

©Copyright 2024  
Jyoti Arvind Singh

A thesis  
submitted in partial fulfillment of the  
requirements for the degree of

University of Washington

Committee:

Program Authorized to Offer Degree:

University of Washington

**Abstract**

Meta path of Thoughts: Verbalized Meta paths in Heterogeneous Graph as Contextual Augmentation to LLM

Jyoti Arvind Singh

Chair of the Supervisory Committee:

Ankur Teredesai

School of Engineering & Technology

Heterogeneous Graph Neural Networks (HGNNs) excel in capturing graph topology and structural information. However, they are ineffective in processing textual components present in nodes and edges and thus producing suboptimal performance in downstream tasks such as node classification. Additionally, HGNNs lack in their explanatory power and are considered black box. Although, Large Language Models (LLMs) are good at processing textual information, however, utilizing them for tasks like node classification can be non-trivial since it is difficult to identify the ideal graphical context and present it in a form suitable for LLMs to consume effectively. We introduce a framework that combines the strength of both models by leveraging the context obtained through meta paths, which are generated during training of HGNNs. This approach enables understanding of complex and indirect relationships between different types of nodes. Our novel framework enhances the prediction accuracy of HGNNs and transparency of their decision-making process through natural language explanations provided by LLMs. We demonstrate that our proposed framework outperforms FAST-GTN, an HGNN tailored for heterogeneous graph data, on two network datasets, with improvements in F1 Score from 0.81 and 0.6 of the baselines to 0.9 and 0.91, respectively. Furthermore, the efficacy of framework in generating explanations datasets (DBLP citation graph and Goodreads graph dataset), with improvements in F1 Score from 0.81 and and 0.66 of the baselines to 0.9 and 0.91, respectively. Furthermore, the efficacy of the framework in generating explanations has been evaluated through human evaluation, considering metrics such as helpfulness and factual correctness.

## Acknowledgement

I am deeply grateful to those whose support and contributions were essential to the completion of this work. Special thanks to my advisor, Dr. Ankur Teredesai, for his unwavering guidance, insightful feedback, and belief in my abilities. I also appreciate Yihui Chong, current Research Scientist at DSO National Laboratories, Singapore, for the generous grant and support, which were crucial for the success of this research. Finally, my heartfelt thanks to my research and life partner, Harshvardhan Solanki, current Applied Scientist at Amazon, for his invaluable feedback and research contribution throughout my thesis as well as to Amazon Science for providing the compute resources necessary for this study.

|          |                                                                               |           |
|----------|-------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>INTRODUCTION</b>                                                           | <b>6</b>  |
| 1.1      | Motivation . . . . .                                                          | 7         |
| 1.2      | Problem Definition . . . . .                                                  | 7         |
| 1.3      | Challenges . . . . .                                                          | 7         |
| 1.4      | Key Contributions . . . . .                                                   | 8         |
| <b>2</b> | <b>Technical Background</b>                                                   | <b>10</b> |
| 2.1      | Heterogeneous Graph [21] [20] . . . . .                                       | 10        |
| 2.2      | Meta-paths [22] [20] . . . . .                                                | 10        |
| 2.3      | Metapath Instance [23] . . . . .                                              | 11        |
| 2.4      | HGNNs and FastGTN [7] . . . . .                                               | 11        |
| 2.5      | Explainability of Supervised Machine Learning [24] . . . . .                  | 11        |
| 2.6      | GNNs Explainability [25][26][12] . . . . .                                    | 11        |
| 2.7      | Use of LLMs for model interpretability and natural language explanations [28] | 12        |
| 2.8      | Evaluation: Node Classification Task . . . . .                                | 12        |
| 2.8.1    | Micro-F1 Score . . . . .                                                      | 12        |
| 2.8.2    | Macro-F1 Score . . . . .                                                      | 13        |
| <b>3</b> | <b>Approach</b>                                                               | <b>14</b> |
| 3.1      | Meta-path Generation and Selection . . . . .                                  | 14        |
| 3.1.1    | Meta-path Computation . . . . .                                               | 14        |
| 3.1.2    | Top-K Meta-path Selection . . . . .                                           | 15        |
| 3.2      | Meta-path Verbalization using LLM . . . . .                                   | 15        |
| 3.3      | Contextual Representation Construction . . . . .                              | 15        |
| 3.3.1    | Sampling Meta-path Instances . . . . .                                        | 15        |
| 3.3.2    | Integrating Verbal Interpretations . . . . .                                  | 15        |
| 3.3.3    | Incorporating Textual Attributes . . . . .                                    | 15        |
| 3.4      | Semantic Similarity for Context Enrichment . . . . .                          | 16        |
| 3.4.1    | Semantic Similarity Computation . . . . .                                     | 16        |
| 3.4.2    | Similarity Enhanced Contextual Representation . . . . .                       | 16        |
| 3.5      | Chain-of-Thought Instructions for LLM . . . . .                               | 16        |
| 3.5.1    | Step-by-Step Reasoning with CoT Prompts . . . . .                             | 16        |
| 3.5.2    | Prediction and Explanation Generation . . . . .                               | 16        |
| 3.6      | Evaluation . . . . .                                                          | 16        |

|          |                                         |           |
|----------|-----------------------------------------|-----------|
| <b>4</b> | <b>Experiments</b>                      | <b>21</b> |
| 4.1      | Dataset Description . . . . .           | 21        |
| 4.2      | Hyperparameter Configurations . . . . . | 22        |
| 4.3      | Progression of Experiments . . . . .    | 24        |
| 4.4      | Explanation Generation . . . . .        | 25        |
| 4.5      | Summary . . . . .                       | 25        |
| <b>5</b> | <b>Results and Analysis</b>             | <b>26</b> |
| 5.1      | Dataset Overview . . . . .              | 26        |
| 5.2      | Node Classification Results . . . . .   | 26        |
| 5.3      | Results Analysis . . . . .              | 28        |
| 5.4      | Summary . . . . .                       | 29        |
| <b>6</b> | <b>Conclusion and Future Work</b>       | <b>30</b> |
| 6.1      | Conclusion . . . . .                    | 30        |
| 6.2      | Future Work . . . . .                   | 30        |

LIST OF FIGURES

|     |                                                                    |    |
|-----|--------------------------------------------------------------------|----|
| 3.1 | Steps involved in the process . . . . .                            | 17 |
| 3.2 | Verbalization of metapath: Prompt template . . . . .               | 17 |
| 3.3 | Augmented Prompt in the template mentioned in Figure 3.2 . . . . . | 18 |
| 3.4 | Example generation: Verbalization of metapath . . . . .            | 19 |
| 3.5 | Prompt template for getting prediction and explanation . . . . .   | 20 |

## LIST OF TABLES

|     |                                                                                                                                  |    |
|-----|----------------------------------------------------------------------------------------------------------------------------------|----|
| 4.1 | Dataset statistics for DBLP and Goodreads. . . . .                                                                               | 22 |
| 4.2 | LLM inference Hyperparamters . . . . .                                                                                           | 22 |
| 4.3 | Meta-Path Hyperparameters . . . . .                                                                                              | 22 |
| 4.4 | Fast-GTN training hyperparameters . . . . .                                                                                      | 22 |
| 4.5 | Hyperparameters in homogeneous context . . . . .                                                                                 | 23 |
| 4.6 | Dataset: Label Types and Distribution . . . . .                                                                                  | 23 |
| 5.1 | Comparison of different approaches on DBLP and Goodreads datasets. . . . .                                                       | 27 |
| 5.2 | Comparison of Factual Correctness and Helpfulness Ratings between Different Approaches for DBLP and Goodreads Datasets . . . . . | 28 |

Heterogeneous graphs (HGs) have gained prominence for their ability to model complex relationships across diverse domains such as natural language processing [1][2], recommendation systems [3][4], community detection [5][6], and biological networks. By representing real-world data through diverse node and edge types, HGs effectively capture intricate interactions and dependencies [7]. To leverage this capability, Heterogeneous Graph Neural Networks (HGNNs) have been developed to perform tasks such as node classification and link prediction. These models rely on meta-paths for aggregating information and generating meaningful embeddings tailored to downstream tasks.

Despite their utility, HGNNs face significant challenges, including a lack of explainability and reliance on manually constructed meta-paths. The construction of these meta-paths demands considerable domain expertise and computational resources, making it a bottleneck for their widespread adoption. Furthermore, the interpretability of HGNN predictions is limited due to their complex, non-linear transformations, which diminishes trust in critical applications like healthcare and finance [8]. Existing techniques for enhancing explainability, such as gradient-based [9][10][11] and perturbation-based approaches [12][13][14], have yet to fully address the unique challenges posed by heterogeneous graph data.

Simultaneously, Large Language Models (LLMs) have demonstrated exceptional capabilities in tasks requiring textual reasoning, such as summarization, question answering, and natural language generation [15][16][17][18]. However, their integration with structured graph data remains an open research area. LLMs excel in processing and generating text, while HGNNs specialize in structural representation learning. Aligning these complementary strengths offers a unique opportunity to enhance both predictive accuracy and interpretability.

This work introduces a novel framework that combines the semantic and textual reasoning capabilities of LLMs with the structural learning strength of HGNNs. By automating the use of meta-path-based contexts and leveraging LLMs for explainable insights, the framework addresses limitations in existing approaches. This integration not only improves classification performance but also provides intuitive, human-readable explanations, thereby advancing the practical utility of heterogeneous graphs in high-stakes domains.

## 1.1 Motivation

The increasing adoption of heterogeneous graphs across various fields highlights their potential to represent diverse and complex relationships. Traditional Graph Neural Networks (GNNs), while successful for homogeneous graph data, lack the adaptability needed to handle the structural diversity of HGs. To bridge this gap, HGNNs, such as Graph Transformer Networks (GTNs) [19] and FastGTNs [20], enable the automatic transformation of HGs into meta-path-based subgraphs for task-specific applications. Despite these advancements, the reliance on manually designed meta-paths remains a limitation, as it significantly influences model performance and scalability.

Explainability is another critical challenge for HGNNs. Their "black-box" nature raises concerns about trust and accountability, particularly in sensitive domains. Existing explainability frameworks have limited applicability to HGNNs due to the inherent complexity of heterogeneous graph data [8]. On the other hand, LLMs have shown promise in generating interpretable outputs but lack the ability to directly utilize structured graph information.

Motivated by these gaps, this research aims to develop a framework that integrates LLMs and HGNNs, leveraging the strengths of both paradigms. By incorporating meta-path contexts alongside semantic and textual information, the proposed approach enhances predictive performance and generates interpretable results. This effort addresses existing challenges while opening new possibilities for applying heterogeneous graphs in critical areas such as healthcare, finance, and scientific discovery.

## 1.2 Problem Definition

We address the problem of developing an open-source natural language explanation framework for heterogeneous graphs. This framework leverages meta-path-aware contextual understanding to ensure explanations are factually correct, contextually relevant, and grounded in the graph's complex structure and semantic relationships. To achieve this goal, we define and solve the following sub-problems:

1. **Extract and Verbalize Significant Meta-paths:** Develop methods to identify and verbalize the most influential meta-paths extracted from the GTN framework, transforming them into a format interpretable by large language models (LLMs).
2. **Enable LLMs to Capture Complex Relationships:** Equip LLMs to understand and leverage the complex relationships encoded by these meta-paths for generating accurate and meaningful explanations.
3. **Enhance Node Classification with Contextual Information:** Integrate additional relevant context, such as semantic similarities and textual attributes, to improve the accuracy of node classification and the transparency of LLM-driven explanations.
4. **Perform Qualitative Evaluation of Explanations:** Conduct qualitative evaluations of explanations generated by the framework on heterogeneous graph datasets to measure their factual correctness, helpfulness, and contextual relevance.

## 1.3 Challenges

We highlight the challenges involved in solving the problem of building a natural language explanation framework for heterogeneous graphs.

The first challenge lies in handling the large and sparse nature of the datasets. Both the DBLP and Goodreads datasets comprise tens of thousands of nodes and edges, with relationships that are heterogeneous in nature. For instance, the DBLP dataset includes diverse node types such as authors, papers, and conferences, connected by multiple types of edges, while the Goodreads dataset features books, authors, publishers, and more. This sparsity, coupled with the large-scale data, makes identifying meaningful meta-paths computationally expensive and complex.

The next challenge pertains to the heterogeneous and textually heavy nature of the datasets. Nodes such as *Papers* in DBLP and *Books* in Goodreads are associated with extensive textual attributes like abstracts and descriptions. These attributes, while rich in information, require significant preprocessing and semantic understanding to integrate effectively into the framework. Transforming this unstructured textual data into representations that can complement graph-based learning models is non-trivial.

Another significant challenge is the extraction and verbalization of meta-paths. Meta-paths must not only be identified but also transformed into human-readable forms that can be processed by large language models (LLMs). Ensuring the meta-paths are both accurate and semantically meaningful for interpretability is essential but requires domain expertise and computational resources.

A further challenge involves enabling LLMs to capture complex relationships encoded by the meta-paths. Aligning the structured data of heterogeneous graphs with the unstructured reasoning capabilities of LLMs requires sophisticated prompt engineering and contextual enrichment. This alignment is crucial to ensure that the LLM can provide explanations that are both factually correct and contextually relevant.

Finally, the task of qualitative evaluation of explanations presents its own difficulties. Evaluating the helpfulness, factual correctness, and relevance of explanations requires human judgment and is inherently subjective. Creating standardized evaluation criteria and ensuring consistency across diverse datasets like DBLP and Goodreads is a resource-intensive task that requires careful design.

## 1.4 Key Contributions

Our primary contributions could be summarized as follows:

1. **Advanced Explanation Framework:** Specifically designed for heterogeneous graph-structured data to address the challenges of explainability. Provides contextually relevant, factually accurate, and easily interpretable explanations rooted in the graph’s intricate structure and semantic relationships.
2. **Enhanced Node Classification Method:** Achieves significant improvement in node classification performance, surpassing the baseline FastGTN model (State-of-the-Art) on heterogeneous node classification tasks. Demonstrates higher F1 scores, showcasing the effectiveness of the proposed approach in leveraging meta-path-based context and semantic information.
3. **Dynamic Prompt Generation Library:** Automates the creation of natural language prompts by extracting and verbalizing meta-path types from heterogeneous graphs. Enriches prompts with additional context derived from meta-path instances and semantically similar nodes, significantly enhancing the effectiveness of LLMs in both prediction and explanation tasks.

These contributions collectively establish a robust framework for integrating heterogeneous graph structures with large language models, advancing the state-of-the-art in both predictive performance and interpretability.

The rest of the thesis is organized as follows: in chapter 2, we introduce the various techniques that we employ towards solving this problem and the evaluation metrics that are used. Chapter 3 discusses the basic approaches and methods employed to address the different challenges mentioned earlier. In chapter 4, we explain the experimental setup along with the dataset description and describe various hyperparameters and experiments performed in this work. Chapter 5 summarizes the experimental results. Finally in chapter 6, we outline future work and conclude the thesis.

## 2.1 Heterogeneous Graph [21] [20]

A directed graph  $G = (V, E, T_v, T_e)$  is defined as a heterogeneous graph if each node  $v \in V$  and each edge  $e \in E$  have corresponding type mapping functions  $\tau_v(v) : V \rightarrow T_v$  and  $\tau_e(e) : E \rightarrow T_e$ . This type of graph can be described using a set of adjacency matrices  $\{A_t\}_{t=1}^{|T_e|}$  or a tensor (i.e.,  $A \in \mathbb{R}^{|V| \times |V| \times |T_e|}$ ), where  $A_t \in \mathbb{R}^{N \times N}$  represents the adjacency matrix for the  $t$ -th edge type, and  $|V| = N$ . In this context,  $A_t[i, j]$  indicates the weight of the edge of type  $t$  from node  $j$  to node  $i$ . When there is only one type of node and one type of edge, i.e.,  $|T_v| = 1$  and  $|T_e| = 1$ , the graph is referred to as a homogeneous graph.

## 2.2 Meta-paths [22] [20]

In the context of heterogeneous graphs, a meta-path refers to a sequence of edges that connect nodes through multiple types of relations. Formally, a meta-path can be represented as:

$$\mathcal{P} = v_1 \xrightarrow{t_1} v_2 \xrightarrow{t_2} \dots \xrightarrow{t_{n-1}} v_n,$$

where  $v_1, v_2, \dots, v_n$  denote the nodes in the sequence, and  $\xrightarrow{t_i}$  represents an edge of type  $t_i$  from node  $v_i$  to node  $v_{i+1}$ . Here,  $t_i$  specifies the edge type of the edge  $(v_i, v_{i+1})$  along the meta-path.

A meta-path can also be described as a composite relation  $R = t_1 \circ t_2 \circ \dots \circ t_n$  between node  $v_1$  and  $v_{n+1}$ , where  $R_1 \circ R_2$  denotes the composition of relations  $R_1$  and  $R_2$ . Given the composite relation  $R$  or the sequence of edge types  $(t_1, t_2, \dots, t_n)$ , the adjacency matrix  $A_P$  of the meta-path  $P$  is derived from the multiplication of adjacency matrices as  $A_P = A_{t_n} \cdots A_{t_2} A_{t_1}$ .

The concept of a meta-path includes multi-hop connections and introduces new graph structures represented by adjacency matrices. For instance, the meta-path Author-Paper-Conference (APC),

$$A \xrightarrow{AP} P \xrightarrow{PC} C$$

which can be represented as  $A \xrightarrow{AP} P \xrightarrow{PC} C$ , results in an adjacency matrix  $A_{APC}$  by multiplying  $A_{AP}$  and  $A_{PC}$ .

## 2.3 Metapath Instance [23]

An instance  $p$  of the metapath  $P$  is a particular sequence of nodes  $\langle n_1, n_2, \dots, n_l \rangle$  within the network that adheres to the pattern specified by  $P$ . Multiple instances can correspond to a single metapath.

For example, in a citation network, consider the metapath "Author-Paper-Author" (APA). Various sequences of nodes, such as  $A_i P_i A_j$ , can fit this metapath pattern, making them instances of the "APA" metapath.

## 2.4 HGNNs and FastGTN [7]

Heterogeneous graph neural networks (HGNNs) are advanced models designed to handle the complexity of heterogeneous graphs (HGs), which consist of diverse types of nodes and edges. Unlike traditional graph neural networks, HGNNs account for the semantic and structural heterogeneity inherent in HGs. They use deep learning techniques to learn embeddings from node attributes and interactions, effectively capturing the rich, multi-faceted information within these graphs. This capability makes HGNNs particularly powerful for tasks requiring a nuanced understanding of complex networks. The training methodology [20] employed by Graph Transformer Networks (GTN) and FastGTN involves the transformation of original graphs into multiple subgraphs referred to as metapaths above. This transformation is integral to the learning process, where node representations are further derived through convolution operations conducted on top of these metapaths. Learnt node representations on top of these metapaths are further used for downstream tasks like node classification. Additionally, the methodology used in FastGTN also provides the weighting/attention-score on how it influences the node prediction. They also demonstrated that meta-paths predefined by domain knowledge consistently receive top rankings from FastGTNs, further validating their influence through the use of attention scores.

## 2.5 Explainability of Supervised Machine Learning [24]

Consider a supervised task  $T$  with the aim of learning a mapping from  $X$  to  $Y$ , and a model  $M$  trained for this task. Given a set of  $(x, y)$  pairs  $\subseteq (X, Y)$  and the model  $M$ , generate an explanation  $e$  from a given set  $D_E$  such that  $e$  "explains" the prediction  $\hat{y} = M(x)$ .

These explanations can be either local to a single test input  $(x, y)$  or global when they explain prediction over a specific dataset  $D' \subseteq (X, Y)$ . Further, the explanation can be generated either post-hoc (i.e., after the model training) or ante-hoc where the model itself is self-interpretable, i.e., it explains its predictions.

Local explanations only consider the local neighborhood of the given data instance while global explanations are concerned about the model's overall behavior and thus, search for patterns in the model's predictions. In this work, we focus on Local explanations.

## 2.6 GNNs Explainability [25][26][12]

The survey [25] presents a comprehensive overview of GNN explainability techniques. There exists a research gap for techniques that generate natural language explanations that are

crucial to the practical application of GNN utilizing metapaths, which are crucial for interpreting heterogeneous graph neural networks (HGNNs). Metapaths [27] represent sequences of node types and edge types in heterogeneous graphs, offering a pathway to understand complex relationships and interactions within the data.

Most current explainability methods discussed in [25] focus on visual or quantitative interpretations, such as highlighting important nodes and edges or providing importance scores. These methods, while useful, fall short of offering intuitive, human-readable explanations that describe the reasoning behind a model’s prediction in natural language. The integration of metapaths into generative natural language explanations would bridge this gap by translating the intricate relationships within heterogeneous graphs into comprehensible narratives, enhancing interpretability for non-expert users.

## 2.7 Use of LLMs for model interpretability and natural language explanations [28]

The work presented in [28] proposes the opportunity to rethink LLMs for interpretability with a more ambitious scope. LLMs can elicit more elaborate explanations than the previous generation of interpretable ML techniques. While previous methods have often relied on restricted interfaces such as saliency maps, LLMs can communicate directly in expressive natural language.

## 2.8 Evaluation: Node Classification Task

The evaluation of the proposed framework focuses on assessing its performance in the node classification task for the Goodreads dataset. We compare the *Micro-F1* and *Macro-F1* scores of various approaches, ranging from baseline models to the proposed methods that incorporate meta-path-based contextual information with examples provided to the Large Language Model (LLM). The evaluation highlights the effectiveness of our framework in enhancing prediction accuracy by leveraging meta-path-aware contexts and strategically designed examples.

### 2.8.1 Micro-F1 Score

The *Micro-F1 Score* is calculated by aggregating the contributions of all classes to compute a single metric, treating the classification problem as a binary classification task for each instance. This metric is particularly useful for datasets with class imbalances as it gives equal weight to every instance. The formula for the Micro-F1 Score is as follows:

$$\text{Micro-F1} = \frac{2 \cdot \text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \quad (2.1)$$

where:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{c \in C} \text{TP}_c}{\sum_{c \in C} (\text{TP}_c + \text{FP}_c)} \quad (2.2)$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{c \in C} \text{TP}_c}{\sum_{c \in C} (\text{TP}_c + \text{FN}_c)} \quad (2.3)$$

Here,  $TP_c$ ,  $FP_c$ , and  $FN_c$  represent the true positives, false positives, and false negatives for class  $c$ , respectively, and  $C$  is the set of all classes.

### 2.8.2 Macro-F1 Score

The *Macro-F1 Score* is calculated by computing the F1 score for each class individually and then averaging these scores. Unlike the Micro-F1 Score, the Macro-F1 Score treats all classes equally, regardless of class imbalance, by weighing each class equally in the final average. The formula for the Macro-F1 Score is as follows:

$$\text{Macro-F1} = \frac{1}{|C|} \sum_{c \in C} F1_c \quad (2.4)$$

where:

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (2.5)$$

and:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \quad (2.6)$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \quad (2.7)$$

We formulate the problem of building an open-source natural language explanation framework for heterogeneous graphs, leveraging meta-path-aware contextual understanding to ensure explanations are factually correct, helpful, and contextually relevant. This framework is designed to utilize the intricate structure and semantic relationships within heterogeneous graphs to improve both node classification accuracy and the interpretability of predictions. Achieving this involves addressing four major subproblems:

1. **Extract and Verbalize Significant Meta-paths:** Identify and verbalize the most influential meta-paths from the FastGTN framework to create human-readable representations.
2. **Enable LLMs to Capture Complex Relationships:** Develop techniques to allow LLMs to effectively utilize the meta-paths for reasoning and explanation tasks.
3. **Enhance Context with Semantic and Textual Information:** Incorporate additional context from semantically similar nodes and textual attributes to enrich predictions and explanations.
4. **Perform Qualitative Evaluation of Explanations:** Conduct rigorous qualitative evaluations to assess the factual correctness, helpfulness, and relevance of the explanations generated by the framework.

In the next few sections, we discuss the non-trivial challenges and detailed methodology for each of these subproblems.

## 3.1 Meta-path Generation and Selection

### 3.1.1 Meta-path Computation

The meta-paths are created during the training phase of FastGTN, a model that transforms heterogeneous graphs into meta-path-based subgraphs for node classification. The detailed procedure for training FastGTN and computing meta-paths is explained in [reference\_to\_paper].

### 3.1.2 Top-K Meta-path Selection

Meta-paths are ranked based on their importance to node classification, as determined by the attention weights computed during FastGTN training. The Top-K meta-path types are selected in descending order of their weights, with further filtering to retain only those starting with the target node type.

## 3.2 Meta-path Verbalization using LLM

To make meta-paths accessible to LLMs, they must be translated into natural language representations. A structured one-shot LLM prompt is designed to generate verbal interpretations of meta-paths. The prompt includes node types, edge types, and the meaning of relationships, guiding the LLM to produce clear, logical descriptions. Figures 3.2, 3.3 and 3.4 illustrate the prompt templates, augmented prompts, and sample outputs of meta-path verbalization. This process bridges the gap between structured graph data and unstructured text that LLMs can process effectively.

**Example:** Sample meta-path instance for P-A-P:

”Attention is All You Need - A. Vasvani - NIPS 2017 - Dekel - Online Learning with a Hint”

Verbalization: The paper ”Attention is All You Need” is connected to the paper ”Online Learning with a Hint” through their authors ”A. Vasvani” and ”O. Dekel” and their shared association with the conference ”NIPS 2017”. This connection indicates that the two papers may belong to similar fields of study due to the shared context of the conference and the expertise of their authors.

## 3.3 Contextual Representation Construction

Contextual information for the target node is constructed using the following steps:

### 3.3.1 Sampling Meta-path Instances

Multiple instances of each selected meta-path type are sampled based on their normalized attention weights. This ensures that the most relevant meta-path instances are prioritized for context construction.

### 3.3.2 Integrating Verbal Interpretations

The verbalized meta-paths are applied to all sampled instances using regular expressions, ensuring consistent and meaningful contextual representation.

### 3.3.3 Incorporating Textual Attributes

Nodes participating in the sampled meta-path instances often contain rich textual attributes (e.g., abstracts, descriptions). These attributes are integrated alongside the verbalized meta-paths to provide additional semantic information.

## 3.4 Semantic Similarity for Context Enrichment

The meta-path-derived context is further enriched with semantic information:

### 3.4.1 Semantic Similarity Computation

Top-2 Nodes with highest semantic similarity to the target node are identified using S-Bert embeddings of 600 dimensions. These nodes provide additional relevant textual content to complement the meta-path context.

### 3.4.2 Similarity Enhanced Contextual Representation

By combining meta-path instances, textual attributes, and semantically similar nodes, the constructed context becomes richer and more informative for LLM reasoning tasks.

## 3.5 Chain-of-Thought Instructions for LLM

To ensure effective reasoning and explanation generation, the following steps are employed:

### 3.5.1 Step-by-Step Reasoning with CoT Prompts

Chain-of-Thought (CoT) prompting involves presenting the LLM with step-by-step instructions to connect the contextual information to the target node.

### 3.5.2 Prediction and Explanation Generation

The LLM uses the provided context and CoT instructions to generate accurate node classifications and clear, human-readable explanations. Figure 3.5 provides an example of CoT-style instructions tailored for these tasks. This approach addresses the outlined subproblems and provides a robust framework for enhancing the interpretability and predictive power of heterogeneous graph-based models.

## 3.6 Evaluation

To validate the proposed methodology, we evaluate its effectiveness in addressing the outlined subproblems. The evaluation is conducted on heterogeneous graph datasets, such as DBLP and Goodreads, focusing on tasks like node classification and explanation generation. Performance metrics, including micro-F1 and macro-F1 scores, are used to assess the accuracy and class-wise balance of node classification.

For explanation tasks, qualitative evaluation by human annotators measures the factual correctness, helpfulness, and contextual relevance of the generated explanations. The proposed framework is compared against baseline models, such as FastGTN, and alternative LLM-based approaches using varying levels of contextual information.

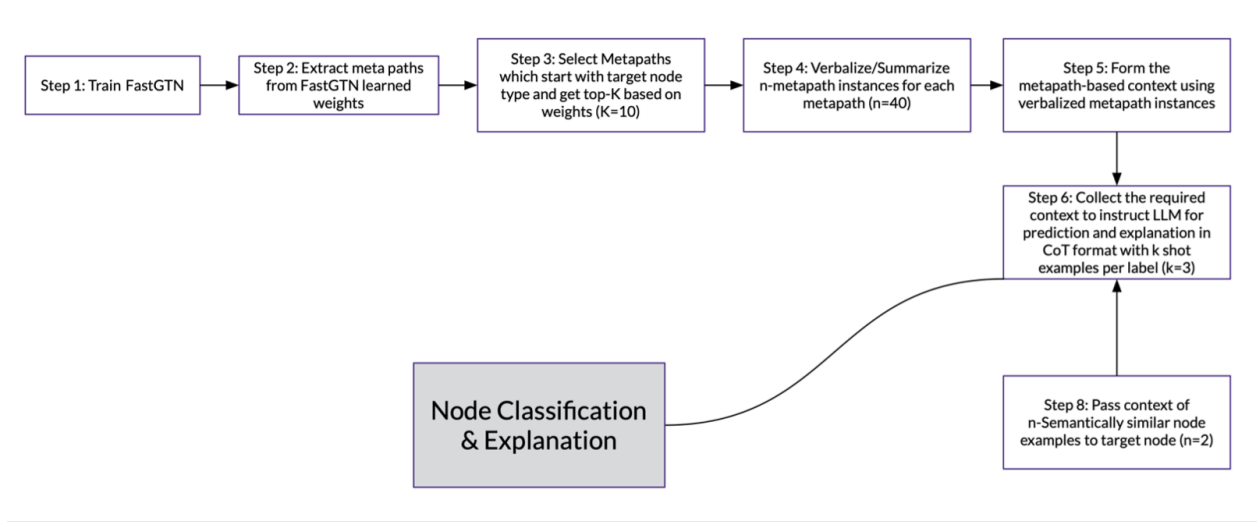


Figure 3.1: Steps involved in the process

Figure 2. Verbalization of metapath: Prompt template:

**Metapath Summarization Prompt**

Human: You are a human whose goal is to generate the meaning of different meta-paths in a heterogeneous graph network. You are provided with all the node types and edge-type and the underlying relation between them.

Please find the node types enclosed inside the XML tags `<node-type>`,`</node-type>`  
`<node-type>node-type</node-type>`

Edge-type and their underlying relation are enclosed between XML tags `<edge-relation>`,`</edge-relation>`  
`<edge-relation>edge-relation</edge-relation>`

Please find one meta-path summarization example below: meta-path-summarization-example-1  
 Can you generate the meaning of the following meta-paths enclosed inside XML tags `<meta-path>`,`</meta-path>`  
`<meta-path>meta-path</meta-path>`

Write the meaning inside xml tags `<meaning>` `</meaning>`. Enclose the meaning of every metapath individually into xml tags `<meaning>`,`</meaning>`.

Figure 3.2: Verbalization of metapath: Prompt template

Figure 3. Augmented Prompt in the template mentioned in Figure 2.

```
Prompt After Augmenting

Human: You are a human whose goal is to generate the meaning of different meta-paths in a heterogeneous graph network. You are provided with all the node types and edge types and the underlying relation between them.

Please find the node types enclosed inside the XML tags <node-type>,</node-type>
<node-type>
P: paper
A: author
C: conference
</node-type>

Edge-type and their underlying relation are enclosed between XML tags <edge-relation>,</edge-relation>
<edge-relation>
PC:=PC (P stands for bcd, S stands for efg) : Paper bcd is published in conference efg
CP:=CP (C stands for efg, P stands for bcd) : Conference efg has paper bcd as one of its submission
PA:=PA (P stands for bcd, A stands for hij) : Paper bcd has been authored by hij
AP:=AP (A stands for hij, P stands for bcd) : Author hij has authored paper bcd
PP:=PP (P stands for bcd, P stands for klm) : Paper bcd has referenced another paper klm
</edge-relation>

Please find one meta-path summarization example below:
BAB (B stands for xyz, A stands for abc, B; stands for def) : Book xyz is related to book def as they are both written by the same author abc. They might be similar semantically or have a similar theme

Can you generate the meaning of the following meta-paths enclosed inside XML tags <meta-path>,</meta-path>
<meta-path>
PAP; (P stands for NbrnT, A stands for P3fAb, P; stands for nFbm0)
PP;A (P stands for NbrnT, P; stands for P3fAb, A stands for nFbm0)
PP;C (P stands for NbrnT, P; stands for P3fAb, C stands for nFbm0)
PA (P stands for NbrnT, A stands for P3fAb)
PC (P stands for NbrnT, C stands for P3fAb)
PP;P;; (P stands for NbrnT, P; stands for P3fAb, P;; stands for nFbm0)
PP; (P stands for NbrnT, P; stands for P3fAb)
PCP; (P stands for NbrnT, C stands for P3fAb, P; stands for nFbm0)
PP;AP;; (P stands for NbrnT, P; stands for P3fAb, A stands for nFbm0, P;; stands for HnKYa)
</meta-path>

Write the meaning inside xml tags <meaning> </meaning>. Enclose the meaning of every metapath individually into XML tags <meaning>,</meaning>.

Assistant:
```

Figure 3.3: Augmented Prompt in the template mentioned in Figure 3.2

**Figure 4.** Example generation: Verbalization of metapath

**Output Example for Summarizing Meta-path**

<meaning>PP; (P stands for NbrnT, A stands for P3fAb, P; stands for nFbm0): Paper NbrnT is related to paper nFbm0 through author P3fAb. This meta-path suggests that both papers are likely to share thematic or research similarities due to the common authorship.</meaning>

<meaning>PP;A (P stands for NbrnT, P; stands for P3fAb, A stands for nFbm0): Paper NbrnT references paper P3fAb, which is authored by nFbm0. This path could indicate that the research in paper NbrnT is influenced or builds upon the work done by author nFbm0 in paper P3fAb.</meaning>

<meaning>PP;C (P stands for NbrnT, P; stands for P3fAb, C stands for nFbm0): Paper NbrnT references paper P3fAb, which is published in conference nFbm0. This meta-path might suggest a thematic or subject matter connection between the two papers, focused on the themes of the conference.</meaning>

<meaning>PA (P stands for NbrnT, A stands for P3fAb): Paper NbrnT has been authored by P3fAb. This indicates a direct authorial relationship where P3fAb contributes to the content and research findings of the paper NbrnT.</meaning>

<meaning>PC (P stands for NbrnT, C stands for P3fAb): Paper NbrnT is published in conference P3fAb. This meta-path suggests that the paper's subject is relevant to the themes or focus areas of the conference.</meaning>

<meaning>PP;P;; (P stands for NbrnT, P; stands for P3fAb, P;; stands for nFbm0): Paper NbrnT references paper P3fAb, which in turn references another paper nFbm0. This extended chain of references may indicate a progression or evolution of ideas and research across the three papers.</meaning>

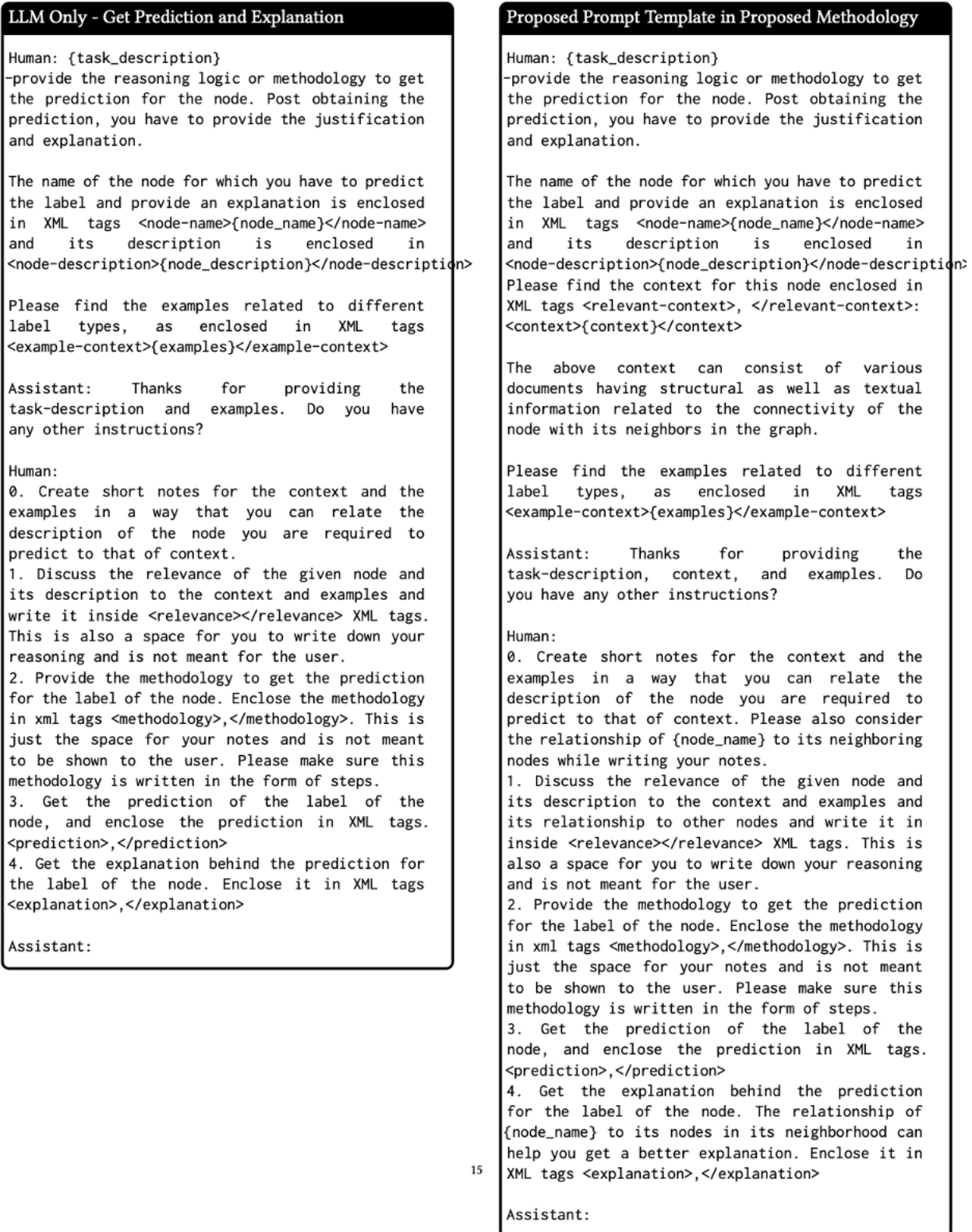
<meaning>PP; (P stands for NbrnT, P; stands for P3fAb): Paper NbrnT references paper P3fAb. This relationship implies a direct citation or intellectual linkage, suggesting that NbrnT builds upon or discusses the research presented in P3fAb.</meaning>

<meaning>PCP; (P stands for NbrnT, C stands for P3fAb, P; stands for nFbm0): Paper NbrnT is published in conference P3fAb, and conference P3fAb features another paper nFbm0. This suggests a thematic linkage between the two papers based on their common selection for the same conference.</meaning>

<meaning>PP;AP;; (P stands for NbrnT, P; stands for P3fAb, A stands for nFbm0, P;; stands for HnKYa): Paper NbrnT references paper P3fAb, which is authored by nFbm0 who also authored paper HnKYa. This complex meta-path suggests a deep thematic or research linkage potentially due to the repeated involvement of author nFbm0.</meaning>

Figure 3.4: Example generation: Verbalization of metapath

Figure 5. Prompt template for getting prediction and explanation



15

Figure 3.5: Prompt template for getting prediction and explanation

This chapter discusses the experimental setup and methodology used to evaluate the proposed framework for node classification and explanation generation tasks. We describe the datasets, the choice of LLM, and the progressive series of experiments, highlighting the incremental changes made to refine the model’s performance. The experiments start with baseline methods and evolve to leverage Meta path-aware contexts and semantic information.

## 4.1 Dataset Description

We used two publicly available heterogeneous graph datasets:

- **DBLP:** A citation network comprising papers (P), authors (A), and conferences (C) as node types, with field of study as labels.
- **Goodreads:** A books dataset with nodes such as books (B), authors (A), publishers (P), formats (F), and language codes (L), with genres as labels.

Both datasets were filtered to include three target labels: Machine learning, computer networking, and theoretical computer science in DBLP Dataset and Fiction, non-fiction, and romance in GoodReads Dataset.

The statistics of the filtered datasets, including node and edge types, are summarized in Table 4.1.

| Dataset   | Node Types                                                                          | Edge Types                                                                                                                                                                                                        |
|-----------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DBLP      | Author: 17,567<br>Conference: 1,562<br>Paper: 6,000                                 | Author-Paper: 18,810<br>Conference-Paper: 5,687<br>Paper-Author: 18,810<br>Paper-Conference: 5,687<br>Paper-Paper: 261                                                                                            |
| Goodreads | Author: 4,452<br>Book: 3,600<br>Publisher: 1,790<br>Format: 26<br>Language code: 47 | Author-Book: 4,762<br>Book-Author: 4,762<br>Book-Language code: 2,040<br>Language code-Book: 2,040<br>Book-Publisher: 2,665<br>Publisher-Book: 2,665<br>Book-Book: 32<br>Book-Format: 2,762<br>Format-Book: 2,762 |

Table 4.1: Dataset statistics for DBLP and Goodreads.

## 4.2 Hyperparameter Configurations

In our experiments, we utilized Claude 2.1 as the LLM, configured with a temperature of 0 (deterministic behavior), top-P of 1 (considering all token probabilities), and a maximum of 500 tokens to sample. For meta-path context, we employed 10 types of metapaths with a total of 40 instances. Fast-GTN training was conducted using three layers, a dropout of 0.3, and hidden dimensions of 128, with learning rates of 0.01 for DBLP and 0.001 for Goodreads. Homogeneous contexts were incorporated up to three hops, with neighbor distributions adjusted across 1-hop, 2-hop, and 3-hop levels. The datasets featured target node types of papers and books, with label distributions equally balanced across categories such as machine learning and fiction.

Table 4.2: LLM inference Hyperparamters

| LLM choice | temperature | top_P | max_tokens_to_sample |
|------------|-------------|-------|----------------------|
| Claude 2.1 | 0           | 1     | 500                  |

Table 4.3: Meta-Path Hyperparameters

|                             |    |
|-----------------------------|----|
| Number of type of metapaths | 10 |
| Total metapath instances    | 40 |

Table 4.4: Fast-GTN training hyperparameters

| Dataset    | Num_layers | Dropout | Num_channels | hidden_dim | lr    |
|------------|------------|---------|--------------|------------|-------|
| heightDBLP | 3          | 0.3     | 4            | 128        | 0.01  |
| Goodreads  | 3          | 0.3     | 8            | 128        | 0.001 |

Table 4.5: Hyperparameters in homogeneous context

| <b>Count of neighbours</b>     | <b>1-hop</b> | <b>2-hop</b> | <b>3-hop</b> |
|--------------------------------|--------------|--------------|--------------|
| Homogeneous context upto 1 hop | 40           | 0            | 0            |
| Homogeneous context upto 2 hop | 20           | 20           | 0            |
| Homogeneous context upto 3 hop | 13           | 13           | 14           |

Table 4.6: Dataset: Label Types and Distribution

| <b>Dataset type</b> | <b>Target node type</b> | <b>Labels</b>                                                                    | <b>Label distribution</b> |
|---------------------|-------------------------|----------------------------------------------------------------------------------|---------------------------|
| DBLP                | paper                   | machine_learning,<br>theoretical_com-<br>puter_science, com-<br>puter_networking | Equally distributed       |
| Goodsreads          | books                   | fiction, non-fiction, ro-<br>mance                                               | Equally distributed       |

## 4.3 Progression of Experiments

The experiments are structured in the following manner:

### **Experiment 1: FastGTN (Baseline)**

The baseline experiment used the FastGTN model to generate predictions. FastGTN identifies influential meta-paths and uses them for node classification. The predictions were evaluated using micro-F1 and macro-F1 scores, serving as a benchmark for subsequent experiments.

### **Experiment 2: LLM with 5-shot Examples Per Label**

To evaluate the raw capability of the LLM without additional context, we provided 5-shot examples for each label in the prompt. This experiment helped ensure no information leakage and served as a control to understand the LLM’s standalone performance on node classification.

### **Experiment 3: LLM with Context up to 1-hop Neighbor**

In this experiment, we introduced 1-hop neighbor context, ignoring the heterogeneity aspect of the graph along with 5-shot random examples per label. For each target node, information from its immediate neighbors was incorporated into the prompt. This experiment demonstrated the impact of local graph topology on prediction accuracy.

### **Experiment 4: LLM with Context up to 2-hop Neighbor**

The context was expanded to include 2-hop neighbors of the target node, representing a slightly broader view of the graph. Like Experiment 3, 5-shot random examples per label were used. This step assessed whether increasing the context depth improved predictions or introduced noise.

### **Experiment 5: LLM with Context up to 3-hop Neighbor**

Further extending the context, this experiment included 3-hop neighbors. While this provided an even broader context, it also posed challenges in terms of information saturation and noise. This experiment evaluated the trade-offs of using deeper graph contexts.

### **Experiment 6: LLM with Meta path as Context**

In this experiment, we incorporated Meta path-based contextual information. The Top-K meta-paths were selected from FastGTN based on their influence on node classification. Each meta-path was verbalized into natural language, and instances of these meta-paths were sampled to create a rich, structured context. Alongside this, 5-shot random examples per label were included in the prompt. This experiment assessed the effectiveness of Meta path-aware context.

## Experiment 7: LLM with Meta path and Hybrid Examples

Building upon Experiment 6, this setup combined Meta path-based context with a mix of semantically similar and random examples. Specifically:

- **2-shot semantic examples:** Top 2 Nodes with highest semantic similarity to the target node were included to enhance contextual relevance.
- **3-shot random examples:** 3-shot random examples were selected to maintain diversity and robustness in the context.

This hybrid approach aimed to maximize both the quality of predictions and the interpretability of explanations.

### 4.4 Explanation Generation

Alongside node classification, explanations generated by the LLM were evaluated qualitatively:

- **Human Evaluation:** Human annotators rated explanations on clarity, relevance, and depth (helpfulness) and factual correctness (accuracy and consistency).
- **Ranking Methodology:** Explanations from different approaches (homogeneous context, Meta path-aware context, etc.) were compared for the same predictions. Annotators ranked explanations to identify the most preferred approach.

Label definitions for "helpful" and "factual correctness" categories were provided to evaluators to ensure consistency and reliability in evaluation. These qualitative metrics offered insights into how effectively the proposed framework enhanced explanation quality.

### 4.5 Summary

This chapter outlined the experiments conducted to evaluate the proposed methodology, progressively introducing contextual information and refining the approach. Starting from baseline methods, the experiments demonstrated the incremental impact of incorporating graph topology, Meta path-aware contexts, and semantic similarity on both node classification and explanation generation tasks. The next chapter discusses the results and insights gained from these experiments.

This chapter presents the results of our experiments on the DBLP and Goodreads datasets. We evaluate various approaches for node classification tasks and explanation generation, comparing their performance based on Micro-F1 and Macro-F1 scores. The results are followed by a detailed analysis that highlights key observations and the impact of different modifications in the methodology.

## 5.1 Dataset Overview

The DBLP and Goodreads datasets were chosen due to their diverse characteristics and their relevance to node classification tasks in heterogeneous graphs. The DBLP dataset consists of research papers and their associated metadata, while the Goodreads dataset contains book reviews and ratings, representing two distinct domains for evaluating graph-based learning methods.

## 5.2 Node Classification Results

The performance of different approaches for node classification on the DBLP and Goodreads datasets is summarized in Table 5.1. We compare a series of models, ranging from the Fast-GTN baseline to more sophisticated models that incorporate graph context and example-based learning.

The node classification performance on the DBLP dataset is summarized as follows:

### Fast-GTN Baseline

The Fast-GTN model served as the baseline for both datasets. On DBLP, Fast-GTN achieved a Micro-F1 score of 0.81 and a Macro-F1 score of 0.82. For Goodreads, the scores were lower, with a Micro-F1 of 0.667 and a Macro-F1 of 0.672. The differences can be attributed to the dataset characteristics, with DBLP likely due to heavier Textual attributes and more complex relationships in Goodreads Dataset .

| Dataset   | Approach                                                                                                        | Micro_f1       | Macro_f1       |
|-----------|-----------------------------------------------------------------------------------------------------------------|----------------|----------------|
| DBLP      | Fast-GTN                                                                                                        | 0.81           | 0.82           |
|           | Only LLM with 5 shot examples per label                                                                         | 0.767          | 0.793          |
|           | LLM with context up to one-hop and 5 shot random examples per label                                             | 0.721          | 0.745          |
|           | LLM with context up to two-hop and 5 shot random examples per label                                             | 0.78           | 0.81           |
|           | LLM with context up to three-hop and 5 shot random examples per label                                           | 0.799          | 0.825          |
|           | LLM with metapath as context and 5 shot random examples per label                                               | 0.876          | 0.882          |
|           | <b>LLM with metapath as context and 2-shot semantic examples per label and 3-shot random examples per label</b> | <b>0.90259</b> | <b>0.90533</b> |
| Goodreads | Fast-GTN                                                                                                        | 0.667          | 0.672          |
|           | Only LLM with 5 shot examples per label                                                                         | 0.771          | 0.791          |
|           | LLM with context up to one-hop and 5 shot random examples per label                                             | 0.84           | 0.853          |
|           | LLM with context up to two-hop and 5 shot random examples per label                                             | 0.843          | 0.856          |
|           | LLM with context up to three-hop and 5 shot random examples per label                                           | 0.855          | 0.8606         |
|           | LLM with metapath as context and 5 shot random examples per label                                               | 0.877          | 0.886          |
|           | <b>LLM with metapath as context and 2-shot semantic examples per label and 3-shot random examples per label</b> | <b>0.904</b>   | <b>0.909</b>   |

Table 5.1: Comparison of different approaches on DBLP and Goodreads datasets.

### LLM with 5-Shot Examples

When we tested the LLM with 5-shot examples per label, the performance on DBLP decreased slightly, achieving a Micro-F1 of 0.767 and a Macro-F1 of 0.793. For Goodreads, the scores were improved, with a Micro-F1 of 0.771 and a Macro-F1 of 0.791. This indicates that while LLMs can utilize a limited number of examples effectively, more advanced strategies incorporating graph structure are necessary for further improvements. This also ensures that there was almost no inherent information leak happening with LLM.

### Incorporating Graph Context

Progressively adding graph context improved the performance across both datasets: - 1-hop context led to a Micro-F1 of 0.721 and Macro-F1 of 0.745 on DBLP, and a Micro-F1 of 0.84 and Macro-F1 of 0.853 on Goodreads. - 2-hop context continued to enhance performance, achieving 0.78 and 0.81 on DBLP, and 0.843 and 0.856 on Goodreads. - 3-hop context further boosted performance, reaching a Micro-F1 of 0.799 and a Macro-F1 of 0.825 on DBLP, and 0.855 and 0.8606 on Goodreads. However, the improvements became marginal with deeper contexts.

### Meta-Path Context

The inclusion of meta-path context resulted in the most significant improvements. On DBLP, the model with meta-path context achieved a Micro-F1 of 0.876 and a Macro-F1 of 0.882, while on Goodreads, the scores increased to 0.877 (Micro-F1) and 0.886 (Macro-F1).

### Best Results with Combined Context and Examples

The highest performance across both datasets was achieved with the model that incorporated meta-path context, 2-shot semantic examples, and 3-shot random examples. For DBLP, this resulted in a Micro-F1 of 0.90259 and a Macro-F1 of 0.90533, and for Goodreads, the model reached a Micro-F1 of 0.904 and a Macro-F1 of 0.909.

# Explanation Generation Results

In addition to node classification, we evaluated the proposed model’s ability to generate explanations for its decisions. The best-performing model, which combined meta-path context with semantic examples, also excelled in explanation generation. These explanations were rated as more factual and helpful compared to those generated by other methods. The model’s ability to generate contextually rich and semantically relevant explanations aligns with the improvements observed in the classification tasks.

| <b>Dataset</b>                                                        | <b>% of labels with factual correctness rating as 3</b> | <b>% of labels with Helpfulness rating as 3</b> | <b>% of labels with Helpfulness rating as 2</b> |
|-----------------------------------------------------------------------|---------------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| <b>DBLP</b>                                                           |                                                         |                                                 |                                                 |
| LLM with homogeneous context                                          | 95.12                                                   | 10                                              | 90                                              |
| LLM with metapath as well as semantic examples (proposed methodology) | 97.5                                                    | 30                                              | 70                                              |
| <b>Goodreads</b>                                                      |                                                         |                                                 |                                                 |
| LLM with homogeneous context                                          | 94.23                                                   | 19.23                                           | 76.92                                           |
| LLM with metapath as well as semantic examples (proposed methodology) | 96.15                                                   | 38.46                                           | 57.69                                           |

Table 5.2: Comparison of Factual Correctness and Helpfulness Ratings between Different Approaches for DBLP and Goodreads Datasets

## 5.3 Results Analysis

The results demonstrate that the combination of graph-based context (especially meta-path context) and semantic examples provides the best performance across both node classification and explanation generation tasks. Key observations from the analysis include:

- **Graph Context:** Incorporating hop-based graph context progressively improves the model’s performance, with diminishing returns at deeper hops.
- **Meta-Path Context:** The inclusion of meta-path context significantly boosts performance, showing that the model benefits from learning the relationships between nodes through path-based structures.
- **Example Variability:** Combining context from semantically similar node with n-shot random examples along with metapath based context enhances the model’s ability to generalize and classify nodes accurately, as evidenced by the highest performance with 2-shot semantic and 3-shot random examples.
- **Explanations:** The combination of rich context and varied examples also results in better explanations, making the model more interpretable and practical for real-world applications.

Overall, the results confirm the effectiveness of combining graph context and example-based learning strategies for node classification in heterogeneous graphs. These methods not only improve classification accuracy but also enhance the quality of explanations, making them both more factual and helpful.

## 5.4 Summary

The experiments on the DBLP and Goodreads datasets have shown that models incorporating meta-path context and semantic examples perform significantly better than traditional approaches like Fast-GTN. By leveraging graph-based structures and enriching the examples with semantically meaningful context, our model achieves superior performance in both classification tasks and explanation generation. These findings underline the importance of context and example diversity in improving both accuracy and interpretability in graph-based machine learning tasks.

## 6.1 Conclusion

In this work, we introduced a novel framework for leveraging meta-path-aware contextual information to enhance node classification and explanation generation for heterogeneous graphs. The framework successfully integrates Large Language Models (LLMs) with heterogeneous graph neural networks (HGNNs), providing an innovative approach to addressing challenges in both prediction accuracy and interpretability.

The experimental results on DBLP and Goodreads datasets demonstrated the significant advantages of the proposed methodology. By incorporating metapath-based context, semantic similarity, and hybrid examples, we achieved state-of-the-art performance in node classification tasks, surpassing Fast-GTN and homogeneous-context LLM approaches. Furthermore, the explanations generated by the framework were rated highly in terms of factual correctness and helpfulness, showcasing the ability to make HGNN decisions more interpretable and user-friendly.

This study highlights the importance of combining structural and textual information in heterogeneous graphs and demonstrates the power of LLMs when provided with enriched, well-constructed prompts.

## 6.2 Future Work

This study highlights the importance of combining structural and textual information in heterogeneous graphs and demonstrates the power of LLMs when provided with enriched, well-constructed prompts.

While the focus of this work has been on node classification, the proposed explainability framework has the potential to be extended to other graph machine learning tasks. The following directions outline future research opportunities:

- 1. Explainability for Link Prediction:**

Link prediction tasks aim to predict whether an edge exists between two nodes in a graph. Incorporating meta-paths and semantic context into LLM prompts can provide intuitive explanations for predicted links, such as "These two nodes are likely connected because they share similar neighbors or belong to related communities." This could be especially valuable in applications like social network analysis or recommender systems.

## 2. **Explainability for Graph Classification:**

Graph classification involves predicting labels for entire graphs rather than individual nodes. By extending the current framework, LLMs can generate explanations about why a graph belongs to a specific class, leveraging global structural features and meta-path-based insights. For example, in molecular graph analysis, explanations could highlight critical substructures responsible for a chemical property.

## 3. **Dynamic Context for Temporal Graphs:**

Temporal or dynamic graphs evolve over time, adding complexity to tasks like temporal node classification or event prediction. The explainability framework can be adapted to include temporal patterns and sequences, offering time-sensitive insights such as "This node's behavior changed due to interactions with its neighbors in recent time steps."

## 4. **Improved Scalability for Large Graphs:**

As graph sizes increase, sampling techniques could be refined to identify the most relevant subgraphs or meta-paths for explanation. This would ensure the framework remains computationally efficient while providing high-quality insights.

## 5. **User-Driven Explanation Customization:**

Future work can explore customizable explanations tailored to different audiences. For instance, technical users might prefer detailed, meta-path-focused explanations, while non-technical users might benefit from simplified, high-level descriptions.

## 6. **Evaluation Metrics for General Explainability:**

Developing domain-specific and task-specific evaluation metrics for explanations is crucial to standardizing the assessment of explainability across graph ML tasks. This would ensure consistency and provide benchmarks for future advancement.

- [1] Jianming Zheng et al. “Heterogeneous graph neural networks to predict what happen next.” In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 328–338.
- [2] Siyong Xu et al. “Topic-aware heterogeneous graph neural network for link prediction.” In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 2261–2270.
- [3] Yitong Pang et al. “Heterogeneous global graph neural networks for personalized session-based recommendation.” In: *Proceedings of the fifteenth ACM international conference on web search and data mining*. 2022, pp. 775–783.
- [4] Huance Xu et al. “Global context enhanced social recommendation with hierarchical graph neural networks.” In: *2020 IEEE international conference on data mining (ICDM)*. IEEE. 2020, pp. 701–710.
- [5] Cunchao Tu et al. “A unified framework for community detection and network representation learning.” In: *IEEE Transactions on Knowledge and Data Engineering* 31.6 (2018), pp. 1051–1065.
- [6] Yixiang Fang et al. “Effective and efficient community search over large heterogeneous information networks.” In: *Proceedings of the VLDB Endowment* 13.6 (2020), pp. 854–867.
- [7] Chuxu Zhang et al. “Heterogeneous graph neural network.” In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 793–803.
- [8] Hao Yuan et al. “On explainability of graph neural networks via subgraph explorations.” In: *International conference on machine learning*. PMLR. 2021, pp. 12241–12252.
- [9] Jost Tobias Springenberg et al. “Striving for simplicity: The all convolutional net.” In: *arXiv preprint arXiv:1412.6806* (2014).
- [10] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks.” In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.
- [11] Phillip E Pope et al. “Explainability methods for graph convolutional neural networks.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 10772–10781.

- [12] Zhitao Ying et al. “Gnnexplainer: Generating explanations for graph neural networks.” In: *Advances in neural information processing systems* 32 (2019).
- [13] Dongsheng Luo et al. “Parameterized explainer for graph neural network.” In: *Advances in neural information processing systems* 33 (2020), pp. 19620–19631.
- [14] Thorben Funke, Megha Khosla, and Avishek Anand. “Hard masking for explaining graph neural networks.” In: (2020).
- [15] Yiming Tan et al. “Evaluation of ChatGPT as a question answering system for answering complex questions.” In: *arXiv preprint arXiv:2303.07992* (2023).
- [16] Tianyi Zhang et al. “Benchmarking large language models for news summarization.” In: *arXiv preprint arXiv:2301.13848* (2023).
- [17] Steven Moore<sup>1</sup> et al. “Check for updates Empowering Education with LLMS-The Next-Gen Interface and Content Generation.” In: *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings*. Springer Nature. 2023, p. 32.
- [18] Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. “LLM Based Generation of Item-Description for Recommendation System.” In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 2023, pp. 1204–1207.
- [19] Seongjun Yun et al. “Graph Transformer Networks.” In: *Neural Information Processing Systems (NeurIPS)*. arXiv:1911.06455. 2019. DOI: 10.48550/arXiv.1911.06455.
- [20] Seongjun Yun et al. “Graph Transformer Networks: Learning meta-path graphs to improve GNNs.” In: *Neural Networks* 153 (2022), pp. 104–119.
- [21] Chuan Shi et al. “A survey of heterogeneous information network analysis.” In: *IEEE Transactions on Knowledge and Data Engineering* 29.1 (2016), pp. 17–37.
- [22] Xiao Wang et al. “Heterogeneous graph attention network.” In: *The world wide web conference*. 2019, pp. 2022–2032.
- [23] Yanbin Jiang et al. “Incorporating metapath interaction on heterogeneous information network for social recommendation.” In: *Frontiers of Computer Science* 18.1 (2024), p. 181302.
- [24] Nadia Burkart and Marco F. Huber. “A Survey on the Explainability of Supervised Machine Learning.” In: *CoRR* abs/2011.07876 (2020). arXiv: 2011.07876. URL: <https://arxiv.org/abs/2011.07876>.
- [25] Jaykumar Kakkad et al. “A survey on explainability of graph neural networks.” In: *arXiv preprint arXiv:2306.01958* (2023).
- [26] Thomas Schnake et al. “Higher-order explanations of graph neural networks via relevant walks.” In: *IEEE transactions on pattern analysis and machine intelligence* 44.11 (2021), pp. 7581–7596.
- [27] Rana Hussein, Dingqi Yang, and Philippe Cudré-Mauroux. “Are meta-paths necessary? Revisiting heterogeneous graph embeddings.” In: *Proceedings of the 27th ACM international conference on information and knowledge management*. 2018, pp. 437–446.
- [28] Chandan Singh et al. “Rethinking Interpretability in the Era of Large Language Models.” In: *arXiv preprint arXiv:2402.01761* (2024).