DYNA-SKILL: Dynamic Self-Prompting Knowledge Graphs for Improving Logical Reasoning in Language Models

Anonymous submission

Abstract

Despite recent advances in large language models (LLMs), logical reasoning remains a challenging area, particularly for complex, multi-step reasoning in open-domain contexts. To address this, we introduce the Custom Graph Dataset, a novel graph-based knowledge resource designed to enhance LLMs' reasoning capabilities. Using a Self-Prompting mechanism, our approach automatically generates both predefined and dynamic relations, creating a dual-triple structure (Head-Relation-Tail and Tail-Dynamic Relation-Additional Tail) that supports richer multi-step reasoning. This Self-Prompting-driven process captures a broad and adaptable range of logical connections, combining predefined relational knowledge with dynamically generated, context-specific relations. Experimental results demonstrate that models fine-tuned on this dataset significantly outperform both baseline and control models, particularly on reasoning-intensive benchmarks like Commonsense QA, Riddle Sense, and ARC Challenge. Notably, the dataset includes 133 unique dynamic relations, such as *Analogous*, *Contextual*, and *Complementary*, which contribute to nuanced, context-sensitive reasoning. While general-purpose data offers benefits for some tasks, our findings validate that a targeted, logic-specific dataset can substantially improve LLMs' reasoning skills. This work underscores the potential of flexible, Self-Prompting-generated knowledge structures to advance LLM reasoning capabilities, suggesting future directions in combining structured and unstructured data to optimize inference.

Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in a range of natural language processing tasks, including question answering, summarization, and commonsense reasoning (Brown 2020). Despite these advances, logical reasoning remains a challenging frontier, as LLMs often struggle to perform complex, multi-step inferences, especially in open-domain and contextually rich scenarios(Bender et al. 2021). To address these limitations, researchers have increasingly turned to structured knowledge representations, such as commonsense knowledge graphs, which organize facts and relations to guide LLMs through more structured reasoning processes.

One prominent example of such a knowledge resource is ATOMIC(Sap et al. 2019; Hwang et al. 2021), a commonsense knowledge graph structured to support "if-then" reasoning across various human-centered scenarios. Following ATOMIC, COMET(Bosselut et al. 2019) introduced an approach to automatically generate commonsense knowledge by fine-tuning LLMs on predefined relational templates. However, both ATOMIC and COMET are limited by their reliance on manually defined relation types and a relatively narrow scope of head categories. This dependence on fixed relations restricts the knowledge graph's flexibility and limits the types of logical inferences it can support.

In this work, we propose a novel approach to automatically generating a graph-structured knowledge base that extends beyond the constraints of predefined relations. By leveraging a Self-Prompting mechanism(Li et al. 2022), our approach dynamically generates new relations and connections between events, resulting in a more expansive and nuanced dataset that captures a wider variety of logical relationships. This process not only enables the inclusion of predefined relations but also allows for the automatic generation of dynamic relations, enriching the dataset's logical structure without manual intervention. Each data point is represented as a dual-triple structure: (Head - Relation - Tail) and (Tail - Dynamic Relation - Additional Tail), providing a foundation for richer, multi-step inferences across diverse domains.

To evaluate the effectiveness of our generated dataset in enhancing logical reasoning, we conduct experiments on well-established benchmarks, including ARC Challenge(Clark et al. 2018), Commonsense QA(Talmor et al. 2018), HellaSwag(Zellers et al. 2019), QASC(Khot et al. 2020), and Riddle Sense(Lin et al. 2021). Additionally, we compare our dataset against a control dataset (CC News) to illustrate its specific contribution to logical reasoning, beyond general language understanding. Our experiments show that models fine-tuned on our dataset outperform baseline models in logical reasoning tasks, suggesting that a dataset specifically designed for logic and inference significantly improves model performance in these areas. Our dataset not only enables flexible and dynamic relation generation but is also scalable in terms of both size and domain, making it adaptable to various logical reasoning scenarios.

The main contributions of our work are as follows:

(i) We introduce a method for automatically generating a graph-based knowledge dataset that balances both predefined and dynamically generated relations, thereby increasing the flexibility and adaptability of the knowledge base.

(ii) We develop a dual-triple structure (Head-Relation-Tail and Tail-Dynamic Relation-Additional Tail) that supports multi-step inferences and encompasses a broader spectrum of logical relationships.

(iii) Through experiments on multiple benchmarks, we demonstrate that our dataset enhances logical reasoning performance in LLMs, surpassing both baseline and control models, and validating the effectiveness of our approach.

Related Work

Our work distinguishes itself by combining the strengths of structured commonsense knowledge graphs (such as ATOMIC and COMET)(Sap et al. 2019; Bosselut et al. 2019; Hwang et al. 2021) with the dynamic generation capabilities of self-prompting models, addressing the rigidity of predefined relation types. By leveraging self-prompting to generate novel relations within a structured graph, we provide a flexible and adaptable dataset tailored to logical reasoning, a gap not fully addressed by previous works in commonsense knowledge or open-domain reasoning.

Commonsense Knowledge Graphs

ATOMIC ATOMIC (Sap et al. 2019) introduced one of the first large-scale commonsense knowledge graphs tailored to support "if-then" reasoning. This knowledge base captures a range of human-centered scenarios, structuring data into categories such as intentions, reactions, and effects to provide contextually rich, structured commonsense knowledge. While ATOMIC has been instrumental in enabling commonsense reasoning in LLMs, it heavily relies on manual curation and predefined relation types. This limits its scope and adaptability when applied to open-domain logical reasoning tasks, where unforeseen relations might emerge.

COMET COMET (Bosselut et al. 2019; Hwang et al. 2021) advanced ATOMIC by utilizing transformer-based models to automatically generate knowledge for predefined relational templates. By fine-tuning on ATOMIC and ConceptNet, COMET can predict commonsense relations for new events, expanding the knowledge base without direct human intervention. However, COMET is similarly constrained by its reliance on fixed relations; the model is only able to generate knowledge within the scope of existing relation types, making it less adaptable to novel or dynamically evolving scenarios. Our work addresses this limitation by enabling LLMs to generate flexible, previously undefined relations, thereby broadening the scope of inference capabilities.

Self-Prompting for Flexible Relation Generation

The concept of Self-Prompting as a method to enable LLMs to generate contextually appropriate prompts and answers on-the-fly has shown potential for open-domain question answering (QA). "Self-Prompting Large Language Models for Zero-Shot Open-Domain QA"(Li et al. 2022) introduces a framework where LLMs autonomously generate questions and answers in a multi-step process, allowing the model to dynamically adapt to new information without predefined prompts. We adapt this self-prompting technique to our graph dataset by allowing LLMs to dynamically create new relations based on the content, bypassing the limitations of fixed relation types. This approach enables a dynamic relation-generation mechanism that enhances logical flexibility and adaptability in our dataset.

Method

In this study, we aim to create a graph-based knowledge dataset that enhances logical reasoning capabilities in language models. Using a Self-Prompting approach(Li et al. 2022) and leveraging the GPT-4-turbo API(Achiam et al. 2023), we automatically generate relational data points structured as Head - Relation - Tail - Dynamic Relation - Additional Tail. Specifically, the GPT-4-turbo API is employed to generate each component—Head, Tail, Dynamic Relation, and Additional Tail—ensuring a high level of contextual relevance and diversity in the data. Figure 1 provides an overview of the overall framework, showing the sequential steps in generating and utilizing this dataset. We then convert these data points into readable sentences to fine-tune language models. The methodology consists of the following steps:

1. **Head-to-Tail Generation**

   Head Definition: We define various Head entities across multiple categories to represent the main subjects of each logical reasoning event. These Heads are categorized by topics such as Social Interaction, Physical Entities, Events, and Causal Relations to cover a broad range of commonsense scenarios that facilitate logical reasoning. Specifically, we categorize Heads into Social-Interaction Relations, which includes activities related to daily life and social interactions, such as "Education," "Household," and "Relationship Management"; Physical-Entity Relations, encompassing everyday objects and tools like "Tools," "Vehicles," and "Appliances"; Event-Centered Relations, focusing on events like "Festivals," "Weddings," and "Public Gatherings"; and Causal Relations, which encompasses causes and conditions like "Economic Events," "Technological Failures," and "Climate Events." Additional categories include Causal Chain, Temporal Relations, Duration, Frequency, Direction and Movement, Conditional Relations, Necessary and Sufficient Conditions, Hierarchical Relations, Part-Whole Relations, and Quantitative Relations. Each of these categories captures unique logical structures and interactions, diversifying the logical contexts that models might encounter and providing a solid basis for a wide range of logical reasoning tasks.

   Relation Definition: Each Head category is associated with predefined relations, which guide the generation of Tail elements and ensure consistency across generated data points. These relations include context-specific types tailored to each category. Drawing upon insights from previous works like ATOMIC and COMET, we expand the range of predefined relations, enabling a richer and more varied relational structure than in prior studies. For example, Social-Interaction Relations include relations such as *xIntent* (intention behind an action),

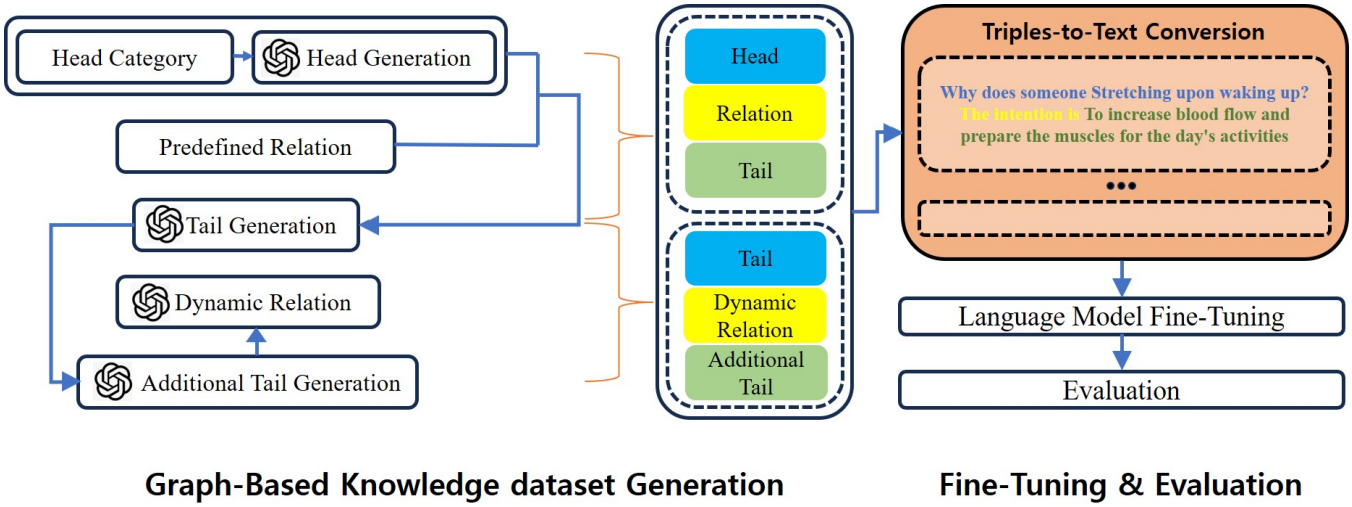

Figure 1: Overall Framework for Graph-Based Knowledge Dataset Generation and Fine-Tuning. In the dataset generation phase, Heads are generated from predefined categories, followed by Tail generation based on predefined relations and Heads. Additional Tails are then generated based on the Tail, and a dynamic relation is defined to capture the relationship between the Tail and Additional Tail, enabling multi-step logical connections. These elements are structured into a dual-triple format (Head-Relation-Tail and Tail-Dynamic Relation-Additional Tail), which is subsequently converted into natural language sentences for fine-tuning. In the fine-tuning and evaluation phase, these text-converted triples are used to fine-tune language models, which are evaluated on logical reasoning tasks to assess the effectiveness of the dataset.

xNeed (prerequisites for an action), and *oEffect* (impact on others), which help capture the interpersonal and motivational aspects of social interactions, allowing us to model complex social dynamics and motivational reasoning beyond the scope of previous commonsense knowledge graphs. Physical-Entity Relations involve relations like *ObjectUse* (the typical use of an object), *AtLocation* (where an object is typically found), and *CapableOf* (actions an object can perform), which describe the functional and situational properties of physical entities, adding layers of contextual knowledge essential for practical reasoning about objects. Event-Centered Relations include *IsAfter* (what happens after an event), *HasSubEvent* (sub-events within a main event), and *Causes* (what leads to an event), which support temporal and causal reasoning that builds upon, but goes beyond, the fixed relational templates found in models like COMET. Causal Relations involve causal connections captured through *Cause and Effect* and *Causal Chain*, describing cause-and-effect sequences that clarify outcome dependencies, enhancing the model's ability to perform complex inference that mirrors real-world reasoning patterns. For other categories, relations are specified based on context, such as *Temporal Sequence* in Temporal Relations, *If-Then Statements* in Conditional Relations, *Part-Whole Relations* for entities with internal structures, and *Quantities and Measures* in Quantitative Relations, allowing us to model temporal dependencies, conditional reasoning, and compositional structures within entities.

Each relation is associated with specific prompts to guide the generation of Tail responses. For instance, a *xIntent* relation for a social action Head may ask, "What is the possible intention behind this action?" These relations and prompts allow us to capture a rich variety of logical connections, including cause-effect relationships, hierarchical structures, and conditional dependencies. By building upon and extending the relation types defined in existing works, we provide a more extensive and versatile relational framework that enhances the depth and flexibility of the generated knowledge, making it particularly well-suited for logical reasoning tasks.

2. **Tail-to-Additional Tail and Dynamic Relation Generation**

   Additional Tail Generation: To expand upon the initial Tail, we use a Self-Prompting approach to generate an Additional Tail related to the existing Tail. This process deepens the logical connections by prompting the model with questions about further related actions or events.

   Dynamic Relation Generation: We dynamically define the relationship between the Tail and Additional Tail by prompting the model with "What kind of relationship does 'additional tail' have with 'tail'?" This method allows the model to generate previously undefined relations, adding flexibility to the dataset by incorporating novel and context-specific connections.

3. **Dual-Triple Structure: (Head - Relation - Tail) and (Tail - Dynamic Relation - Additional Tail)**

   Triple Separation: Each data point is structured as two distinct triples: (Head - Relation - Tail) and (Tail - Dynamic Relation - Additional Tail). This dual-triple structure enables multi-step reasoning by connecting events in layered logical relationships.

Multi-Layered Logical Representation: The dual-triple structure allows us to express complex, multi-step relationships beyond simple fact-based connections, enabling the language model to learn deeper logical reasoning capabilities.

4. **Text Conversion of Triples for Language Model Fine-Tuning**

   Triple-to-Text Conversion: After generating the (Head - Relation - Tail) triples, we convert each triple into a natural language sentence using a function designed to adapt each relation type into a specific sentence structure. For example, a triple such as (Head: "PersonX makes coffee", Relation: "xIntent", Tail: "to help") would be converted to "Why does someone make coffee? The intention is to help."

   Conversion Process: A custom function processes each triple according to its relation type, producing readable sentences. This function ensures that each triple is expressed as a coherent and contextually relevant sentence that is easy for the language model to interpret.

   Storing and Preparing Data for Fine-Tuning: The converted text data is saved line-by-line in a text file, which serves as the input for fine-tuning. This conversion enables the dataset to be directly utilized in fine-tuning language models, enhancing their logical reasoning capabilities through structured, narrative-like training data.

5. **Fine-Tuning Language Models on Converted Text Data**

   Fine-Tuning Setup: We fine-tune BERT, RoBERTa, DeBERTa, and DistilBERT models(Kenton and Toutanova 2019; Liu 2019; He et al. 2020; Sanh 2019) using the converted text data. Each model is trained to enhance its logical reasoning capabilities with our dataset, which provides explicit logical connections.

   Comparison with Control Dataset: To evaluate the specific contribution of our dataset to logical reasoning, we compare the performance of models fine-tuned on our custom dataset with those fine-tuned on a control dataset (CC News), which is expected to have limited impact on logical reasoning. By observing that models trained on CC News show a smaller improvement in logical reasoning tasks compared to those trained on our dataset, we demonstrate that our dataset effectively enhances reasoning capabilities in a way that general text data cannot. This comparison underscores the value of our graph-structured knowledge in fostering deeper inference abilities.

Experiment

Datasets

Custom Graph Dataset: The primary dataset used in this study is a graph-based knowledge dataset generated through the Self-Prompting method as described in the Methodology section. This dataset is structured with triples in the format (Head - Relation - Tail) and (Tail - Dynamic Relation - Additional Tail), which are then converted to natural language sentences. We utilize approximately 100,000 sentences for fine-tuning each model, aiming to assess the impact of this structured dataset on logical reasoning capabilities.

CC News Dataset: To evaluate the specific contribution of our custom dataset, we use a control dataset, CC News, primarily oriented towards general language understanding. Given its lack of a specific logical reasoning structure, the CC News dataset is expected to have limited impact on logical reasoning skills. Approximately 100,000 sentences from this dataset are used to ensure a consistent dataset size, allowing for a fair comparison with our Custom Graph Dataset.

Models

To examine the impact of the datasets across various model architectures, we fine-tune and evaluate four pre-trained language models with different capacities and configurations. We use BERT (Kenton and Toutanova 2019), a foundational transformer encoder known for its strong performance across diverse NLP tasks, and RoBERTa (Liu 2019), an optimized variant of BERT that incorporates improvements in pre-training strategies for enhanced performance. Additionally, we employ DeBERTa (He et al. 2020), which builds on BERT with disentangled attention mechanisms for more effective language understanding, and DistilBERT (Sanh 2019), a smaller and faster version of BERT, allowing us to assess the dataset's impact on lightweight models with fewer parameters.

Evaluation Tasks

To measure the effectiveness of the Custom Graph Dataset in enhancing logical reasoning, we evaluate each fine-tuned model across five established commonsense and reasoning benchmarks. These benchmarks include the ARC Challenge (Clark et al. 2018), a multiple-choice science exam dataset assessing complex reasoning abilities, and Commonsense QA (Talmor et al. 2018), a benchmark designed to evaluate commonsense knowledge through multiple-choice questions. We also test the models on HellaSwag (Zellers et al. 2019), a task where models must select the most plausible continuation of a given situation, testing commonsense and situational reasoning. Additionally, we use QASC (Khot et al. 2020), a question-answering benchmark focused on explanations and logical inference, as well as Riddle Sense (Lin et al. 2021), a dataset consisting of riddles that require lateral thinking and contextual reasoning for problem-solving.

Fine-Tuning and Evaluation Procedure

Fine-Tuning: Each model is fine-tuned using approximately 100,000 sentences from both the Custom Graph Dataset and the CC News Dataset. The fine-tuning process follows a standard language modeling objective, optimizing the models to understand and utilize the structural patterns and relationships within each dataset.

Evaluation Metrics: Model performance on each task is evaluated using accuracy, which serves as a straightforward metric for assessing each model's ability to select the correct answer. This enables direct comparison of the effectiveness

of different datasets and configurations in enhancing logical reasoning capabilities.

Comparison Settings

Two-Tiered Comparison Approach: Our experiment is structured as a two-tiered comparison to isolate and validate the contribution of the Custom Graph Dataset:

1. **Baseline vs. Custom Dataset Comparison:** The first comparison assesses logical reasoning improvements achieved by fine-tuning on the Custom Graph Dataset compared to baseline models without additional fine-tuning. This comparison allows us to test if the Custom Graph Dataset specifically enhances logical reasoning abilities beyond the baseline.
2. **Custom Dataset vs. Control Dataset Comparison:** In the second comparison, we evaluate models fine-tuned on the Custom Graph Dataset against those fine-tuned on the CC News Dataset. This comparison is designed to confirm that the logical reasoning improvements are specific to the Custom Graph Dataset, rather than arising from general language fine-tuning on an unrelated dataset. By showing superior performance in logical reasoning tasks with the Custom Graph Dataset, we highlight the unique benefits of a reasoning-specific dataset.

Consistent Dataset Size: Both datasets contain approximately 100,000 sentences. This setup ensures that any performance differences are attributable to the dataset's logical structure and content rather than its size.

Hypotheses

- **H1:** Using 100,000 sentences in the Custom Graph Dataset is sufficient to produce a notable enhancement in logical reasoning performance, validating the dataset's scalability for targeted inference tasks.
- **H2:** Models fine-tuned on the Custom Graph Dataset will outperform both baseline models and those fine-tuned on the CC News Dataset in logical reasoning tasks, demonstrating that a dataset specifically designed for logical inference provides unique benefits.

Result

Performance Comparison Across Baseline and Custom Dataset Fine-Tuned Models

Table 1 provides a performance comparison between baseline models and models fine-tuned on the Custom Graph Dataset across various reasoning benchmarks. The results demonstrate that fine-tuning with the Custom Graph Dataset significantly enhances logical reasoning performance across several tasks. Notably, models fine-tuned with the Custom Graph Dataset generally outperform their baseline counterparts, especially on tasks like ARC-Challenge, Commonsense QA, and Riddle Sense. This improvement highlights the Custom Graph Dataset's effectiveness in fostering logical inference capabilities, suggesting that the structured and relational nature of this dataset is beneficial for tasks requiring nuanced reasoning.

The Custom Graph Dataset's structured relations and diverse set of connections allow models to build multi-step inferences and comprehend more complex scenarios, which are less accessible through baseline training alone. However, certain tasks, such as QASC and HellaSwag, show only marginal improvements, indicating that additional contextual knowledge may still play a role in specific types of inference.

Comparison Between Models Fine-Tuned on Custom Dataset vs. CC News Dataset

In Table 2, we compare models fine-tuned on the Custom Graph Dataset against those fine-tuned on the CC News Dataset to isolate the unique impact of our dataset on logical reasoning tasks. The results indicate that the Custom Graph Dataset provides a distinct advantage for logical reasoning, as models fine-tuned with it generally achieve higher scores on tasks like ARC-Challenge, Commonsense QA, and Riddle Sense compared to models fine-tuned on CC News.

Key insights from this comparison include the following observations. BERT fine-tuned on the Custom Graph Dataset outperforms BERT fine-tuned on CC News in tasks such as ARC-Challenge and Riddle Sense, affirming the Custom Graph Dataset's relevance in tasks that require logical inference and structured reasoning. Similarly, RoBERTa fine-tuned on the Custom Graph Dataset achieves better results than its CC News counterpart on Commonsense QA and HellaSwag, suggesting that the relational variety in the Custom Graph Dataset enhances the model's reasoning abilities across situational and commonsense contexts. Furthermore, DeBERTa fine-tuned on the Custom Graph Dataset consistently outperforms DeBERTa fine-tuned on CC News across most tasks, with notable improvements in HellaSwag and QASC. This pattern implies that the Custom Graph Dataset enhances the model's logical inference abilities more effectively than general-purpose text data.

This comparison further validates the hypothesis that a reasoning-specific dataset like the Custom Graph Dataset offers unique advantages over a general language dataset when targeting logical reasoning skills. While the CC News Dataset contributes to broader linguistic and contextual understanding, it lacks the structured relational information needed for complex logical inferences.

Qualitative Analysis of Dynamic Relations

The Custom Graph Dataset incorporates a diverse set of dynamically generated relations, adding flexibility to the model's reasoning capabilities. By filtering out relations that appear fewer than ten times, we identified 133 unique dynamic relations, which occur a total of 49,998 times throughout the dataset. The most frequently occurring relation was *Causal*, appearing 24,825 times, but as this is a pre-existing relation, we excluded it from the analysis of novel dynamic relations. Figure 2 shows the top 20 dynamic relations ranked from the 2nd to the 21st most frequent, with types like *Analogous*, *Sequential*, *Contextual*, and *Complementary* appearing most frequently. These relations support nuanced, multi-step reasoning by creating contextually rich

	ARC-Challenge(Clark et al. 2018)	Commonsense QA(Talmor et al. 2018)	HellaSwag (Zellers et al. 2019)	QASC (Khot et al. 2020)	Riddle Sense (Lin et al. 2021)
BERT (Kenton and Toutanova 2019)	22.61	18.76	24.59	11.12	19.59
BERT_Custom (Kenton and Toutanova 2019)	**25.77**	**20.64**	**24.60**	**11.56**	**20.37**
RoBERTa (Liu 2019)	**25.43**	19.00	24.69	**13.82**	16.69
RoBERTa_Custom (Liu 2019)	23.72	**22.52**	**25.44**	11.66	**19.78**
DeBERTa (He et al. 2020)	23.04	19.08	24.84	11.99	**21.25**
DeBERTa_Custom (He et al. 2020)	**25.09**	**20.39**	**25.62**	**12.74**	18.51
DistilBERT (Sanh 2019)	**25.77**	18.84	24.76	12.53	**21.84**
DistilBERT_Custom (Sanh 2019)	23.55	**19.49**	**25.71**	**13.07**	17.60

Table 1: Comparison of baseline models and models fine-tuned with the Custom Graph Dataset on various reasoning benchmarks, including ARC-Challenge, Commonsense QA, HellaSwag, QASC, and Riddle Sense. Bold values indicate the highest accuracy achieved for each task within each model type. Results show that models fine-tuned with the Custom Graph Dataset generally outperform baseline models across multiple reasoning tasks, especially on tasks focused on logical inference and commonsense reasoning. This improvement highlights the Custom Graph Dataset's contribution to enhancing logical reasoning capabilities in language models, as compared to general baseline performance.

	ARC-Challenge(Clark et al. 2018)	Commonsense QA(Talmor et al. 2018)	HellaSwag (Zellers et al. 2019)	QASC (Khot et al. 2020)	Riddle Sense (Lin et al. 2021)
BERT_Custom (Kenton and Toutanova 2019)	**25.77**	**20.64**	24.60	11.56	**20.37**
BERT_CC_News (Kenton and Toutanova 2019)	24.83	19.33	**24.72**	**13.50**	18.41
RoBERTa_Custom (Liu 2019)	23.72	**22.52**	**25.44**	11.66	**19.78**
RoBERTa_CC_News (Liu 2019)	**26.88**	21.05	25.19	**12.63**	17.92
DeBERTa_Custom (He et al. 2020)	25.09	**20.39**	**25.62**	**12.74**	**18.51**
DeBERTa_CC_News (He et al. 2020)	**25.60**	19.66	24.36	11.66	17.53
DistilBERT_Custom (Sanh 2019)	23.55	**19.49**	**25.71**	**13.07**	17.60
DistilBERT_CC_News (Sanh 2019)	**25.34**	18.59	25.15	12.42	**20.67**

Table 2: Comparison between models fine-tuned with the Custom Graph Dataset and models fine-tuned with the CC News Dataset on reasoning benchmarks. Bold values highlight the best performance between the two fine-tuning datasets for each model and task. This table provides insights into the specific benefits of each dataset for different logical reasoning tasks, demonstrating the added value of the Custom Graph Dataset in most cases, especially in logic-centric benchmarks.

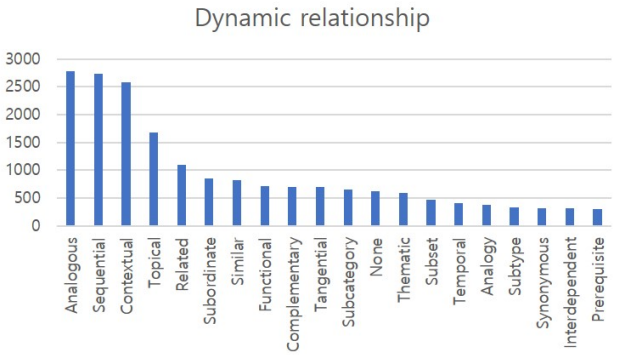

Figure 2: Distribution of the top 20 most frequent dynamic relationship types, ranked from 2nd to 21st in frequency. The chart highlights the prevalence of various dynamic relations, with Analogous, Sequential, and Contextual relations appearing most frequently. This visual demonstrates the diversity and frequency of relationship types generated in the dataset, providing insight into the relational variety beyond standard predefined relations.

connections between concepts. These dynamic relations offer models additional relational context, enabling them to make logical inferences that extend beyond standard, predefined relational structures.

Hypothesis Validation

Our experimental findings align with the following hypotheses:

- **H1 Validation:**The use of 100,000 sentences from the Custom Graph Dataset is sufficient to produce a measurable enhancement in logical reasoning performance, supporting the dataset's scalability and effectiveness even at moderate sizes.
- **H2 Validation:** Fine-tuning on the Custom Graph Dataset provides a greater boost in logical reasoning performance compared to models fine-tuned on the CC News Dataset or left at baseline, confirming the dataset's efficacy in enhancing inference skills.

Discussion

The experimental results demonstrate the substantial impact of our Custom Graph Dataset on logical reasoning tasks across various language model architectures(Kenton and Toutanova 2019; Liu 2019; He et al. 2020; Sanh 2019), highlighting the benefits of a dynamically generated, graph-structured dataset for enhancing inference capabilities. However, our study also reveals specific limitations, suggesting areas for refinement and directions for future research.

Our two-tiered experiment design provided clear insights into the effectiveness of our Custom Graph Dataset for logical reasoning. In the first comparison, models fine-tuned on the Custom Graph Dataset consistently outperformed baseline models across logical reasoning benchmarks(Li et al. 2022; Clark et al. 2018; Talmor et al. 2018; Zellers et al. 2019; Khot et al. 2020; Lin et al. 2021), including Commonsense QA, Riddle Sense, and ARC-Challenge, confirming that a targeted, structured dataset offers meaningful improvements. This enhancement can be attributed to the dataset's structured knowledge and diverse set of dynamically generated relations that facilitate multi-step and causal reasoning. After filtering out low-frequency relations, the dataset retained 133 unique dynamic relations across 49,998

instances, with frequent relations such as *Causal*, *Analogous*, and *Contextual*. These relations enable richer, context-sensitive inference patterns, underscoring the importance of a relationally dense and flexible dataset.

In the second comparison, between the Custom Graph Dataset and the CC News dataset, models fine-tuned on the Custom Graph Dataset demonstrated superior performance in tasks requiring logical inference, particularly in ARC-Challenge, Commonsense QA, and Riddle Sense. However, on tasks such as QASC and HellaSwag, models fine-tuned on the CC News dataset occasionally achieved comparable or marginally better results. This suggests that, while our dataset enhances logical reasoning in specific areas, general language data may still contribute valuable background knowledge for certain types of inference. These findings imply that a hybrid approach combining structured reasoning-specific datasets with general text data could optimize model performance across varied reasoning tasks. Future studies might explore integrating these two data types to develop models that leverage both comprehensive linguistic context and explicit logical structures.

An important observation is the variability in performance gains across model architectures. Larger models, such as RoBERTa and DeBERTa, benefitted more from fine-tuning with the Custom Graph Dataset than smaller models like DistilBERT. This suggests that higher-capacity models may better capture and utilize the complex structures in the dataset, supporting the hypothesis that model architecture plays a crucial role in leveraging the full potential of structured datasets for reasoning. For smaller models, distilled or simplified versions of the structured data might yield more practical benefits, especially given computational constraints.

The qualitative analysis of dynamic relations further reveals the unique impact of our dataset, where newly generated relation types facilitate nuanced, multi-step reasoning. Relations beyond predefined categories, such as *Causal*, *Analogous*, and *Contextual*, allow for flexibility in understanding complex concept interactions not present in traditional datasets. However, the dynamic generation of relations also introduces challenges in consistency, as these relations may not always align precisely with specific inference requirements. Refining the mechanisms for dynamic relation generation could enhance the precision and relevance of generated relations, potentially leading to further performance gains.

Limitations and Future Work

Despite the strengths of the Custom Graph Dataset, one limitation of this study is the dataset size. While 100,000 instances demonstrated efficacy, this scale may not capture the full spectrum of logical relations needed for more complex reasoning tasks. As future work, we aim to expand the dataset to 300,000 instances to determine if a larger dataset provides further gains in logical reasoning performance. This expansion will also allow us to conduct a more rigorous comparison with an equivalently scaled 300,000-instance CC News dataset, facilitating a balanced evaluation of structured logical data against general-purpose text. Such a comparison could yield further insights into how dataset size affects reasoning performance.

Additionally, our current analysis is focused on logical reasoning tasks; the transferability of these gains to other domains, such as knowledge retrieval or fact-based reasoning, remains unexplored. Future studies could investigate the applicability of structured, reasoning-specific datasets across a broader range of tasks. Moreover, as larger models showed enhanced benefits with structured data, future research could explore strategies for smaller models, such as using distilled or simplified versions of the structured dataset, to enable similar gains at lower computational costs.

Finally, the results reinforce the distinct advantages of a logic-specific dataset for reasoning tasks. While general text data is beneficial for broad contextual understanding, it lacks the targeted support for explicit reasoning provided by structured datasets. This finding underscores the need for focused datasets designed to enhance logical inference capabilities in language models, suggesting a promising path for advancing reasoning-specific dataset development and applications in the field of NLP.

Conclusion

In this study, we introduced and assessed the Custom Graph Dataset, a novel graph-based knowledge resource crafted to enhance logical reasoning capabilities in language models. Utilizing a Self-Prompting approach, our dataset generates dynamically defined relations, bridging the limitations of predefined relational types by enabling both standard and context-specific connections. This unique approach led to a dual-triple structure that captures complex, multi-step inferences, which proved advantageous across multiple reasoning-intensive benchmarks.

Our experiments demonstrated that models fine-tuned on the Custom Graph Dataset consistently outperformed both baseline models and those fine-tuned on a general-purpose control dataset, CC News. This trend was particularly evident in tasks such as Commonsense QA, Riddle Sense, and ARC-Challenge, validating the hypothesis that a logically structured, relation-rich dataset significantly boosts inference skills. By introducing 133 unique dynamic relations, including *Analogous*, *Contextual*, and *Complementary*, we equipped language models with the flexibility needed for nuanced, context-sensitive reasoning. The consistent improvement observed in causal and commonsense reasoning tasks underscores the importance of dedicated datasets in advancing complex inference skills.

The scalability and adaptability of the Custom Graph Dataset make it suitable for a broad range of logical reasoning applications, offering a flexible foundation for further advancements in structured knowledge representation. By advancing reasoning-focused datasets and optimizing automatic relation generation, we move closer to bridging the gap between general language understanding and robust logical reasoning, paving the way for future models capable of sophisticated multi-step inference.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 6384–6392.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8082–8090.

Li, J.; Wang, J.; Zhang, Z.; and Zhao, H. 2022. Self-prompting large language models for zero-shot open-domain QA. *arXiv preprint arXiv:2212.08635*.

Lin, B. Y.; Wu, Z.; Yang, Y.; Lee, D.-H.; and Ren, X. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Sanh, V. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.