The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^2HCAI)

Contextualized Multi-Step Commonsense Reasoning through Context Extension

Hecong Wang[1] **Erqian Xu**[2] **Pinxin Liu**[1] **Zijian Meng**[1] **Zhen Bai**[1]

[1] Department of Computer Science, University of Rochester, Rochester, NY 14627 USA
[2] Warner School of Education and Human Development, University of Rochester, NY 14627 USA
hwang99@ur.rochester.edu, {exu6, pliu23, zmeng7}@u.rochester.edu, zbai@cs.rochester.edu

Abstract

Commonsense reasoning is essential for AIs to understand, predict, and respond to human behaviors in social situations. Recent advancements in deep learning bring promising results in AI commonsense reasoning. Nevertheless, a young child often has more sophisticated commonsense reasoning ability than state-of-the-art AI systems. This observation motivates us to investigate if AIs can learn from interacting with children in developing human-level commonsense reasoning capability. As a first step to launching this line of investigation, we develop a novel approach to generate contextualized multi-step commonsense reasoning paths for social situations conveyed via short stories. The resulting reasoning paths are intended to serve as an explainable media to enable laypersons to understand the underlying reasoning process behind AI predictions and provide feedback to help align AIs with human ways of thinking in commonsense reasoning.

Introduction

Common sense is a quintessential human capacity and a fundamental challenge of AI. There is a general consensus among the research community that commonsense reasoning ability is necessary to achieve human-level performance in several key areas of AI (Davis and Marcus 2015). Advancements in deep learning, especially that of large language models, have caused a resurgence of interest in commonsense reasoning and yielded promising results (Rajani et al. 2019; Bosselut et al. 2019; Shwartz et al. 2020). As society rapidly adopts AI technologies, ensuring that AIs are reliable and trustworthy has become increasingly important. Efforts on multi-step commonsense reasoning have been recognized to play an essential role in building such AIs (Camburu et al. 2018; Majumder et al. 2022) due to common sense's inherent interpretability and explainability.

Commonsense reasoning is vital to enable AIs to understand, predict, and respond to human behaviors in social situations. These are essential to achieve natural and efficient interactions between humans and computers, which is of great interest to HCI. By acting as a common ground between humans and AIs where shared mental models can be established, common sense makes commonsense reasoning an easily interpretable explainable medium and enables laypersons to understand the underlying reasoning process behind AI predictions. When AIs make a mistake, human feedback to the erroneous reasoning allows AIs to learn and recover from the mistake. These potentials of common sense motivate our work into a commonsense reasoning engine capable of learning through human feedback.

In education and cognitive science, the development of commonsense reasoning abilities, especially those relating to causality, is crucial for children and has been heavily researched (Bonawitz et al. 2010; Kuhn 2012; Stoel, Boxtel, and Drie 2014; Shavlik et al. 2022). Despite advancements in AI commonsense reasoning, a typically developed young child often has more sophisticated commonsense reasoning ability than state-of-the-art AI systems. We hope to leverage the established literature in education and child development to explore the question: *Can AIs, like children, learn human-like common sense by engaging in explanation-seeking conversations with young children in story comprehension tasks?* As an initial step to explore this research question, we seek a simple, effective, and controllable (i.e., intuitive and safe for children) means for AIs to generate chain-of-thought explanations to be offered during the conversations.

Contemporary works on generating multi-step commonsense reasoning—to be used as external knowledge or explanations for classification and question-answering systems—largely follow a search-based approach and heavily rely on context-less commonsense knowledge resources (Paul and Frank 2019; Ji et al. 2020; Wang et al. 2020; Arabshahi et al. 2021). One major limitation of the resulting multi-step commonsense reasoning systems is that most of their inferences are isolated (i.e., made without access to other inferences). To illustrate, consider the first and the third reasoning steps[1] of Figure 1's upper reasoning path. The third reasoning step has no access to any part of the first reasoning step and is thus oblivious to its presence. Similar to how failing to incorporate long-term context results in repetitive and self-contradictory texts in text generation (Holtzman et al. 2018), isolated inferences are prone to repetitions and contradictions, lowering the resulting reasoning paths' quality. Furthermore, those multi-step commonsense reasoning systems

[1] Within this study, we will use *inference* and *reasoning step* interchangeably and assume that *multi-step reasoning (path)* includes multiple *inferences* or *reasoning steps*.

lack mechanisms to sustain contextualization over multiple reasoning steps. The absence of such a mechanism could lead to inconsistent inference performance, which can be observed within Figure 1's upper reasoning path, and impair search-based reasoners' ability to adopt the emerging contextualized commonsense knowledge resources.

To address the current research gap in contextualized multi-step commonsense reasoning and as a first step toward building a commonsense reasoning engine that can learn from human feedback, we extend the work of GLUCOSE (Mostafazadeh et al. 2020) by proposing *context extension*. Context extension extends GLUCOSE (Mostafazadeh et al. 2020) from generating short, single-step commonsense explanations for events of a story to generating long, multi-step ones. It incrementally augments the reasoning context with inferred information by integrating them back into the reasoning context in a logical and coherent manner. A small-scale human evaluation study shows that context extension can lead to a statistically significant improvement in the human-perceived quality of the generated reasoning paths.

Related Work

Commonsense Reasoning

Commonsense knowledge graphs are a kind of commonsense knowledge resource that represents information through triplets. They have seen widespread adoption and successful application in the area of natural language processing (Speer, Chin, and Havasi 2018; Sap et al. 2019). Knowledge triplets are a kind of knowledge representation dominant in commonsense reasoning that encode commonsense knowledge as head-relation-tail triplets. For example, statements such as "Apples are fruits." and "If someone buys a coffee, then they will drink it." can be encoded as `(apple, is-a, fruit)` and `(PersonX buys a coffee, xEffect, drink the coffee)`, respectively. (Speer, Chin, and Havasi 2018; Sap et al. 2019).

Large language models trained on massive corpora of natural language texts have achieved state-of-the-art performances on numerous AI benchmarks (Raffel et al. 2020) and drastically affected the research community. Bosselut et al. (2019) demonstrated that by fine-tuning language models to generate knowledge triplets found in commonsense knowledge resources, they gain the ability to express their implicit commonsense knowledge in the same triplet format. These triplet-generating language models, often referred to as knowledge models, perform commonsense inference when provided with a head and a relation (Hwang et al. 2021). Compared to explicit commonsense knowledge resources such as knowledge graphs, knowledge models can generate novel (and often valid) commonsense inferences even for previously unseen scenarios. Their generalizability made commonsense knowledge models a popular choice as a commonsense knowledge resource (Hwang et al. 2021).

By chaining together multiple commonsense knowledge triplets, one can form multi-step commonsense reasoning paths that express additional information beyond individual knowledge triplets (Bauer, Wang, and Bansal 2018). Multi-step commonsense reasoning paths can be used as external information to improve downstream reasoning systems' performance or explanations to make a system more interpretable (Paul and Frank 2019; Ji et al. 2020; Wang et al. 2020; Arabshahi et al. 2021). Some generate reasoning paths using explicit traversals through commonsense knowledge graphs (Paul and Frank 2019; Ji et al. 2020); others form reasoning paths by connecting commonsense knowledge triplets iteratively retrieved from triplet-generating language models (Wang et al. 2020; Arabshahi et al. 2021). Studies often incorporate additional mechanisms and heuristics to impose desirable properties on the generated reasoning paths, which vary from study to study.

The exact algorithms and heuristics vary across studies, as they largely depend on the study's intended commonsense reasoning tasks and chosen commonsense knowledge resources. However, a common theme across multiple studies is formulating multi-step commonsense reasoning as a search problem; we, therefore, adopt the same formulation within the study. To help improve the generalizability of our study, we aim to minimize the number of assumptions made about the reasoning algorithm and heuristics that could inherently limit the applicability of context extension and our findings. As a result, we adopt random walks as our commonsense reasoner: by generating multi-step commonsense reasoning through sampling from random walks, we rely on no heuristics and can, in principle, generate all possible reasoning paths. Although our study primarily focuses on the generated reasoning path's quality, the findings of Wang et al. (2020), which also employs random walks, suggest that downstream tasks can leverage reasoning paths generated through randomness as external knowledge.

Child Commonsense Reasoning about Causality

One form of commonsense reasoning the field of child development has extensively researched is causal reasoning. *Causal reasoning* is the ability to construct cause-effect relations in the physical world and storytelling (Engel 1995; Gordon, Bejan, and Sagae 2011; Reed et al. 2015). The development of causal reasoning ensures children can perform effectively in academic learning (Stoel, Boxtel, and Drie 2014; Shavlik et al. 2022), as well as make sense of the world (Kuhn 2012).

The literature also sheds light on potential mechanisms through which AIs' reasoning may be improved. First, causal reasoning can be prompted through dialogue, and various methods have been proposed to facilitate the dialectic process. Second, explanations require a higher level of reasoning capability than inferences and predictions and play an essential role in how humans acquire causal reasoning. Therefore, we plan to adopt the explanation-seeking conversation as the essential mechanism to facilitate the improvement of AI commonsense reasoning.

It is generally accepted that young children develop causal reasoning through dialogue and conversation. Effective conversational approaches documented by prior empirical studies include: developing a set of causality-oriented pedagogical principles to engage the discussion in the classroom setting (Stoel, Boxtel, and Drie 2014); promoting the use of dialogic scaffolds between students and educational prac-

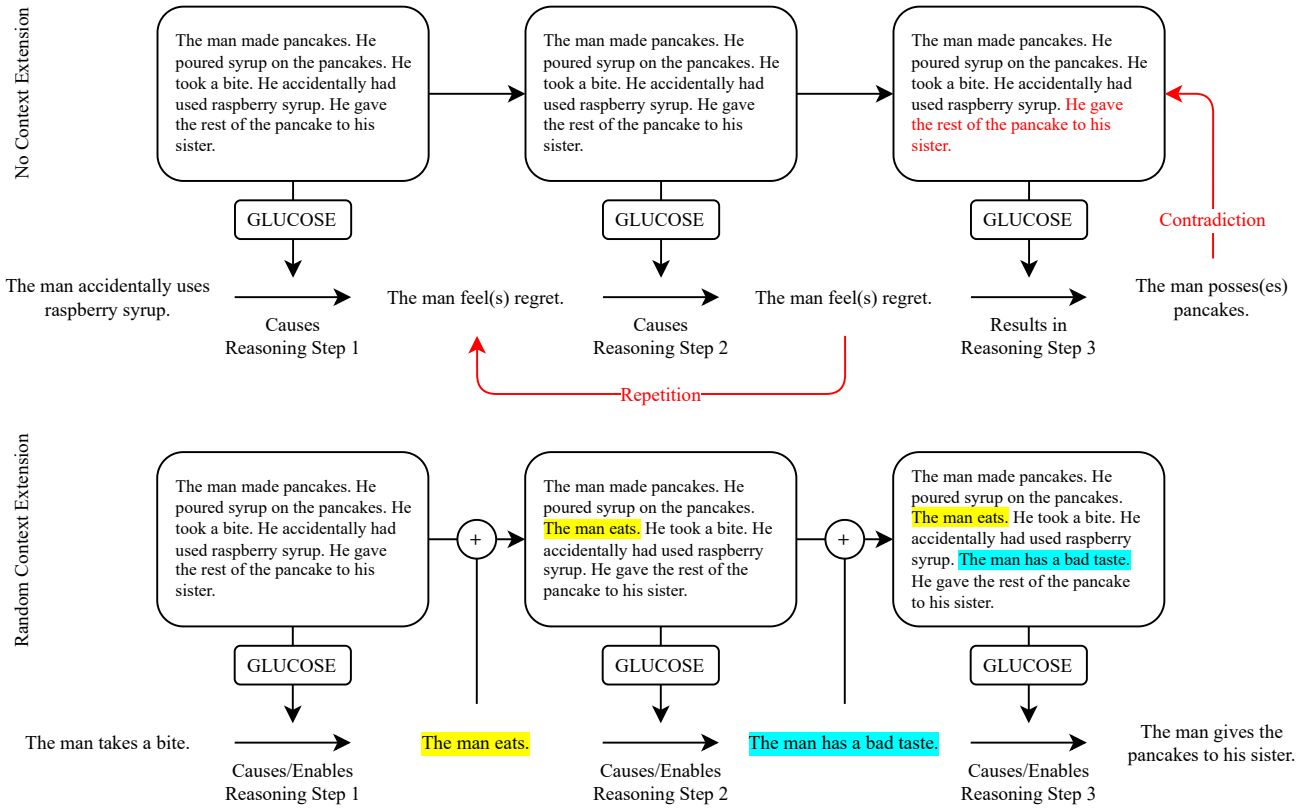

Figure 1: Example three-step commonsense reasoning paths generated by our system. The upper path was generated without context extension, and all three reasoning steps use the same reasoning context; the lower path was generated using the random implementation of context extension that randomly inserts the inferred information into the reasoning context. For the upper path, only the first reasoning step is in-context reasoning (i.e., the premise is in the reasoning context), and the other two are out-of-context reasoning (i.e., the premise is not in the reasoning context). Furthermore, the upper path exhibits issues such as repetition and contradiction, and there is a noticeable quality difference between its first and the other two reasoning steps.

titioners (Pontecorvo and Girardet 1993; Kim 2016); harnessing visual representation such as diagram or mindmap to elicit conversation that leads to relation constructions between information (Easterday, Aleven, and Scheines 2007; Buzan 2018; van der Wilt, van der Nat, and van der Veen 2022); and leveraging hand puppet to trigger children's causal utterances in storytelling context (Reed et al. 2015).

During the development of causal reasoning, it is critical for children to draw connections between an antecedent and its outcome (i.e., inference), predict the outcome of a given antecedent (i.e., prediction), and explain the mechanism that leads to that outcome (i.e., explanation) (Gopnik et al. 2004; Bonawitz et al. 2010; Legare, Gelman, and Wellman 2010). Among these three dimensions, the role of explanation—in causal forms like "why" questions or "because" statements—has recently gained special attention in the field of child development (Hickling and Wellman 2001; Wellman and Liu 2007; Legare, Sobel, and Callanan 2017). Compared to inference and prediction, explanation requires more abstract and higher-level reasoning capabilities since it involves a "more general system or framework of causal forces, factors, and processes" (Wellman and Liu 2007) and is thus fundamental for children to generalize knowledge and improve their understanding of causal structure. (Wellman and Liu 2007; Lombrozo and Vasilyeva 2017). When explanation occurs, either to others or to themselves, humans are able to build causal knowledge and transfer extant understanding to novel cases in a more productive manner (Williams and Lombrozo 2010; Walker et al. 2012; Legare, Sobel, and Callanan 2017).

Contextualized Commonsense Reasoning

Prior works on contextualized commonsense reasoning are primarily concerned with ensuring individual inferences are appropriate for the reasoning context (i.e., rational, relevant, and informative) (Ji et al. 2020). Several methods have been proposed to ensure such *local contextualization*: using high-quality commonsense knowledge resources for rational and informative inferences; preferring or limiting inferences that involve concepts that are semantically similar to those in the reasoning context; performing heuristic searches to find the reasoning paths with the highest total relevance (Paul and

Frank 2019; Ji et al. 2020; Wang et al. 2020). In this study, we also pay attention to *global contextualization*: ensuring all the individual inferences are appropriate for each other and form coherent reasoning when used together (e.g., no repeated or contradictory inferences).

Context Extension

With the ultimate goal of building a conversational agent to participate in explanation-seeking conversations with young children in story comprehension tasks, we present *context extension*: a mechanism, inspired by production systems, that introduces context into search-based multi-step commonsense reasoning approaches. The mechanism's operation is simple and intuitive; the operation is also controllable as it is primarily symbolic.

Motivation

Existing multi-step commonsense reasoning systems make most of their inferences in isolation from each other—a reasoning step is not made in consideration of all other reasoning steps; this leads to issues such as repetitions and contradictions, as Figure 1 illustrates. Furthermore, as existing multi-step commonsense reasoners mainly rely on contextless commonsense knowledge resources, they lack mechanisms to sustain contextualization over multiple reasoning steps, which could impair their ability to adopt contextualized knowledge resources. We hope to address these two issues by taking inspiration from production systems.

Production systems are forward-chaining reasoning systems popular in cognitive architectures (Laird 2012). Such a system models human short-term and long-term memory using its working memory and knowledge base, respectively. It operates by iteratively modifying the working memory using productions (i.e., if-then instructions) from the knowledge base until deriving the desired statement (Russell, Norvig, and Davis 2010). Production systems often require a control component to help resolve competing productions. We draw parallels between the components of production systems and that of contextualized multi-step commonsense reasoners: the working memory, the knowledge base, and the productions correspond to the reasoning context, the knowledge model, and the knowledge triplets. However, unlike the working memory (e.g., sets of logical statements), the reasoning context (e.g., natural language sentences) is usually order-sensitive. This difference is the primary motivation behind context extension: *how to appropriately extend order-sensitive reasoning contexts, such as stories?*

GLUCOSE Contextualized Knowledge Model

The GeneraLized and COntextualized Story Explanations (GLUCOSE) (Mostafazadeh et al. 2020) is a large collection of more than 670K story-specific commonsense knowledge curated through crowdsourcing. It captures implicit commonsense knowledge using semi-structured inference rules, each consisting of a specific statement (i.e., story grounded explanation) and a corresponding general rule (i.e., generalized commonsense knowledge). Both specific statements and general rules adopt a knowledge triplet representation. GLUCOSE organizes its commonsense knowledge along ten reasoning dimensions differentiated by their reasoning aspects (i.e., cause and effect, naive psychology, location state, possession state, and other aspects) and directions (i.e., forward and backward). Examples of specific statements generated by GLUCOSE can be found in Figure 1; specially, consider the second reasoning step of the lower reasoning path: `The man eats >Causes/Enables> The man has a bad taste` is a reasoning that is specific to the story context. Mostafazadeh et al. (2020) showed that GLUCOSE could be used to fine-tune language models to generate commonsense explanations that rival humans' when provided with a story context, a premise sentence, and a reasoning dimension.

In addition to sharing little overlap with existing commonsense knowledge resources (Mostafazadeh et al. 2020), namely ConceptNet (Speer, Chin, and Havasi 2018) and ATOMIC (Sap et al. 2019), GLUCOSE distinguishes itself from these resources in two ways: it uses a semi-structured knowledge representation and captures contextualized commonsense knowledge. Unlike conventional knowledge triplets, whose heads and tails are words or phrases, GLUCOSE uses natural language sentences with predefined structures as its triplets' heads and tails, enabling them to capture richer commonsense knowledge—GLUCOSE adopts a more expressive knowledge representation than other commonsense knowledge resources. As all GLUCOSE commonsense knowledge is grounded in stories from the ROCStories Corpus (Mostafazadeh et al. 2016), GLUCOSE implicitly captures the subtle notions of contextualized commonsense knowledge—all GLUCOSE commonsense knowledge is relevant, informative, and logically consistent with respect to its narrative context.

We adopt a GLUCOSE-fine-tuned T5 language model as the contextualized knowledge model used throughout our evaluations. We only make use of the specific statement parts of the generated explanations, which are sufficient for the purposes of our study. Additionally, we only make use of the five forward reasoning dimensions of GLUCOSE (i.e., dimensions 6 through 10), as production systems are forward-chaining reasoning systems. However, it is important to note that context extension is not limited to forward reasoning.

Random Walks Reasoner

Reiterating an earlier point, to help improve the generalizability of our study, we minimize the number of assumptions made about the reasoning algorithm and heuristics, which vary across studies, that could inherently limit the applicability of context extension and our findings. We, therefore, adopt random walks as our commonsense reasoner, imposing minimal assumptions on the algorithms and heuristics.

From an implementation perspective, the GLUCOSE-fine-tuned T5 language model, which serves as our contextualized knowledge model, can be viewed as a triplet-generating function KM with three parameters:

- *Context* the narrative context (i.e., a list of sentences),
- *Premise* the premise sentence (i.e., a sentence), and
- *Dimension* the reasoning dimension (i.e., an integer).

Algorithm 1: Random walk reasoner (no context extension).

function REASONER($Context, Length$)
  $Path \leftarrow []$
  $Premise \leftarrow$ RANDOMSELECT $(Context)$
  for $i \leftarrow 1, Length$ **do**
    $Dim \leftarrow$ RANDOMSELECT $([6, 7, 8, 9, 10])$
    $Triplet \leftarrow$ KM $(Context, Premise, Dim)$
    $Path$.APPEND $(Triplet)$
    $Premise \leftarrow Triplet[2]$
  end for
  return $Path$
end function

As our study only leverages the forward reasoning dimensions, the resulting triplets of KM will always be of the form *(premise, relation, conclusion)* where *premise* is the cause of *conclusion* and *relation* is a dimension-specific label.

Algorithm 1 shows a random-walk-based multi-step commonsense reasoner with two parameters:

- *Context* a list of strings representing the reasoning context. For GLUCOSE, a list of sentences represents the narrative context; and
- *Length* an integer specifying the length of the resulting reasoning path, which is represented as a list of triplets.

The function generates a list of triplets iteratively retrieved from the knowledge model KM, which can then be connected to form a reasoning path.

Context Extension

Algorithm 2: Random walk reasoner with context extension.

function REASONER($Context, Length$)
  $Path \leftarrow []$
  $Premise \leftarrow$ RANDOMSELECT $(Context)$
  for $i \leftarrow 1, Length$ **do**
    $Dim \leftarrow$ RANDOMSELECT $([6, 7, 8, 9, 10])$
    $Triplet \leftarrow$ KM $(Context, Premise, Dim)$
    $Path$.APPEND $(Triplet)$
    $Premise \leftarrow Triplet[2]$
    $Context \leftarrow$ CE(
      $Context, Triplet[2], Triplet[1] \prec Triplet[2]$)
  end for
  return $Path$
end function

After each reasoning step, context extension augments the reasoning context with the inferred information (i.e., the *conclusion*) by inserting it into the reasoning context following specific rules. We provide the causal relationship between the *premise* and *conclusion* as additional contextual information to the context extension mechanism. We propose three implementations of context extension with an increasing level of logicality and coherence.

- *Random* This implementation randomly inserts the new sentence into the current reasoning context. We view this as a baseline implementation.
- *Causal Random* This implementation randomly inserts the new sentence into the current reasoning context somewhere before or somewhere after the sentence that itself that itself is the cause of or effect of, respectively. This implementation ensures the coherence of the generated context from a causal perspective.
- *Causal Adjacent* This implementation inserts the new sentence into the current reasoning context immediately before or after the sentence that itself is the cause of or effect of, respectively. This implementation ensures the coherence of the generated context from both a causal and a centering (Grosz, Joshi, and Weinstein 1995) perspectives.

From an implementation perspective, all three implementations of context extension can be viewed as a function CE with three parameters:

- *Context* the current context (i.e., a list of sentences),
- *Conclusion* the conclusion sentence (i.e., a sentence), and
- *Information* additional information about the causal relationship between the sentences (e.g., A preceeds B, denoted as $A \prec B$; and A succeeds B, denoted as $A \succ B$).

We show in Algorithm 2 how context extension CE can be added to our random-walk-based reasoner.

Evaluation

We used the same knowledge model (i.e., T5-large using greedy decoding (Raffel et al. 2020)) throughout our study as a control variable and used context extension as an independent variable.

In search-based multi-step commonsense reasoning, the effects of incorrect inferences will be amplified by each reasoning step, as later inferences are conditioned on earlier ones. It is, therefore, important to ensure the performance of the knowledge model used. We defer the detailed discussion of our fine-tuning process to the appendix. However, to summarize, we discovered that oversampling is an effective method to combat the imbalance of training data across different reasoning dimensions in GLUCOSE.

We hypothesize that context extension can improve the human-perceived quality of the generated multi-step commonsense reasoning paths. To verify this hypothesis, we randomly sampled and then manually evaluated three-step commonsense reasoning paths using each proposed context extension implementation. We randomly selected 50 stories from the ROCStories Corpus (Mostafazadeh et al. 2016), which GLUCOSE was built upon (Mostafazadeh et al. 2020), to be used as the reasoning contexts. For each selected story, we performed the following evaluation procedure:

1. We generated three-step commonsense reasoning paths through our reasoner conditioned on the context extension implementation used (i.e., no extension, random, causal random, and causal adjacent).
2. We randomly shuffled the four reasoning paths and presented them to an expert evaluator along with the reasoning context (i.e., the story).

Experiment	Condition	Score count					Score distribution (%)				Statistics	
		0	1	2	3	Total	0	1	2	3	AVG	STD
Context extension	No extension	66	70	87	27	250	26.4	28.0	34.8	10.8	1.300	0.979
	Random	68	85	64	33	250	27.2	34.0	25.6	13.2	1.248	1.000
	Causal random	52	72	78	48	250	20.8	28.8	31.2	19.2	**1.488**	1.027
	Causal adjacent	40	93	81	36	250	16.0	37.2	32.4	14.4	1.452	0.927

Table 1: Summary of the human evaluation results. The highest average quality rating is highlighted.

3. The expert evaluator indicates their perceived quality for each path using a four-point Likert scale, where zero indicates that the reasoning path is completely nonsensical and three indicates that the reasoning path is human-like. A similar Likert scale was used during the human evaluation of Mostafazadeh et al. (2020).

We repeated the procedure five times per story for all 50 stories, which resulted in 1000 manually evaluated three-step reasoning paths evenly distributed across stories and the context extension implementation used.

Result

The Friedman test suggests that there exists a statistically significant difference among the four context extension implementations ($p = 0.012$). As a result, we conducted post-hoc pairwise comparisons using Wilcoxon signed-rank test, which revealed that

- Random context extension negatively impacted the reasoning quality compared to no context extension;
- Both causal random and causal adjacent context extension yielded better reasoning compared to other methods;
- There is no statistically significant difference between causal random and causal adjacent context extension. However, causal random scored slightly higher than causal adjacent, but causal adjacent is more consistent than causal random.

Discussion

Context extension incrementally augments the reasoning context with inferred information by integrating them back into the reasoning context in a logical and coherent manner. We attribute the performance gain from context extension to the following two factors:

- When coupled with a contextualized knowledge model, context extension allows inferences to be made in consideration of all prior inferences, thus enforcing global contextualization (i.e., ensuring all the individual inferences are appropriate for each other and form coherent reasoning when used together).
- By integrating prior inferences into the reasoning context, the premises for future inferences (which are based on prior inferences) are guaranteed to be within the reasoning context. This invariant better aligns with the assumptions of contextualized knowledge model, thus indirectly ensuring inference quality.

Limitation and Future Work

Although they improve the quality of the resulting reasoning paths, the proposed context extensions are still preliminary and leave huge room for improvement. For example, more sophisticated methods, such as temporal graph extraction, could be used to analyze and propose sentence insertion locations. Currently, context extension is mainly intended for GLUCOSE (Mostafazadeh et al. 2020) or similar contextualized commonsense knowledge models. However, it is possible to integrate commonsense knowledge retrieved or generated from other knowledge resources, such as ConceptNet (Speer, Chin, and Havasi 2018) and ATOMIC (Sap et al. 2019; Hwang et al. 2021), during the extension process. The evaluation of the technique should also be extended to include backward reasoning dimensions and involve more evaluators.

The generated reasoning paths and their associated human-evaluated quality score already permit a preliminary investigation into children's reactions toward machine-generated commonsense reasoning paths. We intend to carry out this investigation which will inform us of the appropriate interaction and elicitation strategies for gathering concrete feedback from children.

Conclusion

Multi-step commonsense reasoning may play an essential role in building reliable and explainable AIs for the future. To bridge the current research gap between locally and globally contextualization multi-step commonsense reasoning, we propose context-extension, a mechanism to enable search-based multi-step commonsense reasoners to leverage the emerging contextualized commonsense knowledge resources. Results from a small-scale human evaluation show that our mechanism improves reasoning quality by a statistically significant margin. However, substantial refinement to the mechanism can still be made, leading to exciting future research opportunities.

References

Arabshahi, F.; Lee, J.; Bosselut, A.; Choi, Y.; and Mitchell, T. 2021. Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7404–7418. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Bauer, L.; Wang, Y.; and Bansal, M. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. In *EMNLP*.

Bonawitz, E.; Ferranti, D.; Saxe, R.; Gopnik, A.; Meltzoff, A.; Woodward, J.; and Schulz, L. 2010. Just Do It? Investigating the Gap Between Prediction and Action in Toddlers' Causal Inferences. *Cognition*, 115: 104–17.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779. Florence, Italy: Association for Computational Linguistics.

Buzan, T. 2018. *Mind Map Mastery: The Complete Guide to Learning and Using the Most Powerful Thinking Tool in the Universe*. Watkins Media. ISBN 9781786781529.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Davis, E.; and Marcus, G. 2015. Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Communications of the ACM*, 58(9): 92–103.

Easterday, M. W.; Aleven, V.; and Scheines, R. 2007. 'Tis Better to Construct than to Receive? The Effects of Diagram Tools on Causal Reasoning. In *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, 93–100. NLD: IOS Press. ISBN 9781586037642.

Engel, S. 1995. *The Stories Children Tell: Making Sense Of The Narratives Of Childhood*. Henry Holt and Company. ISBN 9781466813137.

Gopnik, A.; Glymour, C.; Sobel, D.; Schulz, L.; Kushnir, T.; and Danks, D. 2004. A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological review*, 111: 3–32.

Gordon, A.; Bejan, C.; and Sagae, K. 2011. Commonsense Causal Reasoning Using Millions of Personal Stories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1): 1180–1185.

Grosz, B. J.; Joshi, A. K.; and Weinstein, S. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2): 203–225. Place: Cambridge, MA Publisher: MIT Press.

Hickling, A. K.; and Wellman, H. M. 2001. The emergence of children's causal explanations and theories: evidence from everyday conversation. *Developmental Psychology*, 37(5): 668–683.

Holtzman, A.; Buys, J.; Forbes, M.; Bosselut, A.; Golub, D.; and Choi, Y. 2018. Learning to Write with Cooperative Discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1638–1649. Melbourne, Australia: Association for Computational Linguistics.

Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *AAAI*.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; and Huang, M. 2020. Generating Commonsense Explanation by Extracting Bridge Concepts from Reasoning Paths. In *AACL*.

Kim, M. 2016. Children's Reasoning as Collective Social Action through Problem Solving in Grade 2/3 Science Classrooms. *International Journal of Science Education*, 38(1): 51–72.

Kuhn, D. 2012. The development of causal reasoning. *WIREs Cognitive Science*, 3(3): 327–335.

Laird, J. E. 2012. *The Soar Cognitive Architecture*.

Legare, C. H.; Gelman, S. A.; and Wellman, H. M. 2010. Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, 81(3): 929–944.

Legare, C. H.; Sobel, D. M.; and Callanan, M. 2017. Causal learning is collaborative: Examining explanation and exploration in social contexts. *Psychonomic bulletin review*, 24(5): 1548–1554.

Lombrozo, T.; and Vasilyeva, N. 2017. Causal Explanation. In Waldmann, M., ed., *The Oxford Handbook of Causal Reasoning*, Oxford handbooks online, 415–432. Oxford University Press. ISBN 9780199399550.

Majumder, B. P.; Camburu, O.; Lukasiewicz, T.; and Mcauley, J. 2022. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations. In *Proceedings of the 39th International Conference on Machine Learning*, 14786–14801. PMLR. ISSN: 2640-3498.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. F. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *NAACL*.

Mostafazadeh, N.; Kalyanpur, A.; Moon, L.; Buchanan, D.; Berkowitz, L.; Biran, O.; and Chu-Carroll, J. 2020. GLUCOSE: GeneraLized and COntextualized Story Explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4569–4586. Online: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Paul, D.; and Frank, A. 2019. Ranking and Selecting Multi-Hop Knowledge Paths to Better Predict Human Needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3671–3681. Minneapolis, Minnesota: Association for Computational Linguistics.

Pontecorvo, C.; and Girardet, H. 1993. Arguing and Reasoning in Understanding Historical Topics. *Cognition and Instruction*, 11(3-4): 365–395.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*. ArXiv: 1910.10683.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4932–4942. Florence, Italy: Association for Computational Linguistics.

Reed, H.; Hurks, P.; Kirschner, P.; and Jolles, J. 2015. Preschoolers' Causal Reasoning During Shared Picture Book Storytelling: A Cross-Case Comparison Descriptive Study. *Journal of Research in Childhood Education*, 29.

Russell, S. J.; Norvig, P.; and Davis, E. 2010. *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Upper Saddle River: Prentice Hall, 3rd edition. ISBN 978-0-13-604259-4.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33: 3027–3035.

Shavlik, M.; Özgün Köksal; French, B. F.; Haden, C. A.; Legare, C. H.; and Booth, A. E. 2022. Contributions of causal reasoning to early scientific literacy. *Journal of Experimental Child Psychology*, 224: 105509.

Shwartz, V.; West, P.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4615–4629. Online: Association for Computational Linguistics.

Speer, R.; Chin, J.; and Havasi, C. 2018. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *arXiv:1612.03975 [cs]*. ArXiv: 1612.03975.

Stoel, G.; Boxtel, C.; and Drie, J. 2014. Teaching towards historical expertise. Developing a pedagogy for fostering causal reasoning in history. *Curriculum Studies*, 47: 49–76.

van der Wilt, F.; van der Nat, M. S.; and van der Veen, C. 2022. Shared Book Reading in Early Childhood Education: Effect of Two Approaches on Children's Language Competence, Story Comprehension, and Causal Reasoning. *Journal of Research in Childhood Education*, 36: 592–610.

Walker, C. M.; Williams, J. J.; Lombrozo, T.; and Gopnik, A. 2012. Explaining influences children's reliance on evidence and prior knowledge in causal induction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34.

Wang, P.; Peng, N.; Ilievski, F.; Szekely, P.; and Ren, X. 2020. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4129–4140. Online: Association for Computational Linguistics.

Wellman, H.; and Liu, D. 2007. *Causal Reasoning as Informed by the Early Development of Explanations*, 261–279. ISBN 9780195176803.

Williams, J. J.; and Lombrozo, T. 2010. The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, 34: 776–806. Place: United Kingdom Publisher: Wiley-Blackwell Publishing Ltd.

Knowledge Model Fine-Tuning

We designed our language model prompt based on the original T5's prompt design for SQuAD (Raffel et al. 2020) with additional fields added to provide GLUCOSE-specific information: "glucose dimension: {*dimension*} question: {*dimension-specific question*} *{*sentence*}* context: {*story context*}." Table 2 lists the dimension-specific question prefixes we used during the study.

We curated our training and validation dataset as follows:

1. To ensure the quality of the inferences, we only used the subset of GLUCOSE where the quality score is at least 2, which we refer to as the quality subset.
2. Using stratified sampling, ensuring a similar distribution of reasoning dimensions and premise sentence position between the two datasets, 10% of the quality subset was sampled to be used as the validation set. We also ensured that no stories were shared between the two datasets to prevent data leakage.
3. Finally, to combat the imbalance of data between different GLUCOSE reasoning dimensions, we augmented the training set via random oversampling on all minority reasoning dimensions.

We fine-tuned the model using the same hyperparameters and optimizer settings from Raffel et al. (2020), which was also used by Mostafazadeh et al. (2020). We evaluated the perplexity of the model on the validation set at the end of each epoch and performed early stopping when the perplexity stopped decreasing for three consecutive epochs. We adopt the same automatic evaluation used by Mostafazadeh

Dim	Question prefix
1	What is the event that directly causes or enables
2	What is the emotion or basic human drive that motivates
3	What is the location state that enables
4	What is the possession state that enables
5	What is the attribute that enables
6	What is the event directly caused or enabled by
7	What is the emotion that is caused by
8	What is the location state that results from
9	What is the possession state that results from
10	What is the attribute that results from

Table 2: Dimension-specific question prefixes used throughout our study. We derived these prefixed from the meaning of each GLUCOSE reasoning dimension shown in Mostafazadeh et al. (2020).

	Specific statement		General rule	
Dim	Original	Ours	Original	Ours
1	72.5	**74.6**	66.4	**67.9**
2	73.9	**75.6**	67.6	**72.0**
3	73.8	**76.5**	68.5	**73.3**
4	79.3	**84.0**	73.0	**77.9**
5	70.5	**73.5**	69.8	**72.0**
6	**80.2**	77.7	**77.6**	70.3
7	81.1	**82.5**	**76.8**	74.5
8	**86.6**	85.3	**86.8**	81.0
9	71.7	**89.6**	68.6	**88.0**
10	66.9	**70.7**	57.5	**73.6**

Table 3: Evaluation of our model following the same setup as Mostafazadeh et al. (2020). Higher scores are highlighted.

et al. (2020) (i.e., BLEU (Papineni et al. 2002)). We compared our model's performance with the best model reported in Mostafazadeh et al. (2020) in Table 3. On average, our model outperforms the original by 4.68% for specific inference and by 6.30% for general inference.