

---

# G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering

---

Xiaoxin He<sup>1</sup>    Yijun Tian<sup>2</sup>    Yifei Sun<sup>1</sup>    Nitesh V. Chawla<sup>2</sup>    Thomas Laurent<sup>3</sup>

Yann LeCun<sup>4,5</sup>    Xavier Bresson<sup>1</sup>    Bryan Hooi<sup>1</sup>

{xiaoxin, xaviercs, bhooi}@comp.nus.edu.sg, sunyifei@u.nus.edu  
{yijun.tian, nchawla}@nd.edu, tlaurent@lmu.edu, yann@cs.nyu.edu

<sup>1</sup>National University of Singapore    <sup>2</sup>University of Notre Dame    <sup>3</sup>Loyola Marymount University  
<sup>4</sup>New York University    <sup>5</sup>Meta AI

## Abstract

Given a graph with textual attributes, we enable users to ‘chat with their graph’: that is, to ask questions about the graph using a conversational interface. In response to a user’s questions, our method provides textual replies and highlights the relevant parts of the graph. While existing works integrate large language models (LLMs) and graph neural networks (GNNs) in various ways, they mostly focus on either conventional graph tasks (such as node, edge, and graph classification), or on answering simple graph queries on small or synthetic graphs. In contrast, we develop a flexible question-answering framework targeting real-world textual graphs, applicable to multiple applications including scene graph understanding, common sense reasoning, and knowledge graph reasoning. Toward this goal, we first develop our Graph Question Answering (GraphQA) benchmark with data collected from different tasks. Then, we propose our *G-Retriever* approach, which integrates the strengths of GNNs, LLMs, and Retrieval-Augmented Generation (RAG), and can be fine-tuned to enhance graph understanding via soft prompting. To resist hallucination and to allow for textual graphs that greatly exceed the LLM’s context window size, *G-Retriever* performs RAG over a graph by formulating this task as a Prize-Collecting Steiner Tree optimization problem. Empirical evaluations show that our method outperforms baselines on textual graph tasks from multiple domains, scales well with larger graph sizes, and resists hallucination.<sup>1</sup>

## 1 Introduction

**Graphs and Large Language Models (LLMs).** The advent of LLMs has significantly shaped the artificial intelligence landscape. As these models are applied to increasingly diverse tasks, their ability to process complex structured data will be increasingly vital. In particular, in our interconnected world, a significant portion of real-world data inherently possesses a graph structure, such as the Web, e-commerce, recommendation systems, knowledge graphs, and many others. Moreover, many of these involve graphs with textual attributes (*i.e.*, *textual graphs*), making them well-suited for LLM-centric methods. This has spurred interest in combining graph-based technologies, particularly graph neural networks (GNNs), with LLMs to enhance their capabilities in understanding and modeling graphs [13, 20, 37].

---

<sup>1</sup>Our codes and datasets are available at: <https://github.com/XiaoxinHe/G-Retriever>.

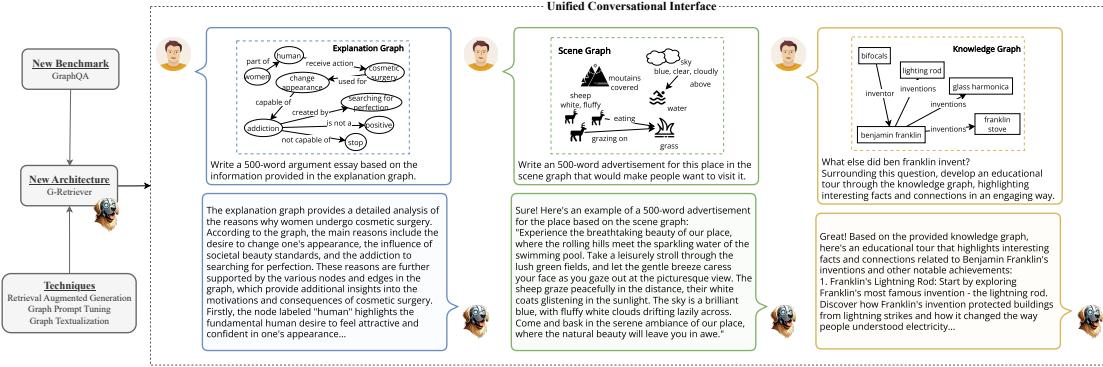


Figure 1: We develop a flexible question-answering framework targeting real-world textual graph applications. Presented here are examples showcasing the model’s adeptness in handling generative and creative queries in practical graph-related tasks: common sense reasoning, scene understanding, and knowledge graph reasoning, respectively.

**The Present Work: Enabling ‘Chat With Your Graph’.** While existing works integrate LLMs and GNNs in various ways, they mostly focus on conventional graph tasks such as node, edge and graph classification [6], or answering simple questions on small or synthetic graphs [37]. In contrast, we develop a flexible question-answering framework targeting real-world textual graph applications. This framework enables users to ‘chat with their graph’ through a conversational interface, suitable even for complex and creative questions about graphs. As shown in Figure 1, our *G-Retriever* shows effective graph understanding abilities, while also incorporating the human-friendliness of conversational interfaces.

**The Need for a Comprehensive GraphQA Benchmark.** Question answering (QA) is a fundamentally important task in natural language processing, serving as a key benchmark for assessing LLMs and providing a unified interface for various capabilities. Despite extensive research in QA, there is a lack of a comprehensive and diverse benchmark specifically for the graph modality. Our study addresses this by introducing our GraphQA benchmark, targeting real-world graph applications: common sense reasoning, scene understanding, and knowledge graph reasoning (refer to Figure 2). This benchmark is vital for measuring the capability of a model in answering a wide range of questions about graphs across diverse applications.

**New Architecture for GraphQA.** To enable effective and efficient graph QA, even on large graphs, we propose *G-Retriever*, a new framework combining the strengths of GNNs, LLMs, and RAG (Figure 3). Next, we will discuss the motivation, strengths, and details of our model.

**Tackling Hallucination in Graph LLMs.** LLMs are prone to hallucination, a phenomenon where the generated content is factually inaccurate or nonsensical [10]. We validate the presence of this issue in graph settings. In particular, we employ a baseline method that adapts MiniGPT-4 [49] to graphs, where a frozen LLM interacts with a trainable GNN that encodes graph data as a soft prompt. Our findings, shown in Table 1, indicate that hallucination, an important problem in text-based LLMs, is also prevalent in Graph LLMs. This may be attributed to the baseline’s inability to recall the entire graph structure from a single graph embedding, leading to the generation of incorrect nodes or edges during the QA task. In contrast, by employing RAG for direct information retrieval from the actual graph, our *G-Retriever* mitigates this issue, as substantiated by Table 1.

**Enhancing Scalability and Efficiency in Graph LLMs.** Recent research endeavors have explored translating graphs into natural language, such as by flattening nodes and edges into a text sequence, enabling their processing by LLMs for graph-based tasks [48]. However, this method faces critical scalability issues. Converting a graph with thousands of nodes and edges into a text sequence results in an excessive number of tokens, surpassing the input capacity of many LLMs. An alternative of truncating the graph text sequence to fit the LLM’s input token limit leads to loss of information and response quality. In contrast, *G-Retriever* overcomes these issues with its RAG component, which allows for effective scaling to larger graphs by selectively retrieving only relevant parts of the graph.

Table 1: Observation and mitigation of hallucination: comparative analysis of baseline graph LLMs and our *G-Retriever*.

|                           |   |
|---------------------------|---|
| User                      | <p>What kind of animal is in the bushes?<br/>Answer the question and list the nodes or edges from the graph that support your answer.</p>   |
| LLM + Graph Prompt Tuning | <p>[Incorrect response with hallucinated nodes and edges]<br/> The animal in the bushes is a deer.<br/> Nodes:<br/> * Deer (node 1)<br/> * Bushes (node 2)<br/> Edges:<br/> * Deer → Bushes (edge 1)<br/> * Deer → Grass (edge 2)<br/> * Bushes → Grass (edge 3)</p>  |
| <i>G-Retriever</i> (Ours) | <p>[Correct response with accurate node and edge references]<br/> Based on the information provided in the graph, the animal in the bushes is an elephant.<br/> The nodes that support this answer are:<br/> * Node 1: The elephant is mentioned in the first line of the graph.<br/> The edges that support this answer are:<br/> * The edge connects the elephant to the bushes, indicating that the elephant is in the bushes.<br/> Therefore, the answer to the question is: The animal in the bushes is an elephant.</p> |

**Tailoring the RAG Approach to Graphs.** Existing RAG methodologies, primarily designed for simpler data types, fall short in handling the complexities of graphs [5]. Hence, we introduce a new retrieval method tailored for graphs. Notably, we formulate subgraph retrieval as a Prize-Collecting Steiner Tree (PCST) optimization problem, enabling a principled and effective solution. This also allows us to return a subgraph relevant to the generated text as an output of our approach, improving explainability.

The contributions of this paper are outlined as follows:

- **New Benchmark for GraphQA.** We introduce a diverse benchmark targeted at real-world graph QA applications, filling a crucial research gap.
- **Enabling ‘Chat with a Graph’.** *G-Retriever* is a flexible graph QA framework suitable even for complex and creative questions. It can be efficiently fine-tuned to enhance graph understanding via soft prompting.
- **Advanced Graph Retrieval Techniques.** We pioneer the integration of RAG with graph LLMs to address the issues of hallucination and lack of scalability, and propose a new graph retrieval method based on PCST.
- **Empirical Findings.** We demonstrate the efficiency and effectiveness of *G-Retriever* in multiple domains, and also present the significant finding of hallucination in graph LLMs.

## 2 Related Work

**Graphs and Large Language Models.** A significant body of research has emerged at the intersection of graph-based techniques and Large Language Models [13, 20, 26, 37, 46]. This exploration spans diverse aspects, ranging from the design of general graph models [15, 21, 33, 39, 43], and multi-modal architectures [19, 41] to practical applications. Noteworthy applications include fundamental graph reasoning [2, 44, 48], node classification [3, 4, 6, 9, 28, 32, 42], graph classification/regression [27, 47], and leveraging LLMs for knowledge graph-related tasks [12, 24, 34].

**Retrieval-Augmented Generation (RAG).** The concept of Retrieval-Augmented Generation, initially proposed by Lewis *et al.* [17], has gained increased attention for its ability to mitigate the issue of

hallucination within LLMs and enhance trustworthiness and explainability [5]. Despite its success in language-related tasks, the application of retrieval-based approaches to general graph tasks remains largely unexplored. Our research is the first to apply a retrieval-based approach to general graph tasks, marking a novel advancement in the field and demonstrating the versatility of RAG beyond language processing.

**Parameter-Efficient Fine-Tuning (PEFT).** The field of LLMs has witnessed significant advancements through various parameter-efficient fine-tuning techniques. These methodologies have played a crucial role in refining LLMs, boosting their performance while minimizing the need for extensive parameter training. Notable among these techniques are prompt tuning, as introduced by Lester *et al.* [16], and prefix tuning, proposed by Li and Liang [18]. Furthermore, methods like LoRA [8], and the LLaMA-adapter [45], have been influential. These advancements in PEFT have laid the foundation for the development of sophisticated multimodal models. Prominent examples in this domain include MiniGPT-4 [49], LLaVA [22], and NExT-Chat [38]. There are also emerging efforts in applying PEFT to graph LLMs, such as GraphLLM [2] for basic graph reasoning tasks and GNP [34] for multi-option QA on knowledge graphs.

### 3 Formalization

This section establishes the notation and formalizes key concepts related to textual graphs, language models for text encoding, and large language models and prompt tuning.

**Textual Graphs.** A textual graph is a graph where nodes and edges possess textual attributes. Formally, it can be defined as  $G = (V, E, \{x_n\}_{n \in V}, \{x_e\}_{e \in E})$ , where  $V$  and  $E$  represent the sets of nodes and edges, respectively. Additionally,  $x_n \in D^{L_n}$  and  $x_e \in D^{L_e}$  denote sequential text associated with a node  $n \in V$  or an edge  $e \in E$ , where  $D$  represents the vocabulary, and  $L_n$  and  $L_e$  signify the length of the text associated with the respective node or edge.

**Language Models for Text Encoding.** In the context of textual graphs, language models (LMs) are essential for encoding the text attributes associated with nodes and edges, thereby learning representations that capture their semantic meaning. For a node  $n$  with text attributes  $x_n \in D^{L_n}$ , an LM encodes these attributes as:

$$z_n = \text{LM}(x_n) \in \mathbb{R}^d, \quad (1)$$

where  $z_n$  is the output of the LM, and  $d$  is the dimension of the output vector.

**Large Language Models and Prompt Tuning.** LLMs have introduced a new paradigm for task-adaptation known as “pre-train, prompt, and predict”, replacing the traditional “pre-train, fine-tune” paradigm. In this paradigm, the LLM is first pre-trained on a large corpus of text data to learn general language representations. Then, rather than fine-tuning the model on task-specific labeled data, the model is prompted with a textual prompt that specifies the task and context. Subsequently, the model generates the output directly based on the prompt and the input.

The LLM, parameterized by weights  $\theta$ , takes a sequence of tokens  $X$ , and a prompt  $P$  as input, and generate a sequence of tokens  $Y = \{y_1, y_2, \dots, y_r\}$  as output. Formally, the probability distribution of the output sequence given the concatenated input sequence and prompt, *i.e.*,  $[P; X]$ , is defined as follows:

$$p_\theta(Y|[P; X]) = \prod_{i=1}^r p_\theta(y_i|y_{<i}, [P; X]). \quad (2)$$

Here,  $y_{<i}$  represents the prefix of sequence  $y$  up to position  $i - 1$ , and  $p(y_i|y_{<i}, [P; X])$  represents the probability of generating token  $y_i$  given  $y_{<i}$  and  $[P; X]$ .

Soft prompt tuning eliminates the need for manual prompt design. Given a series of  $p$  tokens  $X = \{x_1, x_2, \dots, x_p\}$ , after being processed by the text embedder, it forms a matrix  $X_e \in \mathbb{R}^{p \times d_l}$ , where  $d_l$  is the dimension of the embedding space. Soft prompts can be represented as parameters  $P_e \in \mathbb{R}^{q \times d_l}$ , where  $q$  is the length of the prompt. The prompt is then concatenated with the embedded input, forming a single matrix  $[P_e; X_e] \in \mathbb{R}^{(q+p) \times d_l}$ . This combined matrix is processed by the self-attention layers in LLM as usual. Training involves maximizing the likelihood of  $Y$  through backpropagation, with gradient updates applied solely to  $P_e$ , while  $\theta$  remains fixed.

Table 2: Summary of datasets used in GraphQA benchmark.

| Dataset           | ExplaGraphs            | SceneGraphs  | WebQSP                             |
|-------------------|------------------------|--|------------------------------------|
| #Graphs           | 2,766                  | 10,000   | 4,737                              |
| Avg. #Nodes       | 5.17                   | 19.13  | 1370.89                            |
| Avg. #Edges       | 4.25                   | 68.44  | 4252.37                            |
| Node Attribute    | Commonsense concepts   | Object attributes (e.g., color, shape, material, activity) | Entities in Freebase               |
| Edge Attribute    | Commonsense relations  | Relations (e.g., actions, spatial relations, comparatives) | Relations in Freebase              |
| Task              | Common sense reasoning | Scene graph question answering                             | Knowledge based question answering |
| Evaluation Matrix | Accuracy               | Accuracy   | Hit@1                              |

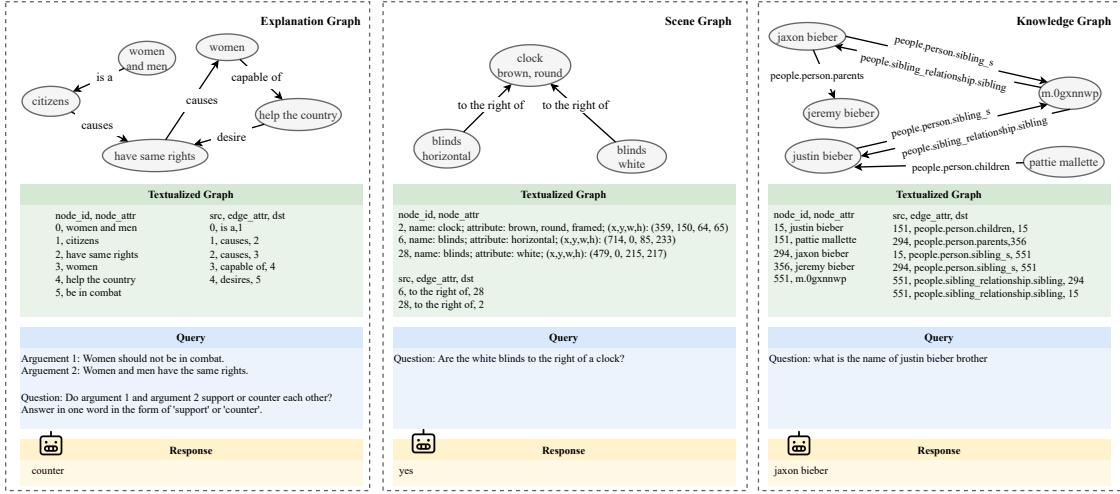


Figure 2: Illustrative examples from the GraphQA benchmark datasets.

## 4 Proposed GraphQA Benchmark

Our GraphQA represents a comprehensive and diverse benchmark for graph question-answering. It is tailored to assess the capabilities of models in answering a wide range of questions about graphs across diverse domains.

### 4.1 Data Format

Each entry in the GraphQA benchmark consists of a textual graph, a question related to the graph, and one or more corresponding answers, as illustrated in Figure 2.

**Textual Graphs.** The textual graph is converted into a natural language format, resulting in a list of nodes and edges, akin to a CSV file format. It is important to note that while multiple methods exist for textualizing a graph, our focus is not on identifying the optimal solution. Instead, we prioritize a straightforward yet empirically effective approach for representing graphs in natural language, facilitating the benchmark’s use in diverse GraphQA scenarios.

**Questions and Answers.** Questions are designed to explore specific elements or relationships within the graph. Answers, residing within the attributes of nodes or edges, often require multi-hop reasoning for accurate identification.

### 4.2 Description of Datasets

The GraphQA benchmark integrates three existing datasets: ExplaGraphs, SceneGraphs, and WebQSP. Table 2 presents the summary statistics of these datasets. It is important to note that these datasets were not originally developed for this work. However, a significant contribution of our research is the standardization and processing of these diverse datasets into a uniform data format suitable for the GraphQA benchmark. These datasets, previously utilized in different contexts, are

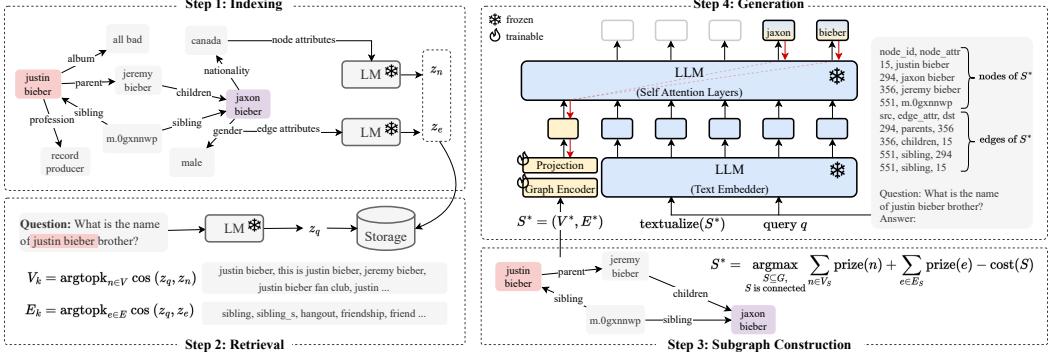


Figure 3: Overview of the proposed *G-Retriever*: 1) Indexing: Graphs are indexed for efficient query processing; 2) Retrieval: The most semantically relevant nodes and edges are retrieved, conditioned on the query; 3) Subgraph Construction: A connected subgraph is extracted, covering as many relevant nodes and edges as possible while maintaining a manageable graph size; 4) Generation: An answer is generated using a ‘graph prompt’, a textualized graph, and the query.

reintroduced with a new focus tailored for GraphQA. For a detailed comparison with the original datasets, see the Appendix B.

**ExplaGraphs** is a dataset for generative commonsense reasoning, focusing on creating explanation graphs for stance prediction in debates. It offers detailed, unambiguous commonsense-augmented graphs to evaluate arguments supporting or refuting a belief. The primary task is to assess whether arguments are supportive or contradictory, using accuracy as the metric. We have converted the triplet-form provided in Saha *et al.* [30] into a standard graph format.

**SceneGraphs**, a visual question answering dataset, includes 10,000 scene graphs. Each graph details objects, attributes, and relations within an image. This dataset challenges users with tasks requiring spatial understanding and multi-step inference. The task is to answer open-ended questions based on a textual description of a scene graph, evaluated on accuracy. We have sampled from the GQA dataset [11] and constructed standard graphs from the provided JSON files.

**WebQSP** is a large-scale multi-hop knowledge graph QA dataset consisting of 4,737 questions. It was proposed by Yih *et al.* [40] and, following Luo *et al.* [25], utilizes a subset of Freebase, encompassing facts within 2-hops of entities mentioned in the questions. The task involves answering questions that require multi-hop reasoning. Given the possibility of multiple answers for the same question, the hit@1 metric is used to assess the precision of the top returned answer.

## 5 G-Retriever

In this section, we introduce *G-Retriever*, a new architecture tailored for GraphQA, which integrates the strengths of GNNs, LLMs, and RAG. To allow efficient fine-tuning while preserving the LLM’s pretrained language capabilities, we freeze the LLM and use a soft prompting approach on the output of the GNN. Our RAG-based design mitigates hallucinations through direct retrieval of the graph, while allowing our approach to scale to graphs exceeding the LLM’s context window size. To adapt RAG to graphs, we formulate subgraph retrieval as a PCST optimization problem. This approach also allows us to enhance explainability by returning the retrieved subgraph.

*G-Retriever* comprises four main steps: indexing, retrieval, subgraph construction and generation, as depicted in Figure 3. The implementation details of each step are elaborated in the following sections.

### 5.1 Indexing

We initiate the RAG approach by generating node and graph embeddings using a pre-trained LM. These embeddings are then stored in a nearest neighbor data structure.

To elaborate, consider  $x_n \in D^{L_n}$  as the text attributes of node  $n$ . Utilizing a pre-trained LM, such as SentenceBert [29], we apply the LM to  $x_n$ , yielding the representation  $z_n$ :

$$z_n = \text{LM}(x_n) \in \mathbb{R}^d, \quad (3)$$

where  $d$  denotes the dimension of the output vector. Similar preprocessing steps are applied to edges. Refer to Figure 3, Step 1 for an illustrative representation.

## 5.2 Retrieval

For retrieval, we employ the same encoding strategy to the query to ensure consistent treatment of textual information:

$$z_q = \text{LM}(x_q) \in \mathbb{R}^d. \quad (4)$$

Next, to identify the most relevant nodes and edges for the current query, we use a k-nearest neighbors retrieval approach. This method yields a set of ‘relevant nodes/edges’ based on the similarity between the query and each node or edge. The retrieval operation is defined as:

$$\begin{aligned} V_k &= \text{argtopk}_{n \in V} \cos(z_q, z_n) \\ E_k &= \text{argtopk}_{e \in E} \cos(z_q, z_e), \end{aligned} \quad (5)$$

where  $z_n$  and  $z_e$  are the embeddings of node  $n$  and edge  $e$ , respectively. We use the cosine similarity function,  $\cos(\cdot, \cdot)$ , to measure the similarity between the query representation and the node/edge embeddings. The argtopk operation retrieves the top-k elements based on this similarity, providing a set of nodes  $V_k$  and edges  $E_k$  considered most relevant to the query. See Step 2 of Figure 3.

## 5.3 Subgraph Construction

This step aims to construct a subgraph that encompasses as many relevant nodes and edges as possible, while keeping the graph size manageable. This approach offers two key benefits: Firstly, it helps to filter out nodes and edges that are not pertinent to the query. This is crucial because irrelevant information can overshadow the useful data, potentially diverting the focus of the subsequent LLM from the information of interest. Secondly, it enhances efficiency; by keeping the graph size manageable, it becomes feasible to translate the graph into natural language and then input it into the LLM for processing. The Prize-Collecting Steiner Tree algorithm [1] serves as our primary method for identifying such optimally sized and relevant subgraphs. See Step 3 in Figure 3.

**Prize-Collecting Steiner Tree (PCST).** The PCST problem aims to find a connected subgraph that maximizes the total prize values of its nodes while minimizing the total costs of its edges. Our approach assigns higher prize values to nodes and edges more relevant to the query, as measured by cosine similarity. Specifically, the top  $k$  nodes/edges are assigned descending prize values from  $k$  down to 1, with the rest assigned zero. The node prize assignment is as follows:

$$\text{prize}(n) = \begin{cases} k - i, & \text{if } n \in V_k \text{ and } n \text{ is the top } i \text{ node,} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

Edge prizes are assigned similarly.

The objective is to identify a subgraph,  $S^* = (V^*, E^*)$ , that optimizes the total prize of nodes and edges, minus the costs associated with the size of the subgraph:

$$S^* = \underset{\substack{S \subseteq G, \\ S \text{ is connected}}}{\text{argmax}} \sum_{n \in V_S} \text{prize}(n) + \sum_{e \in E_S} \text{prize}(e) - \text{cost}(S), \quad (7)$$

where

$$\text{cost}(S) = |E_S| \times C_e, \quad (8)$$

and  $C_e$  denotes a predefined cost per edge, which is adjustable to control the subgraph size.

The original PCST algorithm is designed for node prizes only. However, given the significance of edge semantics in certain scenarios, we adapt the algorithm to accommodate edge prizes as follows: Consider an edge  $e$  with a cost  $C_e$  and a prize  $P_e$ . If  $C_e > P_e$ , it can be treated as a reduced edge cost of  $C_e - P_e$ . However, if  $P_e > C_e$ , negative edge costs are not allowed in the original algorithm. Our

solution involves replacing edge  $e$  with a ‘virtual node’  $v_e$ , connected to both endpoints of  $e$ . This virtual node is assigned a prize of  $P_e - C_e$ , and the cost of the two new edges leading to the virtual node is set to zero. This modification effectively mirrors the original problem, as including edge  $e$  in the original graph is analogous to including the virtual node in the modified graph. Finally, we optimize the PCST problem using a near-linear time approach [7].

#### 5.4 Answer Generation

**Graph Encoder.** Let  $S^* = (V^*, E^*)$  represent the retrieved subgraph. We use a graph encoder to model the structure of this graph, specifically using a standard Graph Attention Network (GAT) [36]. Our approach for encoding the retrieved subgraph is defined as follows:

$$h_g = \text{POOL}(\text{GNN}_{\phi_1}(S^*)) \in \mathbb{R}^{d_g}, \quad (9)$$

Here, POOL denotes the mean pooling operation, and  $d_g$  is the dimension of the graph encoder.

**Projection Layer.** We incorporate a multilayer perceptron (MLP) to align the graph token with the vector space of the LLM:

$$\hat{h}_g = \text{MLP}_{\phi_2}(h_g) \in \mathbb{R}^{d_l}, \quad (10)$$

where  $d_l$  is the dimension of the LLM’s hidden embedding.

**Text Embedder.** To leverage the text-reasoning capabilities of LLMs, we transform the retrieved subgraph  $S^*$  into a textual format. This transformation involves flattening the textual attributes of the nodes and edges, as illustrated in the green box in Figure 2. We refer to this operation as  $\text{textualize}(\cdot)$ . Subsequently, we combine the textualized graph with the query to generate a response. Let  $x_q$  denote the query; we concatenate it with the textualized graph  $\text{textualize}(S^*)$ . We then map the result to an embedding  $h_t$  using a text embedder, which is the first layer of a pretrained and frozen LLM:

$$h_t = \text{TextEmbedder}([\text{textualize}(S^*); x_q]) \in \mathbb{R}^{L \times d_l}, \quad (11)$$

where  $[;]$  represents the concatenation operation, and  $L$  is the number of tokens.

**LLM Generation with Graph Prompt Tuning.** The final stage involves generating the answer  $Y$  given the graph token  $\hat{h}_g$ , acting as a soft prompt, and the text embedder output  $h_t$ . These inputs are fed through the self-attention layers of a pretrained frozen LLM, with parameter  $\theta$ . The generation process is represented as follows:

$$p_{\theta, \phi_1, \phi_2}(Y|S^*, x_q) = \prod_{i=1}^r p_{\theta, \phi_1, \phi_2}(y_i|y_{<i}, [\hat{h}_g; h_t]), \quad (12)$$

where  $[\hat{h}_g; h_t]$  concatenates the graph token  $\hat{h}_g$  and the text embedder output  $h_t$ . While  $\theta$  is frozen, the graph token  $\hat{h}_g$  receives gradients, enabling the optimization of the parameters of the graph encoder  $\phi_1$  and the projection layer  $\phi_2$  through standard backpropagation.

## 6 Experiments

### 6.1 Experiment Setup

In the indexing step, we use SentenceBert [29] as the LM to encode all node and edge attributes. In the generation step, we use the open-source Llama2-7b [35] as the LLM for all experiments. We use GAT [36] as the graph encoder. Additional details are provided in Appendix A.1.

### 6.2 Model Configurations

In our experiments, we consider three model configurations: 1) *Inference-Only*: Using a frozen LLM for direct question answering; 2) *Frozen LLM + Prompt Tuning (PT)*: Keeping the parameters of the LLM frozen and adapting only the prompt; 3) *Tuned LLM*: Fine-tuning the LLM with LoRA.

Table 3: Performance comparison across ExplaGraphs, SceneGraphs, and WebQSP datasets. This table presents mean scores and standard deviations (mean  $\pm$  std) for different configurations, including Inference-Only, Frozen LLM , and Tuned LLM settings, across the three datasets. A dashed entry indicates the non-applicability of the ‘Question-Only’ method for SceneGraphs, where questions cannot be answered without the textual graph.

| Setting                    | Method                    | ExplaGraphs         | SceneGraphs         | WebQSP              |
|----------------------------|---------------------------|---------------------|---------------------|---------------------|
| Inference-Only             | Question-Only             | 0.5704              | –                   | 0.6136              |
|                            | Textual Graph + Question  | 0.5812              | 0.3548              | 0.4195              |
| Frozen LLM + Prompt Tuning | Prompt Tuning             | $0.5876 \pm 0.0032$ | $0.6851 \pm 0.0011$ | $0.4975 \pm 0.0049$ |
|                            | <i>G-Retriever</i>        | $0.8696 \pm 0.0129$ | $0.8614 \pm 0.0080$ | $0.6732 \pm 0.0076$ |
|                            | $\Delta$ Prompt Tuning    | $\uparrow 47.99\%$  | $\uparrow 22.97\%$  | $\uparrow 35.32\%$  |
| Tuned LLM                  | LoRA                      | $0.8741 \pm 0.0186$ | $0.8594 \pm 0.0062$ | $0.6174 \pm 0.0012$ |
|                            | <i>G-Retriever + LoRA</i> | $0.8768 \pm 0.0274$ | $0.9077 \pm 0.0104$ | $0.7011 \pm 0.005$  |
|                            | $\Delta$ LoRA             | $\uparrow 0.31\%$   | $\uparrow 5.62\%$   | $\uparrow 13.56\%$  |

Table 4: Average number of tokens and nodes before and after implementing retrieval.

| Dataset     | Before Retrieval (Avg.) |         | After Retrieval (Avg.)    |                          |
|-------------|-------------------------|---------|---------------------------|--------------------------|
|             | # Tokens                | # Nodes | # Tokens                  | # Nodes                  |
| SceneGraphs | 1,396                   | 19      | 235 ( $\downarrow 83\%$ ) | 5 ( $\downarrow 74\%$ )  |
| WebQSP      | 100,627                 | 1,371   | 610 ( $\downarrow 99\%$ ) | 18 ( $\downarrow 99\%$ ) |

### 6.3 Main Results

Table 3 demonstrates the effectiveness of our method under various configurations across three datasets: ExplaGraphs, SceneGraphs, and WebQSP.

**Inference-Only Performance:** *G-Retriever* surpasses all inference-only baselines. Notably, the integration of textual graphs did not consistently yield positive results, as observed in WebQSP. This inconsistency might be attributed to the complexity and potential noise of the textualized graph, which could hinder the LLM’s direct reasoning capabilities.

**Frozen LLM + Prompt Tuning Performance:** *G-Retriever* outperforms traditional prompt tuning across all datasets, with an average performance increase of 35%, underscoring the effectiveness of our approach in leveraging frozen LLMs for improved performance.

**Tuned LLM Performance:** The combination of our method with LoRA (*i.e.*, *G-Retriever + LoRA*) achieves the best performance, outperforming standard LoRA fine-tuning by 5.62% on SceneGraphs and 13.56% on WebQSP.

### 6.4 Efficiency Evaluation

The efficiency of our approach is highlighted by the data presented in Table 4. Implementing our graph-based retrieval led to a significant reduction in the average number of tokens and nodes across datasets. Specifically, for SceneGraphs, an 83% reduction in the average number of tokens was observed. In WebQSP, which features a larger graph size, there was a remarkable decrease of 99% in both tokens and nodes. These substantial reductions demonstrate the efficiency of our method, underscoring its potential in managing large-scale graph data.

### 6.5 Mitigation of Hallucination

To evaluate hallucination, we instructed the models to answer graph-related questions, specifically identifying supporting nodes or edges from the graph. We manually reviewed 100 responses from both our method and the baseline (*i.e.*, LLM with graph prompt tuning), verifying the existence of the nodes and edges referenced in the model’s output within the actual graph. Table 5 shows that

Table 5: Quantitative comparison of hallucination on the SceneGraphs dataset.

|                    | Baseline | <i>G-Retriever</i> |
|--------------------|----------|--------------------|
| Valid Nodes        | 31%      | 77%                |
| Valid Edges        | 12%      | 76%                |
| Fully Valid Graphs | 8%       | 62%                |

Table 6: Ablation study on the WebQSP dataset.

| Method                | Hit@1               | $\Delta_{G\text{-}Retriever}$ |
|-----------------------|---------------------|-------------------------------|
| w/o Graph Encoder     | $0.5963 \pm 0.0127$ | $\downarrow 11.42\%$          |
| w/o Projection Layer  | $0.6576 \pm 0.0083$ | $\downarrow 2.32\%$           |
| w/o Textualized Graph | $0.5414 \pm 0.0082$ | $\downarrow 19.58\%$          |
| w/o Node Retrieval    | $0.6658 \pm 0.0029$ | $\downarrow 1.10\%$           |
| w/o Edge Retrieval    | $0.5837 \pm 0.0068$ | $\downarrow 13.29\%$          |

*G-Retriever* significantly reduces hallucinations by 54% compared to the baseline (see Appendix C for details).

## 6.6 Ablation Study

In this ablation study, we assess the individual impact of key components within our pipeline, structured into two parts:

**Architecture Design:** We evaluate the roles of the graph encoder, projection head, and textualized graph. The study indicates performance drops when any of these components are removed, with the graph encoder and textualized graph showing declines of 11.42% and 19.58%, respectively. This demonstrates their complementary effects in representing the graph in both textual and embedded formats.

**Retrieval Mechanism:** We also assess the impact of various retrieval strategies, including scenarios where either node retrieval or edge retrieval is omitted. Notably, the exclusion of node retrieval leads to a minor decrease in performance by 1.10%, while omitting edge retrieval results in a more substantial decline of 13.29%.

Further details are available in Appendix A.3. Additionally, we present further studies on our framework in the appendix: it is robust to the choice of graph encoders (see Appendix A.4) and benefits from the increased scale of LLMs (see Appendix A.5).

## 7 Conclusion

In this work, we introduce a new GraphQA benchmark for real-world graph question answering and present *G-Retriever*, an architecture adept at complex and creative queries. Experimental results show that *G-Retriever* surpasses baselines in textual graph tasks across multiple domains, scales effectively with larger graph sizes, and demonstrates resistance to hallucination.

**Limitations and Future Work:** Currently, *G-Retriever* employs a static retrieval component. Future developments could investigate more sophisticated RAG where the retrieval is trainable.

## Acknowledgment

XB is supported by NUS Grant ID R-252-000-B97-133.

## References

- [1] D. Bienstock, M. X. Goemans, D. Simchi-Levi, and D. Williamson, “A note on the prize collecting traveling salesman problem,” *Mathematical programming*, vol. 59, no. 1-3, pp. 413–420, 1993.
- [2] Z. Chai *et al.*, “Graphilm: Boosting graph reasoning ability of large language model,” *arXiv preprint arXiv:2310.05845*, 2023.
- [3] Z. Chen *et al.*, “Exploring the potential of large language models (llms) in learning on graphs,” *arXiv preprint arXiv:2307.03393*, 2023.
- [4] Z. Chen *et al.*, “Label-free node classification on graphs with large language models (llms),” *arXiv preprint arXiv:2310.04668*, 2023.
- [5] Y. Gao *et al.*, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint arXiv:2312.10997*, 2023.
- [6] X. He, X. Bresson, T. Laurent, A. Perold, Y. LeCun, and B. Hooi, “Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning,” *arXiv preprint arXiv:2305.19523*, 2023.
- [7] C. Hegde, P. Indyk, and L. Schmidt, “A nearly-linear time framework for graph-structured sparsity,” in *International Conference on Machine Learning*, PMLR, 2015, pp. 928–937.
- [8] E. J. Hu *et al.*, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [9] J. Huang, X. Zhang, Q. Mei, and J. Ma, “Can llms effectively leverage graph structural information: When and why,” *arXiv preprint arXiv:2309.16595*, 2023.
- [10] L. Huang *et al.*, *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*, 2023. arXiv: 2311.05232 [cs.CL].
- [11] D. A. Hudson and C. D. Manning, “Gqa: A new dataset for real-world visual reasoning and compositional question answering,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.
- [12] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen, “Structgpt: A general framework for large language model to reason over structured data,” *arXiv preprint arXiv:2305.09645*, 2023.
- [13] B. Jin, G. Liu, C. Han, M. Jiang, H. Ji, and J. Han, “Large language models on graphs: A comprehensive survey,” *arXiv preprint arXiv:2312.02783*, 2023.
- [14] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [15] B. Lei, C. Liao, C. Ding, *et al.*, “Boosting logical reasoning in large language models through a new framework: The graph of thought,” *arXiv preprint arXiv:2308.08614*, 2023.
- [16] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” *arXiv preprint arXiv:2104.08691*, 2021.
- [17] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.
- [18] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” *arXiv preprint arXiv:2101.00190*, 2021.
- [19] X. Li, D. Lian, Z. Lu, J. Bai, Z. Chen, and X. Wang, “Graphadapter: Tuning vision-language models with dual knowledge graph,” *arXiv preprint arXiv:2309.13625*, 2023.
- [20] Y. Li *et al.*, “A survey of graph meets large language model: Progress and future directions,” *arXiv preprint arXiv:2311.12399*, 2023.
- [21] H. Liu *et al.*, “One for all: Towards training one graph model for all classification tasks,” *arXiv preprint arXiv:2310.00149*, 2023.
- [22] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *arXiv preprint arXiv:2304.08485*, 2023.
- [23] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [24] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, “Reasoning on graphs: Faithful and interpretable large language model reasoning,” *arXiv preprint arXiv:2310.01061*, 2023.
- [25] L. Luo, Y.-F. Li, G. Haffari, and S. Pan, “Reasoning on graphs: Faithful and interpretable large language model reasoning,” *arXiv preprint arXiv:2310.01061*, 2023.

- [26] S. Pan, Y. Zheng, and Y. Liu, “Integrating graphs with large language models: Methods and prospects,” *arXiv preprint arXiv:2310.05499*, 2023.
- [27] C. Qian, H. Tang, Z. Yang, H. Liang, and Y. Liu, “Can large language models empower molecular property prediction?” *arXiv preprint arXiv:2307.07443*, 2023.
- [28] Y. Qin, X. Wang, Z. Zhang, and W. Zhu, “Disentangled representation learning with large language models for text-attributed graphs,” *arXiv preprint arXiv:2310.18152*, 2023.
- [29] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [30] S. Saha, P. Yadav, L. Bauer, and M. Bansal, “Explagraphs: An explanation graph generation task for structured commonsense reasoning,” *arXiv preprint arXiv:2104.07644*, 2021.
- [31] Y. Shi, Z. Huang, S. Feng, H. Zhong, W. Wang, and Y. Sun, “Masked label prediction: Unified message passing model for semi-supervised classification,” *arXiv preprint arXiv:2009.03509*, 2020.
- [32] S. Sun, Y. Ren, C. Ma, and X. Zhang, “Large language models as topological structure enhancers for text-attributed graphs,” *arXiv preprint arXiv:2311.14324*, 2023.
- [33] J. Tang *et al.*, “Graphgpt: Graph instruction tuning for large language models,” *arXiv preprint arXiv:2310.13023*, 2023.
- [34] Y. Tian *et al.*, “Graph neural prompting with large language models,” *arXiv preprint arXiv:2309.15427*, 2023.
- [35] H. Touvron *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [36] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” *arXiv preprint arXiv:1710.10903*, 2017.
- [37] H. Wang, S. Feng, T. He, Z. Tan, X. Han, and Y. Tsvetkov, “Can language models solve graph problems in natural language?” *arXiv preprint arXiv:2305.10037*, 2023.
- [38] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, “Next-gpt: Any-to-any multimodal llm,” *arXiv preprint arXiv:2309.05519*, 2023.
- [39] R. Ye, C. Zhang, R. Wang, S. Xu, and Y. Zhang, “Natural language is all a graph needs,” *arXiv preprint arXiv:2308.07134*, 2023.
- [40] W.-t. Yih, M. Richardson, C. Meek, M.-W. Chang, and J. Suh, “The value of semantic parse labeling for knowledge base question answering,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2016, pp. 201–206.
- [41] M. Yoon, J. Y. Koh, B. Hooi, and R. Salakhutdinov, “Multimodal graph learning for generative tasks,” *arXiv preprint arXiv:2310.07478*, 2023.
- [42] J. Yu, Y. Ren, C. Gong, J. Tan, X. Li, and X. Zhang, “Empower text-attributed graphs learning with large language models (llms),” *arXiv preprint arXiv:2310.09872*, 2023.
- [43] J. Yu, R. He, and R. Ying, “Thought propagation: An analogical approach to complex reasoning with large language models,” *arXiv preprint arXiv:2310.03965*, 2023.
- [44] J. Zhang, “Graph-toolformer: To empower llms with graph reasoning ability via prompt augmented by chatgpt,” *arXiv preprint arXiv:2304.11116*, 2023.
- [45] R. Zhang *et al.*, “Llama-adapter: Efficient fine-tuning of language models with zero-init attention,” *arXiv preprint arXiv:2303.16199*, 2023.
- [46] Z. Zhang, H. Li, Z. Zhang, Y. Qin, X. Wang, and W. Zhu, *Graph meets llms: Towards large graph models*, 2023. arXiv: 2308.14522 [cs.LG].
- [47] H. Zhao *et al.*, “Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning,” *bioRxiv*, pp. 2023–05, 2023.
- [48] J. Zhao *et al.*, “Graphtext: Graph reasoning in text space,” *arXiv preprint arXiv:2310.01089*, 2023.
- [49] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.

## A Experiment

### A.1 Implementation Settings

Experiments are conducted using 2 NVIDIA A100-80G GPUs. Each experiment is replicated four times, utilizing different seeds for each run to ensure robustness and reproducibility.

**Graph Encoder.** We use GAT [36] as the GNN backbone. Our configuration employs 4 layers, each with 4 attention heads, and a hidden dimension size of 1024.

**LLM.** We use the open-sourced Llama2-7b [35] as the LLM backbone. In fine-tuning the LLM with LoRA [8], the `lora_r` parameter (dimension for LoRA update matrices) is set to 8, and `lora_alpha` (scaling factor) is set to 16. The dropout rate is set to 0.05. In prompt tuning, the LLM is configured with 10 virtual tokens. The number of max text length is 512, the number of max new tokens, *i.e.*, the maximum numbers of tokens to generate, is 32.

**PCST.** For retrieval over graphs via PCST, for the SceneGraphs dataset, we select the top  $k$  nodes and edges, setting  $k$  to 3. Here, the cost of edges, denoted as  $C_e$ , is set to 1. Regarding the WebQSP dataset, we set  $k = 3$  for nodes and  $k = 5$  for edges, with the edge cost,  $C_e$ , adjusted to 0.5. For the ExplaGraphs dataset, which is characterized by a small graph size averaging 5.17 nodes and 4.25 edges (as detailed in Table 2), the entire graph can fit in the LLM’s context window size. Consequently, we aim to retrieve the whole graph by setting  $k$  to 0, effectively returning the original graph unaltered.

**Optimization.** We use the AdamW [23] optimizer. We set the initial learning rate at 1e-5, with a weight decay of 0.05. The learning rate decays with a half-cycle cosine decay after the warm-up period. The batch size is 4, and the number of epochs is 10. To prevent overfitting and ensure training efficiency, an early stopping mechanism is implemented with a patience setting of 2 epochs.

### A.2 Details of Model Configurations

In our experiments, we consider three model configurations:

1) *Inference-Only*: Using a frozen LLM for direct question answering, with query (see Figure 4a), or with textual graph and query (see Figure 4b).

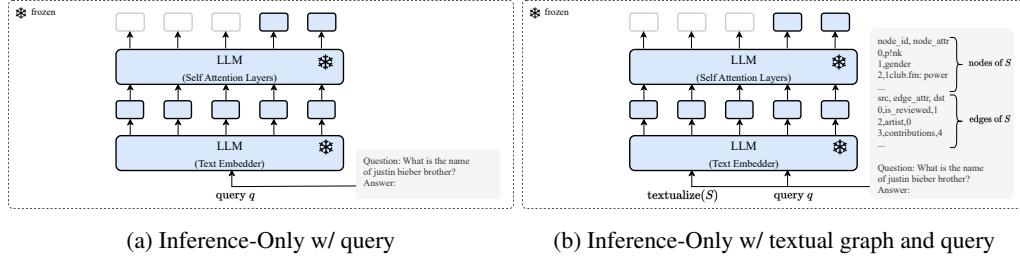


Figure 4: Model configuration 1) *Inference-Only*.

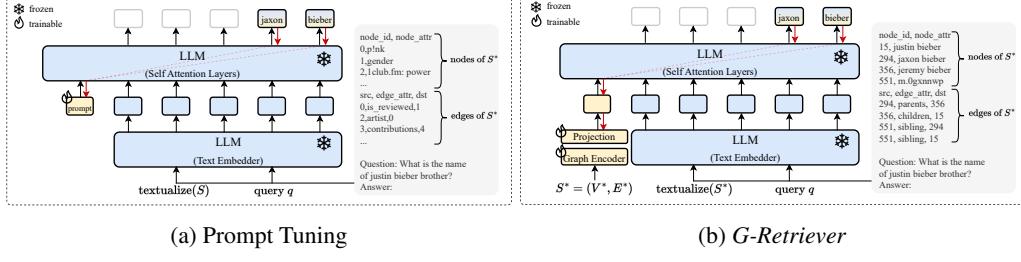
2) *Frozen LLM + Prompt Tuning (PT)*: Keeping the parameters of the LLM frozen and adapting only the prompt. This includes soft prompt tuning (see Figure 5a) and our *G-Retriever* method (see Figure 5b).

3) *Tuned LLM*: Fine-tuning the LLM with LoRA. Refer to Figure 6a for the standard fine-tuning of an LLM for downstream tasks, and see Figure 6b for the combined approach of fine-tuning the LLM with G-Retriever.

### A.3 Details of Ablation Study

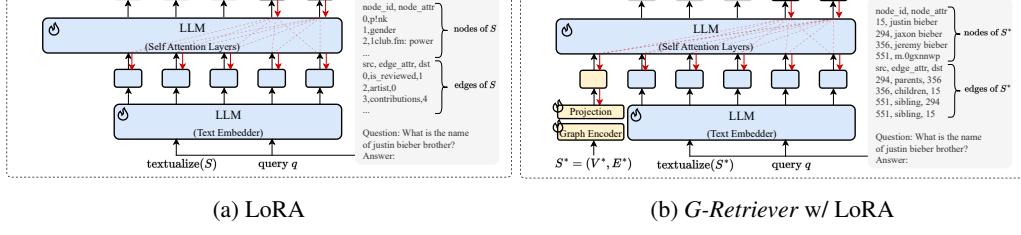
This section illustrates the modifications made to the original architecture in the ablation study, as presented in Figure 7.

**Without Graph Encoder (w/o GraphEncoder):** In this setting, we replaced the graph encoder with trainable soft tokens, setting the number of these virtual tokens to 10.



(a) Prompt Tuning (b) *G-Retriever*

Figure 5: Model configuration 2) *Frozen LLM + Prompt Tuning*.

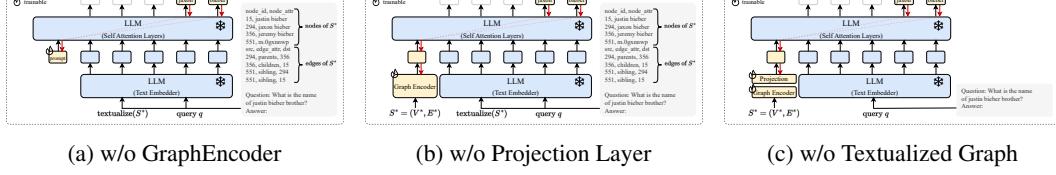


(a) LoRA (b) *G-Retriever w/ LoRA*

Figure 6: Model configuration 3) *Tuned LLM*.

**Without Projection Layer (w/o Projection Layer):** Here, we removed the projection layer following the graph encoder. We configured the output dimension of the graph encoder to be 4096, matching the hidden dimension of Llama2-7b. This allows the output graph token (the yellow token in Figure 7b) to be concatenated directly with the LLM tokens (blue tokens).

**Without Textualized Graph (w/o Textualized Graph):** In this configuration, we modified the textual input to the LLM. Rather than using a combination of the query and the textualized graph, we solely used the query.



(a) w/o GraphEncoder (b) w/o Projection Layer (c) w/o Textualized Graph

Figure 7: Ablation study configurations.

#### A.4 The Choice of Graph Encoder

Table 7: Performance of different graph encoders on the WebQSP and ExplaGraphs datasets.

| Graph Encoder         | WebQSP | ExplaGraphs |
|-----------------------|--------|-------------|
| GCN [14]              | 0.6853 | 0.8290      |
| GAT [36]              | 0.6732 | 0.8696      |
| GraphTransformer [31] | 0.6910 | 0.9016      |

In addition to the GAT [36], we explore other GNNs as the graph encoder, such as GCN [14] and the GraphTransformer [31]. The comparative results of these models on the WebQSP and ExplaGraphs datasets are presented in Table 7.

The results demonstrate that our proposed method exhibits consistent robustness across different graph encoders. Notably, all three encoders – GCN, GAT, and GraphTransformer – demonstrate competitive and closely aligned performance on the WebQSP dataset, with Hit@1 scores of 0.6853, 0.6732, and

0.6910, respectively. However, the performance differentiation becomes more pronounced on the ExplaGraphs dataset, where GraphTransformer exhibits a superior Hit@1 score of 0.9016, followed by GAT and GCN with scores of 0.8696 and 0.8290, respectively. This variation in performance across the datasets highlights the importance of encoder selection based on the specific characteristics and requirements of the dataset.

### A.5 The Choice of LLM

As for the choice of LLM, we considered both Llama2-7b and Llama2-13b. Our experiments demonstrate that stronger LLMs enhance the effectiveness of our method, as shown in Table 8, indicating that it benefits from the increased scale of the LLMs.

Table 8: Performance of different LLMs on the WebQSP dataset.

| LLM   | Llama2-7b | Llama2-13b |
|-------|-----------|------------|
| Hit@1 | 0.6732    | 0.7371     |

## B GraphQA Benchmark

In this section, we detail how our GraphQA benchmark differs from the original datasets, including the specific processing steps we employed. For concrete examples that illustrate the differences between the raw text in the original dataset and in our GraphQA benchmark, please refer to Table 9.

**ExplaGraphs.** The original dataset<sup>2</sup> [30] represents relationships using triplets. We have standardized this format by converting the triplets into a graph representation. Specifically, each head and tail in a triplet is transformed into a node, and the relation is transformed into an edge. Since the test dataset labels are not available, we have utilized only the training and validation (val) datasets from the original collection. We further divided these into training, val, and test subsets, using a 6:2:2 ratio.

**SceneGraphs.** The original GQA dataset is designed for real-world visual reasoning and compositional question answering, aiming to address key shortcomings of previous VQA datasets [11]. It comprises 108k images, each associated with a Scene Graph. In our study, we focus differently on graph question answering; hence, we did not utilize the image counterparts, leveraging only the scene graphs from the original dataset. Additionally, the original dataset describes images using JSON files. We simplified the object IDs to suit our research needs. We randomly sampled 100k samples from the original dataset and divided them into training, validation, and test subsets, following a 6:2:2 ratio.

**WebQSP.** We follow the preprocessing steps from RoG<sup>3</sup> [25]. The original dataset uses a list of triplets format, which we have transformed into our unified graph format. Furthermore, to avoid discrimination between capital and lowercase words, we have converted all words to lowercase. We used the same dataset split as in the original dataset.

## C Hallucination in Graph LLMs

In this section, we present quantitative results regarding hallucinations in the SceneGraphs dataset.

**Baseline.** For our baseline, we adapted MiniGPT-4 [49] to graph contexts. This approach involves a frozen LLM interacting with a trainable GNN that encodes graph data as a soft prompt, denoted as LLM+Graph Prompt Tuning. We focus on graph prompt tuning as the baseline, instead of converting the graph into text, since the textual representation of the graph is large and consistently exceeds the input token limits of LLMs.

**Experiment Design.** We instructed the LLM to answer graph-related questions and to list nodes or edges in the explanation graph that support its answers. Since standard answers for these questions do not exist, allowing the LLM to respond flexibly, it becomes challenging to evaluate its responses.

<sup>2</sup><https://explagraphs.github.io/>

<sup>3</sup><https://huggingface.co/datasets/rmanluo/RoG-webqsp>

Table 9: Comparison of text formats in original datasets and our GraphQA benchmark.

| Dataset     | Original dataset   | GraphQA Benchmark  |
|-------------|--|--|
| ExplaGraphs | (entrainment; capable of; being abused) (being abused; created by; police) (police; capable of; harm) (harm; used for; people) (people; part of; citizens)   | node_id,node_attr\n 0,entrapment\n 1,being abused\n 2,police\n 3,harm\n 4,people\n 5,citizens\n src,edge_attr,dst\n 0,capable of,1\n 1,created by,2\n 2,capable of,3\n 3,used for,4\n 4,part of,5  |
| SceneGraphs | "width": 500, "objects": { "681267": { "name": "banana", "h": 34, "relations": [ {"object": "681262", "name": "to the left of"}, {"w": 64, "attributes": [ "small", "yellow"]}], "y": 55, "x": 248, "681265": { "name": "spots", "h": 16, "relations": []}, "w": 26, "attributes": [], "y": 92, "x": 245, "681264": { "name": "bananas", "h": 50, "relations": [ {"object": "681259", "name": "to the left of"}, {"w": 49, "attributes": [ "small", "yellow"]}], "y": 32, "x": 268, "681263": { "name": "picnic", "h": 374, "relations": []}, "w": 499, "attributes": [ "delicious"], "y": 0, "x": 0, "681262": { "name": "straw", "h": 95, "relations": [ {"object": "681268", "name": "to the right of", "object": "681267", "name": "to the right of"}, {"object": "681253", "name": "to the right of"}, {"w": 15, "attributes": [ "white", "plastic"]}], "y": 55, "x": 402, "681261": { "name": "meat", "h": 27, "relations": [ {"object": "681255", "name": "on", "object": "681255", "name": "inside"}]}, "w": 24, "attributes": [ "small", "brown", "delicious"], "y": 123, "x": 68, "681260": { "name": "rice", "h": 57, "relations": [ {"object": "681255", "name": "on", "object": "681258", "name": "to the left of"}]}, "w": 93, "attributes": [ "piled", "white"], "y": 162, "x": 57, "681269": { "name": "onions", "h": 16, "relations": []}, "w": 24, "attributes": [ "green"], "y": 147, "x": 90, "681268": { "name": "tablecloth", "h": 374, "relations": [ {"object": "681262", "name": "to the left of"}, {"w": 396, "attributes": [ "white"]}], "y": 0, "x": 0, "681256": { "name": "bowl", "h": 99, "relations": [ {"object": "681255", "name": "next to", "object": "681257", "name": "of", "object": "681255", "name": "near", "object": "681256", "name": "to the right of", "object": "681260", "name": "to the right of", "object": "681255", "name": "to the right of"}, {"w": 115, "attributes": [ "full"]}], "y": 184, "x": 178, "681259": { "name": "plantains", "h": 70, "relations": [ {"object": "681264", "name": "to the right of"}, {"w": 45, "attributes": [ "red"]}], "y": 0, "x": 346, "681256": { "name": "spoon", "h": 65, "relations": [ {"object": "681255", "name": "on", "object": "681257", "name": "to the left of", "object": "681255", "name": "in", "object": "681258", "name": "to the left of"}]}, "w": 140, "attributes": [ "large", "metal", "silver"], "y": 196, "x": 0, "681257": { "name": "dish", "h": 81, "relations": [ {"object": "681258", "name": "inside", "object": "681256", "name": "to the right of", "object": "681258", "name": "in", "object": "681255", "name": "to the right of"}, {"w": 108, "attributes": [ "cream colored"]}], "y": 199, "x": 187, "681254": { "name": "meal", "h": 111, "relations": []}, "w": 130, "attributes": [], "y": 121, "x": 58, "681255": { "name": "plate", "h": 138, "relations": [ {"object": "681257", "name": "to the left of", "object": "681254", "name": "of", "object": "681254", "name": "with", "object": "681258", "name": "near", "object": "681258", "name": "to the left of"}]}, "w": 176, "attributes": [ "white", "full"], "y": 111, "x": 30, "681253": { "name": "banana", "h": 30, "relations": [ {"object": "681262", "name": "to the left of"}, {"w": 73, "attributes": [ "small", "yellow"]}], "y": 87, "x": 237, "height": 375} | node_id,node_attr\n 0,banana; attribute: small, yellow; (x,y,w,h): (248, 55, 64, 34)"<br>1,spots; (x,y,w,h): (245, 92, 26, 16)"<br>2,bananas; attribute: small, yellow; (x,y,w,h): (268, 32, 49, 50)"<br>3,picnic; attribute: delicious; (x,y,w,h): (0, 0, 499, 374)"<br>4,straw; attribute: white, plastic; (x,y,w,h): (402, 55, 15, 95)"<br>5,meat; attribute: small, brown, delicious; (x,y,w,h): (68, 123, 24, 27)"<br>6,rice; attribute: piled, white; (x,y,w,h): (57, 162, 93, 57)"<br>7,onions; attribute: green; (x,y,w,h): (90, 147, 24, 16)"<br>8,tablecloth; attribute: white; (x,y,w,h): (0, 0, 396, 374)"<br>9,bowl; attribute: full; (x,y,w,h): (178, 184, 115, 99)"<br>10,plantains; attribute: red; (x,y,w,h): (346, 0, 45, 70)"<br>11,spoon; attribute: large, metal, silver; (x,y,w,h): (0, 196, 140, 65)"<br>12,dish; attribute: cream colored; (x,y,w,h): (187, 199, 108, 81)"<br>13,meal; (x,y,w,h): (58, 121, 130, 111)"<br>14,plate; attribute: white, full; (x,y,w,h): (30, 111, 176, 138)"<br>15,banana; attribute: small, yellow; (x,y,w,h): (237, 87, 73, 30)"<br>src,edge_attr,ds<br>0,to the left of,4\n 2,to the left of,10\n 4,to the right of,8\n 4,to the right of,0\n 4,to the right of,15\n 5,on,14\n 5,inside,14\n 6,on,14\n 6,to the left of,9\n 8,to the left of,4\n 9,next to,14\n 9,in,12\n 9,near,14\n 9,to the right of,11\n 9,to the right of,6\n 9,to the right of,14\n 10,to the right of,2\n 11,on,14\n 11,to the left of,12\n 11,in,14\n 11,to the left of,9\n 12,inside,9\n 12,to the right of,11\n 12,in,9\n 12,to the right of,14\n 14,to the left of,12\n 14,in,13\n 14,with,13\n 14,near,9\n 14,to the left of,9\n 15,to the left of,4\n |
| WebQSP      | [{"FedEx Cup", "sports.sports_award_type.winners", "m.0n1v8cy"}, {"Brandt Snedeker", "sports.sports_award_winner.awards", "m.0n1v8cy"}, {"[FedEx Cup], 'common.topic.article', 'm.08q5wy'}, {"[FedEx Cup], 'common.topic.notable_for', 'g.12559n8g_1'}, {"['Sports League Award Type', 'freebase.type_profile.published', 'Published'], [{"FedEx Cup}, 'common.topic.notable_types', 'Sports League Award Type'], ["m.0n1v8cy"], {"sports.sports_award.award_winner", "Brandt Snedeker"}, {"['Sports League Award Type', 'type.type.expected_by', 'Award'], [{"Sports League Award Type', 'common.topic.article', 'm.06zxtxj'}], [{"2012 PGA Tour", "sports.sports_league_season.awards", "m.0n1v8cy"}, {"['Sports League Award Type', 'freebase.type_hints.included_types', 'Topic'], [{"Sports League Award Type', 'type.type.domain', 'Sports'}, {"m.0n1v8cy}, {"sports.sports_award.award", "FedEx Cup"}, {"['Sports League Award Type', 'freebase.type_profile.strict_included_types', 'Topic'], [{"Sports League Award Type', 'freebase.type_profile.kind', 'Classification'}, {"m.0n1v8cy}, {"sports.sports_award.season", "2012 PGA Tour"}, {"['Sports League Award Type', 'type.type.properties', 'Winners']]}  | node_id,node_attr\n 0,fedex cup\n 1,m.0n1v8cy\n 2,brandt snedeker\n 3,m.08q5wy\n 4,g.12559n8g_\n 5,sports league award type\n 6,published\n 7,award\n 8,m.06zxtxj\n 9,2012 pga tour\n 10,topic\n 11,sports\n 12,classification\n 13,winners\n src,edge_attr,ds\n 0,sports.sports_award_type.winners,1\n 2,sports.sports_award_winner.awards,1\n 0,common.topic.article,3\n 0,common.topic.notable_for,4\n 5,freebase.type_profile.published,6\n 0,common.topic.notable_types,5\n 1,sports.sports_award.award_winner,2\n 5,type.type.expected_by,7\n 5,common.topic.article,8\n 9,sports.sports_league_season.awards,1\n 5,freebase.type_hints.included_types,10\n 5,type.type.domain,11\n 1,sports.sports_award.award,0\n 5,freebase.type_profile.strict_included_types,10\n 5,freebase.type_profile.kind,12\n 1,sports.sports_award.season,9\n 5,type.type.properties,13  |

To address this, we manually examined 100 responses generated by our method and the LLM with graph prompt tuning, verifying whether the nodes and edges referenced in the LLM's output actually exist in the graph.

**Evaluation Metrics.** We assessed the model's faithfulness using three metrics: the fraction of valid nodes (denoted as Valid Nodes), the fraction of valid edges (denoted as Valid Edges), and the fraction of times the entire set of nodes and edges cited was valid (denoted as Fully Valid Graphs).

**Results.** The results, as depicted in Table 5, illustrate the comparative effectiveness of the *G-Retriever* over the baseline LLM+Graph Prompt Tuning method in reducing hallucinations. The LLM+Graph Prompt Tuning approach demonstrated a significantly lower accuracy in referencing graph elements, with only 31% of nodes and 12% of edges being valid, and the entire set of nodes and edges being valid only 8% of the time. In contrast, *G-Retriever* showed substantial improvements: 77% validity

in nodes, 76% in edges, and 62% in the overall validity of referenced node-edge sets. These results underscore the significant reduction in hallucinations with *G-Retriever*, particularly in the challenging task of accurately citing both nodes and edges in graph-based contexts.

## D Demonstrations

We demonstrate the interaction capabilities of G-Retriever with creative questions on different datasets: ExplaGraphs, SceneGraphs, and WebQSP. These examples are showcased in Tables 10, 11, and 12, respectively. Additionally, the examples are visualized in Figure 1.

Table 10: Example of creative chat on ExplaGraphs.

|       |   |
|-------|---|
| Graph | <pre> node_id,node_attr 0,human 1,cosmetic surgery 2,women 3,change appearance 4,addiction 5,searching for perfection 6,stop 7,positive  src,edge_attr,dst 0,receives action,1 2,part of,0 1,used for,3 3,capable of,4 4,created by,5 4,not capable of,6 4,is not a,7 </pre>  |
| User  | Write a 500-word argument essay based on the information provided in the explanation graph.   |
| Model | <p>The explanation graph provides a detailed analysis of the reasons why women undergo cosmetic surgery. According to the graph, the main reasons include the desire to change one's appearance, the influence of societal beauty standards, and the addiction to searching for perfection. These reasons are further supported by the various nodes and edges in the graph, which provide additional insights into the motivations and consequences of cosmetic surgery.</p> <p>Firstly, the node labeled "human" highlights the fundamental human desire to feel attractive and confident in one's appearance. This desire is deeply ingrained in our culture, where beauty is often associated with happiness, success, and social acceptance. As a result, many women feel pressure to conform to societal beauty standards, which can lead to feelings of insecurity and low self-esteem. In response to these feelings, some women may turn to cosmetic surgery as a means of improving their appearance and boosting their self-confidence.</p> <p>Secondly, the node labeled "women" highlights the gendered nature of cosmetic surgery. While men also undergo cosmetic surgery, the majority of cosmetic surgery patients are women. This is likely due to the fact that women are more likely to be exposed to societal beauty standards and are more likely to internalize these standards. As a result, women may be more likely to feel insecure about their appearance and more likely to seek cosmetic surgery as a means of improving their looks.</p> <p>Thirdly, the node labeled "addiction" highlights the potential negative consequences of cosmetic surgery. While some women may initially feel improved confidence and self-esteem after undergoing cosmetic surgery, the graph suggests that this initial boost may not last. In fact, some women may become addicted to the constant pursuit of perfection, leading to a cycle of surgery and dissatisfaction. This can have serious psychological and emotional consequences, including depression, anxiety, and a loss of self-esteem.</p> <p>Fourthly, the node labeled "searching for perfection" highlights the unrealistic expectations that many women have about the results of cosmetic surgery. The graph suggests that women may have unrealistic expectations about the ability of cosmetic surgery to improve their appearance and boost their self-esteem. This can lead to disappointment and dissatisfaction with the results of surgery, as well as a continued pursuit of perfection through additional surgery.</p> <p>Finally, the node labeled "stop" highlights the importance of critical thinking and self-reflection when it comes to cosmetic surgery. The graph suggests that women should be aware of the potential negative consequences of cosmetic surgery and should carefully consider the reasons why they are undergoing surgery. They should also be aware of the unrealistic expectations that they may have about the results of surgery and should be prepared to address any negative emotions or feelings of dissatisfaction that may arise.</p> <p>In conclusion, the explanation graph provides a detailed analysis of the reasons why women undergo cosmetic surgery. The graph highlights the fundamental human desire to feel attractive and confident in one's appearance, the gendered nature of cosmetic surgery, the potential negative consequences of cosmetic surgery, and the importance of critical thinking and self-reflection when it comes to surgery. By understanding these reasons and the potential consequences of cosmetic surgery, women can make informed decisions about whether or not to undergo surgery and can work to address any negative emotions or feelings of dissatisfaction that may arise.</p> |

Table 11: Example of creative chat on SceneGraphs. We utilize GPT-4 to generate an image from the graph, assisting readers in visualizing it

|       |   |
|-------|---|
| Graph | <pre> node_id,node_attr 0,name: head; attribute: black; (x,y,w,h): (386, 185, 8, 4) 1,name: swimming pool; attribute: small, water; (x,y,w,h): (465, 92, 35, 3) 2,name: shore; (x,y,w,h): (49, 85, 436, 108) 3,name: fence; attribute: wired; (x,y,w,h): (11, 158, 167, 46) 4,name: mountains; attribute: covered; (x,y,w,h): (393, 53, 107, 12) 5,name: water; attribute: blue; (x,y,w,h): (0, 50, 499, 154) 6,name: sheep; attribute: walking, grazing; (x,y,w,h): (243, 191, 20, 13) 7,name: sky; attribute: blue, clear, cloudy; (x,y,w,h): (0, 0, 501, 60) 8,name: sheep; attribute: white, fluffy; (x,y,w,h): (409, 176, 13, 10) 9,name: field; attribute: large, green, grassy; (x,y,w,h): (0, 138, 503, 112) 10,name: grass; (x,y,w,h): (0, 84, 500, 166) 11,name: sheep; attribute: white, fluffy; (x,y,w,h): (398, 178, 13, 11) 12,name: sheep; attribute: white, fluffy; (x,y,w,h): (385, 179, 17, 11) 13,name: sheep; attribute: white, fluffy; (x,y,w,h): (364, 181, 14, 10) 14.name: shore; attribute: sandy; (x,y,w,h): (190, 123, 171, 12) src, edge_attr, dst 0,to the left of,11\n 0,to the right of,13\n 3,to the left of,6\n 6,to the right of,3\n 6.grazing on,10\n 6,in,9\n 6,in,10\n 7,above,5\n 8,eating,10\n 8,to the right of,11\n 8,to the right of,12\n 11,to the left of,8\n 11,to the right of,0\n 11,to the right of,13\n 11,to the right of,12\n 11,eating,10\n 12,eating,10\n 12,to the left of,8\n 12,to the right of,13\n 12,to the left of,11\n 13,to the left of,0\n 13,to the left of,11\n 13,to the left of,12\n 13,eating,10\n </pre> |
| Image |   |
| User  | Write an 500-word advertisement for this place in the scene graph that would make people want to visit it.  |
| Model | <p>Sure! Here's an example of a 500-word advertisement for the place based on the scene graph:</p> <p>"Experience the breathtaking beauty of our place, where the rolling hills meet the sparkling water of the swimming pool. Take a leisurely stroll through the lush green fields, and let the gentle breeze caress your face as you gaze out at the picturesque view. The sheep graze peacefully in the distance, their white coats glistening in the sunlight. The sky is a brilliant blue, with fluffy white clouds drifting lazily across. Come and bask in the serene ambiance of our place, where the natural beauty will leave you in awe."</p>   |

Table 12: Example of creative conversation on the WebQSP dataset.

|       |   |
|-------|---|
|       | node_id, node_attr<br>8, benjamin franklin<br>63, glass harmonica<br>98, lightning rod<br>99, stealing god's thunder: benjamin franklin's lightning rod and the invention of america<br>227, benjamin franklin worked with thomas jefferson<br>442, franklin stove<br>525, bifocals<br>706, benjamin franklin drawing electricity from the sky<br>1044, prokop diviš  |
| Graph | src, edge_attr, dst<br>706, visual_art.artwork.art_subject, 8<br>8, base.kwebbase.kwtopic.connections_from, 227<br>8, book.book_subject.works, 99<br>525, law.invention.inventor, 8<br>442, law.invention.inventor, 8<br>98, law.invention.inventor, 1044<br>8, law.inventor.inventions, 98<br>63, law.invention.inventor, 8<br>8, base.argumentmaps.innovator.original_ideas, 442<br>98, law.invention.inventor, 8<br>8, law.inventor.inventions, 525<br>442, base.argumentmaps.original_idea.innovator, 8<br>8, law.inventor.inventions, 442<br>8, law.inventor.inventions, 63  |
| User  | <p>Question: what else did ben franklin invent? Surrounding this question, develop an educational tour through the knowledge graph, highlighting interesting facts and connections in an engaging way.</p>  |
| Model | <p>Great! Based on the provided knowledge graph, here's an educational tour that highlights interesting facts and connections related to Benjamin Franklin's inventions and other notable achievements:</p> <ol style="list-style-type: none"> <li>Franklin's Lightning Rod: Start by exploring Franklin's most famous invention - the lightning rod. Discover how Franklin's invention protected buildings from lightning strikes and how it changed the way people understood electricity. Connection: Franklin worked with Thomas Jefferson on the lightning rod, and their collaboration led to the development of the concept of positive and negative charges.</li> <li>Bifocals: Move on to Franklin's invention of bifocal glasses. Learn how Franklin's vision problems inspired him to create a new type of glasses that allowed people to see both near and far objects clearly. Connection: Franklin's bifocal glasses were a groundbreaking invention that improved the lives of millions of people around the world.</li> <li>Franklin Stove: Explore Franklin's design of the Franklin stove, a cast-iron stove that provided more heat with less fuel. Discover how Franklin's stove revolutionized heating systems and how it helped to reduce fuel costs for households and businesses. Connection: Franklin's stove was an early example of energy efficiency and sustainability, and it paved the way for modern heating systems.</li> <li>Prokop Diviš: Learn about Franklin's connection to Prokop Diviš, a Czech inventor who built on Franklin's work and developed the first electric generator. Discover how Diviš's invention helped to power the Industrial Revolution and how it changed the way people lived and worked. Connection: Franklin's work on the lightning rod inspired Diviš to explore the properties of electricity, and their collaborative efforts helped to advance the field of electrical engineering.</li> <li>Innovator and Original Ideas: Explore Franklin's reputation as an innovator and his ability to come up with original ideas. Discover how Franklin's creativity and curiosity led to many of his inventions and how he was able to see connections between seemingly unrelated things. Connection: Franklin's innovative spirit and ability to think outside the box inspired many other inventors and scientists, including Thomas Edison and Nikola Tesla.</li> <li>Collaboration and Networking: Highlight Franklin's collaborative approach to invention and his ability to network with other inventors and scientists. Discover how Franklin's connections with other innovators helped to advance the field of science and technology. Connection: Franklin's collaborations with other inventors and scientists, such as Thomas Jefferson and Prokop Diviš, helped to accelerate the pace of innovation and pave the way for future breakthroughs.</li> <li>Legacy and Impact: Reflect on Franklin's lasting legacy and the impact of his inventions on modern society. Discover how Franklin's inventions continue to inspire new generations of inventors and scientists and how they have helped to shape the world we live in today. Connection: Franklin's inventions and ideas have had a lasting impact on many fields, including science, technology, and society as a whole. His legacy continues to inspire and motivate people around the world.</li> </ol> <p>By following this educational tour through the knowledge graph, students will gain a deeper understanding of Benjamin Franklin's inventions and their impact on modern society. They will also learn about Franklin's collaborative approach to invention, his innovative spirit, and his ability to think outside the box.</p> |