1), 82–101 (2019). https://doi.org/10.1080/09540091.2018.1454890, https://doi.org/10.1080/09540091.2018.1454890
17. Durall, J., Perrin, S.: Basic Roleplaying: Universal Game Engine. Chaosium Inc. (2023)
18. Gatsakou, C., Bardis, N., Drigas, A.: Role playing vs rpgs as teaching strategies in educational procedure. Technium Soc. Sci. J. **26**, 186 (2021)
19. Gervás, P., Concepción, E., León, C., Méndez, G., Delatorre, P.: The long path to narrative generation. IBM Journal of Research and Development **63**(1), 8:1–8:10 (2019). https://doi.org/10.1147/JRD.2019.2896157
20. Gervás, P.: Computational approaches to storytelling and creativity. AI Magazine **30**, 49–62 (2009)
21. Green, G.M.: Pragmatics and natural language understanding. Psychology Press (1996)
22. Gunning, D.: Machine common sense concept paper. CoRR **abs/1810.07528** (2018), http://arxiv.org/abs/1810.07528
23. Hendrikx, M., Meijer, S., Van Der Velden, J., Iosup, A.: Procedural content generation for games: A survey. ACM Trans. Multimedia Comput. Commun. Appl. **9**(1) (feb 2013). https://doi.org/10.1145/2422956.2422957, https://doi.org/10.1145/2422956.2422957
24. Inie, N., Falk, J., Tanimoto, S.: Designing participatory ai: Creative professionals' worries and expectations about generative ai (2023)
25. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Comput. Surv. **55**(12) (mar 2023). https://doi.org/10.1145/3571730, https://doi.org/10.1145/3571730
26. Jordanous, A.: A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cogn. Comput. **4**(3), 246–279 (2012). https://doi.org/10.1007/s12559-012-9156-1, https://doi.org/10.1007/s12559-012-9156-1
27. Jordanous, A.: Should we pursue SOTA in computational creativity? In: Hedblom, M.M., Kantosalo, A.A., Confalonieri, R., Kutz, O., Veale, T. (eds.) Proceedings of the 13th International Conference on Computational Creativity, Bozen-Bolzano, Italy, June 27 - July 1, 2022. pp. 159–163. Association for Computational Creativity (ACC) (2022), http://computationalcreativity.net/iccc22/papers/ICCC-2022_paper_134.pdf
28. Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.R., Stevens, K., Barhoum, A., Duc, N.M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., Mattick, A.: Openassistant conversations – democratizing large language model alignment (2023)
29. Laws, R., Jackson, S., Bergting, P.: Robin's Laws of Good Game Mastering. Steve Jackson Games, Incorporated (2002), https://books.google.com.uy/books?id=9MANAAAACAAJ
30. Liang, P.P., Zadeh, A., Morency, L.P.: Foundations and trends in multimodal machine learning: Principles, challenges, and open questions (2023)

31. Louis, A., Sutton, C.: Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 708–713. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2111, https://aclanthology.org/N18-2111
32. Martin, L.J., Harrison, B., Riedl, M.O.: Improvisational computational storytelling in open worlds. In: Nack, F., Gordon, A.S. (eds.) Interactive Storytelling. pp. 73–84. Springer International Publishing, Cham (2016)
33. Martin, L.J., Sood, S., Riedl, M.O.: Dungeons and dqns: Toward reinforcement learning agents that play tabletop roleplaying games. In: INT/WICED@AIIDE (2018)
34. Mason, M., Fricker, P., Petersen, S., Willis, L.: Call of Cthulhu: Quick-Start Rules. Chaosium Inc. (2021)
35. Mearls, M., Crawford, J.: Dungeon Master's Guide (5th Edition). Wizards of the Coast (2014)
36. Mearls, M., Crawford, J., Perkins, C.: D&D Basic Rules, Version 1.0. Wizards of the Coast (2018)
37. Mendoza, J.: Gaming intentionally: A literature review of the viability of role-playing games as drama-therapy-informed interventions (2020)
38. Oñate, A., Méndez, G., Gervás, P.: Emolift: Elevator conversations based on emotions. In: International Conference on Innovative Computing and Cloud Computing (2019)
39. Peinado, F., Francisco, V., Hervás, R., Gervás, P.: Assessing the novelty of computer-generated narratives using empirical metrics. Minds and Machines **20**, 565–588 (11 2010). https://doi.org/10.1007/s11023-010-9209-8
40. Peinado, F., Gervás, P.: Transferring game mastering laws to interactive digital storytelling. In: Göbel, S., Spierling, U., Hoffmann, A., Iurgel, I., Schneider, O., Dechau, J., Feix, A. (eds.) Technologies for Interactive Digital Storytelling and Entertainment. pp. 48–54. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
41. Peinado, F., Gervás, P.: Automatic direction of interactive storytelling: Formalizing the game master paradigm. In: Cavazza, M., Donikian, S. (eds.) Virtual Storytelling. Using Virtual Reality Technologies for Storytelling. pp. 196–201. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
42. Grande de Prado, M., et al.: Los juegos de rol en el aula (2010)
43. Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., Yang, D.: Is chatgpt a general-purpose natural language processing task solver? (2023). https://doi.org/10.48550/ARXIV.2302.06476, https://arxiv.org/abs/2302.06476
44. Rameshkumar, R., Bailey, P.: Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5121–5134. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.459, https://aclanthology.org/2020.acl-main.459
45. Riedl, M.O., Young, R.M.: Narrative planning: Balancing plot and character. J. Artif. Int. Res. **39**(1), 217–268 (sep 2010)
46. Riedl, M.O., Bulitko, V.: Interactive narrative: An intelligent systems approach. AI Magazine **34**(1), 67 (Dec 2012). https://doi.org/10.1609/aimag.v34i1.2449, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2449
47. Sap, M., Shwartz, V., Bosselut, A., Choi, Y., Roth, D.: Commonsense reasoning for natural language processing. In: Proceedings of the

58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. pp. 27–33. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-tutorials.7, https://aclanthology.org/2020.acl-tutorials.7

48. Tapscott, A., León, C., Gervás, P.: Generating stories using role-playing games and simulated human-like conversations. In: Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018). pp. 34–42. Association for Computational Linguistics, Tilburg, the Netherlands (Nov 2018). https://doi.org/10.18653/v1/W18-6606, https://aclanthology.org/W18-6606
49. Thue, D., Bulitko, V., Spetch, M.L., Wasylishen, E.: Interactive storytelling: A player modelling approach. In: Artificial Intelligence and Interactive Digital Entertainment Conference (2007)
50. Tychsen, A., Hitchens, M.: The many faces of role-playing games. International journal of role-playing **1**(1), 3–21 (2009)
51. Ward, T.B.: The role of domain knowledge in creative generation. Learning and Individual Differences **18**(4), 363–366 (2008). https://doi.org/https://doi.org/10.1016/j.lindif.2007.07.002, https://www.sciencedirect.com/science/article/pii/S1041608007000842, including Special Issue on Creativity
52. Zhou, P., Zhu, A., Hu, J., Pujara, J., Ren, X., Callison-Burch, C., Choi, Y., Ammanabrolu, P.: An ai dungeon master's guide: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons (2022). https://doi.org/10.48550/ARXIV.2212.10060, https://arxiv.org/abs/2212.10060
53. Zhu, A., Aggarwal, K., Feng, A., Martin, L., Callison-Burch, C.: FIREBALL: A dataset of dungeons and dragons actual-play with structured game state information. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4171–4193. Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://aclanthology.org/2023.acl-long.229