POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Measuring Contextual Adherence in LLM-based Reasoning over Knowledge Graphs

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Riccardo Dell'Oro

Advisor: Prof. Marco Brambilla

Co-advisor: Dr. Antonio De Santis

Academic year: 2025-2026

1. Introduction

When a knowledge-based query is submitted to a Large Language Model (LLM), its response is influenced by two sources of knowledge: parametric knowledge, implicitly encoded in the model's weights during training, and contextual knowledge, explicitly provided in the input prompt. Reliance on parametric knowledge may lead to hallucinations and the inability to address recent or private information. Retrieval-Augmented Generation (RAG) mitigates these issues by incorporating external information into the prompt, improving response accuracy.
However, conflicts may arise when parametric and contextual knowledge contradict each other. If the retrieved information is incorrect, the model may propagate errors; similarly, even when the external context is correct, LLMs may rely on their prior parametric knowledge. Other studies, such as Wu et al. with their ClashEval[4] paper, have shown mixed evidence regarding this "tug-of-war".
Knowledge Graphs (KGs) provide structured and explicit representations of information through entities and relations, and have been integrated into RAG frameworks. This thesis evaluates three KG-based RAG approaches, Graph Constrained Reasoning[2] (GCR), Reasoning on Graphs[1] (RoG), and Think on Graph[3] (ToG), in scenarios where parametric and contextual knowledge conflict. Using the WebQSP question-answering dataset, composed of questions, correct answers, and an associated KG, controlled perturbations are introduced to create modified datasets and assess the performance of these methods. The objective is to analyze the extent to which KG-based RAG approaches enhance reasoning accuracy and support the LLM to adapt to external contextual knowledge.

2. Related Works

In GCR and RoG, an LLM is used to produce paths over the KG, while in ToG it operates as an agent that interactively explores the KG; for GCR and RoG, you can distinguish a path generation phase, and a final answer (FA) generation phase.

2.1. Reasoning on Graph (RoG)

RoG is composed of three modules:

1. Planning Module: a fine-tuned LLM is prompted to generate relation paths useful for answering the given question; a relation

path z is a sequence of relations $r_1 \rightarrow r_2 \rightarrow ... \rightarrow r_l$; the goal is to identify abstract relational structures that could connect question entities to potential answers.
2. Retrieval Module: for each generated relation path z, the system retrieves corresponding reasoning paths W_z from the Knowledge Graph using a constrained breadth-first search; a reasoning path w is a sequence of entities and relations $e_0 \rightarrow r_1 \rightarrow e_1 \rightarrow ... \rightarrow r_l \rightarrow e_l$; these are entity-level instantiations of the "abstract" relation paths, and are generated starting from each entity mentioned in the question.
3. Reasoning Module: the multiple retrieved reasoning paths are provided to a general-purpose LLM, which performs reasoning over them to generate the final answer; this module mitigates noise and incorrect paths by leveraging the LLM's reasoning capability instead of relying on majority voting.

2.2. Graph Constrained Reasoning (GCR)

GCR is composed of three modules:

1. Knowledge Graph Trie Construction: reasoning paths within a fixed number of hops from question entities are retrieved using BFS; these paths are formatted as strings, tokenized, and stored in a KG-Trie, a prefix tree that encodes all valid reasoning paths; the Trie acts as an index to constrain the LLM output during decoding.
2. Graph-Constrained Decoding: a fine-tuned LLM is used to produce reasoning paths and candidate answers; the KG-Trie constrains token generation so that only valid paths in the graph can be produced; once a valid path is completed, decoding switches back to unconstrained mode to generate a hypothesis answer; beam search allows multiple reasoning paths and answers to be generated in parallel.
3. Graph Inductive Reasoning: the multiple reasoning paths and candidate answers are aggregated and fed to a general-purpose LLM which performs inductive reasoning over them to produce a final answer.

Compared to RoG, GCR forces the LLM to produce only valid paths during decoding, rather than validating paths after generation.

2.3. Think on Graph (ToG)

It operates in three phases:

1. Initialization: the LLM extracts topic entities mentioned in the question; these entities serve as starting nodes for reasoning paths.
2. Exploration:
   - Relation exploration: all incoming and outgoing relations from the current frontier entities are retrieved; the LLM selects the most promising relations to answer the question.
   - Entity exploration: entities reachable through selected relations are retrieved, and the LLM selects the most relevant, extended reasoning paths.
3. Reasoning: after each exploration round, the LLM assesses whether the produced paths are sufficient to answer the question; if yes, it produces the answer; otherwise, exploration continues until a maximum depth is reached; if insufficient evidence is found when reaching maximum depth, the LLM switches on its parametric knowledge to produce the final answer.

2.4. ClashEval

ClashEval analyzes the conflict between parametric knowledge and contextual knowledge in LLMs. The authors constructed a dataset of 1,200 QA pairs across five domains (drug dosage, sports records, dates, names, and locations), each paired with a supporting document to use as context. The documents are altered to introduce falsified information (e.g., modifying numerical values, shifting dates, or altering names at different distortion levels). Multiple models are evaluated under three conditions: without context (prior), with correct context and with altered context. Two key metrics are introduced:

- Prior Bias: the probability that a model ignores correct contextual information and sticks to an incorrect prior belief.
- Context Bias: the probability that a model abandons a correct prior belief and follows incorrect contextual information.

Results show that Context Bias generally exceeds Prior Bias, meaning that models tend to follow contextual evidence even when it is incorrect. However, this tendency decreases as perturbation in the context becomes more extreme.

3. Research Questions

This thesis investigates how KG-based RAG influences the balance between parametric knowledge and contextual knowledge in LLMs, especially under contradictory evidence. The study addresses the following research questions:

- **RQ1:** How effectively do KG-based RAG methods make an LLM adhere to contextual information when it contradicts the model's parametric knowledge? How can this trade-off be quantified through Prior Bias and Context Bias?
- **RQ2:** How does adherence to contextual knowledge change as the degree of KG perturbation increases?
- **RQ3:** Which method between GCR, RoG and ToG is more robust and reliable, both in terms of reasoning path generation quality and in terms of final answer accuracy?
- **RQ4:** When KG-based RAG produces a non-adherent answer, to what extent is the error attributable to failures in the path generation phase, versus failures in the final answer generation phase (despite the presence of correct paths)?

4. Dataset Preparation

The experiments are conducted on the WebQSP dataset. Each instance includes a natural language question, question and answer entities, and a KG useful to find the answer. The original test split contains 1,628 questions. A cleaning phase removed structurally inconsistent entries, including cases where answer entities were absent from the graph or unreachable from question entities. After filtering, 1,462 valid questions remained. Questions were manually categorized into five answer types: Locations, Names, Years, Numbers, and Others. Four altered variants of the original dataset were generated: Slight, Significant, Comical, and Uncomp. Alterations were applied to answer entities and propagated consistently throughout the graph. For textual entities (Names, Locations, Others), variations were produced using an LLM at increasing distortion levels. For numerical answers, shifts or multiplicative factors were applied. In the Uncomp variant, answer entities were replaced with their MD5 hashes.

5. Methods

Four experimental settings are evaluated: Plain Question (PQ), Reasoning on Graph (RoG), Graph Constrained Reasoning (GCR), and Think on Graph (ToG). In the PQ setting, questions are submitted directly to the LLM without external knowledge or reasoning paths. This setup serves as a baseline to evaluate the model's performance when relying only on parametric knowledge. For GCR, RoG and ToG, path generation and final answer (FA) generation are analyzed separately to distinguish retrieval failures from reasoning failures. A prediction is classified as:

- Adherent if it matches modified answers.
- Resistant if it matches original answers.
- Incorrect otherwise.

Metrics are computed as proportions over all questions. On the Original dataset, correct answers are labeled as Adherent. To quantify the trade-off between parametric and contextual knowledge, Prior Bias and Context Bias are computed:

- Prior Bias: $Pr(FA = incorrect \mid PQ = incorrect)$
- Context Bias: $Pr(FA = adherent \mid PQ = correct)$

Biases are evaluated only when at least one correct reasoning path is available, separating FA-generation failures from path-generation failures.

Figure 1 summarizes this evaluation flow splitting questions by path availability, PQ correctness, and FA outcome (Adherent/Resistant/Incorrect).

Two models were used for final answer generation: GPT-4.1-standard and GPT-4.1-nano. Experiments with GPT-4.1-nano were conducted on the full cleaned dataset (1,462 questions), while GPT-4.1-standard was evaluated on a subset of 86 questions.

6. Experiments and Results

ToG Think on Graph (ToG) produced unstable results. With GPT-4.1-nano, 77% of the questions resulted in no reasoning path generated; with GPT-4.1-standard, this occurred in 50% of the cases. Due to the high rate of path-generation failures, and the low performance, ToG was not compared with RoG and GCR.

Method	Dataset	Adherence %	Resistance %	Incorrectness %	Prior Bias %	Context Bias %
PQ	Original	58%	0%	42%	-	-
GCR	Original	87%	0%	13%	16%	-
	Slight	64%	10%	26%	-	72%
	Significant	52%	11%	37%	-	59%
	Comical	63%	10%	27%	-	69%
	Uncomp	24%	21%	55%	-	28%
RoG	Original	80%	0%	20%	24%	-
	Slight	50%	21%	30%	-	57%
	Significant	39%	20%	41%	-	46%
	Comical	42%	25%	33%	-	50%
	Uncomp	7%	35%	58%	-	8%

Table 1: Table shows Adherence, Resistance, and Incorrectness metrics for the PQ, GCR, and RoG methods across all datasets. Table also reports the Prior Bias and Context Bias values for the GCR and RoG methods. Answers are classified as adherent if they match the provided context, resistant if they ignore the context but match the ground truth, and incorrect otherwise. Prior Bias (computed only on Original dataset) is the probability that a model ignores correct context and follows an incorrect prior belief; Context Bias (computed only on altered datasets) is the probability that it abandons a correct prior and follows incorrect context. Results refer to the experiments conducted with GPT-4.1-nano model.

Final Answer In the final answer phase, an answer is classified as Adherent if it matches the altered ground truth, Resistant if it matches the original ground truth, or Incorrect otherwise. Table 1 reports the experimental results. Key findings are:

- On the Original dataset, KG-based RAG significantly improves accuracy over PQ.
- Adherence decreases progressively with increased perturbations.
- GCR consistently outperforms RoG (∼15% average improvement).
- GPT-4.1-standard slightly improves Adherence over nano (∼+4%).
- Performance drops sharply on the Uncomp dataset, confirming the role of parametric knowledge when contextual signals become meaningless.

Failures Analysis Failures are divided into:

- Path generation failures: no correct path has been generated.
- Final answer failures: at least one correct path exists, but it is not followed in the FA phase.

Key findings are:

- Most errors (∼70-85%) come from the FA phase, not from missing paths.
- Conditioning Adherence only on questions with at least one correct path increases it by 3-8%.

This shows that even when correct KG evidence is available, models do not always follow it.

Prior Bias and Context Bias Following ClashEval definitions:

- Prior Bias is defined as the number of questions for which the analyzed method (GCR or RoG) gave an Incorrect answer over the number of questions for which PQ gave an Incorrect answer; it is computed only on the Original dataset.
- Context Bias is defined as the number of questions for which the analyzed method (GCR or RoG) gave an Adherent answer over the number of questions for which PQ gave a Correct answer; it is computed only on the modified datasets.

Both biases are computed only considering questions for which at least one correct path has been generated. Figure 1 shows the branches used to compute Prior Bias and Context Bias. Table 1 reports the Prior and Context Bias values for nano model.

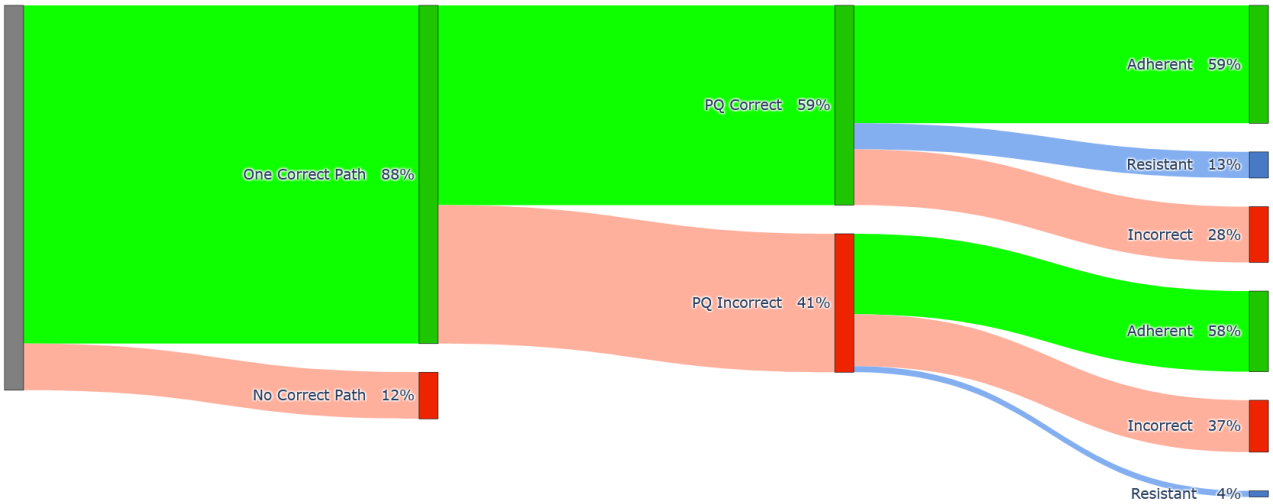

Figure 1: The diagram splits questions by availability of at least one correct reasoning path, PQ correctness (presence of prior knowledge), and final answer outcome (Adherent/Resistant/Incorrect), illustrating the branches used to compute Prior Bias (One correct path → PQ Incorrect → FA Incorrect) and Context Bias (One correct path → PQ Correct → FA Adherent). The diagram refers to the GCR method, using the nano model, on the Significant dataset.

Key findings are:

- Context Bias decreases as perturbation increases.
- GCR shows higher Context Bias than RoG (more likely to follow context).
- Prior Bias is substantially higher than that reported in ClashEval.

This suggests that structured KG context strongly influences model behavior, but not in a reliable way.

6.1. Discussion

RQ1 On altered datasets, Adherence is substantial but not guaranteed: even when at least one correct path exists, Adherence falls in the ∼50-75% band (excluding the Uncomp variant), as reported in Table 1, showing that the model does not always follow KG evidence when it conflicts with parametric knowledge.
The bias analysis shows that Context Bias is generally high but decreases with stronger perturbations, again with an exception on the Comical dataset. Prior Bias is non-negligible on the Original dataset, as shown in Table 1, indicating that even with correct paths available, the model can still fail to override incorrect priors.

RQ2 For both RoG and GCR, Adherence decreases as perturbation grows from Slight to Significant, and collapses on Uncomp. The Comical variant is a consistent exception. Results are shown in Table 1. Overall, stronger perturbations reduce the model's ability to align with contextual knowledge.

RQ3 ToG is not robust in this setting: it fails to generate any path in a large fraction of questions (∼ 77% with nano; ∼ 50% with standard). GCR is the most reliable: it produces only valid paths by construction and has higher final Adherence than RoG (+15% on average), as shown in Table 1. RoG remains competitive on the Original dataset (91% with the standard model), but degrades more steeply under perturbation.

RQ4 Failures are for the most part attributable to the final answer generation phase, not to path generation; from 60% to 86% of errors are attributable to failures in FA. When restricting evaluation to questions with at least one correct path, Adherence improves only by a few percentage points (+3%-8%). This means that a substantial portion of errors remain even when correct evidence is available, indicating that the main bottleneck is the model's ability to correctly use the generated paths.

7. Conclusion

This work investigates the extent to which LLMs adhere to in-prompt information delivered through KG-based RAG methods, even when such information contradicts their parametric knowledge. Starting from the WebQSP dataset, four altered variants were created by modifying correct answers and consistently updating the KG, following an approach inspired by ClashEval. Three KG-based methods (ToG, RoG, GCR) were evaluated against a Plain Question baseline, mainly using GPT-4.1-nano model, with a limited subset tested on GPT-4.1-standard model.

Results show clear differences between the approaches. ToG was found to be unreliable due to frequent failures in generating valid reasoning paths and was excluded from comparative analyses. GCR outperformed RoG, emerging as the most effective method. Analysis of Prior Bias and Context Bias reveals the presence of a discrete alignment with modified KG information, although this effect is not absolute: a Resistance rate of 10-20% suggests that LLMs do not automatically conform to the knowledge presented, even when the prompt explicitly instructs them to do so.

The study presents limitations:

- open-ended evaluation complicates automatic accuracy assessment;
- GPT-4.1-standard was tested on a limited subset due to cost constraints;
- exclusion of ToG reduced the comparative scope.

Future work should analyze ToG failure modes, extend the evaluation to diverse models and architectures, and assess generalizability on alternative datasets.

References

[1] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning, 2024.

[2] Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models, 2025.

[3] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph, 2024.

[4] Kevin Wu, Eric Wu, and James Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence, 2025.