Enhancing Trustworthiness of Large Language Models through Dynamic Knowledge Graph Reasoning

Jongwon Ryu[1], **Mingyu Jeon**[1], **Soomin Chung**[2], **Junyeong Kim**[1]

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, 06974
[2]Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, 08826
[1][fbwhddnjs511, smart2557, junyeongkim]@cau.ac.kr
[2] soomin200@snu.ac.kr

Abstract

Despite significant advances in large language models, many reasoning datasets are still built from a fixed set of predefined relations, manually curated types such as cause, effect, and intent found in knowledge graph datasets such as ATOMIC and COMET. While these predefined relations provide essential structure, the fixed schema limits relational coverage and adaptability to novel contexts. We present DYNA-SKILL, a dual-triple knowledge graph framework that preserves 35 predefined relations consolidated and refined from existing commonsense knowledge graph datasets while augmenting them with 133 additional schema-free dynamic relations generated via a self-prompting mechanism. Each instance consists of two linked triples (Head–Predefined Relation–Tail) and (Tail–Dynamic Relation–Additional Tail) used as independent training samples while retaining linkages for extended reasoning paths. Across reasoning-intensive benchmarks, including CommonsenseQA, RiddleSense, and ARC Challenge, the Hybrid configuration, which combines predefined and dynamically generated relations, achieves performance comparable to or slightly higher than Predefined-only settings and yields up to 3.2% higher accuracy than baseline BERT models. By expanding the relation set from 35 predefined types to a total of 168 relations, DYNA-SKILL enriches relational diversity and improves multi-step logical reasoning, which can enhance performance in real-world scenarios such as complex question answering, multi-document analysis, and causal reasoning, where accurate and adaptable inference is critical.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities across a wide range of natural language processing tasks, including question answering, summarization, and commonsense reasoning [5]. Despite these advances, LLMs continue to struggle with complex, multi-step logical reasoning, particularly in open-domain and contextually rich scenarios [2]. This limitation is partly due to their reliance on implicit knowledge learned during pretraining, without explicit relational structures that facilitate structured inference.

A large proportion of reasoning evaluation datasets are still constructed from a fixed set of manually curated relation types, such as cause, effect, and intent, found in commonsense knowledge graph datasets like ATOMIC [16, 8] and expanded using models such as COMET [4]. While these predefined relations provide essential structure, the fixed schema inherently limits relational coverage and adaptability to novel or context-specific connections. As a result, current reasoning datasets

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

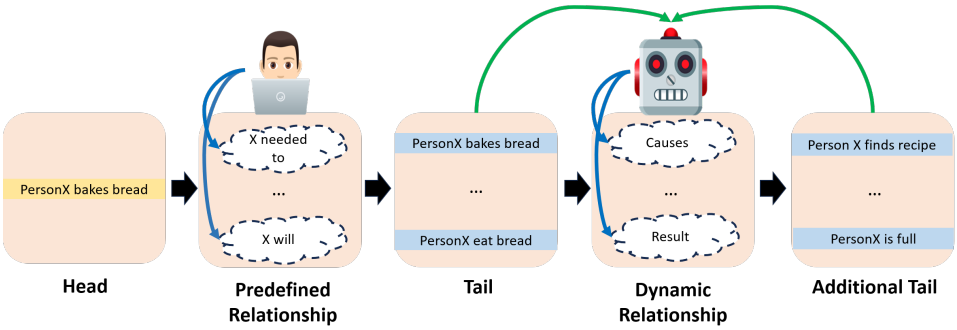

Figure 1: Overview of the Self Prompting Graph Based Knowledge Dataset Generation.

cannot fully support the diverse and dynamic relational patterns required for robust, multi-step inference in real-world applications.

To address these limitations, we propose **DYNA-SKILL**, a Self-Prompting-based approach for automatically generating a graph-structured knowledge dataset that integrates both predefined and dynamically generated relations. As illustrated in Figure 1, our method constructs a dual-triple representation: (Head–Predefined Relation–Tail) and (Tail–Dynamic Relation–Additional Tail). First, Tails are generated using 35 predefined relations, consolidated and refined from existing commonsense knowledge graph datasets. To extend reasoning depth, an Additional Tail is generated based on the Tail, introducing new but logically coherent knowledge. Finally, a Dynamic Relation is inferred between the Tail and Additional Tail, enabling the discovery of 133 schema-free relation types beyond manually curated templates. For example, given the Tail "PersonX bakes bread," our approach may generate the Additional Tail "PersonX finds a recipe" and infer the Dynamic Relation "Causes."

This Self-Prompting-driven process enables LLMs to learn diverse and flexible relational structures, facilitating multi-step inference and contextually adaptive reasoning across various domains. The dual-triple structure serves as a foundation for enhanced logical reasoning, capturing both explicit and implicit connections that conventional knowledge graphs often miss.

We evaluate DYNA-SKILL on five well-established reasoning benchmarks: ARC Challenge [6], CommonsenseQA [19], HellaSwag [20], QASC [10], and RiddleSense [13]. Additionally, we compare against a control dataset (CC News) to isolate the specific contribution of our reasoning-focused dataset beyond general language understanding. Our results show that models fine-tuned on DYNA-SKILL consistently outperform both baseline and control models, particularly in tasks that require multi-step inference.

The main contributions of this work are as follows:

- We introduce a method for automatically generating a graph-based knowledge dataset that integrates 35 predefined and 133 dynamically generated relations, substantially increasing the adaptability and coverage of the knowledge base.
- We develop a dual-triple structure (Head–Relation–Tail and Tail–Dynamic Relation–Additional Tail) that supports multi-step inference and captures a broader range of logical relationships beyond existing commonsense graphs.
- Through experiments on multiple reasoning benchmarks, we demonstrate that DYNA-SKILL significantly enhances LLMs' logical reasoning performance, outperforming both baseline and control models, thereby validating the effectiveness of our approach.

2 Related Work

Our work connects two previously distinct lines of research: (1) reasoning-specific datasets in the form of structured commonsense knowledge graphs, and (2) dynamic relation generation methods such as self-prompting. While knowledge graphs like ATOMIC and COMET provide explicit relational structures, they are restricted by fixed relation schemas. Conversely, dynamic generation methods offer adaptability but lack integration with structured, reasoning-specific graph formats.

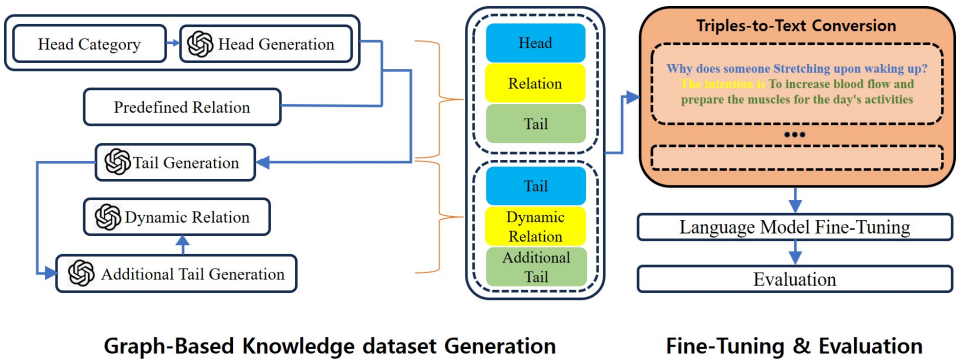

Figure 2: Overview of the DYNA-SKILL framework. The dataset generation phase creates dual-triple knowledge structures by combining predefined and dynamically generated relations. These are converted into natural language for fine-tuning, and models are evaluated on reasoning benchmarks.

DYNA-SKILL combines these strengths by embedding schema-free dynamic relations within a predefined relational framework, creating a flexible yet structured resource for logical reasoning.

2.1 Reasoning-Specific Knowledge Graph Datasets

2.1.1 Knowledge Graph Datasets

ATOMIC [16] is one of the first large-scale commonsense knowledge graphs tailored for "if–then" reasoning. It captures human-centered scenarios through categories such as intentions, reactions, and effects. Its manually curated triples ensure high quality, but the fixed set of relation types limits coverage and adaptability to novel contexts. COMET [4, 8] extends ATOMIC by using transformer-based models to populate predefined relational templates derived from ATOMIC and ConceptNet. Although this automates triple generation, COMET remains bound to its original set of predefined relations, preventing adaptation to unseen relation types. ConceptNet [17] and other large-scale resources such as Freebase [3], DBpedia [11], and YAGO [18] cover a wide range of domains, but their relation inventories are static and schema-bound, which constrains their use for tasks requiring dynamically evolving logical connections.

2.2 Dynamic Relation Generation Methodologies

Self-prompting approaches, such as [12], are not designed to construct reasoning-specific graph datasets. Instead, they dynamically generate contextually relevant prompts and answers in multi-step open-domain QA. While effective for adaptive knowledge acquisition, these methods typically operate without an underlying structured graph, limiting their ability to produce explicit multi-step relational chains for reasoning.

3 Method

In this study, we present DYNA-SKILL, a graph-based knowledge dataset designed to enhance the logical reasoning capabilities of language models. Using a Self-Prompting approach [12] with the GPT-4-turbo API [1], we automatically construct dual-triple knowledge structures in the form of (Head–Predefined Relation–Tail) and (Tail–Dynamic Relation–Additional Tail). Each component, Head, Tail, Dynamic Relation, and Additional Tail, is generated to ensure contextual relevance and relational diversity. Figure 2 illustrates the overall pipeline, from data generation to fine-tuning and evaluation. The following subsections detail each stage of the methodology, and illustrative examples of such dual-triple structures are provided in Table 1.

1. **Head-to-Tail Generation**
   Head Definition: We define *Head* entities across diverse categories to represent the main subjects of logical reasoning events. The categories cover a broad range of commonsense

Algorithm 1 DYNA-SKILL Dataset Generation Process

Input: Predefined relation set R_{pre} (35 types), Head category set C, Large Language Model M
Output: Dual-triple dataset D
Initialize $D \leftarrow \emptyset$
for all category $c \in C$ **do**
  Head Generation: Select category c and generate Head h using M with a category-specific prompt.
  Relation Selection: Choose a predefined relation $r_{pre} \in R_{pre}$ that matches the semantic type of h.
  Tail Generation: Generate Tail $t \leftarrow M(\text{prompt}(h, r_{pre}))$ using a relation-specific template (e.g., "What is the typical use of *Head*?" for *ObjectUse*).
  Additional Tail Generation: Generate Additional Tail $t_{add} \leftarrow M(\text{prompt}(t))$ to extend reasoning depth.
  Dynamic Relation Inference: Infer Dynamic Relation $r_{dyn} \leftarrow M(\text{relation-prompt}(t, t_{add}))$ using a schema-free relation prompt (e.g., "What is the relationship between *Tail* and *Additional Tail*?").
  Append both triples (h, r_{pre}, t) and (t, r_{dyn}, t_{add}) to D.
end for
Text Conversion: For each triple in D, convert to a natural language sentence using predefined mapping rules .
Data Storage: Save the converted sentences to a plain text file for fine-tuning.
return D

scenarios, including:**Social Interaction** (e.g., education, household, relationship management),**Physical Entities** (e.g., tools, vehicles, appliances),**Event-Centered** (e.g., festivals, weddings, public gatherings), **Causal Relations** (e.g., economic events, technological failures, 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 category captures distinct logical structures and interactions, ensuring coverage of diverse reasoning contexts.

Relation Definition: Each *Head* category is associated with predefined relations that guide the generation of Tail elements and ensure consistency across the dataset. These relations include context-specific types tailored to each category. Drawing on insights from prior works such as ATOMIC and COMET, we expand the range of predefined relations to build a richer and more varied relational structure: **Social-Interaction Relations** Examples: *xIntent* (intention behind an action), *xNeed* (prerequisites for an action), *oEffect* (impact on others) These relations capture interpersonal and motivational aspects, enabling reasoning about complex social dynamics. **Physical-Entity Relations** Examples: *ObjectUse* (typical use of an object), *AtLocation* (where an object is typically found), *CapableOf* (actions an object can perform) These describe functional and situational properties essential for practical reasoning. **Event-Centered Relations** Examples: *IsAfter* (what happens after an event), *HasSubEvent* (sub-events of a main event), *Causes* (what leads to an event) These support temporal and causal reasoning beyond fixed templates. **Causal Relations** Examples: *Cause and Effect*, *Causal Chain* These describe outcome dependencies and multi-step cause–effect sequences. **Other Categories** Examples: *Temporal Sequence* (Temporal Relations), *If-Then Statements* (Conditional Relations), *Part-Whole Relations* (compositional structures), *Quantities and Measures* (Quantitative Relations) These model temporal dependencies, conditional logic, and hierarchical structures. Each relation is paired with a specific prompt to guide Tail generation. For example, an *xIntent* relation for a social action Head may use the prompt: *"What is the possible intention behind this action?"* By extending relation types beyond those in existing commonsense graphs, we provide a versatile framework that supports richer logical connections, including cause–effect, hierarchical, and conditional dependencies.

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

Domain	Triple 1 (Predefined Relation)	Triple 2 (Dynamic Relation)
Sports Action	*(Spiking – xNeed – A set or a pass is needed before performing a spike in volleyball.)* **Sentence:** What does someone need before Spiking? A set or a pass is needed before performing a spike in volleyball.	*(A set or a pass is needed before performing a spike in volleyball. – Preparatory – Pulling back the bowstring before releasing an arrow in archery.)* **Sentence:** A set or a pass is needed before performing a spike in volleyball and Pulling back the bowstring before releasing an arrow in archery are connected by Preparatory.
Safety Action	*(Shooting – xNeed – Safety training and proper authorization or permits.)* **Sentence:** What does someone need before Shooting? The prerequisite is Safety training and proper authorization or permits.	*(Safety training and proper authorization or permits. – Complementary – Conducting regular safety audits and inspections.)* **Sentence:** What complements Safety training and proper authorization or permits? It is complemented by Conducting regular safety audits and inspections.
Physical Object	*(Cans – CapableOf – Cans are capable of storing and preserving food or liquids.)* **Sentence:** What is Cans capable of? Cans are capable of storing and preserving food or liquids.	*(Cans are capable of storing and preserving food or liquids. – Functional – Dehydrating fruits and vegetables.)* **Sentence:** Cans are capable of storing and preserving food or liquids and Dehydrating fruits and vegetables are connected by Functional.
Leisure Activity	*(Snorkeling or scuba diving – oWant – Others might want to try snorkeling or scuba diving themselves.)* **Sentence:** What do others want after Snorkeling or scuba diving? Others might want to try snorkeling or scuba diving themselves.	*(Others might want to try snorkeling or scuba diving themselves. – Alternative – Sailing)* **Sentence:** Others might want to try snorkeling or scuba diving themselves and Sailing are connected by Alternative.
Scientific Material	*(Polyester – MadeUpOf – Polyester is a synthetic polymer made primarily from petroleum-derived ethylene glycol and terephthalic acid.)* **Sentence:** What is Polyester made up of? Polyester is a synthetic polymer made primarily from petroleum-derived ethylene glycol and terephthalic acid.	*(Polyester is a synthetic polymer made primarily from petroleum-derived ethylene glycol and terephthalic acid. – Chemical – Synthesis of polycarbonate from bisphenol A and phosgene.)* **Sentence:** Polyester is a synthetic polymer made primarily from petroleum-derived ethylene glycol and terephthalic acid and Synthesis of polycarbonate from bisphenol A and phosgene are connected by Chemical.

Table 1: Examples of Dual-Triple Structures with Natural Language Conversion based on the conversion rules in the text processing script. Each example consists of two linked triples: **(Head–Predefined Relation–Tail)** followed by **(Tail–Dynamic Relation–Additional Tail)**, illustrating how predefined and dynamically generated relations connect to form extended reasoning paths.

Additional Tail Generation: To extend the initial Tail, we apply a Self-Prompting approach to generate an *Additional Tail* that is contextually related to the existing Tail. This step deepens logical connections by prompting the model with targeted questions about further related actions, events, or consequences.

Dynamic Relation Generation: We then determine the relationship between the Tail and the Additional Tail by asking the model: *"What kind of relationship does 'additional tail' have with 'tail'?"* This enables the automatic creation of previously undefined, schema-free relations, thereby enhancing flexibility and incorporating novel, context-specific connections into the dataset.

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 format allows the representation of complex, multi-step relationships that go 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 $(\text{Head}, \text{Relation}, \text{Tail})$ 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"

   is converted into:

   "Why does someone make coffee? The intention is to help."

   Conversion Process: A hybrid-relation function processes each triple according to its relation type, producing readable sentences. This 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 model training, enhancing 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[9, 14, 7, 15] 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 hybrid-relation 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.

4 Experiments

4.1 Datasets

Hybrid-Relation Graph Dataset: Our primary dataset is generated via the Self-Prompting method described in Section 3, combining 35 predefined relations with 133 dynamically generated relations in a dual-triple format. These triples are converted into natural language sentences for model fine-tuning. We use approximately 100,000 sentences for each training run.

Predefined-Only Graph Dataset: A variant of the above dataset containing only the 35 predefined relations, without any dynamically generated relations. The dataset size is matched to the others at approximately 100,000 sentences, allowing a fair comparison to isolate the contribution of dynamic relation generation.

CC News Dataset: A large-scale news corpus used as a control dataset for general-domain fine-tuning. Lacking a reasoning-specific structure, it is expected to have limited impact on logical reasoning performance. We randomly sample 100,000 sentences for size parity with the other datasets.

4.2 Models

We evaluate four transformer-based encoder models with different capacities: BERT [9], RoBERTa [14], DeBERTa [7], and DistilBERT [15]. This selection allows us to measure dataset impact across both large and lightweight architectures.

4.3 Evaluation Tasks

To assess improvements in logical reasoning, we employ five established benchmarks: ARC-Challenge [6], CommonsenseQA [19], HellaSwag [20], QASC [10], and RiddleSense [13]. These

	ARC-Challenge[6]	Commonsense QA[19]	HellaSwag [20]	QASC [10]	Riddle Sense [13]
BERT [9]	22.61	18.76	24.59	11.12	19.59
BERT_CC_NEWS [9]	24.83	19.33	**24.72**	**13.50**	18.41
BERT_Hybrid [9]	**25.77**	20.64	24.60	11.56	**20.37**
BERT_Predifined [9]	25.71	**20.65**	24.58	11.59	20.33
Δ (Hybrid-Predefined)	+0.06	-0.01	+0.02	-0.03	+0.04
RoBERTa [14]	25.43	19.00	24.69	**13.82**	16.69
RoBERTa_CC_NEWS [14]	**26.88**	21.05	25.19	12.63	17.92
RoBERTa_Hybrid [14]	23.72	**22.52**	**25.44**	11.66	**19.78**
RoBERTa_Predefined [14]	23.70	22.49	25.43	11.61	19.77
Δ (Hybrid-Predefined)	+0.02	+0.03	+0.01	+0.05	+0.01
DeBERTa [7]	23.04	19.08	24.84	11.99	**21.25**
DeBERTa_CC_NEWS [7]	**25.60**	19.66	24.36	11.66	17.53
DeBERTa_Hybrid [7]	25.09	**20.39**	**25.62**	**12.74**	18.51
DeBERTa_Predefined [7]	25.11	20.36	25.61	12.71	18.52
Δ (Hybrid-Predefined)	-0.02	+0.03	+0.01	+0.03	-0.01
DistilBERT [15]	**25.77**	18.84	24.76	12.53	**21.84**
DistilBERT_CC_NEWS [15]	25.34	18.59	25.15	12.42	20.67
DistilBERT_Hybrid [15]	23.55	**19.49**	**25.71**	**13.07**	17.60
DistilBERT_Predefined [15]	23.54	19.44	**25.71**	13.05	17.58
Δ (Hybrid-Predefined)	+0.01	+0.05	0.00	+0.02	+0.02

Table 2: Accuracy (%) of each model on five reasoning benchmarks. Bold indicates the best score and underline the second best within each model type. Δ represents the accuracy difference between Hybrid and Predefined-only settings. Hybrid-relation fine-tuning generally achieves competitive or superior results, suggesting that dynamic relation generation contributes to improved logical reasoning performance.

tasks cover diverse reasoning types, including multiple-choice science questions, commonsense inference, situational plausibility, multi-hop QA, and lateral thinking riddles.

4.4 Fine-Tuning and Evaluation Procedure

For each model–dataset pair, we fine-tune using approximately 100,000 training sentences, keeping dataset sizes consistent to control for size effects. Fine-tuning is conducted with a standard language modeling objective, batch size 32, learning rate 2×10^{-5}, and 3 epochs. Evaluation is performed on the benchmark test sets, with accuracy as the primary metric for all tasks.

4.5 Comparison Settings

We conduct a two-tier comparison: i)**Baseline vs. Graph Datasets:** Comparing models fine-tuned on each graph dataset against their unfine-tuned baselines. ii)**Hybrid vs. Predefined vs. CC News:** Comparing reasoning gains from dynamic relations (Hybrid), static relations (Predefined), and general-domain fine-tuning (CC News) to determine the specific contribution of dynamic relation generation. All datasets contain the same number of sentences, ensuring differences are attributable to content rather than size.

5 Result

5.1 Performance Comparison Across Baseline and Hybrid-Relation Fine-Tuned Models

Table 2 compares baseline models with those fine-tuned on the hybrid-relation Graph Dataset across five reasoning benchmarks. Overall, hybrid-relation fine-tuning yields consistent gains over baseline performance, with notable improvements on ARC-Challenge, CommonsenseQA, and RiddleSense. These gains suggest that the dataset's structured, multi-relational design supports more effective multi-step inference and nuanced commonsense reasoning.

While improvements on QASC and HellaSwag are smaller, the results indicate that hybrid-relation fine-tuning still maintains competitive performance, highlighting potential for further enhancement by integrating richer contextual or domain-specific knowledge.

5.2 Comparison Between Hybrid-Relation and CC News Fine-Tuning

Table 2 compares models fine-tuned on the hybrid-relation Graph Dataset with those fine-tuned on the CC News Dataset, isolating the effect of reasoning-specific data. Across most benchmarks, hybrid-relation fine-tuning yields higher scores on tasks such as ARC-Challenge, CommonsenseQA,

and RiddleSense, indicating that structured, multi-relational knowledge directly benefits logical inference.

Model-specific trends further support this conclusion. For example, BERT shows clear gains on ARC-Challenge and RiddleSense when trained on the hybrid-relation dataset, while RoBERTa achieves higher accuracy on CommonsenseQA and HellaSwag, suggesting improved situational and commonsense reasoning. DeBERTa exhibits consistent advantages across tasks, with notable improvements in HellaSwag and QASC, reinforcing the dataset's utility for multi-step inference.

These results validate the hypothesis that reasoning-focused datasets offer advantages over general-purpose corpora for logical reasoning. While CC News improves general language understanding, it lacks the explicit relational structures needed to support complex, stepwise reasoning.

5.3 Comparison Between Hybrid-Relation and Predefined-Only Fine-Tuning

Table 2 also reports results for a *Predefined-Only* variant of our dataset, containing the same 100,000 samples but restricted to the 35 predefined relations without any dynamically generated ones. This comparison isolates the contribution of dynamic relation generation to reasoning performance.

Overall, the performance gap between the Hybrid-Relation and Predefined-Only settings is modest but consistent across several tasks. For example, BERT shows small gains on ARC-Challenge and RiddleSense with Hybrid-Relation training, while RoBERTa benefits slightly on CommonsenseQA and QASC. DeBERTa and DistilBERT also exhibit minor but positive differences in most benchmarks, suggesting that dynamically generated relations introduce additional contextual variety that can support reasoning beyond the coverage of fixed relations.

Although the improvements are not large in absolute terms, their presence across multiple architectures and tasks indicates that dynamic relations add complementary knowledge that predefined schemas cannot fully capture. These results imply that even small increments in relational diversity can compound over multi-step reasoning chains, leading to more robust inference capabilities.

5.4 Qualitative Analysis of Dynamic Relations

The hybrid-relation 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 3 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 connections between concepts. These dynamic relations offer models additional relational context, enabling them to make logical inferences that extend beyond standard, predefined relational structures.

6 Discussion

Our results show that the hybrid-relation Graph Dataset consistently enhances logical reasoning performance across multiple transformer-based architectures [9, 14, 7, 15], validating the benefit of combining predefined and dynamically generated relations in a graph-structured format. The inclusion of 133 schema-free dynamic relations, in addition to 35 predefined types, enables richer multi-step and causal reasoning than fixed-schema datasets alone.

Effectiveness Across Benchmarks

In the first comparison, models fine-tuned on the hybrid-relation dataset outperformed baseline models on reasoning benchmarks [12, 6, 19, 20, 10, 13], particularly CommonsenseQA, RiddleSense, and ARC-Challenge. After filtering low-frequency relations, 133 unique dynamic types remained across 49,998 instances, with frequent categories including *Causal*, *Analogous*, and *Contextual*. These relations provide diverse inference pathways, supporting more flexible reasoning.

Comparison with General-Purpose Data

In the second comparison, the hybrid-relation dataset generally outperformed CC News on reasoning tasks, confirming that explicit relational structure yields unique benefits. Nonetheless, in QASC and HellaSwag, CC News achieved comparable or slightly higher scores, suggesting that broad-domain knowledge can still aid certain forms of inference. This points to the potential of hybrid training strategies that integrate reasoning-specific and general-purpose data.

Model Capacity Considerations

Larger models such as RoBERTa and DeBERTa benefited more from the structured dataset than smaller models like DistilBERT, indicating that model capacity influences the ability to leverage complex relational structures. For resource-limited settings, simplified or distilled variants of the dataset may be necessary to deliver similar benefits.

Qualitative Insights and Challenges

Dynamic relations extend coverage beyond fixed schemas and capture nuanced, context-specific links absent in traditional commonsense graphs. However, automatic generation can produce inconsistencies or overly broad labels. Refining prompt design and incorporating automated validation mechanisms could improve precision and alignment with task requirements.

6.1 Limitations and Future Work

This study used 100K instances for each dataset. While effective, this scale may not fully capture the diversity of logical relations needed for more complex tasks. Future work will expand the dataset to 300K instances and conduct balanced comparisons against equivalently scaled CC News data to assess the interaction between dataset size and reasoning performance. Moreover, the scalability of the approach to significantly larger and noisier real-world datasets, especially those with highly heterogeneous relation types, remains an open challenge that warrants further investigation.

Our evaluation focused on reasoning benchmarks; transferability to other domains, such as fact verification or knowledge retrieval, remains unexplored. Exploring cross-domain applicability, along with model–dataset co-design strategies for smaller architectures, represents an important direction. Finally, while self-prompting allows flexible generation of dynamic relations, ensuring their logical validity remains an open challenge that warrants targeted verification techniques.

7 Conclusion

In this study, we presented the hybrid-relation Graph Dataset, a novel graph-based knowledge resource designed to enhance the logical reasoning capabilities of language models. Built using a Self-Prompting approach, the dataset combines 35 predefined relations with 133 dynamically generated relations, overcoming the limitations of fixed relational schemas. This integration results in a dual-triple structure(Head–Predefined Relation–Tail) and (Tail–Dynamic Relation–Additional Tail)that captures complex, multi-step inferences essential for advanced reasoning.

Experimental results show that models fine-tuned on the hybrid-relation Graph Dataset consistently outperform both baseline models and those fine-tuned on a general-purpose control dataset (CC News), with notable gains on Commonsense QA, Riddle Sense, and ARC-Challenge. The introduction of diverse dynamic relations, such as Analogous, Contextual, and Complementary, equips models with the flexibility to perform nuanced, context-sensitive reasoning. Performance improvements in causal and commonsense reasoning tasks further validate the dataset's effectiveness in strengthening inference skills.

With its scalability and adaptability, the hybrid-relation Graph Dataset offers a robust foundation for a wide range of reasoning-oriented applications. By advancing reasoning-focused dataset construction and refining automatic relation generation, this work contributes to narrowing the gap between general language understanding and sophisticated multi-step logical inference, paving the way for future models capable of more robust and context-aware reasoning.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) and partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.

[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*, 2019.

[5] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[8] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392, 2021.

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[10] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090, 2020.

[11] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.

[12] Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. Self-prompting large language models for zero-shot open-domain qa. *arXiv preprint arXiv:2212.08635*, 2022.

[13] Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*, 2021.

[14] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019.

[15] V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[16] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035, 2019.

[17] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[18] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, 2007.

[19] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[20] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

A Appendix

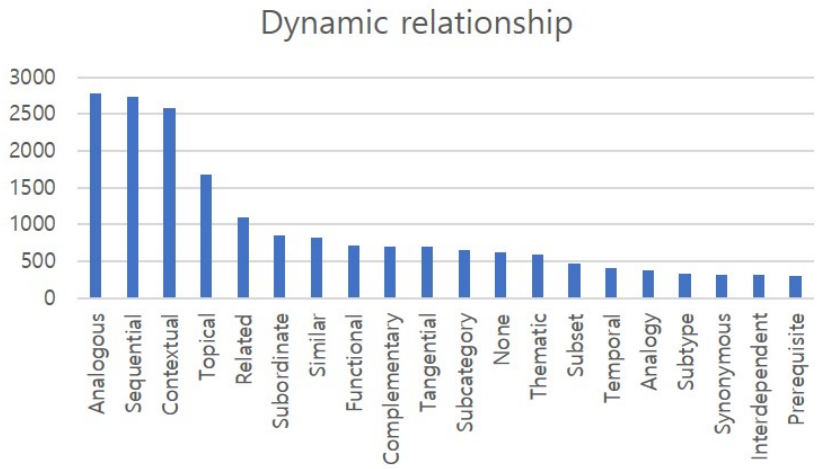

Figure 3: Frequency distribution of the top 20 dynamic relation types (excluding the single most common type). Relations such as *Analogous*, *Sequential*, and *Contextual* occur most frequently, indicating that the self-prompting generation process captures a broad spectrum of context-specific and non-predefined connections. This variety reflects the dataset's ability to extend beyond fixed schemas and enrich multi-step reasoning.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: [NA]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Method

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Method

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [NA]

   Justification: [NA]

   Guidelines: [NA]

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: [NA]

    Guidelines: [NA]

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines: [NA]

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines: [NA]

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines: [NA]

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines: [NA]

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: An LLM was used to refine the sentence.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.