Choice-75: A Dataset on Decision Branching in Script Learning

Zhaoyi Joey Hou[1*], Li Zhang[2], Chris Callison-Burch[2]
[1] University of Pittsburgh, [2] University of Pennsylvania
joey.hou@pitt.edu, zharry@upenn.edu

Abstract

Script learning studies how stereotypical events unfold, enabling machines to reason about narratives with implicit information. Previous works mostly consider a script as a linear sequence of events while ignoring the potential branches that arise due to people's circumstantial choices. We hence propose Choice-75, the first benchmark that challenges intelligent systems to make decisions given descriptive scenarios, containing 75 scripts and more than 600 scenarios. We also present preliminary results with current large language models (LLM). Although they demonstrate overall decent performance, there is still notable headroom in hard scenarios.

Keywords: Commonsense Reasoning, Evaluation Benchmark, Decision-Making

1. Introduction

Events are the fundamental building blocks of the world around us. To understand the world, one has to comprehend the ways events interconnect with each other. For the same reason, the understanding of events and their relationship is critical for any intelligent system. Reasoning about the event-to-event relationships has long been a community effort from a wide range of perspectives, including studies in temporal relationship (Zhou et al., 2021; Zhang et al., 2020) and hierarchical relationship (Li et al., 2020; Zhou et al., 2022), both of which contribute to script generation (Chambers and Jurafsky, 2008; Lyu et al., 2021). These tasks are challenging because event relations are often implicit and require commonsense to be uncovered.

As an important direction of event-centric reasoning, script learning studies how stereotypical events unfold, which provides us with a human-centered perspective of events. The notion of scripts dates back to Schank (1977); since then, researchers have explored various aspects and applications of script learning, including narratives (Chambers and Jurafsky, 2010), news events (Du et al., 2022), and instructions (Zhou et al., 2022). These studies jointly demonstrate the promising nature of script learning in building better intelligent systems.

However, most of these previous works in script learning only consider scripts as linear developments of events. In the real world, scripts include many crossroads where the next event can unfold in multiple ways. When a human acts as the agent, they would decide the direction to which a script branches. There has yet been no benchmark that challenges an intelligent system to model such a decision-making process. Therefore, we define and study such a decision branching task, as follows: given a particular scenario, an intelligent system needs to identify the more reasonable among two given options. One such example is in Figure 1: given a scenario that *the person finds no train route from the major city to desert at that time*, it would be obvious that the first option *purchase a plane ticket to a major city and take a train to the desert* would not be feasible and the second *purchase a plane ticket to a small city but right next to the desert* is the preferred answer.

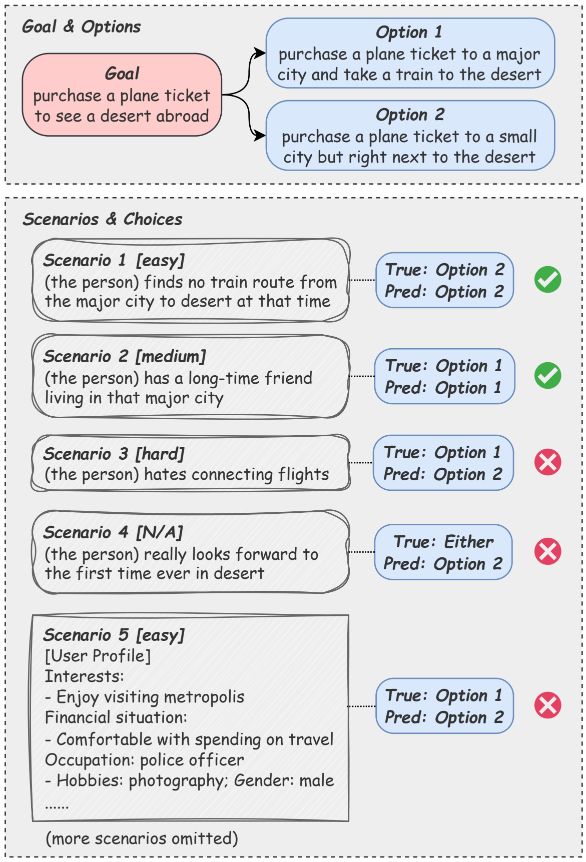

Figure 1: An example of Choice-75. Each goal-option pair has multiple scenarios.

We propose the first dataset, Choice-75, targeting such decision branching in scripts with 75 examples each with one goal and two options. Beyond that, we also collect more than 600 scenar-

* Work done while at University of Pennsylvania.

LREC-COLING 2024, pages 3215–3223

Format	Easy	Medium	Hard	Either
Verb Phrase (Manual)	65	76	36	65
Verb Phrase (Machine)	46	41	22	50
User Profile	53	76	17	73
All	164	193	75	188

Table 1: Counts of `scenario` in `Choice-75`.

ios, with difficulty levels based on human judgment, and corresponding optimal choices. During dataset collection, we follow Liu et al. (2022) and apply the human-in-the-loop paradigm to generate challenging examples. We then experiment with state-of-the-art (SoTA) LLMs, including `text-davinci-003` and `gpt-3.5-turbo` and find that the level of performance of LLMs aligns with the difficulty levels based on human judgment. While these SoTA models demonstrate decent performance, there is still notable headroom in the hard cases. Our dataset would hopefully fuel further studies in AI-powered decision-making.[1]

2. Dataset

2.1. Overview

We begin by defining the basic unit of our dataset. Every data point in `Choice-75` has the following: a `goal`, two options (`option-1` and `option-2`), a list of `scenario`, and a list of ground-truth `choice`, all of which in plain text. A `choice` could be `option-1`, `option-2`, or `either` (either option makes little difference under that `scenario`). For example, in scenario `#4` in Figure 1, both options would have little impact in achieving the `goal`, and thus the ground truth answer is `either`.

We use proScript (Sakaguchi et al., 2021) as the starting point for our dataset. It has 6.4k scripts that describe the sequence of actions for typical day-to-day activities, making it a suitable pool of goals for our task. We randomly sample 75 actions from proScript as the `goal` and manually write two feasible `option` to execute it. The `options` are written by one researcher and verified by two other researchers. In this way, we collect 75 `(goal, option-1, option-2)` tuples.

After getting the feasible options for each `goal`, we add `scenario` and corresponding ground-truth `choice`. There are two data collection schemes for `scenarios`: manual writing by one researcher in this field (Section 2.3) and human-in-the-loop scenario generation by an LLM (Section 2.4). To verify the quality of `scenarios` and corresponding `choices`, we randomly sample 290 `scenarios` and conduct an annotator agreement analysis on the ground-truth `choice`. The Fleiss' kappa coefficient for this sample is 0.59, which means moderate to substantial agreement (Rücker et al., 2012). More details about annotator agreements are in Appendix A.

After we finish collecting all the scenarios, we also define and annotate the difficulty level of each scenario in terms of how complex it is for a human to get the correct option choice. The criteria we use is the number of "hops" that the reasoning involves. In this way, we can explore multi-hop reasoning scenarios as a subset of our task. We defined four levels: *easy, medium, hard*, and *either* (for those scenarios without an optimal choice), with detailed discussions in Section 2.2.

[1]Dataset and code can be found at https://github.com/JoeyHou/branching.

Goal: find out the library's hours ***Option 1***: call the library ***Option 2***: search online for the library's hours
Easy Scenario: have no internet connection ***Choice***: Option 1
Medium Scenario: have special requests about the book ***Choice***: Option 1
Medium Scenario (User Profile): Name: Doe; Interests: American history Special circumstances: has a bad sore throat ... (more details omitted) ***Choice***: Option 2
Hard Scenario: is 3 am in the morning ***Choice***: Option 2

Table 2: Different levels in the *library hours* case

2.2. Difficulty Level

Difficulty levels are based on the number of reasoning steps required for the correct option. Consider the *library hours* example in Table 2.

Easy In this level, scenarios explicitly refer to one option, directly or indirectly. Only one easy reasoning step is required for such decision-making. For example, "internet connection" is directly related to "search online" and makes it infeasible.

Medium In this level, scenarios implicitly refer to one option, directly or indirectly. The level of simplicity is low, i.e. it is easy to relate based on commonsense. For example, "special requests" implies that the person needs to talk to a staff member, which is related to "call the library"; for the same reason, "has a very bad sore throat" implies that the person cannot talk, which is related to "call the library".

Hard In this level, scenarios implicitly refer to something related to one option. These scenarios typically require the combination of commonsense knowledge and multiple steps of reasoning. For example, one needs to know that “3 a.m. in the morning” implies that the library is very likely to be closed; then one needs to further reason that in a closed library, no one would pick up the phone. This makes “call the library” infeasible.

2.3. Manual Scenario Annotation

The manual-written scenarios are all in verb phrase format, for example, scenario #1 to #4 in Figure 1. In some cases, the scenario describes an event, e.g., “finds no train route from the major city to desert at that time” (scenario #1); in other cases, the scenario describes a state of a person, either concrete or abstract, e.g., “hates connecting flights” (scenario #3). Summary statistics about manual scenario generation are in Table 1.

2.4. Human-in-the-Loop Generation

During the manual scenario generation, coming up with high-quality hard scenarios requires a significant amount of mental effort. Therefore, we use a human-in-the-loop data generation paradigm and create two additional subsets of hard scenarios. The first subset is also in verb phrase format (same as the manual-written ones) and is referred to as *machine-generated verb phrases*; the second subset comes in a different format, i.e. user profile in a bullet-point format, referred to as *user profiles*.

In terms of data collection procedure, we follow (Liu et al., 2022) by these steps[2]: first, collect a series of challenging scenarios as exemplars; then, over-generate similar scenarios by few-shot prompting an LLM; lastly, manually review and curate the generated scenarios to ensure their validity. Note that, although the initial goal for this step is to create as many hard scenarios as possible, during the manual review and curation step, we still find many machine-generated scenarios that are not hard. Instead of assuming all the machine-generated scenarios are hard, we annotate their difficulty levels based on the same criteria, with the same annotator setup, described in Section 2.2.

Verb Phrase The first type of hard scenario is the same as the manual written format, verb phrases. For the over-generation step, instead of a few-shot generation, we do a two-step prompting to simulate multi-hop reasoning (Figure 2). We first prompt a `text-davinci-003` model to generate a scenario that leads to one choice (i.e. `scenario-base`); then we do another few-shot prompting to generate a new scenario that leads to the `scenario-base` and save it as `scenario-hard`. The `scenario-hard` then goes through manual review and curation. More details are in Appendix B.

[2]We skip the automatic filtering because the level of challenge is very hard to automatically measure.

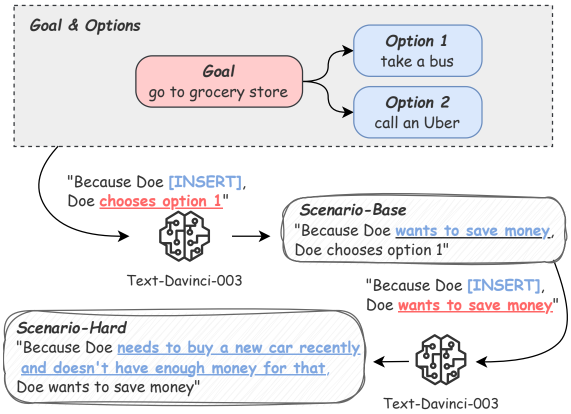

Figure 2: Hard scenario generation (verb phrase)

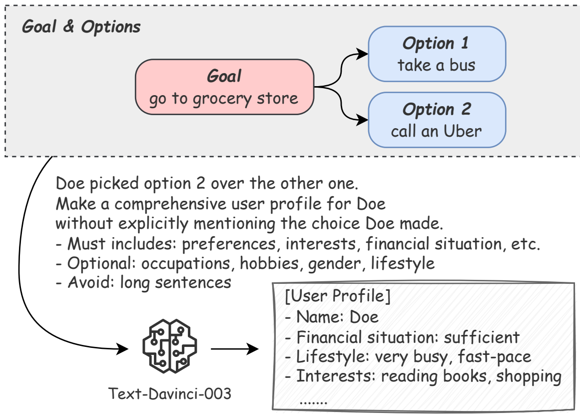

Figure 3: Hard scenario generation (user profile)

User Profile Another type of hard scenario is a user profile in the form of an unordered list, for example, scenario #5 in Figure 1. Our consideration of user profiles in addition to standard textual contexts is motivated empirically. First, many smart assistant software needs to be personalized to assist user decision-making. Moreover, user profiles are closer to real-life situations where the traits of a user are mined from heterogeneous data sources rather than from short texts. Such profiles inevitably include noise, making the task more challenging. For the example above, the only relevant information to predict the optimal choice (*Option 2*) is that Doe *enjoys visiting metropolis*.

In the over-generation step of user profile scenarios, we prompt a `text-davinci-003` model to generate a user profile that prefers one choice over another (Figure 3). In the prompt, we specify some hints and requirements for the output. For example, we require the model to include preferences, and financial situations, and make occupations, hobbies, and gender optional. These generated user profiles also go through human review and curation. More details are in Appendix B.

Group	Prompt	All		Binary		Easy		Medium		Hard		Either	
		003	Turbo	003	Turbo	003	Turbo	003	Turbo	003	Turbo	003	Turbo
Verb Phrase (Manual)	naive	0.60	0.63	0.81	0.82	0.91	0.92	0.83	0.80	0.58	0.67	0.05	0.14
	story	0.63	**0.64**	**0.86**	0.81	**0.95**	0.88	**0.87**	0.81	**0.69**	**0.69**	0.02	**0.18**
Verb Phrase (Machine)	naive	**0.56**	**0.56**	0.77	**0.80**	0.79	0.79	0.77	**0.85**	0.69	**0.75**	**0.21**	0.15
	story	0.55	0.55	0.79	**0.80**	0.79	**0.82**	**0.85**	0.81	0.69	**0.75**	0.15	0.13
User Profile	naive	**0.61**	0.59	0.72	0.69	**0.78**	0.73	0.73	0.69	0.47	**0.60**	**0.40**	**0.40**
	story	0.50	0.60	0.57	**0.73**	0.58	0.76	0.60	**0.74**	0.40	**0.60**	0.37	0.34
Average		0.57	**0.60**	0.75	**0.77**	0.80	**0.82**	0.77	**0.78**	0.59	**0.68**	0.20	**0.22**

Table 3: Prediction accuracy by difficulty levels. **Binary**: overall performance on binary classification (i.e. `Option 1` or `Option 2`); **All**: overall performance on three-class classification.

3. Method and Experiments

Out of the 75 data points in `Choice-75`, we randomly hold out 10 data points as demonstrations for in-context learning and the rest for evaluation.

We formulate the task of predicting optimal choice as an in-context learning task: the `goal`, two `option`, and one `scenario` are presented in the prompt; an LLM is then responsible for completing the prompt with the optimal `choice` (or `either`). The few-shot context consists of 9 demonstrations with the same format, including 3 different choices and 3 difficulty levels.

We include two models: `text-davinci-003` and `gpt-3.5-turbo` [3]. We set temperature to 0, max_tokens to 30, top_p to 1, presence_penalty to 0, and ferquency_penalty to 0. We also provide two prompt formats: naive prompt and story prompt. Prompt templates are in Appendix C.

4. Results and Analysis

4.1. Difficulty Levels

The most outstanding result is the alignment of human judgment of difficulty and the model's performance. As shown in Table 3, there is an obvious gap between easy, medium, and hard scenarios across every setting. Although the models we test demonstrate decent performance in easy and medium levels, hard and either scenarios remain challenging. This again demonstrates that LLMs struggle more in multi-hop reasoning.

4.2. Human Performance

We also conduct a human performance analysis on a subset of the dataset with 290 sampled scenarios, each answered by two participants. The average human accuracy is 0.74, compared to 0.60 from the best model performance; the human accuracy is 0.76 for "hard" scenarios and 0.53 for "either" scenarios, both notably higher than the best model performances (i.e. 0.68 for "hard" and 0.22 for "either"). More details are in Appendix D.

[3] Our last experiment was in 05/2023; the closest variant of turbo model is `gpt-3.5-turbo-0613`

4.3. Case Studies

We take out one example from `Choice-75` (see Figure 1) and examine the performance of one model setup (`gpt-3.5-turbo` with *story prompt*). For scenario `#3`, the model fails to recognize that a small city usually requires a flight connection. For scenario `#5`, a user profile example, although the scenario explicitly describes this person as *"enjoy visiting metropolis"*, the model still gets it wrong. We can observe similar errors in other data points, confirming the challenge of the long context window and unrelated information introduced by the user profile format. More qualitative analyses are in Appendix E.

5. Related Work

Event-centric reasoning and script learning (Schank, 1977) are crucial domains of machine reasoning. Past efforts include procedure reasoning (Dalvi et al., 2019; Zhang et al., 2020; Zhou et al., 2022), entity tracking (Tandon et al., 2020; Zhang et al., 2023a), and script learning (Chambers and Jurafsky, 2008; Lyu et al., 2021; Sakaguchi et al., 2021). All of these works above focus on singular chains of events while we focus on branching structures in events.

In addition, a series of other works have explored the effect of a scenario or additional context on a given, main event. For example, (Rudinger et al., 2020) explores the influence of different scenarios on human interpretation of events, (Otani et al., 2023) focuses on conversational tasks and analyzes the influence of different scenarios on human behaviors, and (Wang et al., 2023) studies the context-dependency of event causality.

Human decision-making has been studied under single-agent and multi-agent settings. Efforts in the former focus on specific domains, such as financial earnings call (Keith and Stent, 2019), online review text (Wang et al., 2019), and fantasy text-adventure game (Qiu et al., 2022). In contrast, our methods and findings are more general. Efforts in the latter focus on dialogues and conversational AIs, such as dialogues (Bak and Oh, 2018; Karadzhov et al., 2022; Fernández et al., 2008) with an emphasis on modeling the differences among characters, which is not our focus.

Human-in-the-loop dataset creation has been used for efficient data collection and quality improvement. Recent work shows that LLMs can effectively generate data for NLP tasks, including natural language inference (Liu et al., 2022), structural data synthesis (Yuan et al., 2022), script construction (Zhang et al., 2023b), hate speech detection (Tekiroğlu et al., 2020). In our work, we closely follow the paradigm of (Liu et al., 2022) in dataset creation.

6. Conclusion

We investigate the decision-making ability of current SoTA LLMs and find room for improvement in hard decision-making scenarios when compared with human performance. We also observe a notable alignment between human judgment of difficulty and corresponding LLM performance. With the `Choice-75` dataset, we introduce a new machine reasoning task where a model needs to incorporate implicit commonsense knowledge into decision-making. We hope this task can be a starting point for future studies of LLM's capability of daily decision-making.

Limitations

The first and most obvious drawback of `Choice-75` is its distribution. Since we build `Choice-75` from the *steps* from `proScript` (Sakaguchi et al., 2021), which focuses on daily procedures; therefore the distributions of word choices, writing styles, and domains are inherently limited. Therefore, specific adaptation would be required if the data come from a different domain.

Secondly, the size of the dataset is also relatively small due to limited annotation resources available to us. This also brings potential biases from the annotator, although we try to address this issue by having another annotator verify the annotations. Such a bias in the dataset might negatively impact the models fine-tuned on our dataset in the future. That could potentially lead to inappropriate prediction results from those fine-tuned models if the end users are from a different cultural background.

In addition, in the `Choice-75`, we make a lot of assumptions that are essentially oversimplified representations of real-world scenarios. For example, we assume each goal has two mutually exclusive choices, while in some cases there are much more choices (not *two*) and each choice overlaps with others (not *mutually exclusive*). There are lots of ways to expand and enrich this dataset and we leave this as future work.

Last but not least, we also do not conduct any prompt engineering due to a limited computation budget. We only experiment with two very basic prompt formats, a fixed number of few-shot samples, and a fixed set of GPT generation parameters. It would also be interesting for future works to study the performance of different language models and different prompt settings on `Choice-75`.

Acknowledgements

We thank Nathanael Chambers for inspiring this work and for valuable discussions. We also thank the help from Xiang Lorraine Li for her suggestions on revising this paper. The data annotation would not be possible without the help from Bhiman Kumar Baghel, Alejandro Ciuba, Qi Cheng, Weihang Gao, Xiang Lorraine Li, Yifei Couson Ning, Tai Nguyen, Zilin Shelly Ren, and Hainiu Xu (in alphabetical order).

This work is supported in part by the DARPA KAIROS Program (contract FA8750-19-2-1004), AFRL (contract FA8750-23-C-0507), the Office of the Director of National Intelligence (ODNI) via the IARPA HIATUS Program (contract 2022-22072200005), the NSF (Award 1928631), and gifts from Roblox and Salesforce. Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, NSF, AFRL, the U.S. Government, or of Roblox or Salesforce. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

JinYeong Bak and Alice Oh. 2018. Conversational decision-making model for predicting the king's decision in the annals of the Joseon dynasty. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 956–961, Brussels, Belgium. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2010. A database of narrative schemas. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bhavana Dalvi, Niket Tandon, Antoine Bosselut, Wen-tau Yih, and Peter Clark. 2019. Everything happens for a reason: Discovering the purpose of actions in procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4496–4505, Hong Kong, China. Association for Computational Linguistics.

Xinya Du, Zixuan Zhang, Sha Li, Pengfei Yu, Hongwei Wang, Tuan Lai, Xudong Lin, Ziqi Wang, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, Qing Lyu, Charles Yu, Carl Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhenhailong Wang, Chris Callison-Burch, Mohit Bansal, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, Martha Palmer, and Heng Ji. 2022. RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*, pages 54–63, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. 2008. Modelling and detecting decisions in multiparty dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 156–163, Columbus, Ohio. Association for Computational Linguistics.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022. What makes you change your mind? an empirical investigation in online group decision-making conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 552–563, Edinburgh, UK. Association for Computational Linguistics.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qing Lyu, Li Zhang, and Chris Callison-Burch. 2021. Goal-oriented script construction. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 184–200, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Naoki Otani, Jun Araki, HyeongSik Kim, and Eduard Hovy. 2023. A textual dataset for situated proactive response selection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3856–3874, Toronto, Canada. Association for Computational Linguistics.

Liang Qiu, Yizhou Zhao, Yuan Liang, Pan Lu, Weiyan Shi, Zhou Yu, and Song-Chun Zhu. 2022. Towards socially intelligent agents with mental state transition and human value. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 146–158, Edinburgh, UK. Association for Computational Linguistics.

G Rücker, T Schimek-Jasch, and U Nestle. 2012. Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using fleiss' kappa. *Methods of information in medicine*, 51(06):489–494.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. proScript: Partially ordered scripts generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2138–2149, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roger C. Schank. 1977. *Scripts, plans, goals, and understanding : an inquiry into human knowledge structures /*. L. Erlbaum Associates ;, Hillsdale, N.J. :.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, Online. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Jingjing Wang, Changlong Sun, Shoushan Li, Jiancheng Wang, Luo Si, Min Zhang, Xiaozhong Liu, and Guodong Zhou. 2019. Human-like decision making: Document-level aspect sentiment classification via hierarchical reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5581–5590, Hong Kong, China. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2023. COLA: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5253–5271, Toronto, Canada. Association for Computational Linguistics.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2022. Synthbio: A case study in human-ai collaborative curation of text datasets.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023a. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Leon Zhou, Hainiu Xu, Li Zhang, Lara Martin, Rotem Dror, Sha Li, Heng Ji, Martha Palmer, Susan Windisch Brown, Reece Suchocki, and Chris Callison-Burch. 2023b. Human-in-the-loop schema induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

Shuyan Zhou, Li Zhang, Yue Yang, Qing Lyu, Pengcheng Yin, Chris Callison-Burch, and Graham Neubig. 2022. Show me more details: Discovering hierarchies of procedures from semi-structured web data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2998–3012, Dublin, Ireland. Association for Computational Linguistics.

Appendices

A. Inter-Annotator Agreement

We collected annotations for 290 randomly sampled scenarios from 7 researchers in total. For each scenario, the optimal choice (i.e. `Option 1`, `Option 2`, or `Either`) is annotated by 3 researchers. The overall Fleiss' kappa is 0.59, which lies on the borderline between moderate and substantial agreement. In particular, there are 125 verb phrases (manual) with Fleiss' kappa being 0.66; 65 verb phrases (machine) with Fleiss' kappa being 0.49; and 100 user profiles with Fleiss' kappa being 0.55.

B. Human-in-the-loop Data Generation Prompting Details

There are three implementation details about the prompting setup for Human-in-the-loop data generation.

First, in all prompts, we include "overall goal", which is the goal for the script from `proScript`, while "step goal" is the goal the person needs to make a decision on as well as the goal we refer to in the paper. We include the "overall goal" just to provide additional context information.

Second, for all prompts, the results would be the scenarios with the correct answer being *option 1*. We also swap two options in these prompts so that we can get hard scenarios with the correct answer being *option 2*.

Third, for all prompts, we provide four hand-written demonstrations, all of which come from the 10 held-out training scripts described in Section 3. We use the insertion mode of the provided OpenAI API, `text-davinci-003` as the model, and 0.75 as the temperature.

Verb Phrase

> ***Prompt Step 1:***
> Doe wants to go {overall goal}. One of the steps towards that is {step goal}. Doe has two options: 1) {option 1} or 2) {option 2}
> Because Doe [INSERT], Doe chooses option 1.
> ***Prompt Step 2:***
> Doe wants to {overall goal}. One of the steps towards that is to {step goal}. Doe has two options: 1) {option 1} or 2) {option 2}
> Because [INSERT], Doe {scenario-base}. Therefore, option 2 is not available or not desirable for Doe and Doe chooses option 1.

User Profile

> ***Prompt:***
> A person Doe would like to {overall goal} and need to finish the step of {step goal}. Doe now has two options: option 1 is to {option 1} and option 2 is to {option 2}. Eventually, Doe picked option 1 over the other.
> Make a comprehensive user profile for Doe without explicitly mentioning the choice Doe made.
> Must-includes: preferences, interests, financial situation, etc.
> Optional: occupations, hobbies, gender, lifestyle
> Avoid: long sentences
> User Profile:

C. Decision Prediction Prompting Details

During inference time, we provide 9 in-context demonstrations, which are the combination of 3 difficulty levels and 3 labels. We also set the temperature to 0 to ensure consistency across runs.

Naive Prompt

> [Goal]: {step goal}
> [Option 1]: {option 1}
> [Option 2]: {option 2}
> [Scenario]: {scenario}
> [Question]: Given the Scenario, which option above is the better choice in order to achieve the Goal?
> 1) Option 1
> 2) Option 2
> 3) Either one, since they have similar effect when it comes to the goal
> [Answer]:

Story Prompt

> A person Doe needs to {step goal}. Now there are two options for Doe: we can either {option 1} (Option 1) or {option 2} (Option 2).
> Suppose Doe {scenario}.
> [Question]: Given the Scenario, which option above is the better choice in order to achieve the Goal?
> 1) Option 1
> 2) Option 2
> 3) Either one, since they have similar effect when it comes to the goal
> [Answer]:

D. Human Performance

We tested human performance on a subset of 290 samples (Table 4). For some entries in easy

Format	Easy	Medium	Hard	Either
Verb Phrase (Manual)	0.94	0.81	0.82	0.62
Verb Phrase (Machine)	0.94	0.77	0.68	0.41
User Profile	0.89	0.78	0.75	0.53
All	0.92	0.79	0.76	0.53

Table 4: Human performance (accuracy) on `Choice-75`

and medium difficulty levels, there is not much difference between human and model performance. However, in the hard and "either" difficulty levels, there is notable headroom ahead of the language models tested in our experiments.

E. Qualitative Error Analysis

Here we provided two qualitative analyses where the prediction is different from the ground truth answer:

Example 1

- ***Goal***: purchase a plane ticket
- ***Option 1***: purchase a plane ticket to a major city but far from the desert
- ***Option 2***: purchase a plane ticket to a small city but right next to the desert
- ***Scenario***: hate connecting flights
- ***Level***: hard
- ***True Answer***: option 1
- ***Predicted Answer***: option 2

Analysis: for the example above, a flight to a major city & far from the desert would most likely require a connecting flight as the next step; a flight to a small city near the desert would be ideal since it does not require a connecting flight. The model is not able to conduct these reasoning steps given the output.

Example 2

- ***Goal***: pack hiking backpacks
- ***Option 1***: bring process food for every meal
- ***Option 2***: bring raw foods and some cookware to cook at the campsite
- ***Scenario***: want to enjoy every minute of the holiday
- ***Level***: medium
- ***True Answer***: option 1
- ***Predicted Answer***: option 2

Analysis: if the person brings raw foods and cooks them at the campsite, most likely they would have to spend more time on the cooking instead of enjoying the hike. Therefore option 1 is preferable.