Understanding Hallucinations in LLMs:
A Graph-Based Reproduction Study

By
YIP Sau Lai
LAM Sum Ying
PENG Muzi

Overview

Paper

Experiments

Motivation (premise)

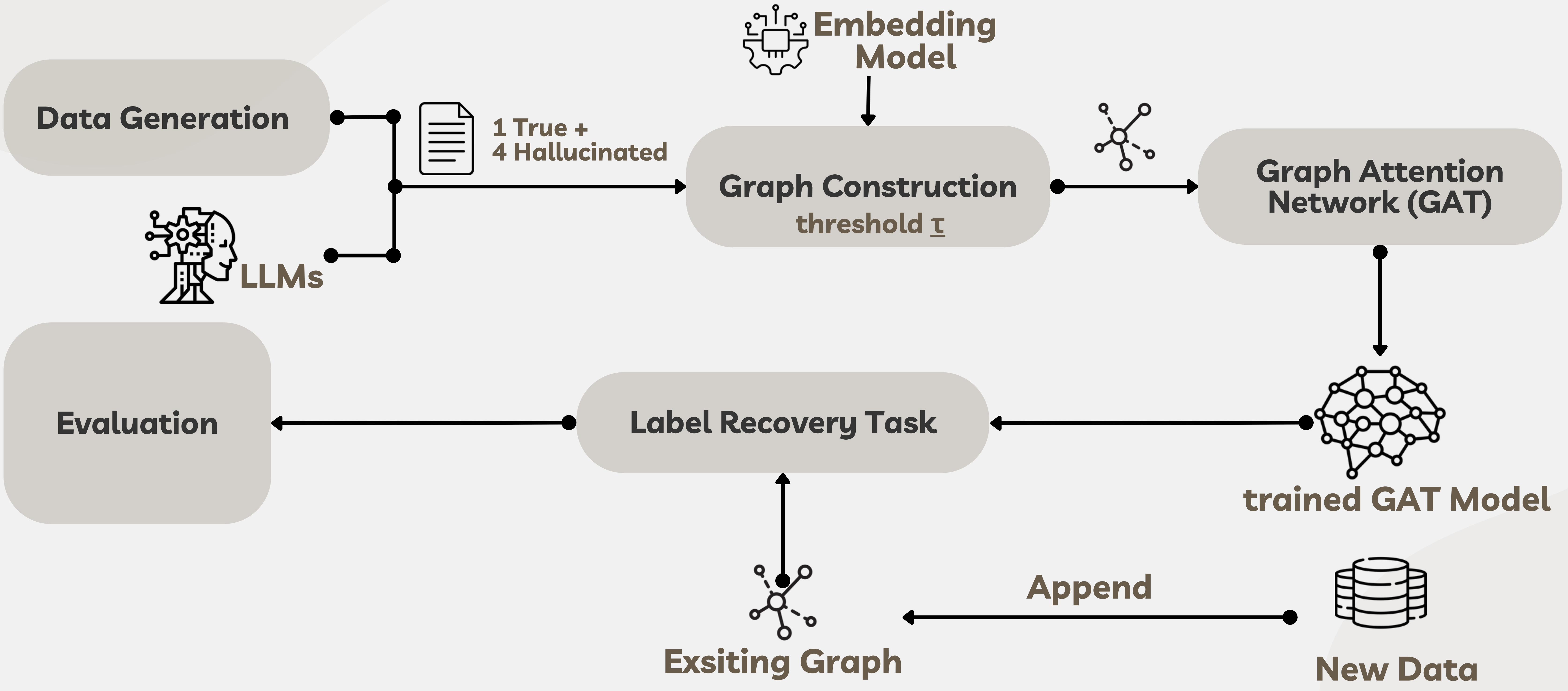
- LLM hallucinations are structured: Hallucinations share characteristics in the latent space.

- Principle of homophily: samples that share text-level characteristics tend to lie closer in the embedding space.

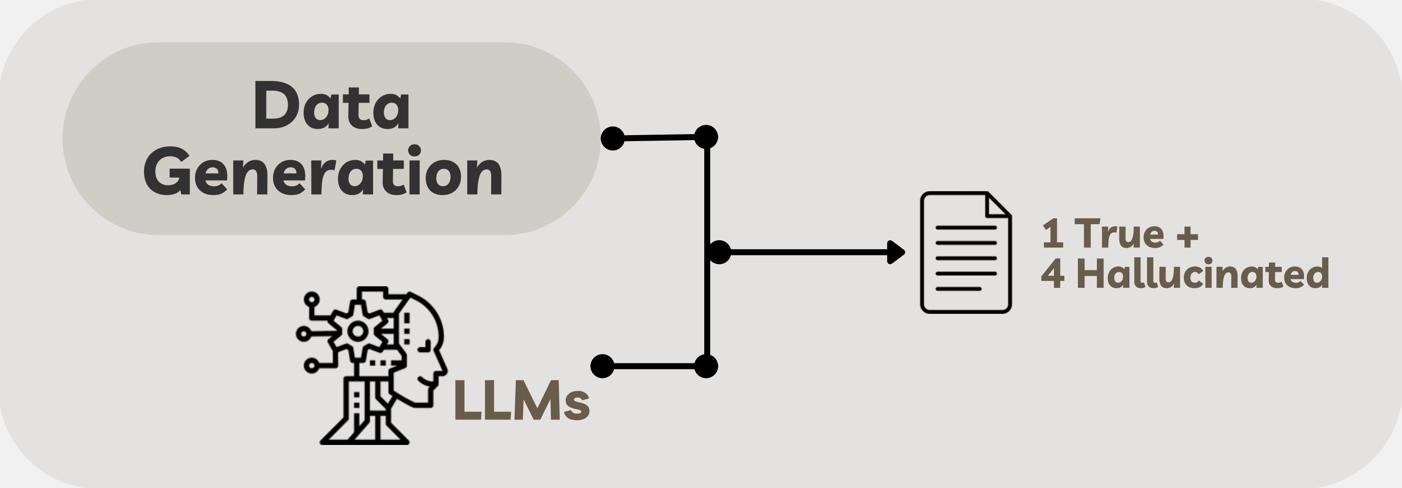
Objectives

- Do LLM-generated hallucinations share characteristics?
- Can we leverage graph structures to identify and learn these characteristics?
- If learned, can we use this knowledge to identify hallucinations among new incoming LLM generations through label recovery?

Methedology



Data Generation



Dataset Composition

for each query, generate:

- 1 true
- 1 true without context.
- 1 true with context.
- 8 hallucinated (misleading)

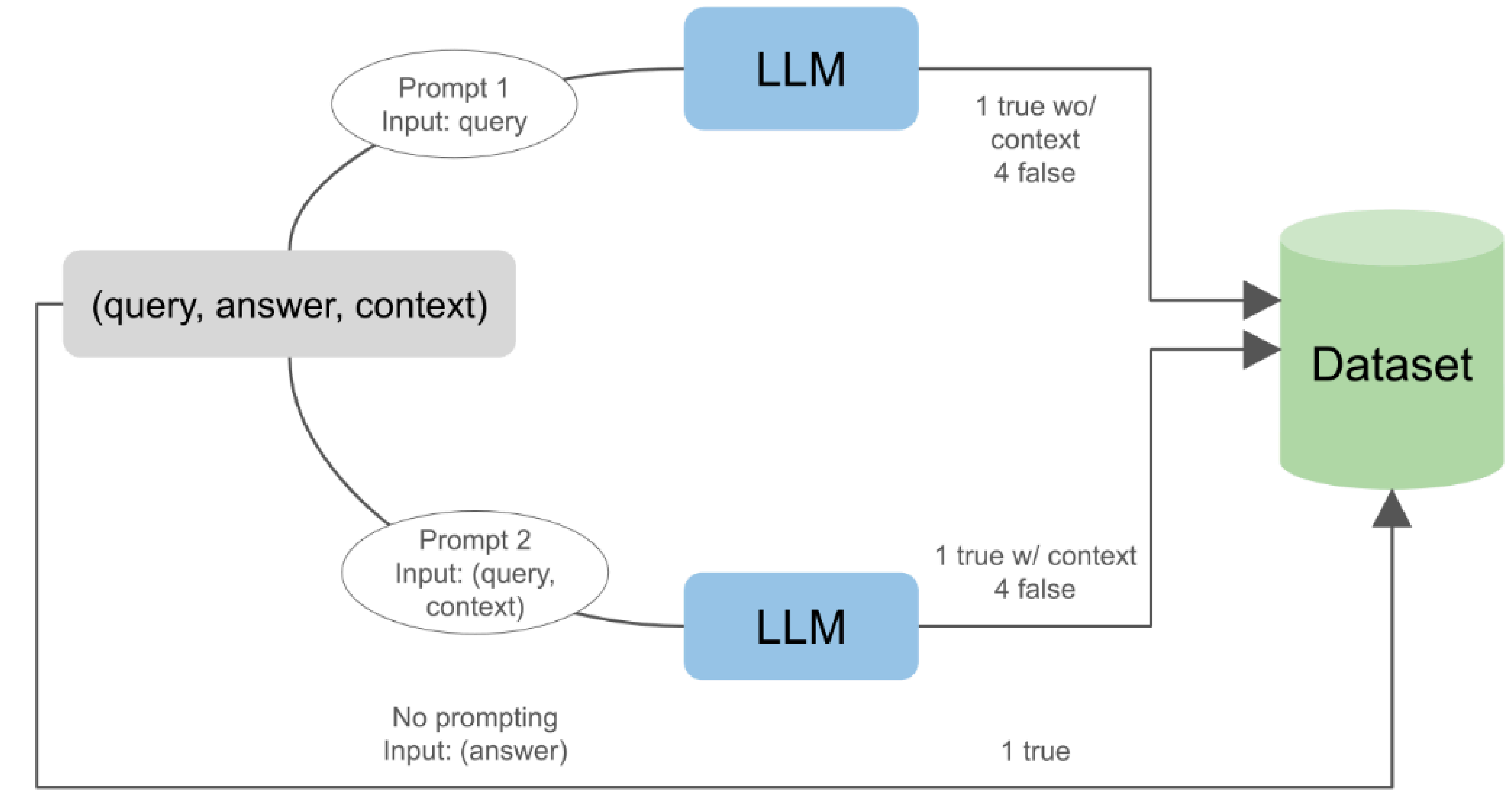
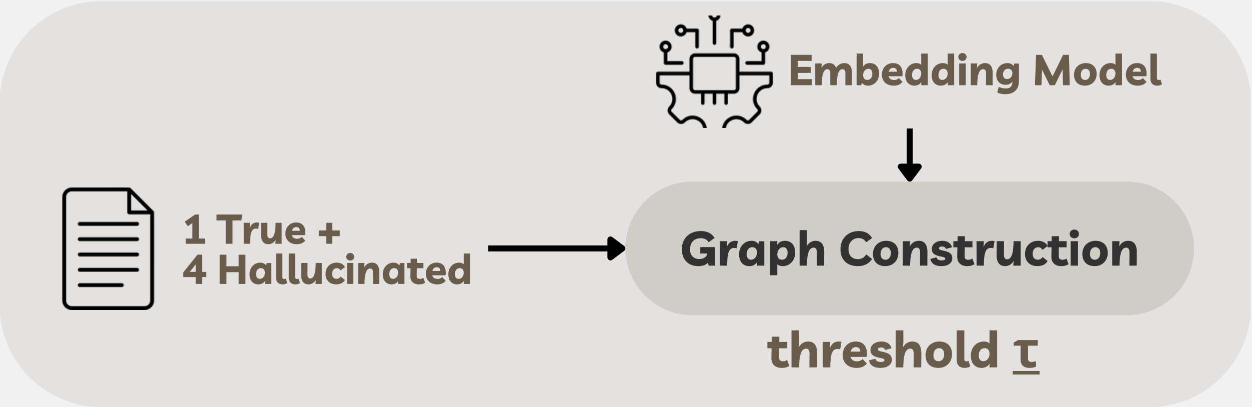
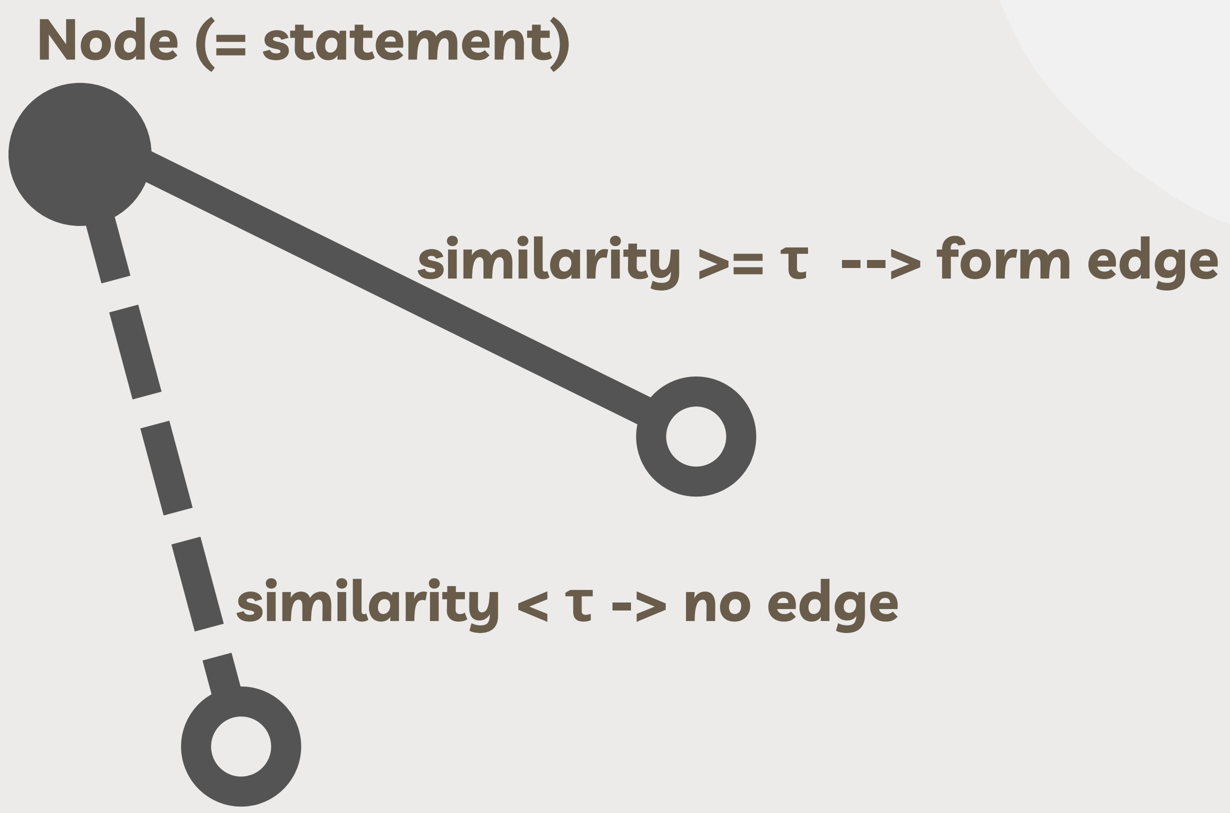


Figure 1: Data generation process.

Graph Construction



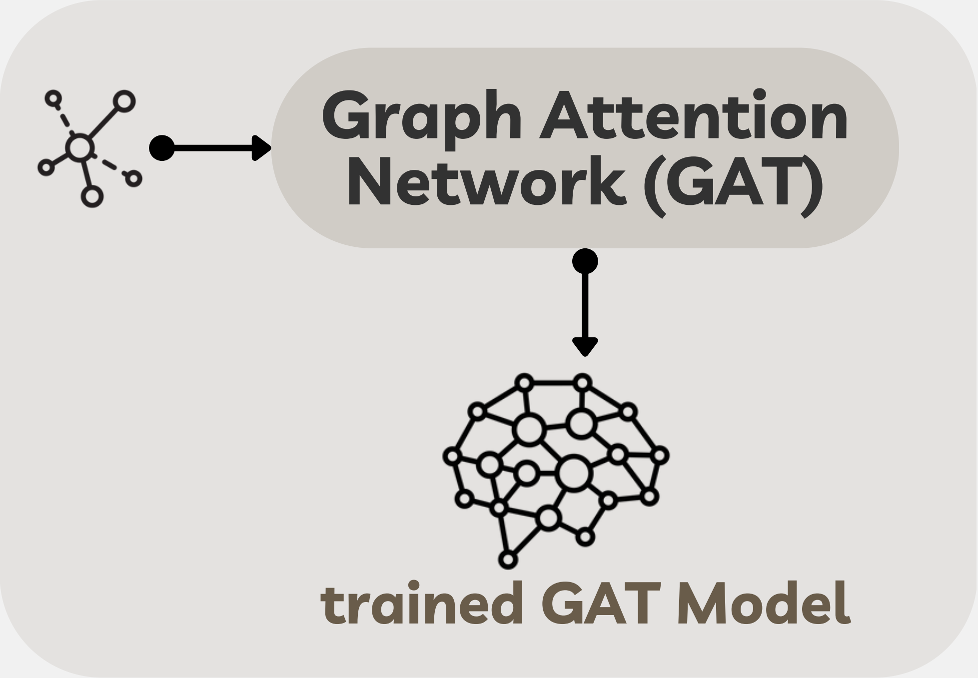
- **Objective**: Create a graph structure to represent relationships between sentences.
- **Node Representation:**
  - Each node corresponds to a sentence from the dataset.
  - Sentence embeddings are generated using **BERT**.
- **Edge Formation:**
  - Use **cosine similarity**; establish an edge if it exceeds a **threshold (τ).**

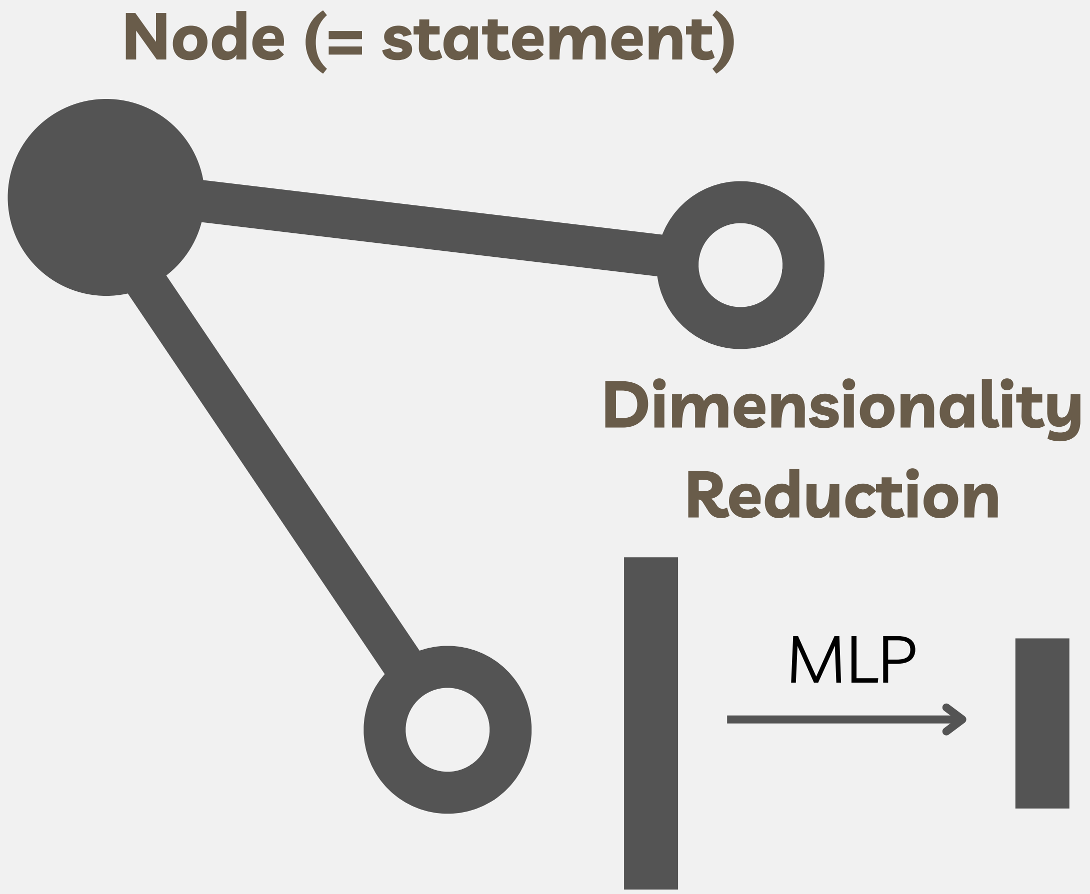


Graph Attention Network (GAT)

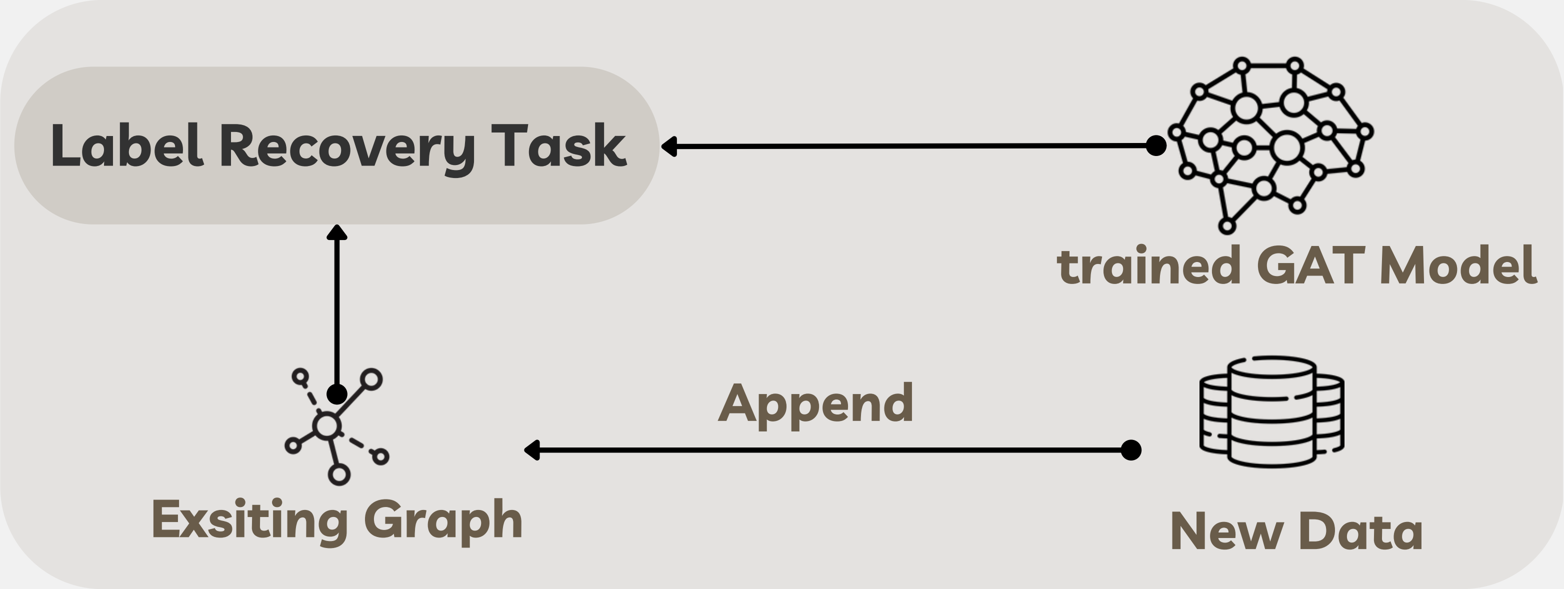
- **Attention Mechanism & Message Passing**
  - Attention scores = AGG(neighbors)
  - Importance ~ similarity

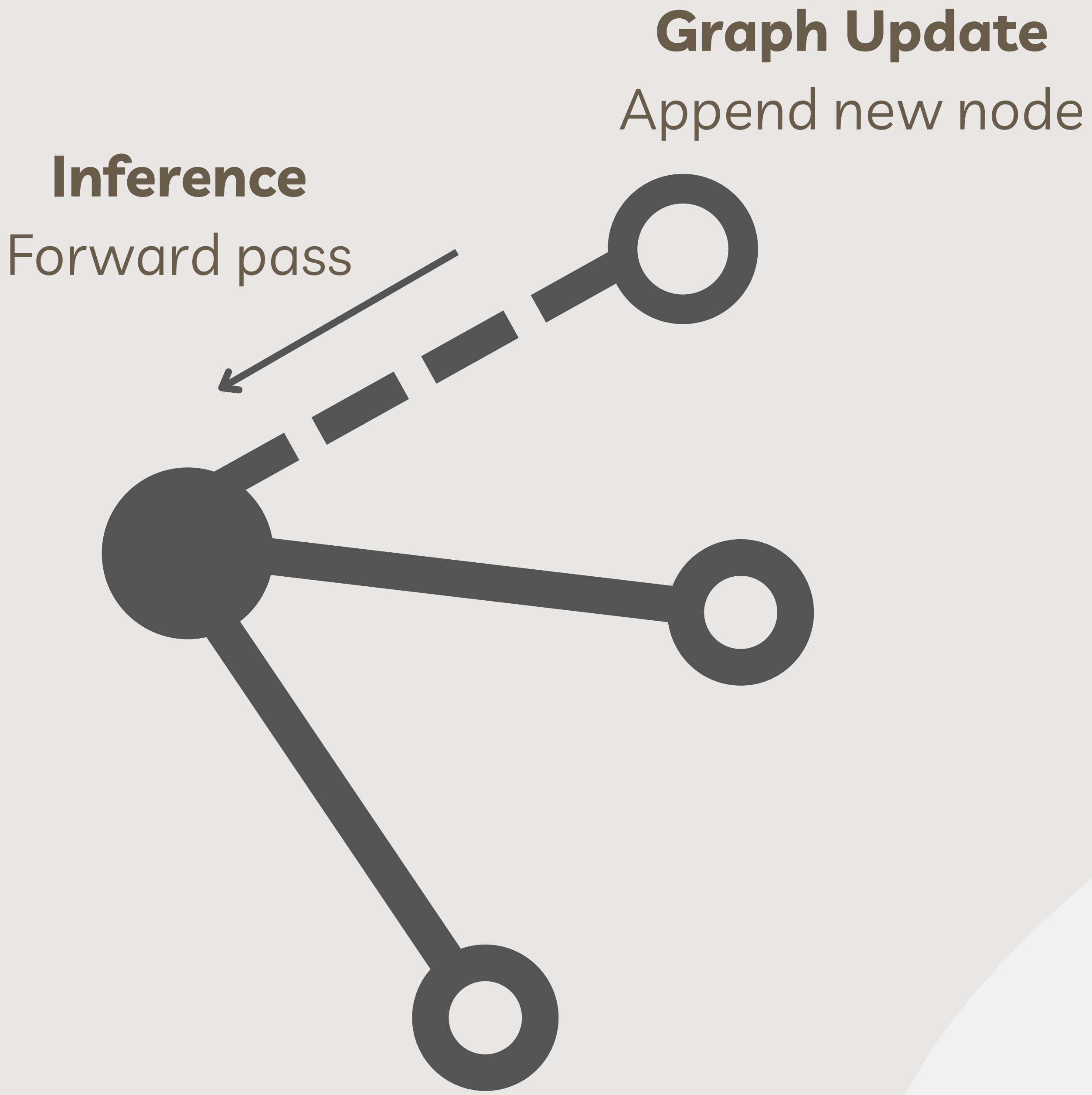
- **Classification Task**
  - Model as an ordinal regression task
  - Output labels ~ degree of hallucination





Label Recovery Task





Experiment Setting

Dataset

MSMARCO-QA

Query
Answer
Context

Llama2

Meta's instruction-tuned

Sentence-level Split

70% - 15% - 15%

Graph

Embedding Model

BERT
(English uncased)

Threshold

0.85

Loss Function

Binary Cross Entropy
(BCE) loss

Metrics

Macro-recall

Accuracy in identifying
individual classes

Macro-precision

Prediction accuracy
per class

AUC-PR

Binary classification
performance

Baselines

DeBERTa-QA

DeBERTa
+
process query-answer pair in unified encoder

MLP-QA

three-layer MLP & ReLU
+
BERT embedding
+
concatenated query-answer

MLP-A

two-layer MLP & ReLU
+
BERT embedding
+
process answer

1. https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli
2. https://huggingface.co/bert-base-uncased

Ablation Study

Split	Model	Recall	Precision	AUC-PR
Train	GAT	**0.5069**	**0.5844**	**0.4153**
	DeBERTa-QA	0.3882	0.5404	0.3517
	MLP-QA	0.3214	0.3880	0.2718
Val	GAT	**0.4972**	**0.5717**	**0.4096**
	DeBERTa-QA	0.3206	0.5059	0.3357
	MLP-QA	0.3150	0.3622	0.2953

Contrastive Learning (CL)

Train	GAT	0.5069	0.5844	0.4153
	CL + GAT	**0.8244**	**0.8281**	**0.7118**
	MLP-A	0.2512	0.3123	0.2014
	CL + MLP-A	0.4286	0.5892	0.3987
Val	GAT	0.4972	**0.5717**	0.4096
	CL + GAT	**0.5305**	0.5438	**0.4212**
	MLP-A	0.2256	0.3110	0.2057
	CL + MLP-A	0.3589	0.4956	0.3278

Conclusion

Q1: *Do LLM-generated hallucinations share characteristics?*

Yes! GAT identifies an underlying structure of the embedding space

GAT vs kNN

Split	Model	Recall	Precision	AUC-PR
Train	GAT	0.5069	0.5844	0.4153
	CL + GAT	**0.8244**	**0.8281**	**0.7118**
	MLP-A	0.2512	0.3123	0.2014
	CL + MLP-A	0.4286	0.5892	0.3987
Val	GAT	0.4972	**0.5717**	0.4096
	CL + GAT	**0.5305**	0.5438	**0.4212**
	MLP-A	0.2256	0.3110	0.2057
	CL + MLP-A	0.3589	0.4956	0.3278
	kNN	0.2434	0.1895	0.2494

Result

Conclusion

Q2: *Can we leverage graph structures to identify and learn these characteristics?*

Yes!

GAT >> kNN

Result

Test Set Performance

	Recall	Precision	AUC-PR
CL + GAT	**0.5142**	0.5430	**0.4057**
GAT	0.4830	**0.5603**	0.3887
CL + MLP-A	0.3727	0.5122	0.3419

Conclusion

Q3: *If learned, can we use this knowledge to identify hallucinations among new incoming LLM genera tions through label recovery?*

Yes!

Comparaitive performance in test set.

Benchmarks

FEVER

Method	Recall	Precision	Label Accuracy
CL + GAT	0.7079	**0.4712**	0.6471
UNC-NLP	**0.7091**	0.4227	**0.6821**

UNC-NLP: best

CL + GAT: comparative

SelfCheckGPT

Method	Sentence-level (AUC-PR)	
	NonFactual	Factual
Random	0.7296	0.2704
LLM + BERT Scores	**0.8196**	**0.4423**
CL + GAT	0.7799	0.4002

LLM + BERT Score： LLM-generated statements

CL + GAT: suffer from small size

Conclusion

Comparative performance without accessing search-base methods

Our Experiments

Setting

Threshold

&

Embedding Model

Data

LLM

&

Dataset

Experiment 1

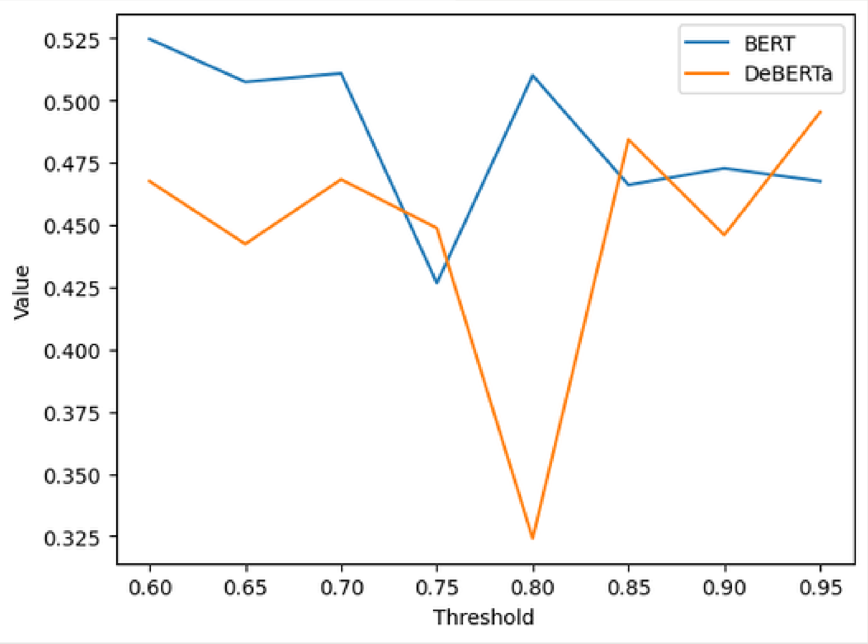
Threshold & Embedding Model

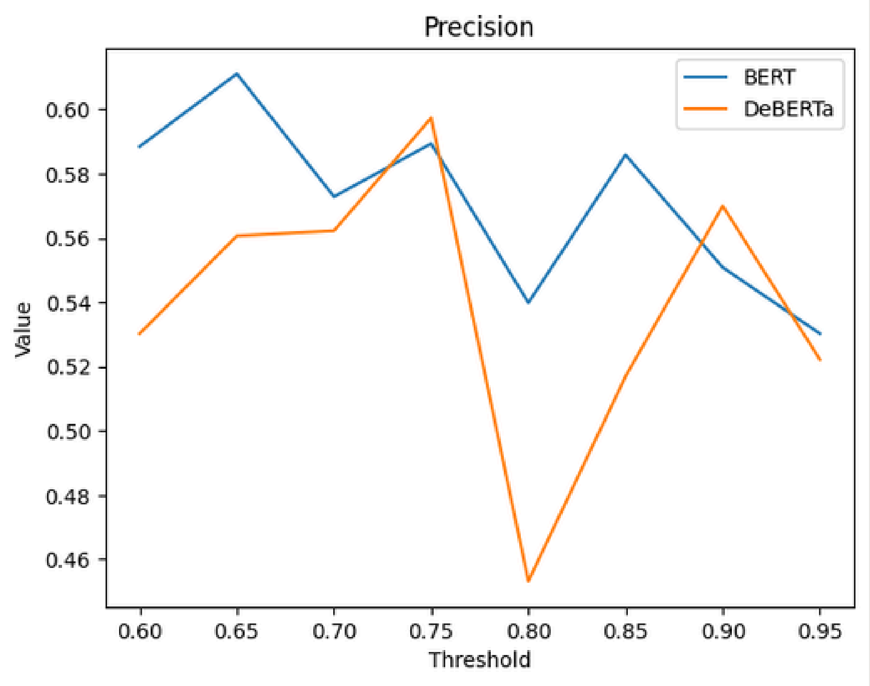
Embedding Model	Threshold	Recall	Precision	AUC-PR
BERT (CL+ GAT) (60% data)	0.60	**0.5246**	0.5884	**0.4334**
	0.65	0.5074	**0.6110**	0.4305
	0.70	0.5108	0.5728	0.4246
	0.75	0.4266	0.5893	0.3863
	0.80	0.5100	0.5398	0.4118
	0.85	0.4660	0.5858	0.4025
	0.90	0.4727	0.5508	0.3968
	0.95	0.4675	0.5302	0.3894
DeBERTa (CL+ GAT) (60% data)	0.60	0.4675	0.5302	**0.3894**
	0.65	0.4423	0.5606	0.3555
	0.70	0.4682	0.5622	0.3679
	0.75	0.4487	**0.5973**	0.3633
	0.80	0.3241	0.4532	0.2888
	0.85	0.4842	0.5169	0.3736
	0.90	0.4459	0.5699	0.3710
	0.95	**0.4952**	0.5221	0.3732

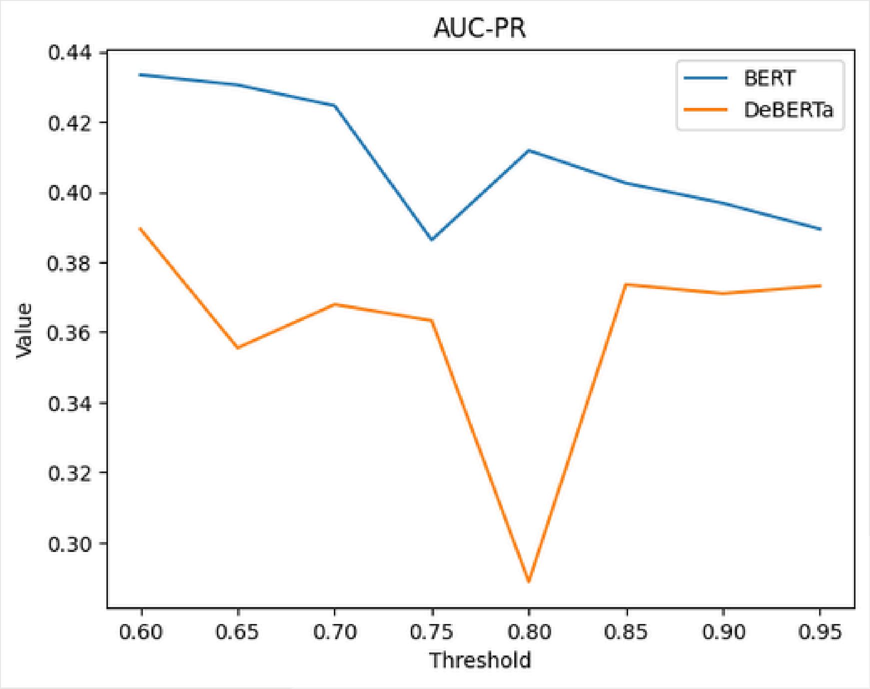
Optimal threshold depend on data

"As MLP-A is trained solely on answers, and DeBERTa uses different embeddings, potentially leading to a distribution shift"

Yes! Performance of DeBERTa is worse and less stable







Experiment 2

New LLM - generated by QWen2.5-14B

	Recall	Precision	AUC-PR
Qwen2.5-14B	0.5434	0.6240	0.4210
Llama2-13B	0.4660	0.5858	0.4025

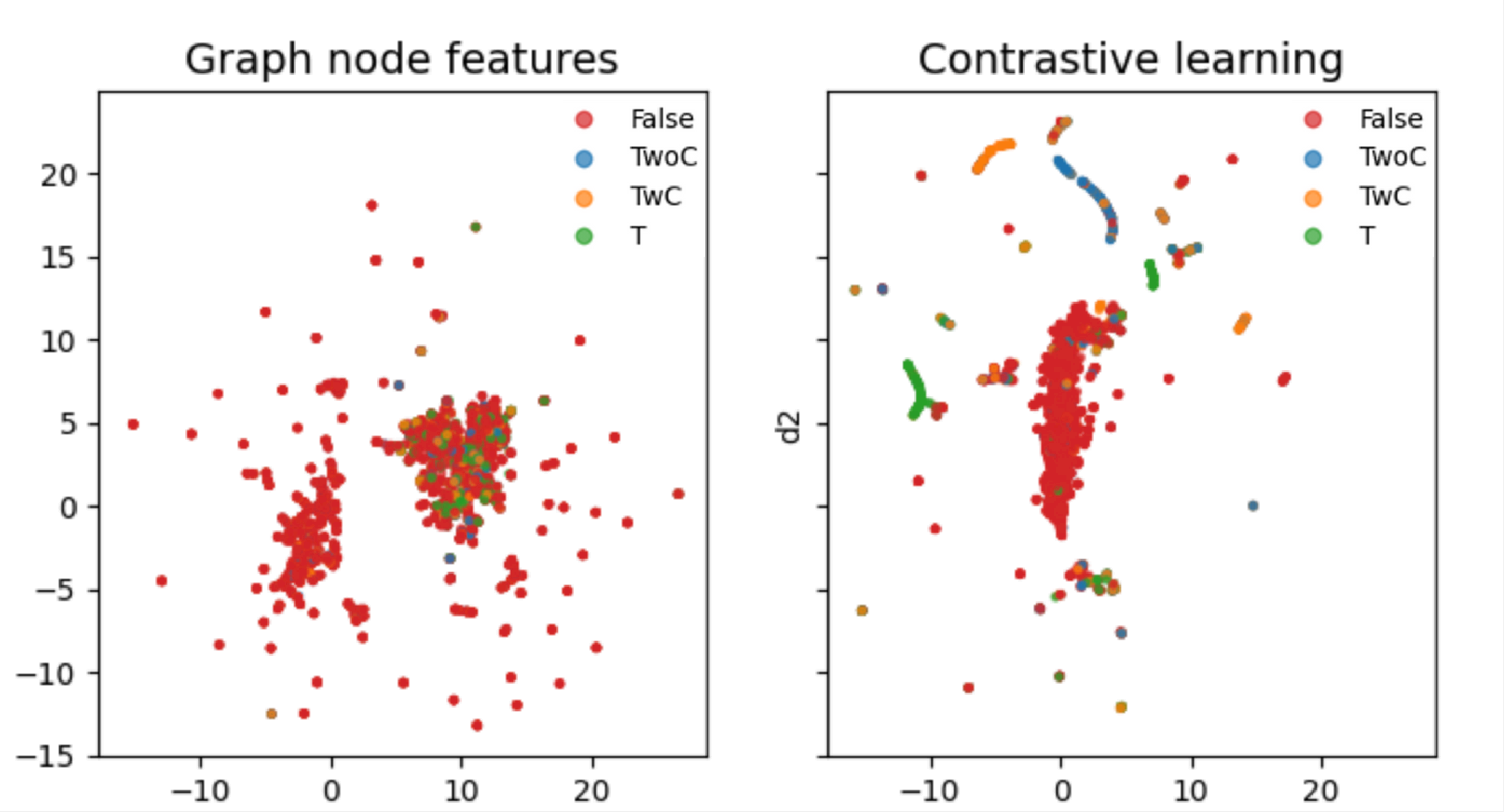
- **Objective:** Conduct experiments to enhance data generation by using different Large Language Models (LLMs).
- **Model:** Qwen2.5-14B
  - A state-of-the-art LLM designed for versatile text generation tasks, known for its ability to produce coherent and contextually relevant responses.
- **Results**: Qwen2.5-14B outperforms MSMARCO-QA, showing promise in data generation with opportunities for further optimization.

Experiment 3

New Dataset

SciQ dataset: It contains 13,679 crowdsourced science exam questions about Physics, Chemistry and Biology. An additional paragraph with supporting evidence for the correct answer is provided.

	Recall	Precision	AUC-PR
Sci_Q	0.3353	0.4454	0.3030
MSMARCO-QA	0.4660	0.5858	0.4025



Conclusion

- LLM-generated hallucinations share characteristics.
- GAT is potential in LLM hallucination detection.

Thank you!