Ontologically Faithful Generation of Non-Player Character Dialogues

Nathaniel Weir, Ryan Thomas, Randolph d’Amore, Kellie Hill, Benjamin Van Durme, and Harsh Jhamtani

1Johns Hopkins University, Baltimore, MD, USA
2Microsoft Semantic Machines, Redmond, WA, USA
3Microsoft Gaming, Redmond, WA, USA

nweir@jhu.edu, hjhamtani@microsoft.com

Abstract

We introduce a language generation dataset grounded in a popular video game. **KNUDGE** (KNowledge Constrained User-NPC Dialogue GEneration) requires models to produce trees of dialogue between video game characters that accurately reflect quest and entity specifications stated in natural language. **KNUDGE** is constructed from side quest dialogues drawn directly from game data of Obsidian Entertainment’s *The Outer Worlds*, leading to real-world complexities in generation: (1) utterances must remain faithful to the game lore, including character personas and backstories; (2) a dialogue must accurately reveal new quest details to the human player; and (3) dialogues are large trees as opposed to linear chains of utterances. We report results for a set of neural generation models using supervised and in-context learning techniques; we find competent performance but room for future work addressing the challenges of creating realistic, game-quality dialogues.

1 Introduction

Player interactions with non-player characters (NPCs) in role-playing games (RPGs) often serve to flesh out backstories while allowing the player to progress through engaging quest storylines [Onuczko et al., 2007]. Figure 1 shows a dialogue turn, taken from *The Outer Worlds* [Obsidian Entertainment, 2019], an RPG famous for its writing. A key challenge in creating NPC dialogues is that they should serve coherent narratives: utterances must faithfully reflect quest structure and game lore—characters, histories, and entity relationships. Dialogues are often purposely designed to start/end quests according to granular specifications (e.g., if player says option A then it starts quest B; if player says option C, then the NPC says D, which is important for completing the quest...) and to serve a storytelling role, actively espousing to the player details about the game world. NPC interactions often take the form of complex trees that can have dozens of nodes, and creating these branching structures according to the many specifications of dialogue authoring can be time-consuming for game designers [Caropreso et al., 2012] and cost companies millions of dollars (see §A). This motivates the pursuit of tools for automatically generating dialogue trees.

However, there is a lack of realistic benchmarks to train and evaluate models for this purpose. van Stegeren and Theune, 2020 highlight that game text corpora should come from real, professionally written games; most research that explores game dialogue relies on crowdsourced or academically-curated text, which is not representative of the highly game- and context-sensitive text of real dialogues. Moreover, related work on game dialogue [Urbanek et al., 2019a; van Stegeren and Mysliwiec, 2021], story generation [Akoury et al., 2020; Chen and Gimpel, 2021], and knowledge conditioning for task-oriented dialogue agents [Choi et al., 2018; Mazaré et al., 2018; Feng et al., 2020] does not address complex dialogue trees and interweaving narratives found in deployed RPGs.

**KNUDGE** will be publicly available upon publication.

---

1https://en.wikipedia.org/wiki/The_Outer_Worlds
To address this dearth of realistic benchmarks for game dialogue authoring, we introduce **KNUDE**: KNowledge-constrained User-NPC Dialogue GEneration, a set of dialogue trees (in English) extracted from *The Outer Worlds*, an existing video game, and paired with granular ontological constraints. **KNUDE** contains 159 dialogues from all 45 side quests in *The Outer Worlds*. It contains 4.7K utterances and an average of 1.3K input constraint tokens per tree. For each quest, we annotate each turn in the dialogues with relevant grounding information—quest- and lore-related natural language (NL) support facts pulled from fan-written wikis. Such fine-grained support fact annotations are useful for training models to generate game dialogues grounded in quest specifications and game lore.

To the best of our knowledge, ours is the first dataset that consists of a set of real game-quality NPC dialogues paired with granular quest and biographical specifications consistent with a well-formed game ontology.

Using **KNUDE** as a test bed, we devise the task of knowledge-constrained NPC dialogue generation (Figure 2). The complex input specifications and limited training data target a realistic development scenario in which a designer is working on a new, partially written game. For this task, we introduce a model class, termed DialogueWriters, that leverages LMs such as GPT [Brown et al., 2020] to generate dialogue trees given input constraints. To address the challenges of long specification passages and branching tree structures, we introduce techniques for prompt construction, retrieved exemplar munging, and tree structure representation. To encourage the use of the ontology to produce engaging dialogue, we experiment with mechanisms to select relevant knowledge, taking advantage of the rich, node-level annotations of **KNUDE**.

Finally, we prescribe evaluation protocols that test whether models can reflect game ontology constraints in addition to generating fluent and coherent dialogue. We conduct automatic and human evaluations of utterances and trees generated from specifications for game quests, as well as for a pair of never-before-seen quests written by a professional game designer. Our experiments reveal further room for improvement in aspects such as ontology usage and maintaining coherence.

## 2 Task Definition

As communities seek to develop AI-based writing tools such as Ubisoft’s GhostWriter [Barth, 2023], there arises a need to reconcile the challenge of dialogue authoring with current NLG techniques. An ideal writing tool might allow for a designer to provide the inputs like granular quest information and bios from a game’s lore, and receive a set of generated utterances, or entire trees, to aid in drafting content similarly to GitHub Copilot. As a game ontology can be large, so too might the amount of user-provided knowledge specifications. Our task targets this scenario, for which no such dataset exists to study how to optimize models.

We define the task of knowledge-constrained NPC dialogue generation as the mapping from a set of quest constraint statements \( Q \), biographical constraint statements \( B \), and participant names \( P \) to a dialogue tree \( D \). We consider two task scenarios: **next utterance prediction** from a partial tree, and **full dialogue generation** of trees with branching player options.
fine-tuning, models such as those described in §4 can leverage it for in-context learning.

3 Data

Writing RPG-quality dialogue trees is difficult for human developers for its many interweaving considerations. 1) The tree must serve its quest function, containing input-specified player utterance options, NPC responses (including specified emotions), and pieces of information the player must learn by the end (e.g. the Log Entry in Figure 2). 2) The utterances must be coherent and engaging to the player. 3) The NPC should embody the persona described in their bio passage explaining personality, history, and relationships, and finally 4) To facilitate world building, the NPC should expost details about other entities whenever it is contextually relevant, but should never violate the ontology through contradiction. With these desiderata in mind, we design KNUDGE to pose a similarly multi-faceted challenge to generation models.

KNUDGE contains NPC dialogue trees from all 45 side quests in the Outer Worlds. This RPG is appealing for our investigation because of its large trees, its award-winning writing, and its tendency for entities to appear in many quests in different capacities. Construction of KNUDGE entails gathering information about each quest (Q) and the Outer Worlds entities appearing or referenced during its dialogues (B) (§3.1), and then extracting trees (D) from the game data (§3.2).

3.1 Game Ontology

We acquired dialogue files from the Outer Worlds creators along with permission to release them publicly; we use quest data and game lore from fan wikis, where a quest’s page lists the in-game objectives and journal logs (the framework also allows for using data from official channels). §B provides further details about how data was extracted.

**Quest Information** A quest in The Outer Worlds appears in the player’s journal with a high-level synopsis and a sequence of objectives, each of which contains game logs providing additional details. Active objectives are completed, and new ones introduced, during an NPC dialogue. We assign each objective a walkthrough passage which includes details on the topics, player utterance options, and quest information that the NPC needs to say by the dialogue’s end. A detailed quest anatomy and examples of Q can be found in §C.

**Biographical Information** We associate with each quest and dialogue a set B of biographical passages about entities appearing or referenced during the quest. We extract passages from entities’ fan wiki pages. While some are short, others can be up to 27 sentences, posing a challenge to generation models; often only part of a long biography might be relevant to a given quest. Examples are shown in §D.

3.2 Dialogue Trees

Illustrated in Figure 3 and §E, dialogue trees in The Outer Worlds are complex directed graphs, containing many conditional utterance options depending on the state of the game—e.g. whether the player is of high enough level at some skill to pass a “check.” To extract a more tractable, quest-related subgraph, we 1) identified the nodes that start and end the interaction by watching online playthrough videos, and then 2) traversed the graph from the start node, following only edges without state-related conditions. We then added conditional edges manually depending on quest relevance.

**Annotating Utterance Nodes with Support Facts** We coordinated with English-speaking professional data annotators to label tree nodes with support facts from quest and biography passages.

Figure 4: Overview of dialogue node annotation with support facts from quest and biography passages.

---

3E.g. with 55 points of the Persuade skill, the player can convince Tucker to return to his mother in the Figure 1 quest.
We introduce a set of methods called DialogueWriters for where utterances should be suggested. The average of 1321 constraining tokens and 406 utterance to-edge to a list of candidates n recent" node specified new location branching off a dialogue in K communications (generating candidate utterances given the ontological specification of GPT-3 before factoring in other pieces of context. This allows developers the flexibility to choose details.

**Tree Traversal** We consider language models (LMs) that accept linear input token sequences. We thus devise a traversal mechanism that, at inference time, converts a dialogue subtree into a maximal coverage linear history. For “most recent” node n, we identify the longest possible path from the start node to n only following any given edge once. This produces utterance history \( H = [u_1, \ldots, u_n] \). We feed \( H \) to a next utterance generator trained via supervised or in-context learning.

**Supervised Learning (SL) Models** We fine-tune a T5-large model [Raffel et al., 2020] to generate \( c_i \) given the concatenation \( B, Q, P, H \). We truncate context from the left when required given T5’s 1024-token window (see §G). We also train a Knowledge Selection (KS) version that decodes support knowledge facts before generating the utterance \( c_i \). This factorizes the next utterance generation into a two-step decision process: first selecting one or more facts from \((Q \cup B)\), and second generating the utterance to reflect the selected facts. We thus use KNUDGE’s node-level annotations to train the model to generate the concatenation \([f_1^{(i)}, \ldots, f_m^{(i)}, c_i]\).

**In-Context Learning (ICL) Models** As there is little training data, fine-tuning might not be effective at learning the difficult generation task. As such, we experiment with methods for in-context learning (ICL) with OpenAI’s text-davinci-003 GPT-3 model [Brown et al., 2020]. We inject \( B, Q, P, H \) into a formatted prompt that naturally elicits the next utterance as a continuation of \( H \). Figure 6 depicts this process; full prompts are shown in §G. This creates a zero-shot prompt. When this does not fill out GPT-3’s 4000-token window, we construct a few-shot prompt by adding dialogs from training quests as exemplars, simulating a scenario in which a developer has written a partial set of quests and is working on a new one. We retrieve exemplars using Okapi-BM25 [Jones et al., 2000] with \([B, Q, P]\) as the query string.

As with the SL framework, we also devise an ICL Knowledge Selection (KS) version of the ICL DialogWriter that first decodes one or more support facts before generating an utterance. We elicit this behavior from GPT-3 by augmenting all utterances in the dialogue history with support facts if they have them (see Figure 19). §H provides further training details.

### 4.2 End-to-End DialogueWriters

For scenarios in which a writer wants the DialogWriter to suggest an entire dialogue instead of nodes at a specific position, we propose a second class of methods that generate many nodes and edges all at once. As shown in Figure 5, we propose a series of prompts that leverage the instruction-tuned capabilities of longer-context models like GPT-4-turbo [Achiam et al., 2023] and Vicuna-13B-v1.5-16k [Zheng et al., 2023] to iteratively generate and refine a dialogue.

**Tree Traversal and Prompt Format** The pipeline first feeds \([B, Q, P]\) and a set of BM25-retrieved exemplars to an initial generator prompt. Exemplar outputs contain all nodes and edges in their dialogue, displayed as a list of dictionaries (see §I). We traverse nodes in breadth-first fashion so that all children of a parent appear as close to it as possible. §I describes the logic for when to display edges.

**Revision Cycle** As the resulting graph has no guarantee of

---

**Table 1: KNUDGE Dataset Statistics**

<table>
<thead>
<tr>
<th>Components</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quests</td>
<td>45</td>
</tr>
<tr>
<td>Entities</td>
<td>168</td>
</tr>
<tr>
<td>Characters</td>
<td>81</td>
</tr>
<tr>
<td>Locations</td>
<td>40</td>
</tr>
<tr>
<td>Groups</td>
<td>21</td>
</tr>
<tr>
<td>Items</td>
<td>18</td>
</tr>
<tr>
<td>Creatures</td>
<td>7.4</td>
</tr>
<tr>
<td>Facts per entity</td>
<td>7.6</td>
</tr>
<tr>
<td>Quests per entity</td>
<td>2.0</td>
</tr>
<tr>
<td>Facts per objective</td>
<td>7.6</td>
</tr>
<tr>
<td>Objectives per quest</td>
<td>4.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Counts</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Players</td>
<td>12.3</td>
</tr>
<tr>
<td>Objectives</td>
<td>9.9</td>
</tr>
<tr>
<td>Utterances</td>
<td>1.0</td>
</tr>
<tr>
<td>NPC utterances</td>
<td>16.7</td>
</tr>
<tr>
<td>Utt. tokens per dialogue</td>
<td>406.6</td>
</tr>
<tr>
<td>Facts per dialogue</td>
<td>83.3</td>
</tr>
<tr>
<td>Utterances per dialogue</td>
<td>406.6</td>
</tr>
<tr>
<td>Facts per NPC utterance</td>
<td>1.0</td>
</tr>
<tr>
<td>Entity Facts</td>
<td>0.57</td>
</tr>
<tr>
<td>Objective Facts</td>
<td>0.42</td>
</tr>
</tbody>
</table>

high agreement on a subset selection problem where the total set size is often of the order of hundreds.

### 3.3 Dataset Analysis

Table 1 describes statistics for the extracted dialogue trees and annotations. The dataset contains over 65k utterance tokens and 210k fact tokens (16 tokens per fact). The player has an average of 2.5 utterance options on their turn to speak. 47% of all nodes (57% of NPC nodes) are annotated with at least one fact. NPC utterances have an average of 1.0 facts. Dialogues have on average 9.9 quest facts and 73.3 biographical facts that models must factor into generation. The largest dialogue contains 103 nodes, 130 edges, and 236 facts.

**Comparison with Related Datasets** KNUDGE is the first dataset to contain dialogue trees from an actual RPG annotated with game quest and biography specifications. Table 2 compares our dataset to contemporaries with comparable information. The largest dialogue dataset to contain dialogue trees from an actual RPG annotated with game quest and biography specifications.

<table>
<thead>
<tr>
<th>Components</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quests</td>
<td>45</td>
</tr>
<tr>
<td>Entities</td>
<td>168</td>
</tr>
<tr>
<td>Characters</td>
<td>81</td>
</tr>
<tr>
<td>Locations</td>
<td>40</td>
</tr>
<tr>
<td>Groups</td>
<td>21</td>
</tr>
<tr>
<td>Items</td>
<td>18</td>
</tr>
<tr>
<td>Creatures</td>
<td>7.4</td>
</tr>
<tr>
<td>Facts per entity</td>
<td>7.6</td>
</tr>
<tr>
<td>Quests per entity</td>
<td>2.0</td>
</tr>
<tr>
<td>Facts per objective</td>
<td>7.6</td>
</tr>
<tr>
<td>Objectives per quest</td>
<td>4.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Counts</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Players</td>
<td>12.3</td>
</tr>
<tr>
<td>Objectives</td>
<td>9.9</td>
</tr>
<tr>
<td>Utterances</td>
<td>1.0</td>
</tr>
<tr>
<td>NPC utterances</td>
<td>16.7</td>
</tr>
<tr>
<td>Utt. tokens per dialogue</td>
<td>406.6</td>
</tr>
<tr>
<td>Facts per dialogue</td>
<td>83.3</td>
</tr>
<tr>
<td>Utterances per dialogue</td>
<td>406.6</td>
</tr>
<tr>
<td>Facts per NPC utterance</td>
<td>1.0</td>
</tr>
<tr>
<td>Entity Facts</td>
<td>0.57</td>
</tr>
<tr>
<td>Objective Facts</td>
<td>0.42</td>
</tr>
</tbody>
</table>

We consider language models (LMs) that accept linear input token sequences. We thus devise a traversal mechanism that, at inference time, converts a dialogue subtree into a maximal coverage linear history. For “most recent” node \( n \), we identify the longest possible path from the start node to \( n \) only following any given edge once. This produces utterance history \( H = [u_1, \ldots, u_n] \). We feed \( H \) to a next utterance generator trained via supervised or in-context learning.

**Supervised Learning (SL) Models** We fine-tune a T5-large model [Raffel et al., 2020] to generate \( c_i \) given the concatenation \( B, Q, P, H \). We truncate context from the left when required given T5’s 1024-token window (see §G). We also train a Knowledge Selection (KS) version that decodes support knowledge facts before generating the utterance \( c_i \). This factorizes the next utterance generation into a two-step decision process: first selecting one or more facts from \((Q \cup B)\), and second generating the utterance to reflect the selected facts. We thus use KNUDGE’s node-level annotations to train the model to generate the concatenation \([f_1^{(i)}, \ldots, f_m^{(i)}, c_i]\).

**In-Context Learning (ICL) Models** As there is little training data, fine-tuning might not be effective at learning the difficult generation task. As such, we experiment with methods for in-context learning (ICL) with OpenAI’s text-davinci-003 GPT-3 model [Brown et al., 2020]. We inject \( B, Q, P, H \) into a formatted prompt that naturally elicits the next utterance as a continuation of \( H \). Figure 6 depicts this process; full prompts are shown in §G. This creates a zero-shot prompt. When this does not fill out GPT-3’s 4000-token window, we construct a few-shot prompt by adding dialogs from training quests as exemplars, simulating a scenario in which a developer has written a partial set of quests and is working on a new one. We retrieve exemplars using Okapi-BM25 [Jones et al., 2000] with \([B, Q, P]\) as the query string.

As with the SL framework, we also devise an ICL Knowledge Selection (KS) version of the ICL DialogWriter that first decodes one or more support facts before generating an utterance. We elicit this behavior from GPT-3 by augmenting all utterances in the dialogue history with support facts if they have them (see Figure 19). §H provides further training details.

### 4.2 End-to-End DialogueWriters

For scenarios in which a writer wants the DialogueWriter to suggest an entire dialogue instead of nodes at a specific position, we propose a second class of methods that generate many nodes and edges all at once. As shown in Figure 5, we propose a series of prompts that leverage the instruction-tuned capabilities of longer-context models like GPT-4-turbo [Achiam et al., 2023] and Vicuna-13B-v1.5-16k [Zheng et al., 2023] to iteratively generate and refine a dialogue.

**Tree Traversal and Prompt Format** The pipeline first feeds \([B, Q, P]\) and a set of BM25-retrieved exemplars to an initial generator prompt. Exemplar outputs contain all nodes and edges in their dialogue, displayed as a list of dictionaries (see §I). We traverse nodes in breadth-first fashion so that all children of a parent appear as close to it as possible. §I describes the logic for when to display edges.

**Revision Cycle** As the resulting graph has no guarantee of
We split K with a list of support fact IDs decoded before the utterance.

Table 2: Comparison of knowledge-constrained generation datasets. K

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Writer</th>
<th>Source</th>
<th>Structure</th>
<th>Dialogue Toks / Item</th>
<th>Constraint Toks / Item</th>
<th>Narrative and Bio Constraints</th>
<th>Level of Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>STORIUM</td>
<td>Crowd</td>
<td>Online story writing game</td>
<td>Sequence of scene entries</td>
<td>19k</td>
<td>1.2K</td>
<td>Scene intro, challenge, location, character descriptions</td>
<td>Story</td>
</tr>
<tr>
<td>TVSTORYGEN</td>
<td>Crowd</td>
<td>Fan wikis</td>
<td>TV episode recap</td>
<td>1.8k</td>
<td>25.9K</td>
<td>Brief summary, character bio</td>
<td>Scene Entry</td>
</tr>
</tbody>
</table>

RPG Dialogue Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Writer</th>
<th>Source</th>
<th>Structure</th>
<th>Dialogue Toks / Item</th>
<th>Constraint Toks / Item</th>
<th>Narrative and Bio Constraints</th>
<th>Level of Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIGHT</td>
<td>Crowd</td>
<td>Test game platform</td>
<td>Sequence of utterances</td>
<td>212 (12 utt)</td>
<td>276</td>
<td>Location description, persona statements, held objects</td>
<td>Dialogue</td>
</tr>
<tr>
<td>TorchLight II</td>
<td>Professional</td>
<td>RPG data</td>
<td>Sequence of quest stages with 0 or 1 utterances</td>
<td>157 (3 utt)</td>
<td>24</td>
<td>Quest title, objectives, details</td>
<td>Dialogue</td>
</tr>
<tr>
<td>WoW</td>
<td>Professional</td>
<td>RPG data</td>
<td>NPC-uttered quest description</td>
<td>61 (1 utt)</td>
<td>15</td>
<td>Quest title, objective</td>
<td>Dialogue</td>
</tr>
<tr>
<td>NUDDGE</td>
<td>Professional</td>
<td>RPG data</td>
<td>Complex quest dialogue tree</td>
<td>407 (29 utt)</td>
<td>1.3K</td>
<td>Quest title, objectives, location, logs, walkthrough, entity bio</td>
<td>Utterance</td>
</tr>
</tbody>
</table>

Figure 5: End-to-end DialogueWriter pipeline.

structural integrity nor accomplishing the specifications in Q and B, we append a set of zero-shot instructions to iteratively revise it using the LLM. We first use a revision prompt that encourages using more of the game lore in B. The revision is then fed to an automatic validation script, which checks for various structural inconsistencies such as malformed elements or disallowed edges (full list shown in §Appendix I). The list of identified violations is fed to a “Structure Revision” prompt to correct them. We repeat this validate-generate loop until either no violations remain or 5 revisions are attempted.

As with the Node Writers, we experiment with a knowledge selection mechanism in the form of an extra dictionary field with a list of support fact IDs decoded before the utterance.

5 Experiments

We split NUDDGE into train, development, and test splits based on quests (60/10/30%). Test set B’s contain a combination of seen and unseen entities.

Baselines To measure the effect of node-level knowledge selection (KS), we compare against an ICL model that selects only one statement instead of many. We also compare against an oracle KS ICL model, which conditions on the gold knowledge annotations for the reference utterance. We maximize the number of in-context examples for all ICL ablations; e.g. the no knowledge model’s prompt can have dozens of examples that are quite short. These ablations thus explore the tradeoff between the impact of the number of in-context examples and the presence of ontological statements.

To measure the effect of KS, we compare against SL and ICL no KS models. To measure the effect of conditioning on Q and B, we compare against ablations to the non-KS ICL model: a no knowledge model that conditions only on the participants P and utterance history H, and a quest only model that conditions on P, H, and Q, but not B.

5.1 Next Utterance Prediction (NUP)

To evaluate Node DialogueWriters (§4.1), we generated utterances at each node in the test dialogues conditioned on a subtree composed of the previous (in serial order) nodes and edges. We measured human judgments of NPC desiderata and automatic overlap against gold items and Q and B (§5.1). We then ran studies comparing larger dialogue structures (§5.2).

Results were verified via bootstrap testing.

Human Evaluation In coordination with a data specialist, we conduct human evaluation to examine models’ qualitative NUP performance. 100 generations per model were judged on a 4-point Likert scale for each of four criteria: 1. Coherence: does the utterance follow naturally from the utterances in the history? 2. (Non-)Violation: does the utterance create contradictions with any of the sentences in the ontology? 3. Biography Usage: does the utterance make use of the biographies in B? 4. Quest Usage: does the utterance progress the dialogue according to the quest details in Q? We provide the full set of annotator instructions in §J.

Automatic Evaluation We use re-scaled BERTScore-F1 [Zhang et al., 2020].4 as well as GPT-4-produced scores5

4We found that BLEU [Papineni et al., 2002], shown in §K, does not correlate with desirable outputs.

5GPT-4 correlates with humans with Spearman $\rho = .45, .17, .35, .47$ for the criteria, but gives lower scores. Prompt shown in §K.
We also evaluate the wordings of various facts when generating candidates.

Table 3: NUP human evaluation results for in-context (ICL) and video other qualities like realism and fluency to create a natural Outer Worlds full.

<table>
<thead>
<tr>
<th></th>
<th>Coherence</th>
<th>Violation</th>
<th>Using B</th>
<th>Using Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gold Reference</strong></td>
<td>3.94</td>
<td>3.97</td>
<td>3.50</td>
<td>3.45</td>
</tr>
<tr>
<td>SL-KS</td>
<td>2.52</td>
<td>3.85</td>
<td>2.17</td>
<td>2.09</td>
</tr>
<tr>
<td>ICL-KS</td>
<td>3.78</td>
<td>3.85</td>
<td>3.29</td>
<td>3.45</td>
</tr>
<tr>
<td><strong>KS Variants:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICL-KS-One</td>
<td>3.73</td>
<td>3.80</td>
<td>3.26</td>
<td>3.45</td>
</tr>
<tr>
<td>ICL-KS-Oracle</td>
<td>3.74</td>
<td>3.87</td>
<td>3.23</td>
<td>3.47</td>
</tr>
<tr>
<td><strong>Non-KS Baselines:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SL (no KS)</td>
<td>2.70</td>
<td>3.74</td>
<td>2.35</td>
<td>2.38</td>
</tr>
<tr>
<td>ICL (no KS)</td>
<td>3.88</td>
<td>3.97</td>
<td>3.25</td>
<td>3.43</td>
</tr>
<tr>
<td>ICL-Quest Only</td>
<td>3.79</td>
<td>3.90</td>
<td>3.03</td>
<td>3.21</td>
</tr>
<tr>
<td>ICL-No Knowledge</td>
<td>3.65</td>
<td>3.69</td>
<td>2.76</td>
<td>2.98</td>
</tr>
</tbody>
</table>

Table 3 shows that no model reaches gold coherence nor gold use of B under human evaluation, suggesting room for improvement on both.

Impact of Knowledge Selection: Table 4 shows that KS variants of the ICL model score a point or two higher than non-KS on overlap with B and Q, reflecting that KS effectively selects and cues the infusion of specific facts into generations. The ICL ablations of B and of \{B, Q\} have according drops in overlap and GPT usage score for both fact sets.

All B and Q-conditioned ICL models perform equivalently under all human metrics except (non)violation, at which KS models perform a slight point worse. We can conclude that KS improves the capacity of ICL writers to directly reflect knowledge passages (i.e. by copying spans), at the expense of a slightly higher likelihood of contradictions. This characterization can be appealing to a game developer; they might prefer for the automatic writer to use their own provided wordings of various facts when generating candidates.

Interpretation of Automated Metrics: We note that automatic metrics that check for overlap with reference text will, in isolation, give only a partial picture for evaluating generations in KNUDGE. This can be seen from the low performance of the gold utterances themselves under these metrics. We find that professionally written utterances do not always have high overlap with knowledge statements themselves. Gold utterances also do not score perfectly under human evaluation of Q and B usage, as not every real-world utterance reflects the ontology, and the KNUDGE ontology does not cover the full Outer Worlds game lore. The gold utterances also provide other qualities like realism and fluency to create a natural interaction. This finding undergirds the need to have multiple angles of evaluation: not simply checking for direct overlap, but evaluating for qualitative criteria such as coherence and appropriate ontology use.

In-Context vs Supervised Learning: The SL models have higher B overlap than ICL. This reflects that SL models incoherently copy spans directly from the context (see Figure 8), hence scoring poorly on human and GPT evaluation.

5.2 Case Studies

We conducted two case studies to assess DialogueWriters under different scenarios. In the first, Node DialogueWriters propose a dialogue skeleton to be fleshed out by a human. In the second, End-to-End DialogueWriters propose entire dialogues. In both cases, we provide methods B, P, Q, and one starting utterance. We evaluated the generated trees via the ACUTE-Eval [Li et al., 2019] pairwise comparison protocol.

Skeleton Generation We had models generating 10 rounds of dialogue. At each turn, we generated three candidate nodes and randomly “commit” one to the history, creating a form a 31-node dialogue ’spine’ for further development (see Figure 22). We selected 8 test dialogues from the game with varying quest roles, e.g. starting vs continuing quests. We constructed 8 more test items from 2 totally novel quests, written for us by a professional game designer, that occur in the Outer Worlds universe and contain entities from the original game. We asked human annotators (including the designer) to select which of two trees were preferred for the following criteria: coherence, nonviolation, biography and quest usage analogous to §5.1, and also 5. Content Suggestion: Do the multiple candidates at each turn propose interesting subtrees?

End-to-End Generation For full dialogue generation by End-to-End DialogueWriters, we used ACUTE-Eval to judge B and Q Usage based on desired outcomes and responses, and Effect on the Game State: By the (possibly multiple) ends of the dialogue, has the game state changed according to the desired specifications? e.g. subquests completed/added, specific items obtained, or other characters affected. Given the

Table 4: NUP BertScore for models against gold utterances and statements in B and Q. Results are shown beside the gold utterance’s score.

<table>
<thead>
<tr>
<th></th>
<th>Gold</th>
<th>Bio B</th>
<th>Quest Q</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BS</td>
<td>GPT</td>
<td>BS</td>
</tr>
<tr>
<td><strong>Gold Reference</strong></td>
<td>–</td>
<td>1.72</td>
<td>20.8</td>
</tr>
<tr>
<td>SL-KS</td>
<td>21.3</td>
<td>1.16</td>
<td>26.6</td>
</tr>
<tr>
<td>ICL-KS</td>
<td>25.1</td>
<td>1.81</td>
<td>24.3</td>
</tr>
<tr>
<td><strong>KS Variants:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICL-KS-One</td>
<td>25.2</td>
<td>1.84</td>
<td>23.9</td>
</tr>
<tr>
<td>ICL-KS-Oracle</td>
<td>26.8</td>
<td>1.87</td>
<td>24.7</td>
</tr>
<tr>
<td><strong>Non-KS Baselines:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SL (no KS)</td>
<td>23.5</td>
<td>1.22</td>
<td>24.0</td>
</tr>
<tr>
<td>ICL (no KS)</td>
<td>26.4</td>
<td>1.69</td>
<td>22.8</td>
</tr>
<tr>
<td>ICL-Quest Only</td>
<td>26.7</td>
<td>1.60</td>
<td>21.9</td>
</tr>
<tr>
<td>ICL-No Knowledge</td>
<td>27.0</td>
<td>1.30</td>
<td>22.4</td>
</tr>
</tbody>
</table>
extensive time required for human evaluation of full dialogues, we used GPT-4 following recent work [Nuismith et al., 2023; Liu et al., 2023] that finds “strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences well” [Zheng et al., 2023]. We verified GPT-4’s correlation with a human expert on a set of dialogues, then assessed all test quests in NUDGE using GPT-4; we compared the full end-to-end pipeline with ablated baseline variants.

Case Study Results  Skeleton case study results are provided in Table 5. Performance is measured as the rate at which annotators selected a model’s tree in a pairwise comparison under the 6 criteria listed in §5.2. We find that annotators preferred the trees of the ICL Node Writer most frequently compared to the other models under all criteria except Quest usage.

The end-to-end results are shown in Table 6, using both GPT-4 and Vicuna-13B-16k as the underlying Writer LLM. The GPT-4 writer with KS and Revision loop equals or outperforms all methods a majority of the time at using B and all approaches except ICL at using Q and achieving game effects. The Vicuna-based writer with KS also achieves similar results, though it performs worse with KS than without—highlighting that future work might consider how non-GPT-4 models can best make use of the selection paradigm.

6 Related Work

Si et al., 2021 focus on the task of story continuation through dialogue between characters while modeling the inter-character relations. However, such past work does not concern with the notion of grounding knowledge or quest objectives to be covered in the generated dialog. van Stegeren and Theune, 2020 propose three sources for building NPC dialogue corpora. Their proposed datasets do not contain any grounding annotation and are not accompanied by explicit descriptions of entities and characters. Callison-Burch et al., 2022 explore automatic generation of conversational turns by players of the tabletop RPG, Dungeons and Dragons, in which NPCs serve a very different role in the gameplay. Scheherazade’s Tavern [Aljammaz et al., 2020] augments a pattern-matching-based NPC interaction system with facts the character knows about the game world. More broadly, past work has explored applications of text generation in various tasks such as quest description generation [van Stegeren and Mysliwiec, 2021], dialogue generation [Si et al., 2021], persona-specific agents in text environments [Urbanek et al., 2019b], and new text world generation [Fan et al., 2020; Ammanabrolu et al., 2022].

Past work has pursued dialogue systems that steer the conversation towards a topic [Wu et al., 2019] or a given NL sentence [Sevegnani et al., 2021; Gupta et al., 2022]. Other work in NLG has explored generating outputs with high-level NL specifications such as string item agendas [Kiddon et al., 2016], sets of facts [Orbach and Goldberg, 2020], or author goals [Riedl, 2009]. NUDGE also comprises NL specifications, though they are comparably richer.

7 Conclusion

When dialogue is used to advance a carefully crafted storyline in a video game, it should be both engaging and consistent with the larger narrative. Language models are increasingly capable of producing engaging dialogue, but to date, research on ensuring dialogue’s consistency with underlying knowledge specifications has focused on datasets developed for the sake of experimentation, rather than actual high-quality game data. This paper introduces NUDGE, a dataset of NPC dialogue trees coupled with a relevant game ontology, drawn from the game, The Outer Worlds. In contrast to prior work, NUDGE is based on content created by a high-profile game development studio, thereby exemplifying real-world complexities in NPC dialogue authoring. We pose a knowledge-grounded generation task that mirrors a realistic development scenario with limited training data over a complex ontology of quests and lore. We find that LM-based methods are able to generate fluent dialogue that relates to provided specifications, but they do not match the quality of professional writers, particularly in terms of coherence and use of the game lore. We hope that NUDGE drives the development of new techniques for faithful game dialogue generation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida,


Appendix

A Cost of Authoring NPC Dialogues

Outer Worlds has 10 narrative/design credits\(^4\), which seems to be about average (Fallout has 4-12, while Skyrim has 9). Per salary.com, that's a position with an average salary of $58k per year (and probably more for AAA titles). Given a 2-3 year average development time for AAA titles, that works out to a conservative ballpark estimate of $1.2m for just this one game.

B Dataset Construction Details

B.1 Data Sources

Quest data and walkthrough passages were pulled from the Outer Worlds wiki of Fextralife,\(^9\) a gamer-focused site containing fan-made walkthroughs for many popular RPGs. Game entity biographies were collected from Fandom.\(^10\) The biography passage for a given entity is the same across all quests in which the entity appears in Knudge, and the set of entities is the same for all dialogues in a quest. Passages were segmented into individual sentences via punctuation boundaries. We identified relevant dialogues and their decision points using playthrough videos by the YouTube user, LordMatrim.\(^11\) All wiki articles were written in English by site users.

C Quest Anatomy and Example Items

Figure 7 provides a detailed anatomy of a Knudge quest, combining in-game quest data with corresponding passages from the fan walkthrough. Figure 11 shows example quest items with corresponding game data and walkthrough passages segmented into statements.

In whole, a dialogue's quest passage set \(Q\) contains:

1. The synopsis (1-2 sentences)
2. The in objective(s) active when entering the dialogue (1 sentence), the associated game log (1-2) and walkthrough passage (3-10), \(^12\)
3. The out objective(s) active upon leaving the dialogue, and the associated game log. \(^13\)\(^14\)

D Example Entity Biography Passages

Figure 12 shows example entities from The Outer Worlds with corresponding biographical passages.

\(^4\)https://www.imdb.com/title/tt9417446/fullcredits
\(^9\)https://www.theouterworlds.wiki.fextralife.com
\(^10\)https://theouterworlds.fandom.com
\(^11\)https://www.youtube.com/@l0rdmatrim
\(^12\)For the first dialogue in a quest, we associate the walkthrough passage describing how to obtain the quest.
\(^13\)We do not associate its walkthrough passage, since the NPC should only be expected to convey new objective information that the player will actually see in game.
\(^14\)The dialogue can lead to multiple new active objectives, some optional. If the dialogue concludes the quest, then no leaving objective is associated.

E Example Dialogues Items

Figure 13 depicts a full example input item conveying quest, biographical, and participant specifications. Figure 14, Figure 15, and Figure 16 depict example dialogue trees.

F Comparison with Other Datasets

van Stegeren and Theune; van Stegeren and Mysliwiec, 2020; 2021 consider datasets of publicly-available side quest data from RPGs such as World of Warcraft. However, their datasets vary in dialogue and quest coverage; for WoW their input is just a quest name and objective, and the generation target is a single-turn, few-sentence quest description spoken by an NPC. Their collect data for the game TorchLight II contains quest datapoints with a limited number dialog utterances per quest with no multi-turn interactions or trees.\(^15\) Others of their collected datasets contain complex branching trees but without constraining knowledge. The dialogues of LIGHT [Urbanek et al., 2019a] are more akin to NPC dialogues, though they comprise few-turn linear chains between two characters in self-contained episodes rather than quest-grounded interactions between a player and an NPC serving multiple game purposes. The size of constraining passages on the LIGHT dialogues are also a scale smaller than those of Knudge. The biographical constraints of Knudge are most similar to that of TVSTORYGEN [Chen and Gimpel, 2021], who also pull articles from fandom wiki pages. However, theirs is a story generation dataset where the target is a longform article describing a TV episode.

G Node DialogueWriter Details

Figure 6 depicts an overview of our tree linearization and prompt construction method. Figure 17 shows example seq2seq items used to train/evaluate the T5-based supervised learning DialogueWriters. We list the biographies of participants last so as to truncate them from the context only when all other bio have been removed. Else, biographies are listed in random order (fixed at the onset for full dialogue generation). Figure 18 depicts example prompts shown to GPT-3 based in-context-learning DialogueWriters.

H Model Training

To construct training items, we iterate through the nodes of each gold dialogue tree in a canonical order \([n_1, \ldots, n_t]\), where \(n_1\) is the tree’s start node. We create a separate item with each \(n_i\) as the generation target. We construct the subtree \(S^{(i)}\) comprised of all nodes \([n_1, \ldots, n_{i-1}]\) and all edges between them. We then construct the input/output pair \((Q, B, P, S^{(i)}) \rightarrow n_i\).

Supervised Learning To train SL DialogueWriter models, for every target node in the training quest dialogues, we construct 5 training examples using different random paths to the node. We train the model for 3 epochs using the default arguments from Hugging Face’s example summarization model.

\(^15\)Table 2 describes statistics for the 82 TorchLight II quests that contain both objective annotations and dialogue lines.
training script. T5 models were trained with a batch size of 1 across 8 Quadro RTX 6000 for an average of 5 hours.

In-Context Learning Given a test item, we construct a BM25 index over the training dialogues and use it to construct an n-shot ICL prompt where n depends on the remaining space available in the context window. Few-shot examples are linearized dialogues containing the most possible nodes from the gold tree. Contexts are left-truncated and can start with partial examples.

I End-to-End DialogueWriter Details

Algorithm 1 describes how to construct the list of nodes and edges that represent an entire dialogue for the End-to-End DialogueWriter.

J Human Evaluation Directions

Below, we enumerate the instructions shown to annotators during human evaluation:

Coherence: does the utterance follow naturally from the utterances in the history? (1) Utterance is nonsensical or ill-formed. (2) Utterance is contradictory of previous utterances in the history. (4) Utterance naturally responds to the history.

Violation: does the utterance create contradictions with any of the sentences in the ontology or objective blurbs? (1) Yes, explicitly contradicts sentences (list the ids). (2-3) (gray area). (4) No, utterance is consistent with the ontology.

Using the Bio Facts: does the utterance make use of the bio sentences in the ontology? (1) Utterance is fully generic and/or ignores the ontology completely, could have been generated had the bio facts not been included. (2-3) Utterance shows awareness of ontology, albeit unnaturally or inconsistently. (4) Utterance naturally incorporates one or multiple pieces of ontology.

Using the Objectives: does the utterance progress the dialogue according to the objective sentences in the prompt? (1) Utterance ignores objective, could have been generated had the obj facts not been included. (2-3) Utterance shows awareness of quest objectives, albeit unnaturally or inconsistently. (4) Utterance naturally incorporates one or multiple quest objective statements.

K Automatic Metric Details

Table 7 shows BLEU-4 scores against the same references as in Table 4. Figure 21 shows the prompt used to elicit the GPT-4 judgments shown in Table 4. We sampled judgments using a temperature of 0.3.

L Full Dialogue Evaluation

L.1 Skeleton Evaluation

Figure 22 depicts an example “spine” tree shown to evaluators during the end-to-end dialogue evaluation.

The instructions shown to annotators are as follows:

You will replace each ‘null’ value with either “a” or “b”, depending on which tree between modela and model b performed better under the following criteria:

1. Coherence: do the utterances in the tree create a realistic dialogue between the player character and the NPC?

2. Violations: does the dialogue tree create contradictions with any of the sentences in the ontology or objective blurbs? Does it contradict itself?

3. Using the Game Lore: does the tree faithfully make use of the bio sentences in the ontology, thereby espousing game lore about characters, groups, locations and items?

4. Covering the Objectives: does the dialogue tree play out according to the objective sentences in the prompt?

5. Content Suggestion: through generating multiple candidates at each turn, does the dialogue tree effectively propose potential dialogue subtrees that would espouse interesting content?

6. Engagingness: does the dialogue tree hold your attention and make you want to hear more from the NPC?

Algorithm 1: Edge Traversal Algorithm

Input: Start node \( v_{\text{start}} \), Graph \( G = (V, E) \)
Output: List of nodes and edges \( L \)

Initialize \( L = [] \)
Initialize queue \( Q = [v_{\text{start}}] \)
Initialize set of visited nodes \( S = \emptyset \)

while \( Q \neq \emptyset \) do

\( v_{\text{current}} = Q.\text{pop()} \)
foreach \( e = (v_{\text{from}}, v_{\text{current}}) \in E \) do
    if \( v_{\text{from}} \in S \) then
        Append \( e \) to \( L \)
    end
end

Append \( v_{\text{current}} \) to \( L \)
Add \( v_{\text{current}} \) to \( S \)
foreach \( e = (v_{\text{current}}, v_{\text{to}}) \in E \) do
    if \( v_{\text{to}} \in S \) then
        Append \( e \) to \( L \)
    end
    else
        Add \( v_{\text{to}} \) to \( Q \)
    end
end

Our automatic validator checks for the following structural issues:

1. Edges whose source or target node doesn’t exist
2. Nodes or subgraphs that are disconnected/unreachable from the start node
3. A long, linear sequence of 8 or more nodes without any branching
4. A dialog that is smaller than 5 nodes
5. A player node with multiple outgoing edges (only NPC nodes can have multiple children; in The Outer Worlds, NPCs respond immediately and deterministically given a player utterance).
6. An NPC node with multiple children that are also NPCs (since NPC responses are deterministic)

It provides a list of any identified issues to the structure revision prompt.

<table>
<thead>
<tr>
<th>Gold Reference</th>
<th>Bio</th>
<th>( B )</th>
<th>Quest ( Q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL-KS</td>
<td>2.6</td>
<td>4.9</td>
<td>2.2</td>
</tr>
<tr>
<td>ICL-KS</td>
<td>7.1</td>
<td>8.3</td>
<td>7.5</td>
</tr>
<tr>
<td>KS Variants:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICL-KS-One</td>
<td>6.8</td>
<td>7.1</td>
<td>7.6</td>
</tr>
<tr>
<td>ICL-KS-Oracle</td>
<td>7.2</td>
<td>8.4</td>
<td>7.2</td>
</tr>
<tr>
<td>Non-KS Baselines:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SL (no KS)</td>
<td>2.9</td>
<td>7.4</td>
<td>11.4</td>
</tr>
<tr>
<td>ICL (no KS)</td>
<td>8.6</td>
<td>6.8</td>
<td>6.6</td>
</tr>
<tr>
<td>ICL-Quest Only</td>
<td>7.9</td>
<td>3.4</td>
<td>6.4</td>
</tr>
<tr>
<td>ICL-No Knowledge</td>
<td>6.8</td>
<td>2.2</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 7: NUP BLEU for models against gold utterances and statements in \( B \) and \( Q \). Results for the latter two shown beside the gold utterance’s score.

L.2 Full Graph Evaluation

Table 8 depicts the results of expert human-annotated pairwise judgments between GPT-4-based end-to-end DialogueWriter and ablated (GPT-4-based) baselines. We use these judgments to verify correlation with GPT-4’s own judgments of pairwise preferences. We prompt humans and GPT-4 very similarly; a screenshot of the interface for interface can be found in Figure 9, and the prompt to GPT-4 can be found in Figure ??.

M Qualitative Results

Figure 8 depicts example outputs by models on an NUP example. We highlight cases in which the models succeed at the desiderata that we strive for in KNUDGE: to convey quest and lore specifications naturally through the interaction. However, we see that SL models and ablated ICL models are less successful. We observe that the gold utterance is more infused with desirable information than any generation; it references the quest’s next location and numerous adversaries that the player will run into, while effectively reflecting the NPC’s overprotective parent persona. This highlights a performance gap between neural and human writers to be addressed by future work. We note that not reaching human performance does not preclude DialogueWriters from being useful to writers, as they can still be used to suggest new directions for dialogues to be verified or modified in a human/AI collaborative writing process.
<table>
<thead>
<tr>
<th>E2E vs</th>
<th>ICL-No Know</th>
<th>ICL-Quest</th>
<th>ICL</th>
<th>ICL-KS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>win</td>
<td>loss</td>
<td>win</td>
<td>loss</td>
</tr>
<tr>
<td>Coh.</td>
<td>0.0</td>
<td>31.2</td>
<td>0.0</td>
<td>31.2</td>
</tr>
<tr>
<td>Viol.</td>
<td>25.0</td>
<td>12.5</td>
<td>12.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Use B</td>
<td>56.2</td>
<td>0.0</td>
<td>18.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Use Q</td>
<td>93.8</td>
<td>6.2</td>
<td>18.8</td>
<td>12.5</td>
</tr>
<tr>
<td>Con.</td>
<td>31.2</td>
<td>62.5</td>
<td>37.5</td>
<td>37.5</td>
</tr>
<tr>
<td>Eng.</td>
<td>0.0</td>
<td>12.5</td>
<td>6.2</td>
<td>6.2</td>
</tr>
<tr>
<td>Game St.</td>
<td>93.8</td>
<td>6.2</td>
<td>37.5</td>
<td>37.5</td>
</tr>
</tbody>
</table>

Table 8: ACUTE-Eval human-expert-rated pairwise results (% win/loss cases, tie not reported) between graphs generated by end-to-end DialogueWriter (GPT-4) under the full ICL-KS-plus-revisions pipeline (Figure 5) versus various baselines. Results are over 16 dialogues per comparison.
Figure 7: Anatomy of a Knudge quest. At any time, an Outer Worlds quest has currently active and previously completed objectives. To construct its Knudge representation, the quest’s high-level synopsis, objectives, and associated log entries from the game data are annotated with corresponding walkthrough article passages.
History: > Agnes Needham: Please, you have to help me get my little Tucker back! He ran away and is going to get himself killed! Oh, I just know a raptidon is melting him with acid as we speak! > Player: Your child is missing? Where’d you last see him? > Agnes Needham: Oh, I just knew you were a good person. ‘Agnes,’ I said, ‘this is the man to save your little Tuckie!’ And I was right!

Gold: Agnes Needham: He ran out into the wilderness a few days ago. I warned him about the raptidons, mantisaurs, and marauders - the toxic sulfur pools, and poisonous plants. But he didn’t listen!

ICL-KS: Agnes Needham: He was always such a good boy, but lately he’s been acting so strange. I just don’t understand why he would want to leave Stellar Bay. It’s the only place he’s ever known.

ICL: Agnes Needham: He left a few weeks ago. Said he was going to Amber Heights to join some revolutionary group. I just know he’s going to get himself killed if he’s not careful!

ICL-Quest Only: Agnes Needham: He left a few weeks ago, headed for Amber Heights. He’s going to get himself killed, I just know it!

ICL-No Knowledge: Agnes Needham: He was playing in the street. But then he saw a sweet little kitty and chased it into the sewers! Oh, the horror!

SL-KS: Agnes Needham: If you’re willing to pay handsomely, I’ll even pay you handsomely if you convince him to return to me.

SL: Player: I’ll do everything I can.

History: > Agnes Needham: Please! Won’t you go and find my boy? > Why would Tucker run away?

Gold: Agnes Needham: He’s been pining for an adventure, says he’s tired of living cooped up behind the walls. But he doesn’t understand how dangerous it is out there.

ICL-KS: Agnes Needham: I just wanted to keep him safe! It’s not like he could experience life cooped up in Stellar Bay his whole life. But the second he steps outside the walls, he’s in danger!

ICL: Agnes Needham: Oh, he’s just like his father. Always running off and getting into trouble. I just worry about him so much. I know he’s a grown man, but he’s still my little boy,

ICL-Quest Only: Agnes Needham: You know how teenage boys are. They just do things without thinking. They don’t realize how much their mothers worry about them.

ICL-No Knowledge: Agnes Needham: I just don’t know. He’s been acting strange lately. Ever since that little girl went missing.

SL-KS: Agnes Needham: Because he was scared, and I couldn’t let him get away with it!

SL: Agnes Needham: Because he’s scared of the things he’ll find out there! And I’m scared he’ll get eaten by some vile creature!

History: > Graham Bryant: Or... now, here’s an idea. There’s an old printing press I’ve been trying to get up and running. > Player: ...why?

Gold: Graham Bryant: I have always dreamed of subverting the colony’s periodicals and turning the Board’s own propaganda against them.

ICL-KS: Graham Bryant: I want to use it to print our message. To spread the word of the Iconoclasts to the people of Halcyon.

ICL: Graham Bryant: The printed word is a powerful tool, Captain. One that can be used to enlighten the masses and bring them into the light of the Eternal Truth.

ICL-Quest Only: Graham Bryant: To spread our message. Help it take root in the colony. I’ve had a request for replacement rollers, but they haven’t arrived yet.

ICL-No Knowledge: Graham Bryant: Ah. To spread the word, of course.

SL-KS: Graham Bryant: Because the Van Noys sent me. They didn’t want to pay me until I got the press running again.

SL: Graham Bryant: Well, the Eternal has a different plan. It’s called the Philosophist Plan.

Figure 8: [best viewed in color] Example generations by the in-context learning (ICL) DialogueWriter and supervised learning (SL) DialogueWriter. Models that condition on both biographical and quest passages are more likely to mention relevant quest information while also referencing game entities and their backgrounds without creating inconsistency, incoherence, or incongruity. See §C and §D for full documentation of quests and entities referenced.
Figure 9: Screenshot of interface shown to human annotators when choosing between a pair of generated graphs under various criteria.
Your job is to judge dialogue trees made by a 'writing copilot' for a video game. These dialogues guide a player through game quests by weaving together game details and lore.

Please read the following instructions carefully before evaluating the dialogue trees:

You’ll get details about quests objectives and game entities. Read these carefully as they’ll help you evaluate the dialogues. Remember, the same quests and characters will show up in other tasks.

The quest details will have a quest name, a high-level description, quest objectives active when entering the dialog and new quest objectives that the dialogue should introduce. They also have a walkthrough of what we should expect to happen during the dialogue.

We have included the details from previous steps of the quest for reference, though the dialogue does not need to reference them.

Important: When reading the dialogue, note that if the conversation returns to a node it previously visited, the corresponding character will not repeat the utterance. The conversation will continue on to any new child of the repeated node.

Determine which dialogue is better for 7 criteria:

Coherence: do the utterances in the tree create a realistic dialogue between the player character and the NPC? Make sure that the conversation between the player and the Non-Player Character (NPC) flows naturally and makes sense. Look out for parts that disrupt the flow. Identify nodes or edges that disrupt the flow. Sometimes, a dialogue might be very close to coherent but for a few structural issues that could be easily fixed by a game writer.

Violations: does the dialogue tree create contradictions with any of the sentences in the ontology or objective blurbs? Are there paths through it in which it contradict itself?

Important: it is ok for NPCs to make up information so long as they do not contradict the previous pieces of dialogue or the ontology.

Using the Game Lore: does the tree faithfully make of the bio sentences in the ontology, thereby exposing game lore about characters, groups, locations and items? Do the NPCs act in line with their character’s persona and background?

Important: If the NPCs make up information, it should NOT be considered a good use of the game lore—this criterium is about whether the NPCs use the game lore that they are given. However, if they do make up information, it shouldn’t penalize them unless it creates a contradiction.

Covering the Objectives: does the dialogue tree play out according to the objective sentences in the prompt? Does it cover all the desired options and responses? Does it give the player the chance to learn all they need to know about the next quest objective?

Content Suggestion: through generating multiple player utterance options at various turns, does the dialogue tree effectively propose potential dialogue subtrees that would espouse interesting content? If so, please note the topics in the comments.

Engagingness: does the dialogue tree hold your attention and make you want to hear more from the NPC?

Effect on the Game: By the (possibly multiple) ends of the dialogue, has the game state changed according to the desired specifications (the “blurb” section of current objectives and all details under “Player Should Have Learned”)? E.g. the player, if they chose the right options, has progressed in their current subquest, has acquired relevant items, and/or has achieved a desired effect on other characters.

Important: The dialogue tree may have multiple endings, so make sure to read through all of them before evaluating. Some of them might end the interaction early, which is fine as long as there are other endings that progress the quest.

Here are the game lore and quest details that the dialogue writer was given to write the dialogue:

{lore_and_objectives}

The dialogues are shown as linear sequences of nodes and edges between the nodes. Each node has a unique ID and a list of possible utterances. Each edge has a source node ID and a target node ID. Nodes with multiple outgoing edges are player choice nodes. The dialogue ends when a node has no outgoing edges.

**DIALOGUE 1:**

{dialogue_1}

**DIALOGUE 2:**

{dialogue_2}

Figure 10: Prompt shown to GPT-4 to elicit pairwise judgments for various criteria between two end-to-end generated graphs. To avoid an observed bias for the first dialogue, we prompt GPT-4 twice, swapping the order for the second. On disagreements, we prompt it a third time with its previous two judgments/explanations and ask for a final determination.
### Quest Name: A Family Matter

**Synopsis:** [0] Tucker Needham ran away from Stellar Bay a few weeks ago to join the Iconoclasts in Amber Heights. [1] His mother Agnes is willing to pay handsomely if you can locate her son and convince him to return.

**Walkthrough:** [0] You can begin this quest by talking to Agnes Needham in Stellar Bay, Monarch. [1] Agnes is by the town’s south-east exit, visibly shaken and calling for help. Hear her out and offer to find her son to being the quest.

**Objective 1: Look for Tucker Needham in Amber Heights**

**Game Log:** [0] Amber Heights is the settlement that houses the Iconoclasts on Monarch. [1] If Tucker Needham survived his travels, his mother thinks he’ll be there.

**Walkthrough:** [0] You find South from Stellar Bay and follow the east road. It will take you to Amber Heights. [1] Head up the hill and go into a residence on the left to meet Tucker Needham.

**Objective 2: Convince Tucker to Return Home**

**Game Log:** [0] Now that you’ve found Tucker Needham in Amber Heights, convince him to return home to his mother in Stellar Bay.

**Walkthrough:** [0] Introduce yourself, and then you can mention your surprise that this grown man is the "little boy" that ran away. You’ll earn 7500xp. [1] Explain to him that she made it sound as if he was a boy in danger, [2] and he’ll say she has been overprotective all her life. [3] and he is ready to live his life without her protection. [4] You can persuade (55) or intimidate (55) to expedite things and get him to go back, [5] or you can ask him what he wants to do about it. [6] The last option will have him tell you to report that he is dead. [7] You can express your concern about what that will do to Agnes, [8] and then either ask for something that would prove a body, or reject the proposition. [9] If you persuade him to go back, you’ll get 7500xp and can return to Stellar Bay to see things play out.

**Objective 3: Return to Agnes Needham in Stellar Bay**

**Game Log:** [0] You convinced Tucker Needham to return home to Stellar Bay. [1] Agnes promised a reward for bringing her son back.

**Walkthrough:** [0] You’ll find his mother is still condescending to him, [1] and you can help him by saying he’s a grown man. [2] You’ll get 7500xp. [3] If you stick around and talk to them some more you’ll see Tucker is standing up for himself. [4] You’ll receive 625 Bit Cartridge, Monarch Stellar Industries Reputation and 15000xp.

### Quest Name: The Commuter

**Synopsis:** [0] The Iconoclasts are due to receive a shipment of vital supplies from Carlotta, a sympathizer that resides in Stellar Bay. [1] The meeting is set to occur at the Bayside Terrace warehouse.

**Walkthrough:** [0] The quest can be obtained by asking Graham if there is anything that needs doing. [1] He is trying to get an old printing press running, but the replacement rollers he’d requisitioned haven’t arrived yet. [2] They were supposed to be delivered by Huxley, but she is still recovering and unable to make the delivery. [3] Graham asks you to go meet the supplier in her stead, and to pick up high-capacity data cartridges with the funds left over from the previous shipment. [4] Zora will interject to ask the player to buy food and medicine instead with the leftover money.

**Objective 1: Get the Printing Press Rollers from Carlotta**

**Game Log:** [0] Travel to the warehouse at Bayside Terrace and find Graham’s contact, Carlotta. [1] She should have a shipment for him. Retrieve it. [2] Speak to Carlotta.

**Walkthrough:** [0] Clear out the Sublight squad that is hunting Carlotta [1] Carlotta is behind a locked door to the east. [2] Activate the intercom next to the door to speak to her and she will unlock it. [3] Go inside and speak to her again to obtain the rollers needed to complete the quest, then choose between the high-capacity data cartridges or food and medicine.

**Objective 2: Get High-Capacity Cartridges or Extra Supplies from Carlotta**

**Game Log:** [0] You got extra supplies for Zora (or) You got High-Capacity Data Cartridges for Graham.

**Walkthrough:**

**Objective 3: Return to Graham**

**Game Log:** [0] Bring the needed parts back to Graham at Amber Heights.

**Walkthrough:** [0] Return to Graham and you’ll find him arguing with Zora about the Van Noys, a unit of the Iconoclasts that is MIA. [1] Inform Graham that you got his rollers, and food and medicine if that was your choice. [2] You’ll receive 7500xp and Zora will ask when the next drop is. [3] Inform her that Sanjar has made it illegal to trade with the Iconoclasts.

### Quest Name: Who Goes There

**Synopsis:** [0] The Groundbreaker’s Mardets have a bounty for a criminal on the run in the Groundbreaker’s Back Bays. [1] You’ve agreed to hunt down the unlawful Captain Gunnar MacRedd. [2] Return his lighter to Commandant Sanita to claim the bounty.

**Walkthrough:** [0] Now that you’ve found Tucker Needham in Amber Heights, convince him to return home to his mother in Stellar Bay.

**Objective 1: Hunt Down and Kill Captain McRedd**

**Game Log:** [0] Based on the bounty listing, Captain McRedd was last sighted in the Back Bays. [1] Head there and take him out.

**Walkthrough:** [0] You can find Captain McRedd in the Back Bays area of the Groundbreaker. [1] To get there head down the elevator in the promenade, [2] and you can’t miss him. [3] You can pass a Persuade (40) check to get him to put his gun down, [4] otherwise you’ll have to kill him and all his guards. [5] If you kill him he drops the Unique Weapon: Mongar. [6] You’ll get 6000xp and MacRedd’s Lighter. [7] If you persuaded him, use Perception to note it says "Sanita" on the lighter. [8] MacRedd will mention it was given to him by Sanita in remembrance of a ‘carnal understanding’ they had a few years back.

**Objective 2: Claim the Bounty’s Reward from Comdt. Sanita**

**Game Log:** [0] McRedd gave you his lucky lighter to give to Sanita. [1] Go turn it in to resolve his bounty.

**Walkthrough:** [0] Turn the lighter in to Commandant Sanita to claim the bounty.
That Came To Roseway

Amber Heights was besieged by a gang of pirates who ransacked the town and massacred all its inhabitants. [6] This tragedy was known as "The Amber Heights Massacre". [6]

Rat

Tucker was coddled by his mother from a very young age, [3] the latter insisting that danger lurked around every corner on Monarch. [4] His mother's overprotectiveness extended well into Tucker's adulthood, [5] leading him to seek to be free in any way possible. [6] After hearing Graham Bryant's broadcasts, Tucker left Stellar Bay to be truly free by joining the Iconoclasts at Amber Heights. [7] He is dazzled by Graham's preachings on true unfettered freedom from the corporate way of life and attributes his enthusiasm to his 'childhood trauma'. [8] He is willing to do anything to remain free, even faking his own death to prevent his mother from continuing to send people to look for him.

Entity: Raptidon
Appears in: A Family Matter
Bio: [0] Raptidons are giant catepillar-like creatures that inhabit various planets in Halcyon. [1] They are creatures native to Monarch. [2] however some corporations have illegally imported them to other planets, [3] such as Auntie Cleo who relocated a group of them to Roseway. [4] Raptidons are of corporate interest due to their potential for producing new chemical by-products which, [5] when refined, can be used to create new board approved products.

Entity: Sulfur Pits
Appears in: A Family Matter

Entity: Monarch
Bio: [0] Monarch, previously known as Terra 1, is one of the many moons of the gas giant Olympus and the site of a failed colony. [1] Terra 1 was initially designated as the primary colonization target of the Halcyon system. [2] The Halcyon Holdings Corporate Board had intended to completely terraform the moon, [3] wiping out the local fauna and flora and replacing it with plants and wildlife native to Earth. [4] However, the terraforming process unexpectedly caused the native species to mutate and grow to significantly larger sizes, [5] rendering them more dangerous and severely crippling the colonization effort. [6] Due to the hostile environment which they had created, [7] the Board was forced to enact a Hazard Clause covering the entirety of Terra 1. [8] Public notice of the clause’s issuance was sent to everyone operating on Terra 1 and led to the evacuation of almost all corporations from the moon. [9] However, one corporation took advantage of the chaos of the evacuation to exploit a legal loophole which allowed them to, [10] as the last corporation remaining on the planet, [11] acquire the planet from the Board. [12] This corporation, under the leadership of Sanjar Nandi and Graham Bryant subsequently rebranded itself to Monarch Stellar Industries (MSI), [13] in line with the renaming of the planet to "Monarch". [14] The actions of MSI earned them the ire of the Board, [15] who retaliated by effectively placing the moon under indefinite embargo, [16] refusing to allow legal transit either in or out. [17] The Board aggressively spread propaganda about Monarch to convince the rest of the population that it was both uninhabited and uninhabitable. [18] This has greatly hampered MSI’s attempts to be recognized as a legitimate corporation and is a thorn in the side of its CEO, Sanjar Nandi. [19] Monarch also has an ocean which goes around the moon at the "twilight band".
[20] It is where the colonists and Monarch Stellar Industries farm their saltuna.

Entity: Stellar Bay
Appears in: A Family Matter, Bolt With His Name, Candi's Cradle, Flowers For Sebastian, Herrick's Handiwork, Mr. Pickett's Biggest Game, Passion Pills, The Stainless Steel Rat
Bio: [0] Outside the city walls, the lands were overrun by the native wildlife, as well as marauders and outlaws. [1] Stellar Bay is a company town located on the planet Monarch. It is owned and operated by Monarch Stellar Industries. [2] Stellar Bay is the largest saltuna producer on the Halcyon colony and used to be one of the most important suppliers of this resource.

Entity: Fallbrook
Appears in: A Cysty-Dance With Death, Slaughterhouse Clive, Space-Crime Continuum, Spratkins

Entity: Cascadia
Appears in: Space-Crime Continuum, The Chimerists Last Experiment, The Ice Palace
Bio: [0] Cascadia is an abandoned company town that was owned and operated by Rizzo’s before it withdrew from Monarch. [1] It is now used as a stronghold by the Marauders. [2] The main attraction is the Cascadia Botting Plant and, [3] for those in the know, [4] the Rizzo Secret Laboratory hidden underneath the Rizzo Sweets Shoppe.

Entity: Amber Heights
Appears in: Little Memento, Odd Jobs, Sucker Bait, The Commuter
Bio: [0] Amber Heights is a location on the Monarch Wilderness and the base of operations for the Iconoclasts. [1] The Iconoclasts run the place somewhat like a commune. [2] Amber Heights was once the place of residence of the entire executive dome of Monarch Stellar Industries. [3] It is now in ruins after a massacre in the past. [4] They lived there with their families and it was the company’s operations center on Monarch. [5] Just after The Board approved the evacuation of the planet through the Hazard Clause, Amber Heights was besieged by a gang of pirates who ransacked the town and massacred all its inhabitants. [6] This tragedy was known as "The Amber Heights Massacre". [7] They were secretly assisted by MSI employee, Graham Bryant, who believed that the massacre would aid him in his quest to rid the colony of corporate influence. [8] In 2345, the same Graham Bryant formed the Iconoclasts and settled the group in the deserted town.

Figure 12: Example entity biographies that appear as constraining knowledge in KNUDGE quest dialogs
At the time of the quest, Tucker Needham is a former resident of Stellar Bay who left to join the Iconoclasts. [1] Before the quest A Family Matter, he can be found in Amber Heights. [2] Tucker was coddled by his mother from a very young age, [3] the latter insisting that danger lurked around every corner on Monarch. [4] His mother’s overprotectiveness extended well into Tucker’s adulthood, [5] leading him to seek to be free in any way possible. [6] After hearing Graham Bryant’s broadcasts, Tucker left Stellar Bay to be truly free by joining the Iconoclasts at Amber Heights. [7] He is dazzled by Graham’s preachings on true unfettered freedom from the corporate way of life and attributes his enthusiasm to his ‘childhood trauma’. [8] He is willing to do anything to remain free, even faking his own death to prevent his mother from continuing to send people to look for him.
Figure 14: Full dialogue tree in KNUDGE for motivating example in Figure 2.
Figure 15: Example full dialogue tree for dialogue *who_goes_there_01* in **KNUDGE**.
Figure 16: Example of longer dialogue tree in KNUDGE, containing numerous decision points, cycles and re-entrances.
Sanjar Nandi is the current CEO of Monarch Stellar Industries, based in Stellar Bay. Sanjar was ambitious but his attention to detail at the expense of big-picture thinking hampered his efforts within MSI. This led to negative performance reviews regarding his tendency to pad reports and talks with numbers and data. Despite his poor performance, Sanjar always showed himself to be a loyal employee of the company. Despite Sanjar's best efforts, he has found it extremely challenging to continue operating MSI on Monarch without the backing of the Board. In order to improve the lives of the people he is responsible for, Sanjar has a plan to rejoin the Board through the use of a BOLT-52 form and proof of another corporate presence on Monarch. He is simultaneously working on a plan to reorganize the Board, hoping that his plans are not found out until MSI has been reinstated. The latter started the Iconoclasts, a group dedicated to spreading the word of Philosophism throughout the galaxy, and Sanjar was left in Stellar Bay to run the company and look after the employees who were left behind. He can also tell you more about the planet, that used to be called Terra 1, and the reform that he and Monarch Stellar Industries tried to achieve to give more humane working conditions for everyone within. Celia Robbins is a middle manager for Monarch Stellar Industries and works with Sanjar Nandi at MSI Headquarters in Stellar Bay. Celia has a crush on Sebastian Adams and will buy whatever he has in stock as an excuse to talk to him. Unfortunately her apartment is filling up with exotic creature parts and her neighbors are starting to complain about the smell. She is not concerned that she and Sebastian may not have much to talk about, as everyone else in Stellar Bay either smells like saltuna or are her boss. The Stranger can offer to set her and Sebastian up on a date. DIALOG CONTEXT: Sanjar believes another company may be operating on Monarch illegally. If he can get proof, then he could use that as leverage to get MSI readmitted to the Halcyon Board. Talk to Sanjar after completing BOLT with His Name. The Board fact: The Board maintains a very tense relationship with MSI, owing to MSI’s democratic ideals and their declared ownership of Monarch. The Board fact: Depending on the actions of the Stranger, MSI may be compelled to rebel against the Board's authoritarianism.
Agnes Needham: I warned him. A raptidon would snap him up first chance it got. I just know one's ripped his arm off and is gnawing on his sweet little fingers!

Agnes Needham: He's been pining for an adventure, says he's tired of living cooped up behind the walls. But he doesn't understand how dangerous it is out there.

Player: Why would Tucker run away?

Agnes Needham: I've got some bits saved up for a rainy day. I'll give you every last one if you just bring my Tucker back to me. I won't even be mad at him running off. You tell him, I won't be mad.

Agnes Needham: Well, I... I guess I can't ask you to leave the town walls for free. It is deathly dangerous out there.

Player: Did you say there was a reward involved with this request?

Agnes Needham: Please! Won't you go and find my boy?

Agnes Needham: He ran out into the wilderness a few days ago. I warned him about the raptidons, mantisaurs, and marauders - the toxic sulfur pools, and poisonous plants. But he didn't listen!

Agnes Needham: Don't you tell me to calm down! I promised my boy I'd protect him for always. But how can I keep him safe if he's run away?

Player: Calm down. Take a deep breath, then tell me what happened.

Agnes Needham: Please, you have to help me get my little Tucker back! He ran away and is going to get himself killed! Oh, I just know a raptidon is melting him with acid as we speak.

Player: I heard a raptidon hiss. Take a deep breath, then tell me what happened.

Agnes Needham: Don't you tell me to calm down! I promised my boy I'd protect him for always. But how can I keep him safe if he's run away?

Agnes Needham: We ran out into the wilderness a few days ago. I warned him about the raptidons, mantisaurs, and marauders - the toxic sulfur pools, and poisonous plants. But he didn't listen!

Agnes Needham: Please! Don't you go and find my boy?

Player: Did you say there was a reward involved with this request?

Agnes Needham: No. It's not worth the risk. He's my only child.

Player: I heard a raptidon hiss. Take a deep breath, then tell me what happened.

Agnes Needham: I'll help you find your son.

Figure 18: Example In-Context Learning (ICL) prompt for GPT-3 based DialogueWriter
Agnes Needham fact: Agnes Needham is a resident of Stellar Bay and the mother of Tucker Needham.
Agnes Needham fact: Agnes' overprotective style of mothering has led her son, Tucker Needham, to run away from home so he can experience life.
Agnes Needham fact: Despite Tucker being 42 years old, she still thinks of him as her 'little boy'.
Tucker Needham fact: the latter insisting that danger lurked around every corner on Monarch.
utterance: > Agnes Needham: Oh, thank you for stopping! Everyone acts like nothing's wrong. Like my little boy isn't at risk of being eaten by some vile creature!

Raptidon fact: Raptidons are giant cat/reptile-like creatures that inhabit various planets in Halcyon.
utterance: > Agnes Needham: Please, you have to help me get my little Tucker back! He ran away and is going to get himself killed! Oh, I just know a raptidon is melting him with acid as we speak!

utterance: > Player: Calm down. Take a deep breath, then tell me what happened.

Agnes Needham fact: Agnes' overprotective style of mothering has led her son, Tucker Needham, to run away from home so he can experience life.
utterance: > Agnes Needham: Please! Won't you go and find my boy?
utterance: > Player: Did you say there was a reward involved with this request?

Stellar Bay fact: Outside the city walls, the lands were overrun by the native wildlife, as well as marauders and outlaws.
utterance: > Agnes Needham: I've got some bits saved up for a rainy day. I'll give you every last one if you just bring my Tucker back to me. I won't even be mad at him running off. You tell him, I won't be mad.
utterance: > Player: Why would Tucker run away?

Tucker Needham fact: the latter insisting that danger lurked around every corner on Monarch.
Tucker Needham fact: leading him to seek to be free in any way possible.
Stellar Bay fact: Outside the city walls, the lands were overrun by the native wildlife, as well as marauders and outlaws.
utterance: > Tucker Needham: He’s been pining for an adventure, says he’s tired of living cooped up behind the walls. But he doesn’t understand how dangerous it is out there.

utterance: > Agnes Needham: I warned him. A raptidon would snap him up first chance it got. I just know one’s ripped his arm off and is gnawing on his sweet little fingers!

utterance: > Player: I'll help you find your son.

Figure 19: Example In-Context Learning (ICL) prompt with CoT-style support knowledge selection
Thank you for helping us evaluate our automatic dialog writing system. We will show you a partial dialogue history and a set of possible candidate continuations of the dialogue. Please rate each of the following candidate continuations of the dialogue: (history)

(history)

Please rate each of the following candidate continuations of the dialogue: (utterances)

Please provide a short explanation of each score, highlighting reasons for low coherence/violation scores and/or high bio/objective scores.

For example, cite the part of the history that the utterance is responding to, or the part of the bio/objective that the utterance is using, or explain why it lacks coherence or violates the ontology/objectives.

Your output format is a serialized json item, one per line, one for each utterance. The items should have the following format: {'id': <utterance id>, 'coherence': <coherence score>, 'violation': <violation score>, 'bio': <bio fact usage score>, 'obj': <objective sentence usage score>, 'explanation': <explanation>}). Do not include anything else other than these items in your output. No other lines of text should be in your output.

Figure 20: Example node/edge list prompt item for GPT-4 based End-to-End DialogueWriter

Figure 21: Prompt template for GPT-4 approximation of human quality judgments.
Agnes Needham: Oh, thank you for stopping! Everyone acts like nothing's wrong. Like my little boy isn't at risk of being eaten by some vile creature!

Player: I'll help. What's the problem?

Agnes Needham: He's gone! I heard that he ran off to join some cult in the wilderness. The Iconoclasts, they call themselves.

Player: Amber Heights? What do they have to do with anything?

Agnes Needham: He's with the Iconoclasts. Those mad people who want to overthrow the Board. I knew he'd been listening to their broadcasts... I never imagined he'd join them.

Player: The Iconoclasts are a group of people living in Amber Heights, who are trying to overthrow the corporations.

Agnes Needham: He said something about some people named the Iconoclasts. He heard about them from some broadcasts on the radio. He said he wanted to experience true freedom.

Player: Do you think he survived the journey?

Agnes Needham: I don't know, but if he did, he's probably in Amber Heights. He's 42, but I still think of him as my little boy. Please, will you go look for him?

Player: Yes, I will look for him.

Agnes Needham: Thank you. If Tucker survived the journey, he'll be in Amber Heights. Please, find him and convince him to come home.

Player: I'll do what I can.

Agnes Needham: Thank you so much! I'll pay you handsomely if you find him and convince him to come back. Please, be careful.

END OF DIALOG

Figure 22: Example dialogue tree generated by the in-context learning knowledge selection DialogueWriter from just the input specifications and starting utterance. Human evaluators were tasked with comparing two such trees and choosing which performed better at a set of qualitative performance criteria. Dialogue follows the specification of the motivating example in Figure 2.
Figure 23: Example dialogue tree generated by the GPT-4-based End-to-end knowledge selection DialogueWriter. Tree shown alongside the quest objective details that should be covered.