MOTIVATION.

[question1] For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
[answer1] We seek to create agents that exhibit human-like capabilities such as commonsense reasoning and natural language understanding in interactive and situated settings. In pursuit of this goal, we provide a dataset that enables the creation of learning agents that can build knowledge graph-based world models of interactive narratives.

[question2] Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?
[answer2] It was created by Prithviraj Ammanabrolu and Mark Riedl at the Georgia Institute of Technology.

[question3] Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
[answer3] It was funded by the US's Defense Advanced Research Projects Agency (DARPA) as part of a fundamental science research grant Science of Artificial Intelligence and Learning for Open-world Novelty (SAIL-ON https://www.darpa.mil/program/science-of-artificial-intelligence-and-learning-for-open-world-novelty).

[question4] Any other comments?
[answer4] No.

COMPOSITION.

[question5] What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
[answer5] Each instance of our dataset takes the tuples of $hs_t; a_t; s_{t+1}; r_{t+1}i$ where s_t and s_{t+1} are two subsequent states of a text game with a_t being the action used to transition states and r_{t+1} is the observed reward for some step t. Everything is in text. These are all collected from various text games and examples of instances are found in Appendix A.1.

[question6] How many instances are there in total (of each type, if appropriate)?
[answer6] The training data has 24198 mappings and is collected across 27 games in multiple genres and contains a further 7836 heldout instances over 9 additional games in the test set.

[question7] Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).
[answer7] The dataset is a sample of the larger set of all possible states in each game.

The samples are made to be biased towards states near the walkthroughs required to finish a game.

[question8] What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

[answer8] Data is all in the form of text, either raw or in structured knowledge graph form.

[question9] Is there a label or target associated with each instance? If so, please provide a description.

[answer9] The data has multiple fields, depending on the tasks defined any of them can be used as labels. E.g. the knowledge graph prediction task has the graph field as the target.

[question10] Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

[answer10] Not all games have human readable attributes for objects—when they do not, these are omitted by leaving the attributes fields blank. All other data is present for all instances.

[question11] Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

[answer11] Instances are grouped together by game through the game field.

[question12] Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

[answer12] We provide a training split of 27 games, and a testing split of 9 games. These are selected on the basis of existing works and each split contains a diverse set of games in terms of genre.

[question13] Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

[answer13]

[question14] Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

[answer14] The creation of the dataset depends on the Jericho framework https://github.com/microsoft/jericho but the archival versions themselves do not have any dependencies.

[question15] Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
[answer15] No, all
data is part of games that are already public.
[question16] Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
[answer16] The data is collected from games
containing situations of non-normative language usage—describing situations that fictional characters may engage in that are potentially inappropriate, and on occasion impossible, for the real world such as running a troll through with a sword. Instances of such scenarios are mitigated by careful curation of the games that the data is collected from. The original Jericho framework [Hausknecht et al., 2020]—further verified by us in this work—uses a curated set of games found not to contain extreme examples of non-normative language usage. This is based on manual vetting and (existing) crowd-sourced reviews on the popular interactive narrative forum IFDB https://ifdb.org/.
If the dataset does not relate to people, you may skip the remaining questions in this section.
[SkipA] YES
[question17] Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
[answer17]
[question18] Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.
[answer18]
[question19] Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
[answer19]
[question20] Any other comments?
[answer20]

COLLECTION PROCESS.

[question21] How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from

other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

[answer21] We build off the popular text game simulator Jericho [Hausknecht et al., 2020], we have constructed a dataset dubbed JerichoWorld that maps text game state observations to both the underlying ground truth knowledge graph representations of the game and the set of contextually relevant actions that can be performed in that state.

[question22] What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

[answer22] To collect the $hs_t; a_t; s_{t+1}; r_{t+1}i$ tuples we implement a basic agent that explores the game along a trajectory corresponding to a game walkthrough. Game walkthroughs are texts describing the solutions to games, generally retrieved from the internet, but already part of the Jericho framework. Walkthroughs, however, only present one possible solution to a game and solve all the core puzzles required to complete a game with the maximum possible score. To achieve greater coverage of the game's state space, our data collection agent stops off to explore by executing random valid actions for n steps before resetting to the walkthrough.

[question23] If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

[answer23] Randomly sampled actions are based on a random seed in Python's random package https://docs.python.org/3/library/random.html. We provide a seed and the specific package version.

[question24] Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

[answer24] Only the authors were involved, building on the contributions of the Jericho developers.

[question25] Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

[answer25] This dataset was developed over a period of 6 months, though the games used within date back to the 1970s.

[question26] Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

[answer26] No human subjects were involved, no IRB process was undertaken.

If the dataset does not relate to people, you may skip the remaining questions in this section.

[SkipB] YES

[question27] Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?
[answer27]
[question28] Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
[answer28]
[question29] Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
[answer29]
[question30] If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
[answer30]
[question31] Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
[answer31]
[question32] Any other comments?
[answer32]

PREPROCESSING/CLEANING/LABELING.

[question33] Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
[answer33] Games were decompiled to extract attributes and ground truth knowledge graphs, the creation script is provided in the GitHub repo.
[question34] Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
[answer34] No, raw binary game states were not saved and were converted to human readable text.
[question35] Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

[answer35] Games were decompiled to extract attributes and ground truth knowledge graphs, the creation script will be provided in the GitHub repository.

[question36] Any other comments?

[answer36] No.

USES.

[question37] Has the dataset been used for any tasks already? If so, please provide a description.

[answer37] No.

[question38] Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

[answer38] No.

[question39] What (other) tasks could the dataset be used for?

[answer39] There are many more tasks that can be framed for other challenges related to world modeling from this dataset. Some immediate examples: (1) offline reinforcement learning for game agents through imitation learning—predicting the sequence of actions that finish the game based on walkthroughs and reward information; (2) knowledge graph verbalization, a form of the standard data-to-text natural language processing task [Wiseman et al., 2017], in which we learn to generate text that is conditioned on a knowledge graph; and (3) description generation conditioned on the names of various objects, locations, and characters—with applications in long-form text generation domains such as automated storytelling [Martin et al., 2018, Fan et al., 2019] and procedural generation of interactive narratives [Ammanabrolu et al., 2020a, Walton et al., 2020].

[question40] Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

[answer40] Users should keep in mind that these come from games and can potentially describe non-normative situations.

[question41] Are there tasks for which the dataset should not be used? If so, please provide a description.

[answer41] This dataset should not be used for tasks that involve direct physical interactions with humans, such as robotics

[question42] Any other comments?

[answer42] No.

DISTRIBUTION.

[question43] Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
[answer43] It is open-sourced.
[question44] How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
[answer44] The dataset will be open-sourced at https://github.com/JerichoWorld/JerichoWorld.
[question45] When will the dataset be distributed?
[answer45] It was first released in May 2021.
[question46] Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
[answer46] The dataset will be under an MIT license, this is indicated on the GitHub repository.
[question47] Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
[answer47] No.
[question48] Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
[answer48] No.
[question49] Any other comments?
[answer49] No.

MAINTENANCE.

[question50] Who will be supporting/hosting/maintaining the dataset?
[answer50] Prithviraj Ammanabrolu will be responsible for maintenance.
[question51] How can the owner/curator/manager of the dataset be contacted (for example, email address)?
[answer51] raj.ammanabrolu@gatech.edu or by filing an issue on the GitHub.
[question52] Is there an erratum? If so, please provide a link or other access point.
[answer52] No.
[question53] Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom,

and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?
[answer53] Yes, more games will be added and corresponding data will be collected. Previous versions will be kept for backwards compatibility.
[question54] If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
[answer54] No.
[question55] Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
[answer55] Yes, versions will be archived on the GitHub repository.
[question56] If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
[answer56] They can fork and submit pull requests to the current repository if they wish to extend it—these will be validated in an open-source manner on GitHub via reviews of the extensions.
[question57] Any other comments?
[answer57] No.