![PES University logo]

Dissertation on

"DETECTING ANOMALOUS SELF-CITATIONS AND GEN-AI"

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

UE21CS461A – Capstone Project Phase - 2

Submitted by:

Farhaan Ebadulla	**PES2UG21CS170**
Gaurav BV	**PES2UG21CS175**
H Manoj	**PES2UG21CS182**
K Kruthik	**PES2UG21CS243**

Under the guidance of

Prof. Nazmin Begum
Assistant Professor
PES University

June - Nov 2024

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

'DETECTING ANOMALOUS SELF-CITATIONS AND GEN-AI'

is a bonafide work carried out by

Farhaan Ebadulla	**PES2UG21CS170**
Gaurav BV	**PES2UG21CS175**
H Manoj	**PES2UG21CS182**
K Kruthik	**PES2UG21CS243**

In partial fulfilment for the completion of seventh semester Capstone Project Phase - 2 (UE21CS461A) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period June 2024 – Nov. 2024. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 7th semester academic requirements in respect of project work.

Signature	Signature	Signature
Prof. Nazmin Begum	Dr. Sandesh B J	Dr. B K Keshavan
Designation	Chairperson	Dean of Faculty

External Viva

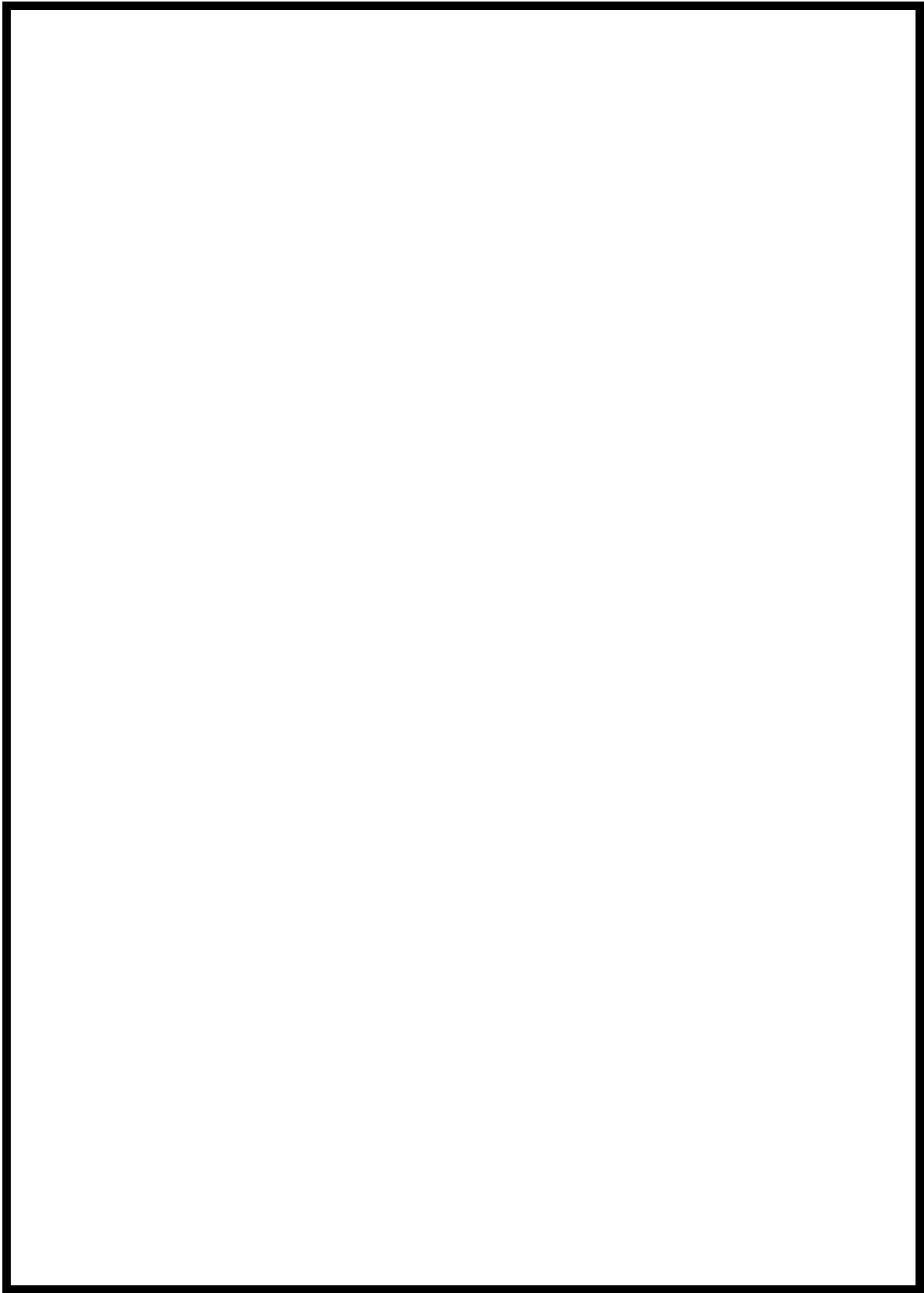**Name of the Examiners**   **Signature with Date**

1. _____   _____
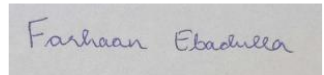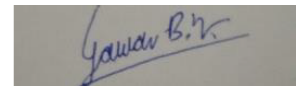
2. _____   _____

DECLARATION

We hereby declare that the Capstone Project Phase - 2 entitled **"Detecting Anomalous Self-Citations Using Relation Networks"** has been carried out by us under the guidance of **Dr. Nazmin Begum, Assistant Professor** and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester June – Nov. 2023. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.
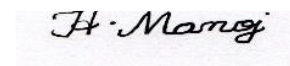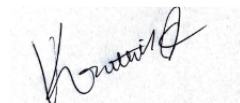
PES2UG21CS170	**Farhaan Ebadulla**	_Farhaan Ebadulla_
PES2UG21CS175	**Gaurav BV**	_Gaurav B.V._
PES2UG21CS182	**H Manoj**	_H. Manoj_
PES2UG21CS243	**K Kruthik**	_Kruthik_

ACKNOWLEDGEMENT

ABSTRACT

Accurate assessment of research impact is of utmost importance because of its gravity in assessing scholarly contributions, which determines the fundings and promotions. While impartial assessments are of utmost need, anomalous self-citation--the practice of inflating one's own research with unmerited citation-may affect the dilution of the importance to be rendered to other research workers' work.

Using citation network analysis, machine learning techniques, and domain-specific knowledge, the study depicts a novel approach to identifying anomalous self-citation patterns.

To this end, it builds a comprehensive citation network through a large academic publication corpus, extracting features always at the author and paper level. This is to develop a predictive model to identify those authors who exercise statistically improbable self-citation levels and distinguish between regular oscillations and persistent self-citation patterns.

The study, thus, evaluates the suitability for various disciplines of approaches for the identification of unusual self-citations, contributing to discussions about the science of evaluation and research ethics and providing a robust approach that will, in the end, work towards cancelling out the skewing effect of excessive self-citation in the assessment of research impact.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In academic research, it is crucial to reference relevant earlier studies. When you cite previous work, you are essentially acknowledging the debt of ideas you owe to others. Moreover, it helps provide context for your own discoveries.

The academic world is concerned about the self-citation phenomena, in which authors cite their own previous publications. Although self-citations are a valid way for researchers to place their most recent work within their body of work, using them excessively or disproportionately might skew how important and impactful one's scholarly contributions are seen to be.

The integrity of research assessment measures is manipulated by anomalous self-citation practices, in which authors consistently and incorrect cite their work. The quality and the impact of the scholarly output of the publication or research paper citation metrics like h-index, journal impact factor and citation counts are used. Important choices about funding distribution, tenure, hiring, and promotion are informed by these variables. However, they can be made unreliable by anomalous self-citation behaviour that artificially inflates the number of citations.

It is necessary to understand and combat such behaviors that artificially increase the number of citations as this helps to uphold the integrity and equity of research publishing. Although certain works have tried to explore the degree and reasons for excessive self-citation, there is still more need to develop efficient detection methods of such behaviours on a much broader scale. It is even more challenging to pinpoint self-citation for the purposes of accentuating one's citation index due to the collaborative cultures and creativity that come in with the disciplines.

This project aims to construct an in-depth structure, as well as the set of tools, for the detection of abnormal self-citation patterns in academic texts. By using such techniques as social network studies, graph theory algorithms, and knowledge of a specific area, we suggest a technique to tell the difference between a real self-citation and one that has been perpetrated to boost an author's citation numbers. The method that we propose here also includes the variety of self-citation strategies which allows for a more nuanced and contextualized analysis of self-citation patterns.

PROBLEM STATEMENT

Researchers rely on the assessment of articles and their citations in the process of academic publishing, which affects several crucial decisions such as funding or the reputation of an author. To determine the quality level of a publication or a citation of a research paper, publication metrics, such as the h-index, journal impact factor, and citation counts among others, are referred to. Generally, there is a problem with the reliability of such measurements. Therefore, the tendency of the authors to exaggerate the degree of relevance of their previous works to the current study often leads to the practice of self-citation. This and other similar opportunistic practices can contribute to an author's exaggerated media coverage which can distort the picture of his or her research output. This issue creates unfair advantages that can hurt researchers who play by the rules with their citations. It also damages the reputation of academic institutions and research organizations. The detection of anomalous self-citation patterns is a complex topic as it involves separating out anomalous self-citations from those which are legit and provide context. The problem is complicated for various of reasons:

•   Due to differences in citation practices, collaborative patterns, and the nature of research in some subjects, self-citations might vary throughout various domains

•   Researchers in the early stages might exhibit higher anomalous self-citations so that they can establish their research paths, while more experienced scholars will have diverse citation patterns

•   Author's might mention the work of their colleagues, co-authorship research team composition can lead to anomalous self-citation

•   Self-citations patterns such as citations rings will reveal themselves over time. Self-citation patterns will change over time, exhibiting brief spikes of self-citations which may not always be an indication of anomalous self-citation

By addressing these problems, we aim to contribute in maintaining the integrity of academic publications. In the end, identifying and reducing anomalous self-citation behaviours will promote a fairer evaluation.

LITERATURE REVIEW

1. Literature Survey

3.1. Paper-1

3.1.1. **Introduction:** This paper introduces NetGAN, a method for generating graphs using random walks while maintaining the structural properties.

3.1.2. Characteristics and Implementation: NetGAN uses an LSTM to handle sequence memory during random walk generation and employs latent space interpolation to smoothly change graph characteristics.

3.1.3. **Features:** NetGAN can generate graphs while maintaining structural properties without explicitly specifying them.

3.1.4. Evaluation: NetGAN's performance is evaluated by link prediction using sampled random walks from the trained generator.

3.2. Paper-2

3.2.1. **Introduction:** The paper proposes a novel approach to evaluate citation credibility by evaluating the stability of citation paths using network perturbations .

3.2.2. Characteristics and Implementation: The approach involves generating a new graph by using network structure information, initializing an array with zeros equal to the number of edges, creating temporary graphs at each iteration, and randomly removing nodes to evaluate network robustness and link stability.

3.2.3. **Features:** The method focuses on randomly removing nodes to assess network stability and credibility of individual links.

3.2.4. Evaluation: The paper highlights the limitations of the proposed method, the textual information can vary a lot , complexity of graph learning, and sensitivity to the sampling strategy. the method is sensitive to how you choose which data to use for the analysis.

3.3. **Paper-3**

3.3.1. **Introduction:** This paper develops a method to identify outlier patterns(motifs) using user queries to find the structures.

3.3.2. Characteristics and Implementation: The method creates user queries with target motifs and search paths to guide searching conditions effectively. The method utilizes sub-graph and meta-path queries to develope a motif discovery algorithm and improve outlier pattern identification accuracy by computing motif similarity from node similarities.

3.3.3. **Features:** The approach aims to improve the understanding of network motifs and outlier patterns.

3.3.4. Evaluation: The paper states limitations such as the dependency on user queries, scalability, and the evaluation metric used.

3.4. **Paper-4**

3.4.1. **Introduction:** This paper develops an algorithm for estimating subgraph concentrations and detecting motifs in various networks.

3.4.2. Characteristics and Implementation: By sampling subgraphs for relative frequency, the algorithm applies a probabilistic sampling technique to estimate subgraph concentrations. To guarantee precise subgraph concentration estimation, analytical adjustments for sample biases are used.

3.4.3. **Features:** The algorithm gives an accurate estimation of subgraph concentrations and identification of network motifs in large, complex networks.

3.4.4. **Evaluation:** The paper mentions constraints in handling very large subgraphs in networks with high node counts and potential sampling biases.

3.5. **Paper-5**

3.5.1. **Introduction:** This paper introduces an algorithms for computing the frequencies of subgraphs in graphs(motifs).

3.5.2. **Characteristics and Implementation**: The algorithm estimates the occurrences of labeled subgraphs(motifs) within a larger graph through the use of a variant of the regularity lemma.

3.5.3. **Features:** The algorithm ensures efficient estimation of the number of induced subgraphs(motifs) with minimal error .\

3.5.4. **Evaluation:** The paper talks about the dependency on the accuracy of approximation, which can affect the precision of the frequency estimate of the subgraph. This paper also

_____

talks about the limitation linked with the applicability of the algorithm on certain types of graphs(heterogenous graphs).

3.6. **Paper-6**

3.6.1. **Introduction:** This paper utilizes a masking mechanism in self-supervised learning for predicting masked nodes and explores heterogeneous graph inputs to enhance gMAE-based learning approaches.

3.6.2. **Characteristics and Implementation:** The paper implements self supervised learning with masking on the Cite Seer dataset and develops an asymmetric encoder-decoder architecture for predicting masked nodes.
It utilizes heterogeneous graph as inputs to enhance the predictive capabilities of the model.
3.6.3. **Features:** The approach aims to identify anomalous self-citations using self-learning graph transformers.
3.6.4. **Evaluation:** The limitation of this methods discussed are quality of the dataset ,computational complexity , the size of the dataset and the quality of the masking process.

3.7. **Paper-7**

3.7.1. **Introduction:** This paper addresses citation prediction, focusing on realistically predicting the average paper citations per year.
3.7.2. **Characteristics and Implementation:** The paper uses HGN, CA, and TE modules to model various object types and interactions accurately. It utilizes heterogeneous network for data representation, which can hold multiple object types and links. The cluster-aware module
models interactions and nodes within research domains, while the text enhancing module performs term mining to improve data quality.
3.7.3. **Features:** The approach aims to improve citation prediction between publications by using cluster-aware and text-enhanced components in heterogeneous graph neural networks.
3.7.4. **Evaluation:** The paper notes the dependency on dataset quality and size, computational complexity, and the effectiveness of knowledge graphs and decoders as limitations.

_____

the limitation of this method is the dataset quality and the effectiveness of the knowledge graphs and the decoders.

3.8. **Paper-8**

3.8.1. **Introduction:** This paper introduces Enhanced Inter-sentence Attention (EIA) for Semantic Textual Similarity task and develops a new multi-head architecture for self-attention for comparing sentence pairs.

3.8.2. **Characteristics and Implementation:** The approach proposes general and specific enhancements in the attention mechanism, leveraging WordNet semantic relations. It integrates original and enhanced attention representations through gated fusion and evaluates performance enhancement on BERT and RoBERTa models, particularly on smaller datasets.

3.8.3. **Features:** The EIA method aims to improve the performance of STS tasks by enhancing inter-sentence attention.

3.8.4. **Evaluation:** The paper mentions limitations such as limited applicability, effectiveness variability, and computational overhead.

3.9. **Paper-9**

3.9.1. **Introduction:** The goal of this paper is to recommend scientific papers using a hybrid technique that combines concepts from collaborative filtering based on a relevance-based language model with content analysis based on probabilistic topic modelling.

3.9.2. **Characteristics and Implementation:** The main aim is to increase the effectiveness of recommendations by utilising latent topics of interest in researcher publications and utilising the social structure of researchers as trustworthy sources of knowledge. The method conducts an experimental research using the DBLP dataset.

3.9.3. **Features:** The paper aims to provide effective scientific paper recommendations by combining content analysis and collaborative filtering techniques.

3.9.4. **Evaluation:** For improved suggestions, the research proposes expanding the method for creating a heterogeneous network to incorporate item qualities as well as structure, but it acknowledges overfitting to noise as a drawback.

3.10. **Paper-10**

3.10.1. **Introduction:** This paper surveys knowledge graph-based recommender systems.

3.10.2. **Characteristics and Implementation:** The paper highlights the importance in generating recommendations utilising knowledge graphs as side information and analyses the difficulties recommender systems face, including data sparsity and cold-start issues.

3.10.3. **Features:** The paper focuses on knowledge graph-based recommender systems and their advantages.

3.10.4. **Evaluation:** The paper does not provide a specific evaluation but discusses the potential of knowledge graph-based approaches for recommender systems.

3.11. **Paper-11**

3.11.1. **Introduction:** This paper develops the Relative Self-Attention Transformer for encoding molecular features in graph data and enhancing Transformer-based models by incorporating atom relations, positional encoding, and neighborhood features.

3.11.2. **Characteristics and Implementation:** The paper implements Graph Transformers for direct operation on molecular graph structures, utilizes relative positional encoding to encode information about relative distances between atoms, and incorporates neighborhood, bond, and distance embeddings to capture physicochemical features comprehensively.

3.11.3. **Features:** The Relative Self-Attention Transformer aims to improve the representation and modeling of molecular features in graph data.

3.11.4. **Evaluation:** The paper notes the dependency on Graph Convolutional Layers, positional information handling as limitations.

3.12. **Paper-12**

3.12.1. **Introduction:** This paper develops GOAt, a local-level Graph Neural Network (GNN) explainer, to address limitations of existing GNN explainers, such as insufficient discriminability and inconsistency on sameclass data samples.

3.12.2. **Characteristics and Implementation:** GOAt expands GNN outputs into scalar products and attributes each product to input features,

demonstrating superior explanation ability using a great deal of testing on both hypothetical and real databases.

3.12.3. **Features:** GOAt aims to provide local explanations for Graph Neural Network (GNN) outputs by attributing each output component to the input features.

3.12.4. **Evaluation:** The paper states that GOAt overcomes limitations like insufficient discriminability and inconsistency compared to existing GNN explainers through extensive evaluations on synthetic and real datasets. However, it mentions similarity with decomposition methods and overfitting to noise as potential limitations.

3.13. **Paper-13**

3.13.1. **Introduction:** This paper utilizes Graph Convolutional Networks (GCNs) for analyzing hyperspectral images (HSI).

3.13.2. **Characteristics and Implementation:** The approach involves spectral-spatial graph construction, dual self-expression attention-based fusion, and clustering. It extracts information using a sliding window and knearest neighbors to construct a graph for the GCN, applies the GCN to extract neighborhood information, fuses the information to construct a final affinity graph, and clusters this graph.

3.13.3. **Features:** The method utilizes GCNs' learning ability to analyze topological relationships between nodes in HSI data.

3.13.4. **Evaluation:** The paper mentions comparison limitations, a single node focus, and model limitations.

3.14. **Paper-14**

3.14.1. **Introduction:** This paper aims to make large language models (LLMs) faster by using Graph Neural Networks (GNNs).

3.14.2. **Characteristics and Implementation:** The paper proposes Linguistic Graph Knowledge Distillation (LinguGKD), which trains a lightweight GNN student model using an LLM teacher model. It involves teacher feature learning via the LLM, extracting semantically-enriched node features, and student feature learning via the GNN model.

3.14.3. **Features:** The approach combines LLMs and GNNs to develop faster language models.

3.14.4. **Evaluation:** The paper mentions similarity with decomposition methods and overfitting to noise as potential limitations.

_____

3.15. **Paper-15**

3.16.1. **Introduction:** This paper evaluates scientific indices and the performance of researchers, aiming to provide a more comprehensive analysis.

3.16.2. **Characteristics and Implementation:** The paper highlights the limitations of existing indices in evaluating researchers' performance and introduces new indices (th-index and tA-index) that offer a more comprehensive and detailed analysis of the quality of publications in the htail.

3.16.3. **Features:** The new indices aim to improve the evaluation of researchers and their scientific impact by addressing limitations of existing indices.

3.16.4. **Evaluation:** The paper notes the lack of consideration for the quality of publications in the h-tail beyond the h-core, only limited information provided about the distribution, difficulty in comparing researchers' work as limitations.

3.16. **Paper-16**

3.17.1. **Introduction:** This paper presents a new way, called the h-index, to solve problems with current evaluation methods like the h-index. It makes evaluations fairer by including self-citations.

3.17.2. **Characteristics and Implementation:** The paper provides realworld examples to demonstrate how self-citations can affect rankings and highlights the importance of considering self-citations in impact evaluation.

3.17.3. **Features:** The dh-index aims to mitigate the bias introduced by self-citations in research impact evaluation.

3.17.4. **Evaluation:** The paper demonstrates the effectiveness of the dhindex in mitigating bias through real-world examples from the field of computer science, but notes limitations such as dependency on specific datasets and ranking services, and the potential for not considering all factors influencing impact assessments.

3.17. **Paper-17**

3.18.1. **Introduction:** This paper compares two datasets and analyzes them to determine the impact of self-citation, aiming to provide a better comprehensive evaluation and highlight the significance of a revised h-index.

3.18.2. **Characteristics and Implementation:** In order to improve the h-index calculation for Google Scholar, the paper quantifies and removes self-citations from the h-index. It examines two datasets to determine how self-citations affect

the h-index: one with scientists who have won awards, and the other with scientists who have not.

3.18.3.   **Features:** The paper introduces a revised h-index calculation by accounting for self-citations.

3.18.4.   **Evaluation:** The paper talks about its limitations such as lack of data sources and the lack of detailed explanation on excluding a citation.

3.18. **Paper-18**

3.19.1.   **Introduction:** This paper proposes a method of recommending similar legal cases with the help of legal knowledge graphs and GNNs

3.19.2.   **Characteristics and Implementation:** The approach involves construction of a knowledge graph from unstructured text data using graph based models along with rule based models. It creates a graph for cases by representing cases by nodes and citations as edges. Latent Dirichlet Allocation topic modeling is used for feature selection, Relational Graph Convolutional Networks are used for citation link prediction and case similarity. They also use DBSCAN clustering as an alternative unsupervised approach to recommend similar cases.

3.19.3.   **Features:** The paper proposes a method for predicting similar cases and citations in the legal domain using knowledge graphs and graph neural networks, incorporating techniques like GNN explainability and continuous knowledge infusion into LLMs.

3.19.4.   **Evaluation:** The paper highlights limitations such as knowledge graph construction challenges, feature selection challenges due to domain-specific terminology, and modeling challenges related to handling long-range dependencies and document structure present in legal texts.

PROJECT REQUIREMENTS SPECIFICATION

1. Data Collection and Preprocessing: Ingest citation data from the Aminer database, retaining essential fields (title, abstract, references).
2. Network Construction and Loop Detection: Constructs a citation network and identifies self-citation loops using graph random walks.
3. Summarization Pipeline:
   o Page-Level Summarization: Summarizes individual pages using BERT.
   o Aggregate Summarization: Summarizes entire papers using Gemini.
4. Citation Context Extraction: Extracts relevant citation context by isolating text around citation markers.
5. Classification Module: Classifies citations as essential or non-essential using LLMs fine-tuned with a rule-based approach.
6. Reporting and Visualization: Provides clear, interpretable reports of detected anomalies for investigation.

_____

System Design

Design Goals

Performance:

•   Handling large-scale citation networks with millions of papers and relations requires optimizations in data processing, graph construction, and model training pipelines. Parallel computing, distributed processing, and efficient data structures and algorithms may be needed to ensure high performance.

Reliability:

•   The system should provide consistent and reliable results, as its outputs may have significant implications for researchers and academic institutions.
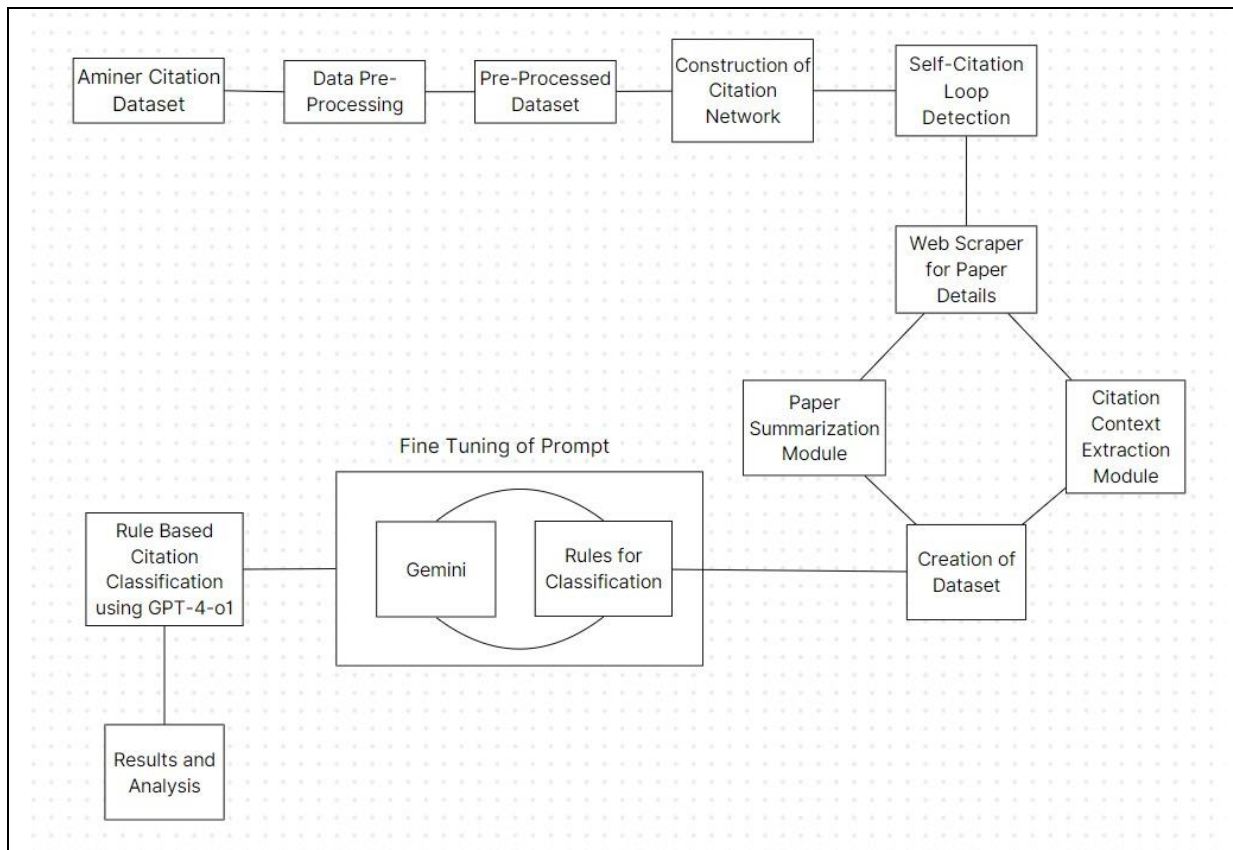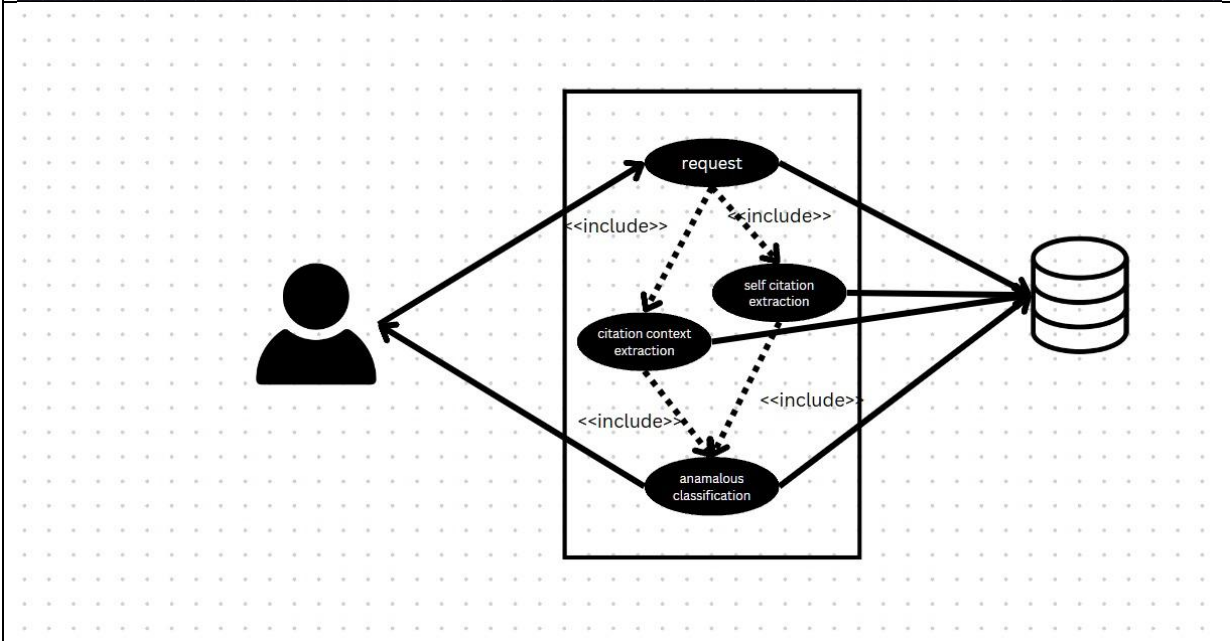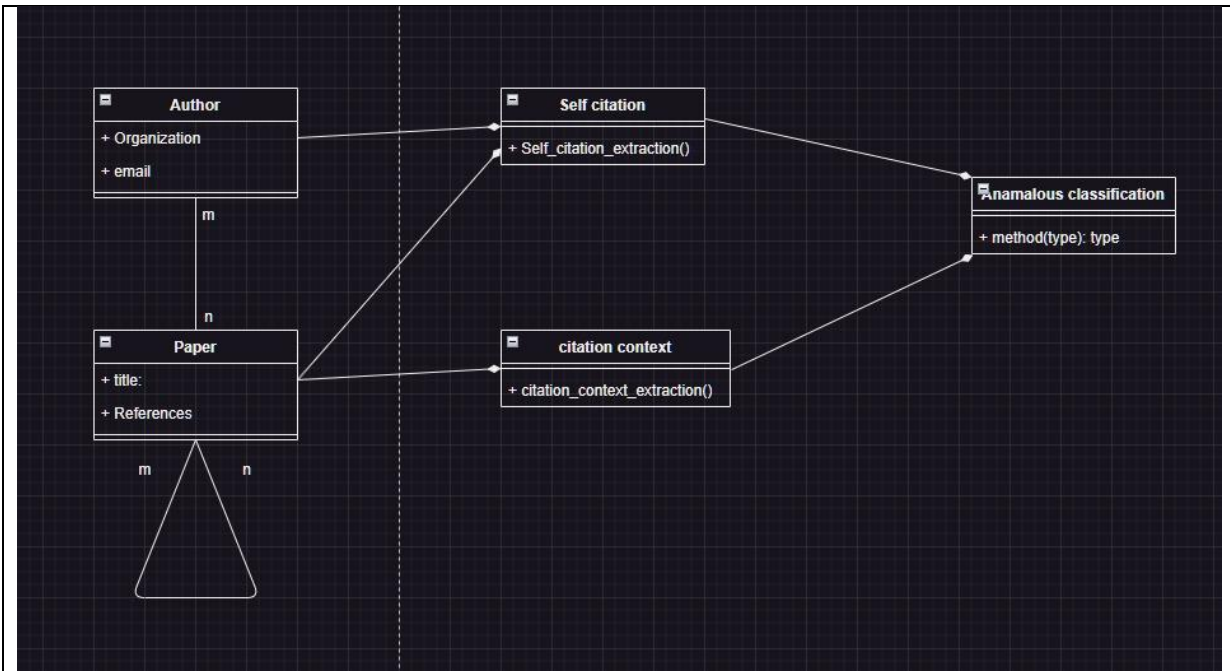
Proposed Methodology

Data Collection and Preprocessing

- Data Source: The dataset originates from Aminer, a large-scale academic citation database containing metadata for scholarly papers, including titles, abstracts, authors, and reference lists.

- Data Filtering and Cleaning:

    o Key Fields Extraction: We keep only the essential fields: paper title, abstract, and references.

    o Noise Reduction: Remove any non-relevant or noisy data, ensuring clean metadata. This step is crucial for accurate citation network modeling.

    o Formatting for Consistency: Standardize field formats to ensure uniformity in references, author names, and citation structure, which simplifies subsequent network analysis.

Citation Network Construction

- Building the Citation Graph:

    o Nodes and Edges: Nodes represent individual papers, and directed edges represent citation relationships. Each directed edge from Paper A to Paper B indicates that Paper A cites Paper B.

    o Self-Citation Definition: Self-citations are identified by author overlap between a citing paper and a cited paper. Self-citation loops are then defined as sequences where an author or group of authors repeatedly cite their own prior work.

- Self-Citation Loop Detection Using Graph Random Walks:

    o Random Walks of Length 3: To identify potential self-citation patterns, perform random walks of length 3 on the graph. These three-step paths allow the detection of citation "loops" where an author may cite themselves multiple times across different papers.

    o Self-Citation Loop Criteria: Loops where all papers have overlapping authorship are flagged, indicating potential self-citation patterns. These loops form the initial pool of citation instances for further analysis.

Web Scraping for Full-Text Access

- Scraping Target Papers in Loops:

    o For each paper identified in a self-citation loop, a scraper retrieves the paper's full text from online sources.

    o This full-text access allows deeper content analysis beyond abstracts and enables comprehensive understanding through summarization.

Summarization Pipeline for Contextual Understanding

_____

- Two-Level Summarization Approach:

Page-Level Summarization with BERT:
Summarize each page of a paper separately to extract key points from the full text.
This segmentation ensures that no critical detail is overlooked, and reduces the load for the second-level summarizer by providing concise summaries per page.

- Aggregated Summarization with Google's Gemini:

After obtaining page-level summaries, apply Gemini to summarize across these summaries.
This step produces a cohesive summary for the entire paper, capturing core arguments, contributions, and contexts in a concise format, suitable for comparing against citation contexts.

Citation Context Extraction
- Extracting Citation Context:

  o For each citation instance, extract four lines of text preceding and following the citation bracket ([]), which is often used to mark references in scientific papers.

  o This "context window" provides enough surrounding text to capture the author's intent and the citation's purpose (whether it supports the current work or simply acknowledges prior work).

  o These citation contexts are key data points for training the classification model.

Definition and Fine-Tuning of Essential and Non-Essential Citation Rules
- Rule Development for Citation Classification:

  o Initially, manually analyze citation instances to develop a set of rules that distinguish essential citations (integral to the paper's argument or results) from non-essential ones (non-crucial self-references or citations that do not add value to the work).

  o These rules cover aspects such as:

    ▪ Importance of Citation: Whether the citation substantiates a key finding or background information.

    ▪ Novelty or Redundancy: If the citation repeats prior work without new insights.

    ▪ Author Self-Referencing: Instances where the author appears to self-cite extensively without clear justification.

- Fine-Tuning the Rules Using Few-Shot Learning with Gemini:

  o Few-shot learning is employed to "teach" Google's Gemini using examples from the dataset.

  o The fine-tuning process involves feeding a curated set of citation contexts labeled manually as essential or non-essential. The model learns to generalize these examples, effectively applying the rules across new, unlabeled contexts.

Classification Pipeline for Anomaly Detection
- Input Preparation:

_____

- o Each input to the model includes the citation context, the summary of the citing paper, and the summary of the cited paper.
- Models Used:
  - o Gemini: Used to fine tune the initial set of rules based on a few anomalous self-citations found manually
  - o GPT-4-o1: Used to classify the self-citations as essential or non-essential based on the rules which were fine-tuned by Gemini

Implementation

Data sampling Steps:
In this data sampling step, we identify the k-hop neighborhood of a given node in a citation network and extract the subgraph containing nodes and their relationships about a particular domain.

Function hop_neighbour(node, k)
    1. Find all nodes within `k` hops from the given `node`:
        a. Use `single_source_shortest_path_length` to calculate the shortest path from `node` to all other nodes in the graph `G`, up to a maximum path length of `k`.
        b. Store these nodes and their distances from `node` in `k_hop_neighborhood`.
    2. Extract the list of nodes from `k_hop_neighborhood`.
    3. Return the list of nodes within `k` hops of `node`.
End Function

Function k_hop_subgraph(G, node, k)
    1. Call `hop_neighbour(node, k)` to get a list of nodes within `k` hops from `node`.
    2. Create a subgraph from `G` that includes only the nodes found in the `k`-hop neighborhood.
    3. Return the subgraph containing nodes within `k` hops of `node` and the edges between them.
End Function

Detecting Self Citation Loops:
This code identifies self-citation loops for each author in a citation network. A self-citation loop occurs when an author or group of authors cities their own work in a cycle, creating a closed path in the citation graph.
1. Calculate adjacency matrix `A` for graph `G`.
2. Compute `A_squared` by multiplying `A` with itself, representing 2-hop paths between nodes.
3. Compute `A_cube` by multiplying `A_squared` with `A`, representing 3-hop paths between nodes.
4. Initialize empty list `self_citing_datas` to store self-citation cycles for each author.
5. For each author `i` in the list of authors:
    a. Find all cycles involving node `i` using the `find_cycles_for_node` function.
    b. Initialize empty list `flag` to store self-citation cycles for author `i`.

    c. For each cycle `j` in the list of cycles:
        i. Locate the index of author `i` within cycle `j`.
        ii. Rotate cycle `j` so that it begins with author `i`, using the `rotate_list` function.
        iii. Append the rotated cycle to `flag`.
    d. Convert `flag` into a DataFrame `df1` with columns ['author', 'source', 'destination'].
    e. Append `df1` to `self_citing_datas`.
6. Concatenate all DataFrames in `self_citing_datas` into a single DataFrame
7. Reset the index of the final concatenated DataFrame and store it back in `self_citing_datas`.
8. Return or use `self_citing_datas` for further analysis.

Summarization Module:
This module summarizes the content of a research paper using a two-level approach. First, each page is summarized using a BART model. Then a higher-level summary of the document is generated by summarizing the combined text.

_____

Initialize and load models:
    1. Load BART summarization model and tokenizer from pretrained models.
    2. Configure Google Gemini API for advanced summarization.
Function summaries(text):
    1. Define prompt for Gemini model to summarize the input `text` in 500 words.
    2. Call Google Gemini model with prompt.
    3. Wait briefly (e.g., 13 seconds) to ensure compliance with API rate limits.
    4. Return the summary text from the response.
Function level1(path):
    1. Open PDF file at the specified `path`.
    2. Initialize empty string `summary` to store combined page summaries.
    3. For each page in the PDF:
        a. Extract text from the page and remove line breaks.
        b. Tokenize text for BART model, truncating to model's maximum length.
        c. Generate a summary for the page using BART, with beam search for higher-quality summaries.
        d. Decode the summary tokens and add it to the `summary` string.
4. Return the final combined `summary` text, which includes summaries from all pages
End Functions

Classification Code:
This module identifies and labels citations in a text as either "Essential" or "Non-Essential" to the passage's main contributions or findings.
Initialize API configurations:
    1. Configure Google Gemini and OpenAI API keys.

Function lol_prompts(text):
    1. If `text` is of type float (e.g., NaN), return an empty string since it is not valid input.
    2. Define the prompt for Google Gemini:
        a. Explain the classification criteria:
           i. "Essential": The citation directly influences the passage's main findings or provides significant insights.
           ii. "Non-Essential": The citation only acknowledges related work without impacting the main analysis
    3. Append the input `text` (i.e., passage containing citation) to the prompt.
    4. Call the Google Gemini model to classify the citation based on the prompt.
    5. Wait for 10 seconds to manage API rate limits.
    6. Return the classification label ("Essential" or "Non-Essential") from the model's response.
End Function

Function anomalous_check(text):
    1. Define prompt for OpenAI GPT-4:
        a. Specify the task of identifying all instances of a reference in the passage.
        b. Classify each instance as either "Essential" or "Non-Essential" based on its contribution to the passage's findings.
    2. Append the input `text` to the prompt.

_____

3. Call OpenAI's GPT-4 model to generate a response based on the prompt.

4. Extract and return the model's classification, providing a single-word response: "Essential" or "Non-Essential".

End Function

Results and Discussion

Through the application of the developed citation network analysis and machine learning techniques, we were able to uncover several insights into the prevalence and characteristics of anomalous self-citation patterns in academic publications. The results of our analysis are summarized below:

a)

1. Citation Context

- **Reference [8]** is mentioned in a passage discussing various bounds on the Price of Anarchy for the KP model and its variants. The reference is cited alongside other works that also explore these bounds.

2. Summary of Paper

- **Paper Title**: *Nash Equilibria in Discrete Routing Games with Convex Latency Functions*

- This paper introduces discrete routing games as a model that combines aspects of the KP and Wardrop models, focusing on non-cooperative routing over parallel links. It examines the Price of Anarchy (PoA), representing the efficiency loss when users act selfishly rather than collaboratively.

- **Key Contributions**:

  - **Fully Mixed Nash Equilibria**: Analyses properties of fully mixed equilibria under conditions like identical users and arbitrary links.

  - **Price of Anarchy Bounds**: Establishes upper bounds for the PoA in discrete routing games.

  - **Pure Nash Equilibria**: Explores single-link strategies and presents the ComputeNash algorithm, which finds these equilibria for arbitrary link configurations.

  - **Complexity of Equilibria**: Investigates the NP-completeness of computing best and worst pure Nash equilibria.

  - **Algorithms**: Proposes the ComputeNash algorithm, which runs efficiently for large configurations using a greedy approach.

  - **Open Problems**: Highlights potential future work, such as finding a Polynomial Time Approximation Scheme (PTAS) for optimizing pure Nash equilibria.

- The study significantly contributes to understanding the interaction between mixed and pure strategies in routing, the PoA under varying conditions, and computational challenges in identifying optimal routes.

_____

3. Results

- **Non-Essential Citation**: The citation to Reference [8] was classified as non-essential.

- **Reasoning**: The reference appears in a list of related works, providing general background on the Price of Anarchy bounds without directly contributing to the passage's primary analysis or critical insights.

4. Explanation

- **Role of Citation**: Reference [8] serves to acknowledge similar studies on PoA bounds in the KP model context without elaborating on specific methods or findings from the cited work. Its role is more supportive than integral, as it does not directly impact the passage's arguments or conclusions.

b)
1. Citation Context

- **Reference [9]** appears in a passage discussing various conjectures and studies related to the Fully Mixed Nash Equilibrium Conjecture within discrete routing games. It is mentioned alongside other references ([4], [8], [17], [19]) as part of a broader discussion on the topic.

2. Summary of Paper

- **Paper Title**: *Nash Equilibria in Discrete Routing Games with Convex Latency Functions*

- This paper presents a discrete routing model combining aspects of the KP and Wardrop models, focusing on non-cooperative routing among parallel links. Key contributions include analysis of the Price of Anarchy (PoA), representing efficiency loss in scenarios where users act independently rather than collaboratively.

- **Key Contributions**:

    o **Fully Mixed Nash Equilibria**: Analyses equilibria where users employ mixed strategies under various conditions, including identical users and arbitrary links.

    o **Price of Anarchy Bounds**: Establishes bounds for PoA in discrete routing games, providing insights into the model's performance under competitive conditions.

    o **Pure Nash Equilibria**: Examines equilibria where users choose single links, introducing an efficient algorithm (ComputeNash) for computing these equilibria.

    o **Best and Worst Pure Nash Equilibria**: Investigates computational challenges in finding optimal equilibria, proving NP-completeness for identifying the best and worst pure Nash equilibria.

    o **Open Problems**: Suggests areas for future research, such as developing a Polynomial Time Approximation Scheme (PTAS) for optimizing pure Nash equilibria.

- This paper enhances understanding of routing efficiency in decentralized systems by examining latency's role in PoA and computational complexities in finding optimal routes.

3. Results

- **Non-Essential Citation**: The citation to Reference [9] is classified as non-essential.

_____

- **Reasoning**: The reference is part of a list of related studies, serving to acknowledge broader research on conjectures related to Fully Mixed Nash Equilibria without contributing directly to the passage's primary analysis.

4. Explanation
- **Role of Citation**: Reference [9] serves an acknowledgment role, recognizing similar studies without directly impacting the main findings. It provides context for the Fully Mixed Nash Equilibrium Conjecture discussion but is not central to the core arguments or conclusions of the passage.

c)

1. **Citation Context**
   Reference [26] appears in a passage discussing the behaviour of algorithms (LocalHeur and DeltaHeur) in a Locally Minimum Cost Forwarding (LMCF) game. The passage specifically examines algorithmic performance in terms of stretch, cost deviation, and node choice, with Reference [26] used to explain the behaviour of the Shortest Path Tree (SPT) at higher powers.

2. **Summary of Paper**
   Paper Title: On a Locally Minimum Cost Forwarding Game
   This paper examines the Locally Minimum Cost Forwarding Game (LMCF) in wireless networks, where nodes selfishly select paths to minimize individual costs. The focus is on determining optimal Nash equilibria under various conditions. Key contributions include:
   - **Existence of Nash Equilibria**: Demonstrates that Nash equilibria exist in LMCF, with minimum spanning trees as a notable example.

   - **Price of Anarchy (PoA) and Stability**: Analyses efficiency losses from selfish behaviour in LMCF, showing a linear PoA under certain cost objectives.

   - **NP-Hardness**: Establishes the NP-hard nature of finding optimal Nash equilibria and introduces the Delta Heuristic to address this.

   - **Simulation Results**: Demonstrates that DeltaHeur generally outperforms the Minimum Spanning Tree (MST), making it useful for controlling network topology in wireless settings.

   The paper emphasizes understanding node behaviour in LMCF as critical for designing efficient wireless networks and positions DeltaHeur as a practical tool for achieving this.

3. **Results**
   - **Essential Citation**: The citation to Reference [26] is classified as essential.
   - **Reasoning**: Reference [26] is integral to explaining a core theoretical concept that supports the analysis of LocalHeur's behaviour in the passage. It provides theoretical justification for SPT's performance in the LMCF game, tying directly to the authors' main analysis on algorithmic behaviour under powered conditions.
   - **Explanation**
     Role of Citation: Reference [26] explains the observed behaviour of LocalHeur compared to DeltaHeur within the LMCF game framework. Its inclusion is essential for understanding LocalHeur's effectiveness, which is central to the primary findings.

_____

Conclusion

In this study, we proposed a method for classing citations into Essential and Non-Essential ones, utilizing advanced language models to evaluate the role of the citation in relation to the principal findings or contributions of the text. By interpreting the intent behind each reference, our method uses the semantics of the passage to identify citations necessary to validate the main argument or method of the paper, rather than those containing no key propositions. As a result, our work can distinguish Essential citations from those that appear as credits to the other paper(s). We demonstrated how our developed model can recognize the importance of citations in academic content. Its application could inform citation practices, and at least identify patterns that could indicate repeated and unreferenced self-citations as a candidate for dubious scholarly integrity. Overall, this classification framework offers a structured approach for analysing citation relevance and can contribute to improved understanding of citation behaviours within academic networks.

Future work

While this method effectively distinguishes essential citations from non-essential ones, several improvements can further refine the approach. One limitation is that the model currently does not account for redundant citations across publication timelines. Ideally, a paper should cite the most recent research relevant to its field. For example, if a paper published in 2015 covers similar content as a paper from 2005, which in turn builds upon a 1999 study, then only the 2015 paper should be cited. Future iterations could incorporate this temporal filtering to avoid redundant or outdated citations. Additionally, the current model does not handle citations that follow certain formats or regex patterns, such as those using "[Suri et al.]" citations. This format is common in specific academic styles, and handling these cases with customized regex-based recognition could improve classification accuracy. Future work could include adapting the method to process such citations effectively, enabling more nuanced and comprehensive citation analyses across diverse academic formats.