Generating Role-Playing Game Quests With GPT Language Models

Susanna Värtinen, Perttu Hämäläinen, and Christian Guckelsberger

***Abstract*—Quests represent an integral part of role-playing games (RPGs). While evocative, narrative-rich quests are still mostly hand-authored, player demands toward more and richer game content, as well as business requirements for continuous player engagement necessitate alternative, procedural quest generation methods. While existing methods produce mostly uninteresting, mechanical quest descriptions, recent advances in AI have brought forth generative language models with promising computational storytelling capabilities. We leverage two of the most successful transformer models, 1) GPT-2 and 2) GPT-3, to procedurally generate RPG video game quest descriptions. We gathered, processed, and openly published a dataset of 978 quests and their descriptions from six RPGs. We fine-tuned GPT-2 on this dataset with a range of optimizations informed by several ministudies. We validated the resulting Quest-GPT-2 model via an online user study involving 349 RPG players. Our results indicate that one in five quest descriptions would be deemed acceptable by a human critic, yet the variation in quality across individual quests is large. We provide recommendations on current applications of Quest-GPT-2. This is complemented by case-studies on GPT-3 to highlight the future potential of state-of-the-art natural language models for quest generation.**

***Index Terms*—Artificial intelligence, computational storytelling, games, generative models, procedural content generation, quests.**

I. Introduction

QUESTS in role-playing games (RPGs) represent explicitly posed, challenging tasks for the player to accomplish. Main quests are vital to progressing in a game while side quests can yield auxiliary rewards to the player. Quests are often narrative-driven and woven into a game's larger story line. At present, most such quests are written by people.

However, players' growing demand for more game content, e.g., in dynamic and open-ended games [1], poses a challenge to human quest designers on both the developer and community side: Writing a large number of quests that are meaningful and of sufficient quality to warrant continuous player engagement requires time, skill, and creativity. To alleviate the quest creation task, designers could either draw inspiration from, or cocreate [2], computationally generated quests and the narratives that communicate their objectives, i.e., quest descriptions. Autonomous computational quest generation methods could moreover enable quests that adapt online to a player's actions, including user-generated content. Next to these practical concerns, we deem it a fascinating scientific question whether high-quality quests can be generated by procedural means.

Existing approaches to procedurally generate quests and their descriptions are lacking as their products are often formulaic and repetitive. Meanwhile, AI research has brought forth novel text-generating language models with powerful computational storytelling capabilities. Arguably the most prominent such model at present is OpenAI's *Generative Pretrained Transformer* (GPT), which has been leveraged to produce various types of realistic, humanlike texts with unprecedented quality, from poetry to fictional news [3], [4], [5].

This article investigates the potential of GPT-2 and GPT-3, the latest two models in the GPT family, to automatically generate quest descriptions for RPGs. By quest descriptions, we denote short texts that explain the quest to the player from the perspective of a quest-giving nonplayable character (NPC). We thus focus on one building block of a larger pipeline, preceded by, e.g., a dynamic quest ingredient generator accounting for the narrative and gameplay context, a dialog generator for the quest giver, and a game logic generator linking the quest's progression to game events and objects.

GPT-3 has more than 100 times more parameters than its predecessor GPT-2 but it cannot be trained or sampled on hardware that players and game studios typically have access to. In this work, we hence focus on fine-tuning GPT-2, based on a custom-made RPG quest description dataset. We have validated the resulting Quest-GPT-2 model both objectively, with training and validation loss as well as conditional perplexity scores, and subjectively via an online user study. To provide indications for the future potential of text generation models, we complement these fine-tuning experiments with case-studies

Manuscript received 20 December 2021; revised 11 July 2022 and 28 October 2022; accepted 8 December 2022. Date of publication 12 December 2022; date of current version 19 March 2024. The work of C. Guckelsberger was supported by the Academy of Finland (AoF) flagship program "Finnish Center for Artificial Intelligence" (FCAI). *(Corresponding author: Christian Guckelsberger.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was implicitly provided based on the guidelines of the Finnish National Board on Research Integrity (TENK).

Susanna Värtinen and Perttu Hämäläinen are with the Department of Computer Science, Aalto University, 02150 Espoo, Finland (e-mail: susanna.vartinen@aalto.fi; perttu.hamalainen@aalto.fi).

Christian Guckelsberger is with the Department of Computer Science, Aalto University, 02150 Espoo, Finland, with the Finnish Center for Artificial Intelligence, FI-00076 Aalto, Finland, and also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NSE1, U.K. (e-mail: christian.guckelsberger@aalto.fi).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TG.2022.3228480.

Digital Object Identifier 10.1109/TG.2022.3228480

on generating quest descriptions with the vanilla GPT-3 model. Our contributions are threefold.

1) A novel and publicly available quest dataset with 978 quests and descriptions from six RPG games.
2) Quest-GPT-2, a fine-tuned variant of GPT-2 to generate RPG quest descriptions, provided the quest as input. The model has been evaluated in a comprehensive user study, involving 349 participants and 500 quest descriptions.
3) A comparison of different language model fine-tuning text formatting techniques, including the use of placeholders for proper nouns and numbers [6] to reduce variance in the Transformer model fine-tuning.

We have made our quest dataset publicly available[1] for use in other creative applications and to support the development of next-generation procedural quest systems.

II. Related Work

Procedural quest generation is a long-lasting challenge in game AI, with related work dating back more than 15 years [7]. We provide a brief, incomplete overview of related work, focusing on the generation techniques and main shortcomings.

Early research on procedural quest generation focused on planning and rule-based approaches. Ware and Young [8] made an interactive narrative adventure game *The Best Laid Plans* that utilizes computational models of intentionality and conflict in controlling its NPCs. Thue et al. [9] have built an interactive storytelling system, Player-Specific Stories via Automatically Generated Events (*PaSSAGE*), which uses player modeling to automatically determine players' preferred styles of play. Si et al. [10] have presented *Thespian*, a framework for creating interactive drama from user-modifiable agents, i.e., characters with different personality styles and action policies.

Some authors have also attempted more emergent, dynamic quest generation methods. McCoy et al. [11] developed the award-winning social puzzle game *Prom Week* that utilizes a "social physics" engine named *Comme il Faut* (CiF). CiF uses character traits, relationships, and desires to influence player–NPC interactions while also utilizing thousands of preprogrammed sociocultural considerations. Guimaraes et al. [12] implemented CiF into the popular RPG *The Elder Scrolls V: Skyrim* [13] as a freely downloadable modification. Many existing quest generation algorithms construct quests based on graphs. Kybartas and Verbrugge [14] used narrative graph rewriting in their REwriting Graphs for Enhanced Narratives (*ReGEN*) system to create complex branching stories. Calvin and Michael [15] leveraged graphs to generate quests for key and lock puzzles in their experimental game *Charbitat*, Pita et al. [16] created dynamically linked quests in persistent multiplayer worlds, and Stocker and Alvin [17] generated nonlinear quests based on implementation-specific rules and natural language. Doran and Parberry [18] analyzed 750 quests from four popular RPGs to identify a common structure to be leveraged in their prototype quest generator through context-free grammars. The latter has been further expanded by Breault et al. [19] in their Creation Of Novel Adventure Narrative (*CONAN*) system. Soares de Lima et al. [20] combine automated planning with evolutionary search guided by story arcs. We note two main shortcomings in the above body of related work. First, the used techniques produce formulaic and repetitive quests and do not generalize well to other games and genres. Second, the generated quests have only been evaluated against computational metrics and quests from existing games but not against players' experiences.

Recent work has overcome these shortcomings through the use of language models for quest generation and user studies for their evaluation. Ammanabrolu et al. [21] fine-tuned GPT-2 for creating quests in the form of cooking instructions in a text-based cooking game. Based on a small user study with 75 participants, they found that the GPT-2 quests were experienced as more valuable and coherent but less surprising and novel than quests produced by random assignment or Markov chains. Most closely related to our work, van Stegeren and Myśliwiec [33] have recently fine-tuned GPT-2 for the generation of quest descriptions told from the perspective of an NPC. Crucially though, they solely use data from the Massively Multiplayer Online RPG (MMORPG) *World of Warcraft* [23]. This is problematic in that such a homogeneous dataset reduces the generalizability of the generator, as supported by the study's authors. Moreover, while MMORPGs contain tens of thousands of quests and thus represent an easy data source, the quests are typically simpler in structure and less varied than their RPG counterparts: Rather than functioning as vehicles for role-playing or captivating story-heavy adventures, they often provide mere busywork for player character progression. Unsurprisingly, their model input only consists of the quest title and objective. Our approach affords more control for integration in a specific game by incorporating more differentiated and essential input information such as the quest-giver, location, involved characters, and quest reward. Van Stegeren and Myśliwiec's user study motivates our use of GPT-2 for quest generation, in that at least some generated descriptions scored higher than user's ratings for human authored texts. This finding must however be taken with a grain of salt, as their study only involved 20 quest descriptions rated by 32 participants, and each corresponding to exactly one quest. Our study in contrast involved 349 participants, rating a total of 500 quest descriptions generated from 50 quests from six RPGs. Our study is thus not only more representative but also allowed us to investigate quality variations in quest descriptions produced from the same quest input.

III. Language Models and the GPT Family

Language modeling and generation has a long history in AI and computational creativity research [24], [25], [26]. Typically, text generation is approached statistically as sampling each *token*—a character, word, or word part—conditional on previous tokens, $c_i \sim p(c_i|c_1 \ldots c_{i-1}; \theta)$, where c_i denotes the ith token in the text sequence, and θ denotes the parameters of the sampling distribution. In this statistical view, the modeling/learning task amounts to optimizing θ based on training data, e.g., to maximize the probabilities of all tokens in the training data conditional on up to N preceding tokens, where N is the *context*

[1] [Online]. Available: https://doi.org/10.17605/OSF.IO/JTQDB

size. Modern language models use deep neural networks to learn the regularities in the data, and θ become the parameters of the network. For the text generation/sampling task, such a neural network takes in a sequence of tokens and outputs the sampling probabilities of each possible next token. There is ample empirical evidence that large enough neural language models can reach beyond memorizing their input and exhibit remarkable creative and intelligent behavior, e.g., in handling novel concepts not included in the training data and only introduced in the prompt.

The GPT model family is based on the transformer neural network architecture introduced in 2017 [4], [27], which is characterized by encoder and decoder blocks as well as a self-attention mechanism. Encoder blocks transforms variable length input data into fixed-sized feature maps, whereas decoder blocks attempt to transform the maps back into the assumed input. The self-attention mechanism relates each input word to each other to establish links between related words, such as names and pronouns, modulating which previous tokens influence each generated token. Transformer models have been proven capable in a wide range of challenging tasks, e.g., generating music and images [28], [29], synthesizing proteins with desired properties [30], and logical and counterfactual reasoning with facts and rules defined using natural language [31]. Most relevant here, they have been shown to produce realistic, humanlike text with unprecedented quality [3], [4], [5].

GPT models are trained with a diverse collection of unlabeled textual data and, optionally, fine-tuned with a small set of task-specific labeled training data. The pretraining allows to encode a large amount of common knowledge and learn long-range dependencies between tokens but fine-tuning has been shown to improve performance on specific tasks considerably [32]. The different models in the GPT family not only differ from each other in terms of the used training data but also notably in scale: GPT-2 has 10 times more parameters than GPT-1, whereas GPT-3 has over one hundred times more parameters than GPT-2 [3], [3], [4]. Training and sampling GPT-3 is at present not possible on the hardware that players and game studios typically have access to. In the rest of this article, we consequently focus on fine-tuning GPT-2, and only use the vanilla GPT-3 model for comparative case-studies on the enhanced capabilities of this more complex model generation.

IV. Training Dataset

We adopt the hypothesis from related work [33] that the data used to pretrain GPT-2 does not contain a sufficient amount of quest examples to facilitate high-quality quest generation without additional fine-tuning based on a separate, specialized dataset. We confirmed this hypothesis by investigating the output of the vanilla GPT-2 model with 744 M parameters, if presented with different quests (cf. Section V). Unfortunately, most of the quest datasets used in previous related work have not been made public, a state of affairs, which is discussed more widely by van Stegeren and Theune [34]. We consequently collected, processed, and published[1] a dataset of 978 quests and quest descriptions from six RPGs to fine-tune GPT-2, and for others to adopt and potentially extend in their projects.

A. Collecting Data

Fine-tuning a language model can require a few thousand examples to produce good results, depending on the task and model size. For instance, GPT-2-774 M has been shown to require around 5000 text samples, when fine-tuning the model for text continuation tasks [35]. Video game descriptions are typically longer than these text samples and we consequently assumed that a dataset of roughly 1000 quests and quest descriptions would suffice for fine-tuning Quest-GPT-2. This is also supported by the observation that GPT variants with more parameters, such as our target model GPT-2-1.5B, are better at learning patterns from few examples [4].

Hand-authoring this amount of quest data for our study would have been too time-intense, hinder comparison to quests in actual games, and introduce the risk of experimenter bias. We consequently decided to use quests from existing RPG games. We collected quests from multiple games for two reasons. First, RPGs from different game series have distinct styles of quest writing, and collecting a diverse set of writing style holds the promise to increase the expressive range of the learned model. Second, we were unlikely to find the required amount of quests in a single, regular RPG. As argued earlier, we discarded MMORPGs as less constrained data source to avoid a negative impact on the quality of our model output.

There are the following two main techniques for obtaining video game texts [34]:

i) extracting text directly from game files;
ii) scraping text from unofficial, fan-curated online sources.

However, game files are often either encrypted or use poorly documented proprietary file formats, whereas fan-written sources, such as online wikis, typically only paraphrase the contents of the in-game texts, e.g., character dialog, instead of directly documenting how they appear to the players.

We consequently focused on (i) and extracted quest texts directly from the game files with modding tools (more detail in Appendix A). We appealed to (ii) by drawing on fan wisdom, selecting the RPG games not only based on quest quality but also based on the presence of high-quality fan wikis and active modding scenes. Information from fan wikis made it easier to retrieve quest data from games files while modding tools allowed us to sidestep the file format and encryption issues.

To obtain a sufficiently large dataset of varied and complex quests, we first collected a total of 878 quest examples from five RPGs. These games share a medieval-esque fantasy setting, which should improve the quality of the model but can also limit the its expressive range. To counteract this, we extended our dataset with one hundred manually written *Minecraft* [39] quests. In total, our dataset comprises 978 quests from six games as summarized in Table I. Additionally, Table II shows how our quest dataset performs on some well-known natural language processing metrics. Overall, all RPGs in our dataset produce similar scores on the depicted metrics: A considerable exception to this is the readability metric, which implies that the *Torchlight II* [40] quest descriptions are more difficult to read than the descriptions from the other RPGs in the dataset. This disparity is likely caused by the fact that fictional names make

TABLE I
QUEST DATASET (978 QUESTS)

Game	Sourcing	Quests
Baldur's Gate [36]	Collected (game files)	100
Baldur's Gate II: Shadows of Amn [37]	Collected (game files)	94
The Elder Scrolls IV: Oblivion [38]	Collected (game files)	215
The Elder Scrolls V: Skyrim [13]	Collected (game files)	389
Minecraft [39]	Written by the authors	100
Torchlight II [40]	collected previously in [34]	80

TABLE II
NATURAL LANGUAGE PROCESSING METRICS (MEAN $\pm$ STDDEV) ON THE QUEST DESCRIPTIONS FROM THE QUEST DATASET (SEE TABLE I)

Game	Readability (Flesch-Kincaid Grade) smaller easier	Syntactics complexity (Dependency distance) larger more complex	Lexical richness (Type-token ratio) larger richer	Word count
Baldur's Gate [35]	3.03 ± 1.61	2.33 ± 0.31	0.73 ± 0.08	99 ± 42
Baldur's Gate II [36]	2.88 ± 1.34	2.18 ± 0.25	0.66 ± 0.08	134 ± 58
The Elder Scrolls IV [37]	3.00 ± 1.65	2.19 ± 0.27	0.66 ± 0.08	143 ± 77
The Elder Scrolls V [13]	2.78 ± 1.53	2.18 ± 0.30	0.71 ± 0.08	105 ± 47
Minecraft [38]	3.36 ± 1.48	2.30 ± 0.28	0.71 ± 0.06	91 ± 29
Torchlight II [39]	4.58 ± 2.15	2.45 ± 0.40	0.74 ± 0.08	79 ± 28
Overall	3.07 ± 1.67	2.23 ± 0.31	0.70 ± 0.08	112 ± 57

up a larger portion of the *Torchlight II* descriptions, relative to their shorter average length.

B. Data Formatting

To generate a quest description, a language model must be given an outline with the desired "ingredients" of a quest as input. We analyzed the collected quests to recognize these ingredients (see Table III). Our quest ingredients align partially with classical narrative analyses in the literature, such as Vladimir Propp's *Morphology of the Folktale* [41]. For example, Propp's definitions of various types of dispatchers and character archetypes bear similarities to our quest-givers. Existing narrative analyses were only of limited use, as they typically span the entire duration of a story while we are more interested in the circumstances at the beginning of a quest.

Not only what information is provided but also how it is laid out is crucial to training a language model: Semantically equivalent pieces of input text can yield wildly different results, likely because some text formats synergize better with the model's pretraining data. We devised and compared three distinct input formats, i.e., quest metadata formats, for representing the quests via their quest ingredients: A highly structured format that resembles XML, later referred to as *XML-like*, a *simple* format that is inspired by *dramatis personae*, i.e., character listings in plays and movie scripts, and a *narrative* format that reads like a small story. The first format, *XML-like* is adopted from Lee [42], who has successfully used a similar format to generate patent claims with GPT-2. Fig. 1 illustrates all three formats based on an example quest.

We devised a generic JSON representation for storing our quests in an organized manner (see Appendix B) and to derive our training data in the three metadata formats. We also hope that storing our quests in a canonical format makes it easier for other researchers to adopt our dataset in their work.

C. Data Processing

While collecting the quest dataset, candidate quests were evaluated by the authors based on the following criteria:

1) novelty and interestingness of narrative and content [43];
2) the existence of clearly defined goals;
3) the length of the quest description.

We excluded quest descriptions that lacked the essential quest ingredients in Table III. As a side-effect, these descriptions were typically very short. We also discarded too long descriptions ($>$256 words), as they might exceed GPT-2's context window that holds 1024 tokens (i.e., roughly 256 English words), resulting in the model forgetting ingredients.

Some candidates did not meet one or multiple criteria and were consequently omitted. Other quests only met these criteria to a limited extent and were consequently manually edited. For instance, quests are usually delivered through sprawling dialog between the player and the quest-giver, not linearly through monolithic pieces of text. As a consequence, quest rewards are commonly discussed after the player has already completed the quest; we had to make some tense changes to accommodate the rewards into the quest descriptions. Moreover, some candidate quests were split into multiple independent quests, as they either

1) involved the quest-giver directing the player to another NPC;
2) or had distinct paths for the player to follow based on their actions in the game.

V. DEVELOPING QUEST-GPT-2

Our text generation example in Fig. 2 demonstrates that GPT-2 can generate some short, rudimentary quest descriptions even without fine-tuning, if one provides few quest examples in the input text. However, the output quality is not convincing. Moreover, quest descriptions typically incorporate many small elements, such as world knowledge, as well as character relationships and archetypes. It is difficult to incorporate those elements into a few quest examples in the input, especially considering the fact that the context window of GPT-2 holds only 1024 tokens, i.e., byte-pair encoded sets of characters. In the following, we describe the process of fine-tuning GPT-2 with our custom dataset into Quest-GPT-2. We made all code publicly available on Github.[2]

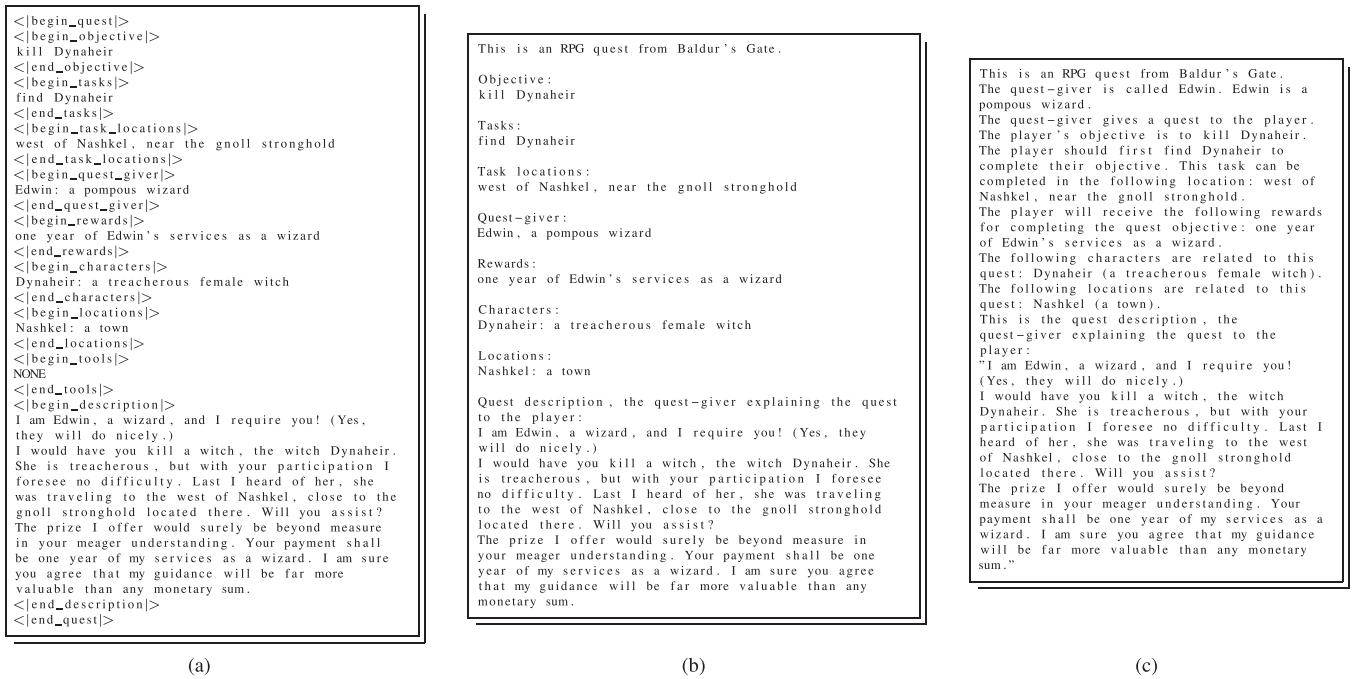
A. Preliminary Fine-Tuning Experiments

We informed the model fine-tuning through a series of quick, small experiments on an Nvidia GTX 1070 8 GB GPU with the two smallest GPT-2 variants (124 M and 355 M parameters) and the *XML-like* quest metadata format. We used the training script from @nshepperd's fork of the official OpenAI GPT-2 Github release, and adopted the default optimizer settings, i.e., Adam with an initial learning rate of 2×10^{-5}. We set the batch size to 1, because larger batch sizes generated out-of-memory exceptions with 8 GB of VRAM.

[2][Online]. Available: https://github.com/svartinen/gpt2-quest-descriptions

TABLE III
QUEST INGREDIENTS IDENTIFIED FROM OUR DATASET

Quest Ingredient	Description	Essential
Quest-giver	The person giving the quest to the player	yes
Objective	The overarching goal of the quest	yes
Tasks	The actions that have to be done to fulfill the objective of the quest	yes
Task locations	The locations where the tasks can be completed	no
Rewards	The rewards given to the player upon the completion of the quest objective	no
Facts	Important facts related to the quest	no
Items	Important items related to the quest	no
Characters	Important characters related to the quest	no
Locations	Some secondary locations related to the quest	no
Groups	Important groups, e.g., factions, related to the quest	no
Enemies	Enemies that the player will face during the quest	no
Description	The quest description shown to the player	yes



(a) (b) (c)

Fig. 1. Comparison of the example quest *Edwin and Dynaheir* from *Baldur's Gate*, expressed in our three proposed quest metadata formats. (a) XML-like. (b) Simple. (c) Narrative.

These early experiments showed promise for generating relatively coherent quest descriptions and even complete quests. We made some small observations in-between adjustments to and repetitions of this setup. First, if the characters have not been explicitly gendered in the metadata, both employed variants of GPT-2 might either choose a binary gender, or randomly flip between male or female pronouns. This behavior was fixed by explicitly including the characters' genders in their descriptions in later experiments. Second, both models displayed signs of overfitting in all experiments, and we consequently employed early stopping later on. Third, the generated descriptions do not always encompass all quest ingredients from the input, and entities might be treated incorrectly. Most strikingly, a character who is referenced multiple times in the input quest outline might appear as several separate people in the output quest description. When comparing the two differently sized GPT-2 variants, the larger GPT-2-355 M produced noticeably more coherent quest descriptions than the smaller GPT-2-124 M while also transmitting the ingredients of the input quest outlines into output quest descriptions more comprehensively. Additionally, the cross-entropy loss for the larger GPT-2-355 M converges noticeably faster toward zero than the loss for the smaller GPT-2-124 M (see Fig. 3).

B. Substituting Proper Nouns and Numbers With Placeholders

To address these consistency issues, we employ the placeholder token technique introduced by Martin et al. [6]: Proper nouns (i.e., unique names) and numbers are replaced in the quest metadata with *placeholder* tokens. The original names and numbers are substituted back into the generated output in a postprocessing step. Fig. 4 displays the example quest

```
objective: kill all creepers
location: woods
quest giver: a butcher
reward: a diamond axe
description: Creepers have taken over the woods! Hunters can't procure game for
me! Kill all creepers! I'll reward you with a diamond axe.

objective: save villagers from a witch
location: a village
quest giver: a villager
reward: 16 emeralds
description: A witch is holding my fellow villagers captive. Someone ought to save
them! Traveler, if you did this task for me, I'd give you 16 emeralds.

objective: kill all zombies
location: caves
quest giver: a villager
reward: 32 golden carrots
description: Zombies are out for blood! Kill all zombies! I'll reward you with 32 golden carrots.
```

Fig. 2. Quest generation with (not fine-tuned) GPT-2-774 M. Here, we provide two full quests as examples (top two). This is followed by a list of ingredients for a new quest (bottom). The system completed the quest description based on this input (in bold).

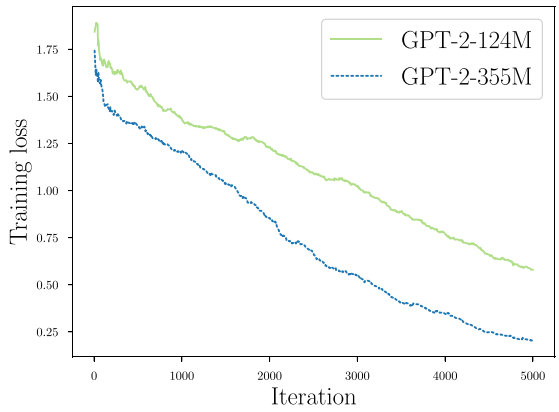


Fig. 3. Cross-entropy loss for our early fine-tuning experiments.

from Fig. 1 in *XML-like* format with placeholders. Generative models like GPT-2 learn complex multivariate probability densities $p(x, y, \ldots)$, which becomes more difficult as the number of variables grows. We assume that names and numbers are independent from other quest content and that the joint distribution can thus be factorized into $p(x, y, \ldots) = p(x)p(y, \ldots)$. We hypothesized that this factorization via placeholders will allow the model to learn content independently from the name and number information that bears no significant meaning.

C. Fine-Tuning Quest-GPT-2

We split the 978 quests in our dataset (see Table I) into training, validation, and test sets with 80:15:5 percent ratios. We used the validation set to mitigate overfitting, and the test set for evaluation against human judgment in our user study (see Section VI). To represent all six source games equally in all sets, the quests were first split proportionally per game, and then combined into the complete training, validation, and test sets. Afterward, we converted the sets into the three proposed quest metadata formats, producing both *raw text* and *placeholder text* for each format for performance comparison.

In contrast to the preliminary experiments, we fine-tuned the largest GPT-2 model with 1.5B parameters. We trained the model six times, once for each combination of metadata format and the two placeholder conditions. We used the same fine-tuning settings as in the preliminary experiments (see Section V-A) for

```
<|begin_quest|>
<|begin_objective|>
kill character_0
<|end_objective|>
<|begin_tasks|>
find character_0
<|end_tasks|>
<|begin_task_locations|>
west of location_0, near the gnoll stronghold
<|end_task_locations|>
<|begin_quest_giver|>
quest_giver: a pompous wizard
<|end_quest_giver|>
<|begin_rewards|>
one year of quest_giver's services as a wizard
<|end_rewards|>
<|begin_characters|>
character_0: a treacherous female witch
<|end_characters|>
<|begin_locations|>
location_0: a town
<|end_locations|>
<|begin_tools|>
NONE
<|end_tools|>
<|begin_description|>
I am quest_giver, a wizard, and I require you! (Yes, they will do nicely.)
I would have you kill a witch, the witch character_0. She is treacherous,
but with your participation I foresee no difficulty. Last I heard of her,
she was traveling to the west of location_0, close to the gnoll stronghold
located there. Will you assist?
The prize I offer would surely be beyond measure in your meager
understanding. Your payment shall be one year of my services as a wizard.
I am sure you agree that my guidance will be far more valuable than any
monetary sum.
<|end_description|>
<|end_quest|>
```

Fig. 4. Example quest in the *XML-like* format with *placeholder text*.

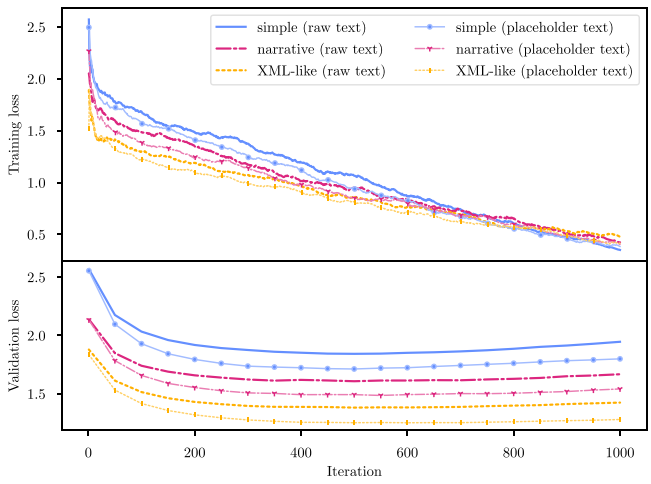


Fig. 5. Fine-tuning results, moving averages of cross-entropy loss.

1000 iterations at most and stopped early once the validation loss increased again. On an Nvidia V100 32 GB GPU, the fine-tuning took ca. 50 min per combination.

Fig. 5 shows the fine-tuning loss. The *placeholder* substitution performs unanimously best in terms of training and validation loss for all metadata formats. Amongst the metadata formats, the *XML-like* format achieves the smallest training and validation loss while the *simple* format performs worst.

Crucially though, comparing metadata formats based on fine-tuning loss only can be misleading: The model might learn repetitive formatting easily without respecting format-independent quest ingredients, thus "masking" the loss values smaller when using heavier formatting. To rule this out, we compared the fine-tuned models with *perplexity*, an established language model metric that measures how well a model can predict each token in a piece of text, with lower values being better. We calculated the

TABLE IV
CONDITIONAL PERPLEXITIES OF THE FINE-TUNED MODELS

Metadata Format	Text Type	Conditional Perplexity
narrative	raw text	10.63
	placeholder text	10.50
simple	raw text	10.95
	placeholder text	10.55
XML-like	raw text	11.05
	placeholder text	10.78

conditional and normalized perplexities of the quest descriptions in the validation set when given a certain quest outline as input. If a model has a low fine-tuning loss but a high conditional perplexity, it most likely predicts the formatting tokens correctly while displaying a high degree of uncertainty with respect to the quest ingredient tokens. The results in Table IV show that *placeholder text* achieves lower perplexity than *raw text* with all three metadata formats, thus supporting our previous findings. While *XML-like* always produces the highest perplexities, the *narrative* format consistently achieves the lowest perplexity regardless of the placeholder use and is thus to be preferred.

Based on these objective metrics, we selected the Quest-GPT-2 model fine-tuned with the *narrative* format and *placeholder text* for the final subjective evaluation.

D. Exploring Quest-GPT-2 Text Generation Settings

We anticipate that even after fine-tuning, many generated quests would not convince a human audience. For example, Fig. 6 shows quest descriptions generated by the fine-tuned model that might be considered somewhat nonsensical by people. Instead of merely sampling the most probable tokens from the output probability distribution, methods such as top-*k* sampling and nucleus sampling have been successfully employed to generate more natural-sounding text [44]. Holtzman et al. [44] have argued that natural language does not maximize probability; humans favor nonobvious language.

As a final step before our user study, we determined the optimal sampling settings for Quest-GPT-2 model inference through four ministudies. The studies were performed among the members of the game AI research group at Aalto University and had three participants on average. We generated 6 to 10 quest descriptions for two quests and each of the below sampling setting configurations, and asked participants to rate the descriptions according to their perceived quality on a 7-point Likert scale. The scale was accompanied with the statement "The quest description fits the quest great." We compared the following sampling setting configurations:

1) nucleus sampling with top-*p* values 0.5, 0.7, and 0.9;
2) top-*k* sampling with top-*k* 40;
3) baseline pure sampling;

with or without the following additional modifiers:

1) temperature: 0.7;
2) repetition penalty: 1.2.

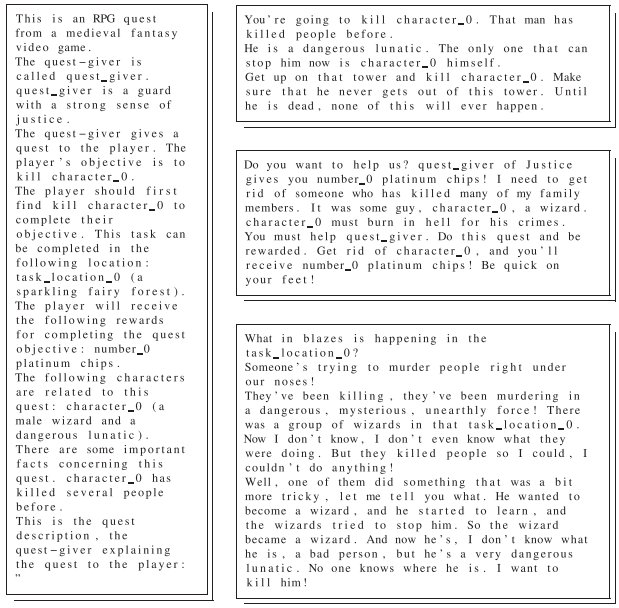


Fig. 6. Quest generation examples after the fine-tuning and before the optimization of sampling settings. Here using aitextgen's default settings. (a) Input quest outline in the *narrative* format with *placeholder text*. (b) Random output quest descriptions generated with the fine-tuned Quest-GPT-2 model.

The first ministudy compared all sampling setting configurations without the additional modifiers, the second one introduced the temperature modifier, the third added in repetition penalty, and the last compared two nucleus sampling configurations, top-*p* values 0.5 and 0.9, to each other with both modifiers and two Likert scale statements "The quest description fits the quest great narratively" and "The quest description fits the quest great in terms of correctness."

It is difficult to balance the narrative quality and the correctness of details: One needs to find the sampling settings that produce an optimal degree of randomness to generate interesting yet sensible quest descriptions. We found that nucleus sampling with top-$p = 0.5$, temperature $= 0.7$, and repetition penalty $= 1.2$ produced the best results with Quest-GPT-2.

E. Rejecting Quest-GPT-2 Outputs

To further improve the model outputs, we implemented two simple heuristic filters that reject bad samples. Both filters exploit our special placeholder tokens (see Fig. 4).

The first filter performs token verification, i.e., it checks whether the special tokens in the output also exist in the input. For instance, the example quest input in Fig. 4 (i.e., lines up to and including the $< |begin\_description| >$) does not include any named groups or related *group_n* tokens. Consequently, the resulting output quest description (i.e., lines after $< |begin\_description| >$) should not contain said tokens either. The second filter complements the first: It checks that important, user-configurable special tokens in the input are present in the output. This filter can ascertain that only outputs are retained, which contain certain desired quest elements, e.g., the output description in Fig. 4 should mention *character_0*.

TABLE V
MEAN NATURAL LANGUAGE PROCESSING METRICS (MEAN ± STDDEV) ON THE GENERATED QUEST DESCRIPTIONS

Game	Readability (Flesch-Kincaid Grade) smaller easier	Syntactics Complexity (Dependence Distance) larger more complex	Lexical Richness (Type-Token Ratio) larger richer	Word Count
Baldur's Gate [35]	2.53 ± 1.20	2.18 ± 0.24	0.78 ± 0.06	90 ± 36
Baldur's Gate II [36]	1.93 ± 1.47	2.13 ± 0.32	0.74 ± 0.07	98 ± 39
The Elder Scrolls IV [37]	2.74 ± 1.18	2.21 ± 0.23	0.71 ± 0.07	127 ± 50
The Elder Scrolls V [13]	2.39 ± 1.37	2.18 ± 0.21	0.73 ± 0.08	104 ± 46
Minecraft [38]	1.32 ± 0.98	2.04 ± 0.25	0.78 ± 0.07	65 ± 22
Torchlight II [39]	2.98 ± 1.06	2.17 ± 0.23	0.72 ± 0.08	95 ± 29
Overall	2.38 ± 1.33	2.17 ± 0.24	0.74 ± 0.08	103 ± 46

VI. EVALUATING QUEST-GPT-2

Writing RPG quest descriptions is usually considered a creative activity, and we thus want Quest-GPT-2 to be a creative system. Assessing creativity however is not easy, and defining *creativity* alone is a source of debate among (computational) creativity researchers [45, p. 77]. Most researchers, however, agree that a creative product must be novel and valuable [46] to be deemed creative. Assessing the novelty of generated artifacts, however, is not straight-forward, as perceived novelty is highly contingent on individual experience [47]. We consequently focus on assessing the quality of the generated quests, and complement ratings with open-ended questions to gather further information on what influenced our participants' assessment. We next present our evaluation methods, describe the results, and, finally, discuss them critically.

A. Experiment Design

We performed a randomized mixed design user study in the form of an online questionnaire in which participants were presented with quests and asked to rate corresponding quest descriptions. We chose a mixed design to obtain ratings on many quest descriptions produced from many quests while avoiding fatigue that could negatively impact response quality.

B. Materials

Participants were presented with a quest from the test set that was set aside during fine-tuning (see Section V-C). For each quest in the test set, we generated 10 quest descriptions with Quest-GPT-2, utilizing the improvements from Sections V-D and V-E. Based on the 50 random quests in the test set (sampled proportionally from each game in our quest dataset as mentioned in Section V-C), we obtained a total of 500 quest descriptions as stimuli in the study. Table V illustrates the same natural language metrics as Table II on the generated descriptions. The generated descriptions are noticeably simpler, i.e., easier to read and shorter, than the original human-authored ones. All quests and quest descriptions are available in a public Open Science Foundation repository.[1]

The quests and their generated descriptions were embedded in an online questionnaire. For improved readability, the quests were presented in the *simple* format [see Fig. 1(b)] without placeholders, instead of the narrative format with placeholders, which was used in fine-tuning Quest-GPT-2.

To keep the individual workload manageable, each participant received five quest descriptions from five randomly sampled test set quests, i.e., 25 quest descriptions in total. To counteract fatigue, the five quests were always presented along with their description instead of interleaving the quests with each other. The presentation order of the quest descriptions for each quest was randomized to avoid order effects.

C. Participants

The study participants were recruited from various RPG subcommunities on Reddit and r/SampleSize, a subcommunity dedicated to (scientific) surveys. The study was advertised toward everyone aged over 18 years with RPG playing experience. We did not offer any incentives for participation.

Overall, 349 respondents participated in the questionnaire, of which 345 responses were retained. We excluded three respondents, as they only provided empty or one-word answers to our free-form questions. Additionally, one respondent was excluded due to being under 18 years old. The gender breakdown of participants was 71.9% male, 20.0% female, 4.9% gender variant/nonconforming, 0.6% other, and 2.6% preferred not to state their gender. 97.1% of participants stated their age, ranging from 18 to 62 years ($M = 28.7$, SD $= 8.1$).

The participants reported their average weekly gaming time as follows: 0.9% played less than an hour, 7.5% 1–4 h, 15.1% 5–8 h, 23.8% 9–12 h, 15.7% 13–16 h, 35.1% more than 16 h, and 2.0% preferred not to say. Regarding the participants' familiarity with RPGs, 35.4% had played *Baldur's Gate*, 30.1% *Baldur's Gate II*, 58.8% *Minecraft*, 58.6% *The Elder Scrolls IV: Oblivion*, 83.2% *The Elder Scrolls V: Skyrim*, 26.7% *Torchlight II*, 76.8% other RPGs, and 0.3% preferred not to say. When asked about other RPG games, the participants listed dozens of Western, Japanese, table-top inspired and MMORPGs, confirming that most participants were avid, experienced RPG fans.

D. Measures

We gathered demographic data on age and gender, as well as player expertise data based on the number of hours spent on playing games per week, and players' favorite RPGs (detailed questions and answer options provided in our public repository[1]). Participants were asked to rate each quest description on a 4-point Likert scale (Strongly Disagree—Strongly Agree), indicating their agreement with the statement "I would be happy to see this quest description in a video game." An even scale was chosen to disallow neutral ratings and support the ratings' interpretation as separating unsuitable (mean rating $\ll 2.5$) from suitable (mean rating $\gg 2.5$) descriptions. We moreover asked the following free-form questions.

- Q1. Which criteria did you use to assess the suitability of each quest description?
- Q2. What upset you most about the unsuitable quest descriptions?
- Q3. What did you like most about the suitable quest descriptions?

The first question was used to understand participant's criteria in assessing quest descriptions, and the last two were used to determine the strengths and weaknesses of the descriptions.

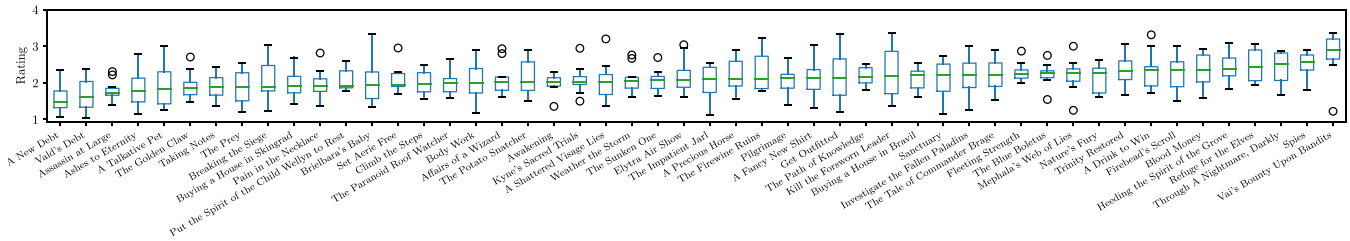


Fig. 7. Box plots of quest description ratings for each of the 50 quests in the test set, sorted by the median in ascending order. Each point represents participants' mean ratings on a quest description produced for the corresponding quest.

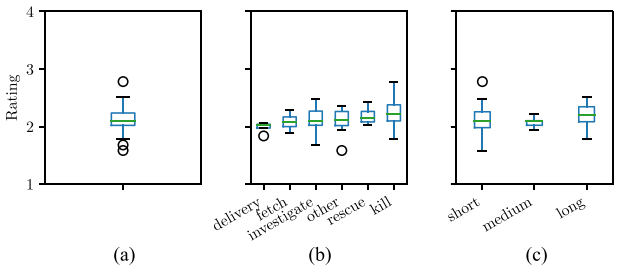


Fig. 8. Box plots of quest description ratings distinguished by quest types. Each point represents the mean rating for a quest in the test set. (a) All quests. (b) Quest types. (c) Quest outline length.

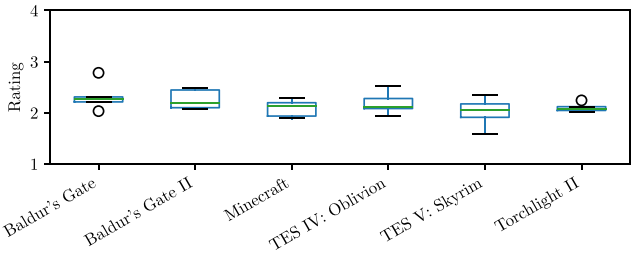


Fig. 9. Box plots of averaged ratings per participant, grouped by game.

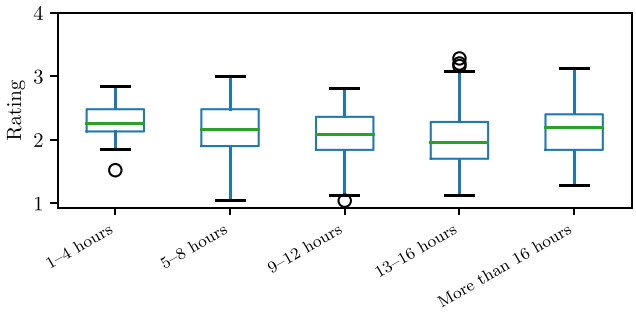


Fig. 10. Box plots of averaged ratings per participant, grouped by their average weekly playtime (groups holding <5% participants were omitted).

E. Procedure

First, the participants were asked to read and agree to an informed consent form. They were then asked to provide details on demographics and expertise. In the main part of the questionnaire, the participants were shown blocks of the following:

1) a random quest;
2) five different descriptions generated for this quest.

After rating all five quest descriptions, they were presented with another quest with the corresponding descriptions. This process was repeated five times, until each participant rated five quest descriptions for five quests. Finally, the participants were given the previously described free-form questions. Each step is illustrated in our public[1] materials.

F. Results

We found strong variations in the perceived quality of quest descriptions (see Fig. 7) within and beyond individual quests. If we interpreted strong deviations from the Likert midpoint as a reliable indicator of suitability, then many quests had a mix of suitable and unsuitable quest descriptions. The median rating over all quests is slightly above 2 and thus below the midpoint of our 4-point Likert scale [see Fig. 8(a)]. We did not find any striking differences in ratings when categorizing quests by their type [see Fig. 8(b)], outline length [see Fig. 8(c)], or the game they originated from (see Fig. 9). Participants generally appear more critical the more they played (see Fig. 10). The exception are those who reported playing for more than 16 h per week, which also includes "hard-core" gamers. We performed a one-way ANOVA to further investigate the effect of reported playtime on the participants' ratings, yielding that differences between the groups are only slightly significant ($F = 2.3$, $p = 0.063$).

Based on participants' rich answers to our free-form questions, we learned that players used various criteria to assess the suitability of the quest descriptions (Question 1). The most often mentioned criteria include correctness in regards to the given quest outline, internal logic as well as coherence, tone and immersiveness. Other common criteria were interestingness, the lack of repetition, grammar, narrative flow, and clear instructions. Sporadically, participants noted humor, the length of the quest description, and the feelings that are evoked while reading the quest descriptions as assessment criteria. There were notable differences in how the participants applied their criteria. In particular, participants were not equally-minded about the importance of criteria, such as grammar, and a small subset of participants' answers indicate that they were lenient with their ratings, as follows:

1) they knew that they were reading AI-generated text ("If these numbers went from 1-10 instead of 1-4, I think they'd get the same ratings, for the most part");
2) they were not native English speakers ("note: I'm not native speaker");

3) or they appreciated the unintentional humor often found within computer-generated text ("They [suitable descriptions] were humorous at times").

Our participants' comments on unsuitable (Question 2) and suitable (Question 3) quest descriptions echoed their assessment criteria. The unsuitable quest descriptions failed and the suitable ones fulfilled them. Unsuitable descriptions were lamented to be nonsensible or illogical, contained unnecessary details, repetition and conflicting information, had poor grammar to the point of "reading "off" as if poorly translated from a Chinese comic," or were simply boring lists of facts. On the contrary, suitable descriptions were found clear, surprising, fun, original, and believable even to the point of being seemingly human-authored, thus supporting that our model marks a step forward in achieving less repetitive and formulaic computer-generated quests. Some participants noted that there were no suitable quest descriptions in their subset, supporting our finding that the descriptions vary greatly in quality.

On a general note, it seems that there is no objective consensus for what makes a good quest description: Some study participants preferred short, no-nonsense descriptions without unnecessary details, whereas others liked longer descriptions laced with in-game lore. Regarding quest objectives, there were participants who would rather only receive hints about what to do, and others who preferred in-dept instructions.

Fig. 11 shows examples of the worst and best rated quest descriptions. In addition to highlighting many of the participants' thoughts on unsuitable quest descriptions, the badly rated descriptions indicate that Quest-GPT-2 sometimes fails to discern different entities from each other even if unique names are substituted with generic placeholders. This behavior is likely inherent to GPT-2, and made worse with complicated relationships between different characters. For instance, Mogrul, the quest-giver of "A New Debt," and Drovas Relvi, Mogrul's debtor in the same quest, are supposed to be different people, yet in the top-most quest description in Fig. 11(a) the quest-giver states that "My name is Mogrul. You might know me as Mogrul, or maybe as Drovas Relvi."

We provide all responses, quantitative and qualitative, as well as the computation of the summary statistics, in anonymized form in our public repository.[1]

Fig. 11. Examples of best and worst rated quest descriptions. (a) Worst rated quest descriptions for the three worst rated quests: *A New Debt*, *Vald's Debt*, and *Assassin at Large*. (b) Best rated quest descriptions for the three best rated quests: *Vai's Bounty Upon Bandits*, *Spies*, and *Through A Nightmare, Darkly*.

G. Discussion

Our results suggest that even the largest variant of GPT-2, fine-tuned on our well-curated dataset, cannot be used to autonomously generate high-quality quest descriptions reliably. This confirms findings in related work [33]. We especially found that Quest-GPT-2 lacks the ability

1) to distinguish between multiple entities;
2) to "glue" quest ingredients well together while not relaying illogical information.

The model's direct successor, GPT-3, has been shown to offer vast, general improvements in text quality [4], and we hypothesize that GPT-3 would handle these two aspects of quest description generation better. To support this hypothesis, we have provided the quest with the worst rated quest description in our experiment, "A New Debt" [see Fig. 11(a)], as input to the vanilla GPT-3 model. In comparison to Quest-GPT-2, the descriptions generated by GPT-3 (see Fig. 12) are noticeably more coherent than the worst rated Quest-GPT-2 descriptions. Given suitable hardware for fine-tuning and tweaks such as our *placeholder text*, we believe that this next generation of models can bring fully autonomous quest description generation within the reach of game developers. We advocate several use-cases for our present model. First, many of the poorly rated quest descriptions outputted by Quest-GPT-2 only contain few issues, such as a single illogical sentence. Therefore, the model could be used as an assistant for co-creative quest writing: A professional RPG writer could first give a rough, simplified quest outline to Quest-GPT-2, and then fill in more complex details into the generated output. Second, Quest-GPT-2 could be used to generate quest ideas: one can supply the starting sentence of a quest outline to generate the rest of the outline and the quest description. Third, Quest-GPT-2 could be used to generate quest descriptions offline which can, after only little human curation, be used in a video game without further changes. This is

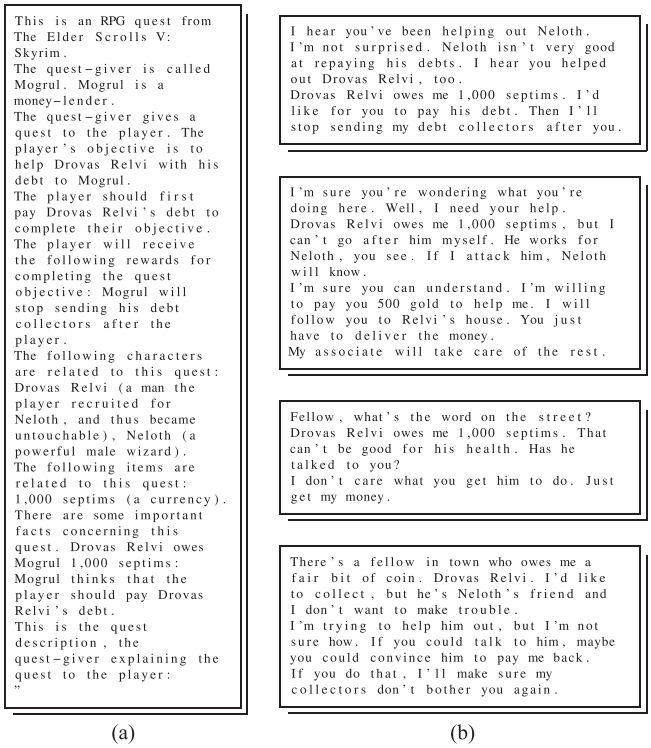


Fig. 12. Quest generation demo with the quest *A New Debt* and GPT-3 (OpenAI API Playground, default text generation settings apart from response length of 700). The quests *Ashes to Eternity*, *Assassin at Large*, and *Vald's Debt* were given as examples to GPT-3 beforehand. (a) Input quest outline in the narrative format with raw text. (b) Four random output quest descriptions generated with GPT-3.

supported by the observation that some quest descriptions were rated highly by people. The curation coefficient, i.e., the ratio of human-acceptable outputs from any given creative system [48] is 0.22, indicating that roughly one in five quest descriptions would be deemed acceptable.

We finally reflect on the limitations of our study. First, we observed both positive and negative bias toward AI-generated text. The former was evident from the participants using lenient ratings as described previously, and the latter was observed from, e.g., one of the participants describing bad experiences with procedurally generated quests from *The Elder Scrolls V: Skyrim* [13]. Such biases are well known when people judge computer-generated artifacts [48]. To alleviate them, we recommend comparing human-written and AI-generated quest descriptions in future studies. Turing-style tests on creative systems have been criticized [49], and we hence suggest to omit any explicit mention of this dichotomy. A second limitation of our study is given by its focus on RPG games with medievalesque fantasy settings. Generalizing our findings to other settings is not advisable, as the model's capacity to generate text on a specific theme depends on the presence of this theme in the original pretraining dataset. A third limitation is given by the gender imbalance, which was inherited from the Reddit communities that participants were recruited from, and should in the future be compensated for via other communities and additional recruitment channels.

VII. Conclusion and Future Work

We have investigated the use of the GPT-2 and GPT-3 language models to generate quest descriptions for RPG games. We built and published a novel quest dataset and employed a strategy for improving learning from limited training data by placeholder substitution similar to that in [6]. We fine-tuned GPT-2 into the quest description generating Quest-GPT-2 model, and conducted an online user study to evaluate its output.

While our results are encouraging, the quality of the generated descriptions varied greatly. Despite the name substitution strategy, Quest-GPT-2 often makes mistakes related to handling a large number of entities, such as characters, groups, and locations. Moreover, Quest-GPT-2 often generates descriptions with questionable logic, repetition, poor grammar, and unnecessary information. While using our model automatically and online is not yet viable, we have proposed three means on how Quest-GPT-2 can already be used by designers offline.

Based on our case-studies on generating quest descriptions with the vanilla GPT-3 model, we hypothesize that the next generation of language models could be fine-tuned with (an extension of) our quest dataset to alleviate the discussed issues. Other potential areas of future work are personalizing quest descriptions for different kinds of RPG players and player characters; replacing our simple heuristic filters with an AI critic for rejecting dissatisfying model outputs as well as using grammar checking tools or other algorithms for improving text quality; and generating other quest-related artifacts, e.g., quest names, journal entries, and dialog trees, in addition to quest descriptions. Moreover, one could investigate expanding the quest generation system to continuous quest lines or multistep quests by including previous quests or quest steps alongside quest ingredients. Bidirectional language models such as BERT [50] could be investigated to provide individual, fill-in suggestions for all quest ingredients, not only the quest descriptions. Finally, we highlight the opportunity for collaborations between games industry and researchers on both, the use of existing datasets to improve new models, and the latter's integration in tools for design-time cocreation.

We encourage researchers and the general public to adopt the techniques presented here and extend our publicly available code and dataset to investigate the future use of large language models for video game quest generation.

Appendix A
Quest Collecting in Detail

We gathered the quests in the following manner. Firstly, the quests from *Baldur's Gate I-II* were extracted by first identifying the quest-giving NPCs by reading *Baldur's Gate* Wiki quest descriptions, then looking for and selecting the relevant game dialog files with Near Infinity, a browser and editor software for games that use the Infinity game engine, and finally using the relevant pieces of dialog to construct proper quest descriptions. Second, the skeletons for *The Elder Scrolls IV-V* quests were first scraped from the Unofficial *Elder Scrolls* Pages in JSON format: Each quest contained information on objective, locations, quest giver, and reward. The final quest descriptions were then

formulated by reading the relevant game files with either *The Elder Scrolls* Construction Set (*The Elder Scrolls IV*) or the Creation Kit (*The Elder Scrolls V*). Finally, the *Torchlight II* quests originally collected by van Stegeren and Theune [34] were in.csv format with the following fields: speaker (quest-giver), text, dialog type, quest name as seen in-game, quest name in game data, quest file, speaker unit type, unit file, and raw quest text. We converted these quests to our JSON schema (Appendix B), cleaned them up, and added any missing, relevant information, such as archetypal character descriptions.

Appendix B
JSON Representation for Quests

```
  "name": "the name of the quest,"
  "objective": "quest objective,"
  "first_tasks": ["a list of tasks that
should be done to fulfill the objective"],
  "first_task_locations": ["a list of lo-
cations corresponding with the tasks, sim-
ilar to the locations field"],
  "quest_giver": {
   "name": "the name or ti-
tle of the quest giver,"
   "description": "a brief, archetypal de-
scription of the quest giver,"
   "location": "the where-
abouts of the quest giver"
  },
  "reward": [a list rewards, a re-
ward is defined {
   "name": "the name of the reward,"
   "description": "a brief, common de-
scription of the reward,"
   "amount": the number of received re-
wards
  }],
  "characters": [(optional) a list of re-
lated characters, a character is de-
fined similarly to the quest giver],
  "enemies": [(optional) a list of re-
lated groups of enemies, mostly
used for declaring a set number of en-
emies for a quest, a group of ene-
mies is defined similarly to a reward],
  "items": [(optional) a list of re-
lated items, e.g tangible items, or even
some more abstract ones like ritu-
als, an item is defined similarly to a re-
ward],
  "groups": [(optional) a list of re-
lated groups, e.g., factions, races, or
creatures, where a group is defined {
   "name": "the name of the group,"
   "description": "a brief, common de-
scription of the group"
  }],
  "locations": [(optional) a list of re-
lated locations, where a location is de-
fined {
   "name": "the name of the location,"
   "description": "a brief, common de-
scription of the location"
  }],
  "tools": ["important facts re-
lated to the quest"],
  "description": "the quest description"
```

Acknowledgment

The authors would like to thank their reviewers for their excellent, helpful feedback. The Aalto University School of Science "Science-IT" project provided the GPT-2 fine-tuning infrastructure.

References

[1] K. Merrick, "Modeling motivation for adaptive nonplayer characters in dynamic computer game worlds," *Comput. Entertainment*, vol. 5, no. 4, pp. 1–32, 2008.
[2] A. Kantosalo and H. Toivonen, "Modes for creative human-computer collaboration: Alternating and task-divided co-creativity," in *Proc. Int. Conf. Comput. Creativity*, 2016, pp. 77–84.
[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, Art. no. 9.
[4] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
[5] R. Dale, "GPT-3: What's it good for?," *Natural Lang. Eng.*, vol. 27, no. 1, pp. 113–118, 2021.
[6] L. Martin et al., "Event representations for automated story generation with deep neural nets," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, pp. 868–875.
[7] B. Kybartas and R. Bidarra, "A survey on story generation techniques for authoring computational narratives," *IEEE Trans. Comput. Intell. AI Games*, vol. 9, no. 3, pp. 239–253, Sep. 2017.
[8] S. G. Ware and R. M. Young, "Intentionality and conflict in The Best Laid Plans interactive narrative virtual environment," *IEEE Trans. Comput. Intell. AI Games*, vol. 8, no. 4, pp. 402–411, Dec. 2016.
[9] D. Thue, V. Bulitko, M. Spetch, and E. Wasylishen, "Interactive storytelling: A player modelling approach," in *Proc. AAAI Conf. Artif. Intell. Interactive Digit. Entertainment*, 2007, pp. 43–48.
[10] M. Si, S. C. Marsella, and D. V. Pynadath, "Thespian: Using multiagent fitting to craft interactive drama," in *Proc. 4th Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2005, pp. 21–28.
[11] J. McCoy, M. Treanor, B. Samuel, A. A. Reed, N. Wardrip-Fruin, and M. Mateas, "Prom week," in *Proc. ACM Int. Conf. Found. Digit. Games*, 2012, pp. 235–237.
[12] M. Guimaraes, P. Santos, and A. Jhala, "CiF-CK: An architecture for social NPCS in commercial games," in *Proc. IEEE Conf. Comput. Intell. Games*, 2017, pp. 126–133.
[13] Bethesda Game Studios, "The Elder Scrolls V: Skyrim," Game [PC], Bethesda Softworks, Rockville, MD, USA, 2011.
[14] B. Kybartas and C. Verbrugge, "Analysis of ReGEN as a graph-rewriting system for quest generation," *IEEE Trans. Comput. Intell. AI Games*, vol. 6, no. 2, pp. 228–242, Jun. 2014.
[15] A. Calvin and N. Michael, "The quest in a generated world," in *Proc. DiGRA*, 2007, pp. 503–509.
[16] J. Pita, B. Magerko, and S. Brodie, "True story: Dynamically generated, contextually linked quests in persistent systems," in *Proc. Conf. Future Play*, 2007, pp. 145–151.
[17] A. Stocker and C. Alvin, "Non-linear quest generation," in *Proc. Int. Florida AI Res. Soc. Conf.*, 2018, pp. 213–216.
[18] J. Doran and I. Parberry, "A prototype quest generator based on a structural analysis of quests from four MMORPGs," in *Proc. Int. Workshop Procedural Content Gener. Games*, 2011, pp. 1–8.

[19] V. Breault, S. Ouellet, and J. Davies, "Let CONAN tell you a story: Procedural quest generation," *Entertainment Comput.*, vol. 38, no. 3, 2021, Art. no. 100 422.
[20] E. Soares de Lima, B. Feijó, and A. L. Furtado, "Procedural generation of quests for games using genetic algorithms and automated planning," in *Proc. Brazilian Symp. Comput. Games Digit. Entertainment*, 2019, pp. 144–153.
[21] P. Ammanabrolu, W. Broniec, A. Mueller, J. Paul, and M. Riedl, "Toward automated quest generation in text-adventure games," in *Proc. Workshop Comput. Creativity Lang. Gener.*, 2019, pp. 1–12.
[22] J. van Stegeren and J. Myśliwiec, "Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation," in *Proc. ACM 16th Int. Conf. Found. Digit. Games*, 2021, pp. 1–8.
[23] Blizzard Entertainment, "World of Warcraft," Game [PC], *Blizzard Entertainment*, Irvine, CA, USA, 2004.
[24] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proc. IEEE*, vol. 88, no. 8, pp. 1270–1278, Aug. 2000.
[25] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Comput. Intell. Mag.*, vol. 9, no. 2, pp. 48–57, May 2014.
[26] R. Loughran and M. O'Neill, "Application domains considered in computational creativity," in *Proc. Int. Conf. Comput. Creativity*, 2017, pp. 197–204.
[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[28] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.
[29] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," 2020, *arXiv:2005.00341*.
[30] A. Madani et al., "ProGen: Language modeling for protein generation," 2020, *arXiv:2004.03497*.
[31] P. Clark, O. Tafjord, and K. Richardson, "Transformers as soft reasoners over language," in *Proc. Int. Joint Conf. Artif. Intell.*, 2020, pp. 3882–3890. [Online]. Available: https://www.ijcai.org/Proceedings/2020/0537.pdf
[32] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," to be published, 2018.
[33] J. van Stegeren and J. Mysliwiec, "Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation," in *Proc. ACM FDG*, 2021, pp. 1–8.
[34] J. van Stegeren and M. Theune, "Fantastic strings and where to find them: The quest for high-quality video game text corpora," in *Proc. Workshop Intell. Narrative Technol.*, 2020, pp. 1–8.
[35] D. M. Ziegler et al., "Fine-tuning language models from human preferences," 2020, *arXiv:1909.08593*.
[36] BioWare, "Baldur's Gate," Game [PC], Interplay Entertainment, Los Angeles, CA, USA, 1998.
[37] BioWare, "Baldur's Gate II: Shadows of Amn," Game [PC], Interplay Entertainment, Los Angeles, CA, USA, 2000.
[38] Bethesda Game Studios, "The Elder Scrolls IV: Oblivion," Game [PC], Bethesda Softworks, Rockville, MD, USA, 2006.
[39] Mojang Studios, "Minecraft," Game [PC], *Mojang Studios*, Stockholm, Sweden, 2011.
[40] Runic Games, "Torchlight II," Game [PC], *Runic Games*, Seattle, WA, USA, 2012.
[41] V. I. Propp, *Morphology of the Folktale*, 2nd ed. Austin, TX, USA: Univ. Texas Press, 1968.
[42] J. S. Lee, "Controlling patent text generation by structural metadata," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 49–62.
[43] P. Gervás, "Computational approaches to storytelling and creativity," *AI Mag.*, vol. 30, no. 3, pp. 49–62, 2009.
[44] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–16.
[45] C. Guckelsberger, "Intrinsic motivation in computational creativity applied to videogames," Ph.D. dissertation, School Elect. Eng. Comp. Sci., Queen Mary Univ. London, London, U.K., 2020.
[46] M. A. Runco and G. J. Jaeger, "The standard definition of creativity," *Creativity Res. J.*, vol. 24, no. 1, pp. 92–96, 2012.
[47] S. McGregor, "Algorithmic information theory and novelty generation," in *Proc. Int. Conf. Comput. Creativity*, 2007, pp. 109–112.
[48] S. Colton and G. Wiggins, "Computational creativity: The final frontier?," in *Proc. 20th Eur. Conf. Artif. Intell.*, 2012, pp. 21–26.
[49] A. Pease and S. Colton, "On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal," in *Proc. AISB Symp. AI Philos.*, 2011, vol. 39, pp. 1–8.
[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2022, pp. 4171–4186.