Can Graph Descriptive Order Affect Solving Graph Problems with LLMs?

Yuyao Ge[1,5], Shenghua Liu[1,5*], Baolong Bi[1,5], Yiwei Wang[2], Lingrui Mei[1,5], Wenjie Feng[3], Lizhe Chen[4], Xueqi Cheng[1,5]

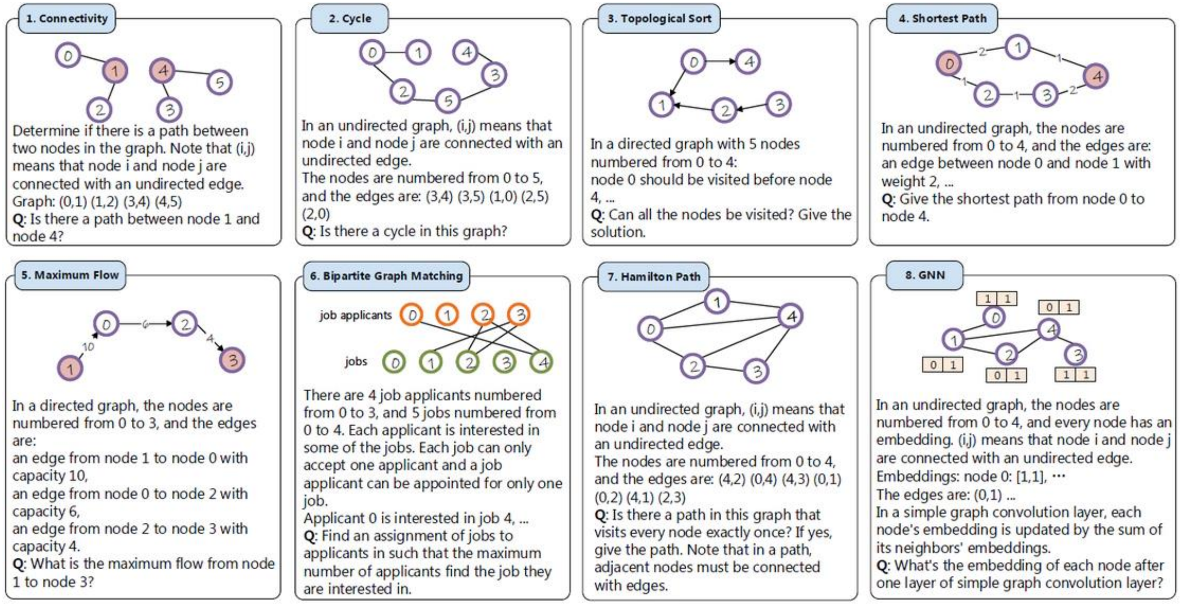
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of California, Merced
[3]University of Science and Technology of China [4]Tsinghua University
[5]University of Chinese Academy of Sciences

ACL 2025 VIENNA

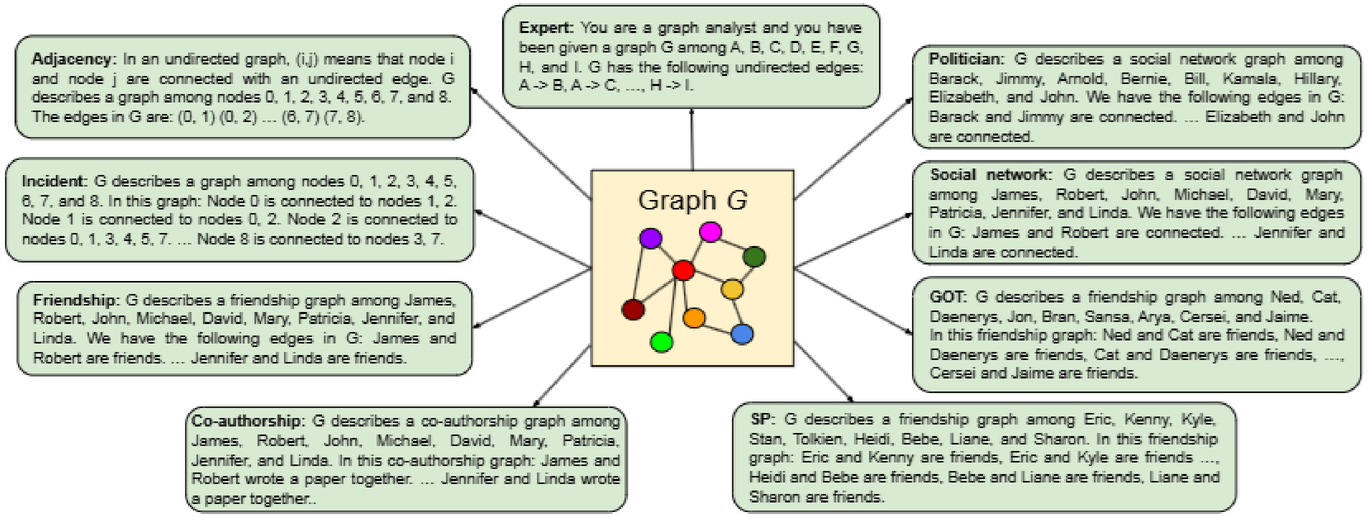
* Corresponding author.

BACKGROUND

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES



Can language models solve graph problems in natural language?, 2023, NIPS Spotlight



Talk like a Graph: Encoding Graphs for Large Language Models, 2024, ACL Findings

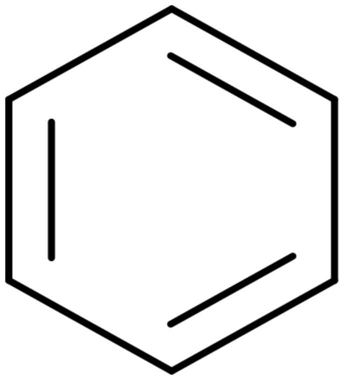
- LLMs Have (Preliminary) Graph Reasoning Abilities
- Appropriate Prompt can help LLM Solve Graph Problems
- Graph Encoding Functions Have Significant Impact on LLM Reasoning
- LLMs Lack a Global Veiw of a Graph

Could the order of graph descriptions be a critical, yet overlooked factor?

MOTIVATION

Why Graph Descriptive Order is So Important?

A Simple Example: For a standard benzene ring, which of the following two descriptions is easier for humans to understand?



Description One

C1-C2, C2-C3, C3-C4, C4-C5, C5-C6, C6-C1

Humans can immediately recognize that this is a benzene ring!

Description Two

C3-C4, C1-C2, C5-C6, C2-C3, C4-C5, C6-C1

Humans need to reorganize the information to understand its structure.

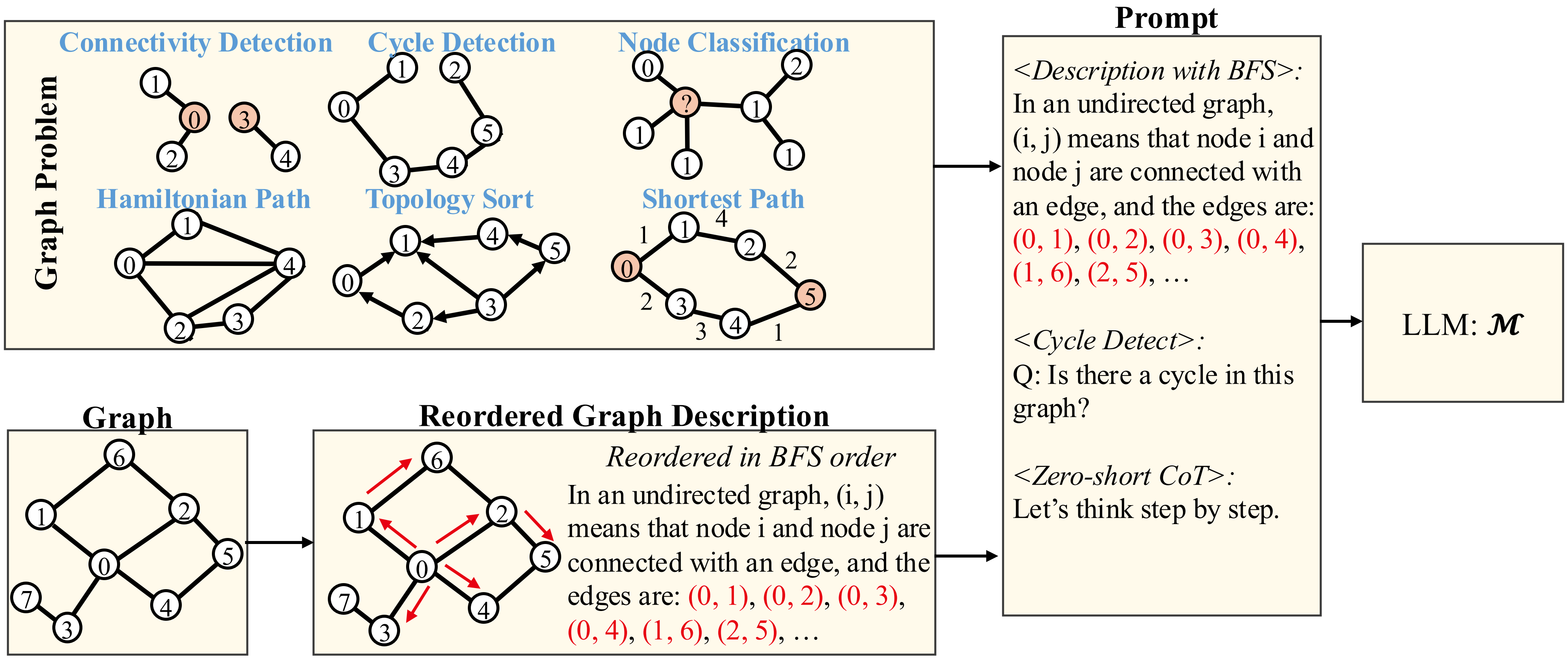
The order of graph description has a significant impact on human understanding of its structure. Does this effect also exist in LLMs?

QUESTIONS

- Does graph description order **affect LLM performance** in solving graph problems?
- Is LLM **robustness to graph description order** consistent across different tasks?
- Are specific graph description orders **better suited for certain graph tasks?**

DESIGN

We designed six types of graph tasks to assess how four graph traversal orders (DFS, BFS, PR, PPR) affect the reasoning performance of six mainstream LLMs.



We organize the descriptions into GraphDO (Graph Description with Order) dataset.

YuyaoGe, Institute of Computing Technology, Chinese Academy of Sciences (ICT)

GraphDO DataSet

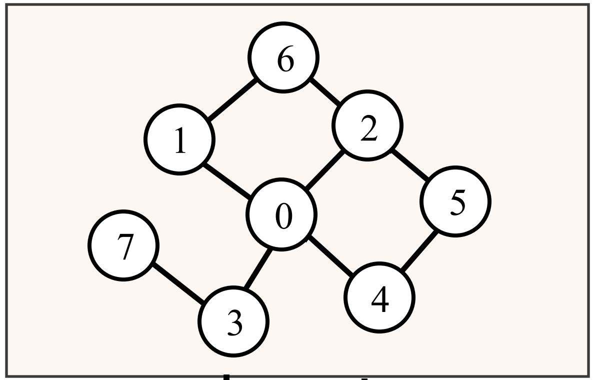
Overview

8,500
Carefully filtered graph cases

6 Types
Topology and graph learning Tasks

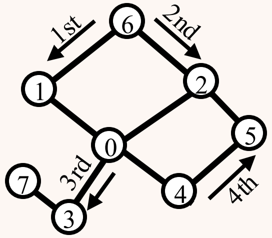
5 Types
Prompting Methods
Ranging from Zero-shot to CoT

Example



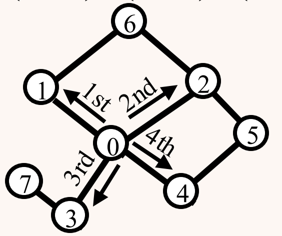
Random Order

In an undirected graph, (i, j) means that node i and node j are connected with an edge, and the edges are:
(6, 1), (6, 2), (0, 3), (4, 5), …



BFS Order

In an undirected graph , (i, j) means that node i and node j are connected with an edge, and the edges are:
(0, 1), (0, 2), (0, 3), (0, 4), …



YuyaoGe, Institute of Computing Technology, Chinese Academy of Sciences (ICT)

Finding 1: Order Significantly Affect Performance

Task	Order	Zero-shot	Zero-shot CoT	Few-shot	CoT	CoT-BAG
Conn.	Random	73.93 (-)	70.71 (-)	81.07 (-)	83.93 (-)	82.14 (-)
	BFS	82.14 (↑11.11)	87.50 (↑23.74)	89.29 (↑10.14)	92.50 (↑10.21)	95.71 (↑16.52)
	DFS	79.29 (↑7.25)	82.14 (↑16.16)	87.14 (↑7.49)	88.21 (↑5.10)	89.29 (↑8.70)
	PR	77.86 (↑5.32)	83.57 (↑18.19)	85.71 (↑5.72)	84.29 (↑0.43)	87.50 (↑6.53)
	PPR	76.79 (↑3.87)	81.07 (↑14.65)	83.93 (↑3.53)	84.64 (↑0.85)	86.07 (↑4.78)
Cycle	Random	51.79 (-)	53.57 (-)	65.36 (-)	75.71 (-)	76.07 (-)
	BFS	55.71 (↑7.57)	56.07 (↑4.67)	79.29 (↑21.31)	86.07 (↑13.68)	86.43 (↑13.62)
	DFS	52.14 (↑0.68)	53.93 (↑0.67)	73.21 (↑12.01)	79.29 (↑4.73)	81.07 (↑6.57)
	PR	55.36 (↑6.89)	56.43 (↑5.33)	70.36 (↑7.65)	80.36 (↑6.14)	83.21 (↑9.39)
	PPR	54.29 (↑4.83)	55.00 (↑2.67)	70.00 (↑7.10)	79.29 (↑4.73)	80.00 (↑5.17)
HamPath	Random	10.71 (-)	15.36 (-)	40.00 (-)	46.07 (-)	45.36 (-)
	BFS	20.00 (↑86.74)	20.71 (↑34.83)	57.86 (↑44.65)	58.57 (↑27.13)	57.14 (↑25.97)
	DFS	33.93 (↑216.81)	37.50 (↑144.14)	67.50 (↑68.75)	63.93 (↑38.77)	59.29 (↑30.71)
	PR	15.00 (↑40.06)	19.29 (↑25.59)	48.93 (↑22.32)	55.00 (↑19.38)	50.00 (↑10.23)
	PPR	16.43 (↑53.41)	18.93 (↑23.24)	50.00 (↑25.00)	53.93 (↑17.06)	50.36 (↑11.02)
TopoSort	Random	28.93 (-)	31.07 (-)	58.21 (-)	56.07 (-)	60.36 (-)
	BFS	43.21 (↑49.36)	40.36 (↑29.90)	67.14 (↑15.34)	61.43 (↑9.56)	65.00 (↑7.69)
	DFS	42.14 (↑45.66)	48.93 (↑57.48)	77.86 (↑33.76)	74.29 (↑32.50)	72.86 (↑20.71)
	PR	35.36 (↑22.23)	35.71 (↑14.93)	71.07 (↑22.09)	58.21 (↑3.82)	65.36 (↑8.28)
	PPR	37.14 (↑28.38)	39.64 (↑27.58)	72.50 (↑24.55)	58.93 (↑5.10)	66.43 (↑10.06)
SPath	Random	20.00 (-)	25.00 (-)	26.07 (-)	38.93 (-)	40.71 (-)
	BFS	35.36 (↑76.80)	42.50 (↑70.00)	45.36 (↑73.99)	67.50 (↑73.39)	65.71 (↑61.41)
	DFS	32.14 (↑60.70)	34.29 (↑37.16)	45.00 (↑72.61)	58.57 (↑50.45)	57.14 (↑40.36)
	PR	30.36 (↑51.80)	43.93 (↑75.72)	38.93 (↑49.33)	43.93 (↑12.84)	48.93 (↑20.19)
	PPR	32.50 (↑62.50)	44.64 (↑78.56)	42.14 (↑61.64)	45.36 (↑16.52)	49.64 (↑21.94)

Sampling	Order	CORA		Pubmed	
		Acc.	Δ	Acc.	Δ
Ego	Random	70.00	-	72.00	-
	BFS	72.00	↑ 2.86	74.00	↑ 2.78
	DFS	71.33	↑ 1.90	77.33	↑ 7.40
	PR	75.33	↑ 7.61	82.67	↑ 14.82
	PPR	73.33	↑ 4.76	77.33	↑ 7.40
Forest Fire	Random	79.33	-	69.99	-
	BFS	82.67	↑ 4.21	74.00	↑ 5.73
	DFS	81.33	↑ 2.52	76.00	↑ 8.59
	PR	83.33	↑ 5.04	76.00	↑ 8.59
	PPR	82.00	↑ 3.36	74.67	↑ 6.69

- **On traditional graph tasks, ordered descriptions result in an improvement of 12% to 70%, while on the node classification task, the improvement ranges from 1.9% to 14.82%**
- **The benefits remain consistent across various prompting strategies.**
- **We hypothesize that LLMs' improved performance with ordered descriptions is due to attention bias.**

Finding 2: Complexity Affects Order Robustness

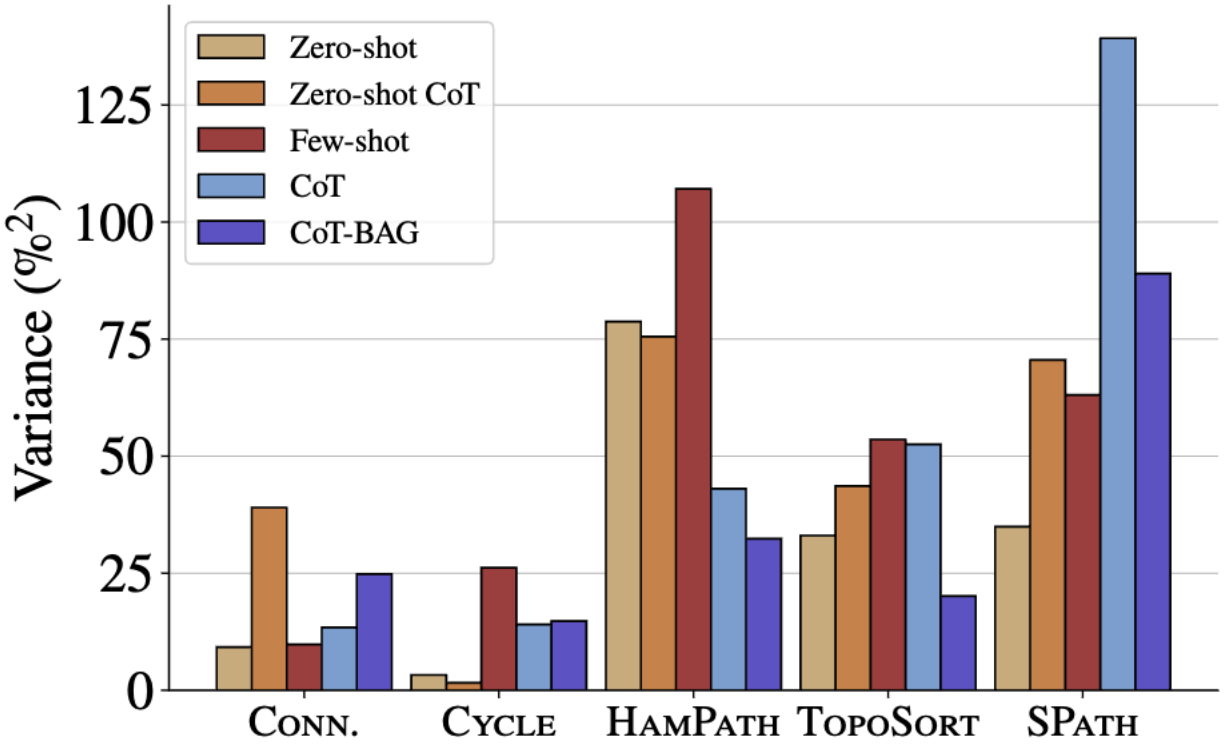
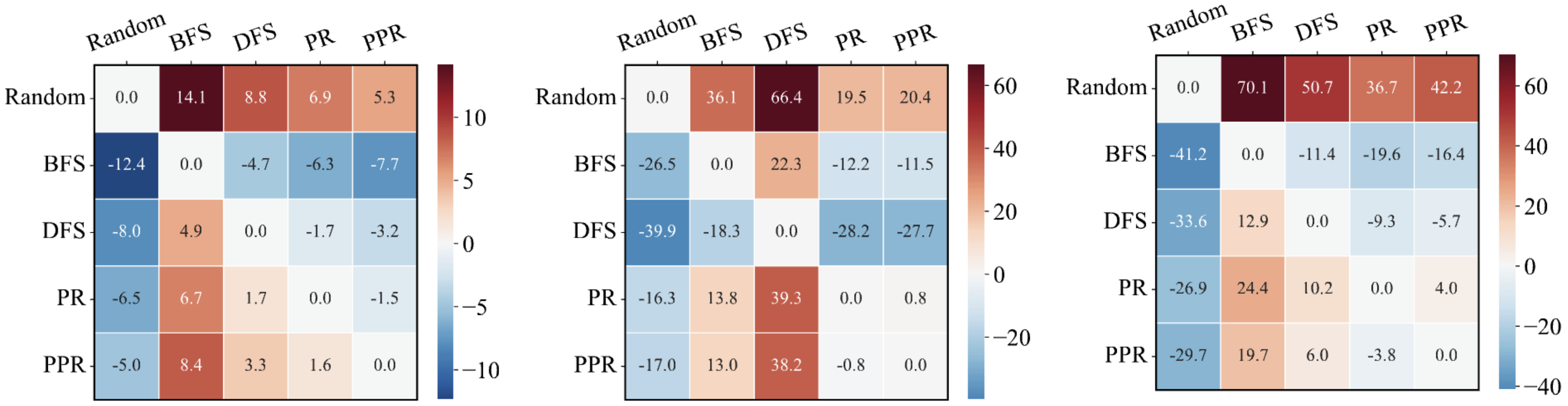


Figure 4: Variance of LLM accuracy across different graph tasks with varying description orders. The variance for each task is computed as $\sigma^2 = \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}}(\mathcal{S}_o - \mu)^2$, where $\mathcal{S}_o$ is the accuracy for order o, μ is the mean accuracy across all orders.

- **Simple tasks (connectivity, cycle) show low variance across orders - inherently robust**
- **Complex tasks (Hamilton path, Shortest path) exhibit high variance - highly sensitive to order**
- **CoT prompting does not eliminate order sensitivity, suggesting fundamental attention bias in LLMs.**

Finding 3: Task Type Determines Best Order



(a) Connectivity (b) Hamilton Path (c) Shortest Path

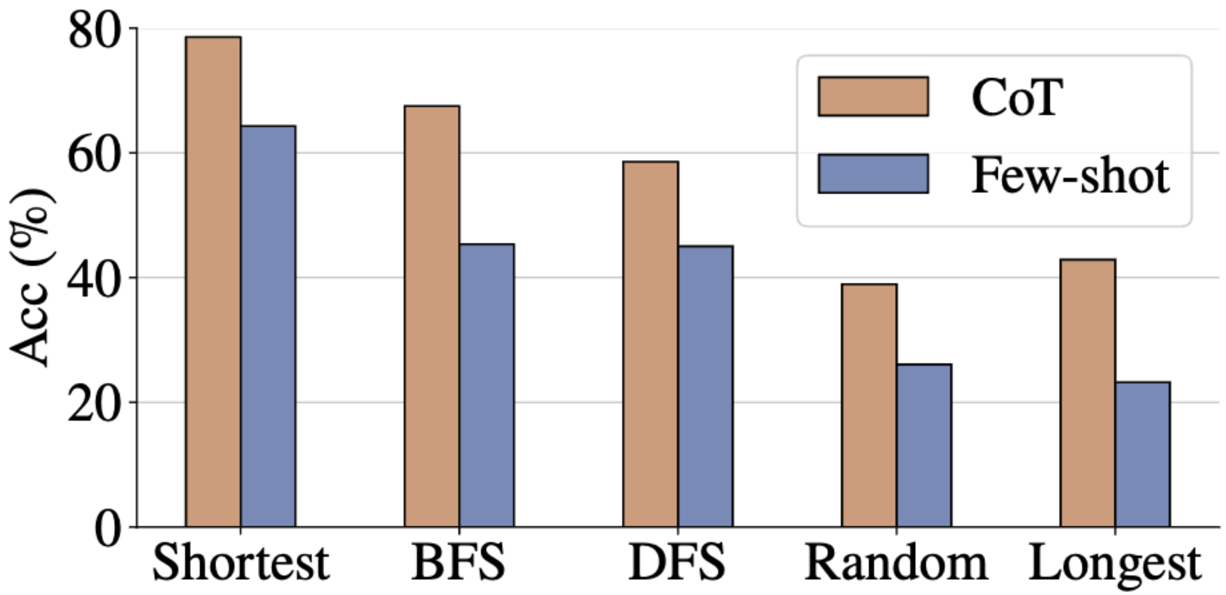
Figure 5: The improvement of average accuracy (calculated as the mean across all prompt types) of the LLM between a graph description in one order (horizontal axis) and its average accuracy on graph descriptions in other orders (vertical axis).

- **BFS excels at local reasoning tasks: connectivity (+14.1%), cycle detection (+12.7%), shortest path (+70.1%)**
- **DFS superior for global exploration: Hamilton path (+66.4% vs random, +22.3% vs BFS)**
- **Probability-based orders (PR/PPR) optimal for node classification tasks.**

Finding 4: Order Improves Graph Understanding

Better graph understanding or just more overlap with the answer?

- **Shortest Path Order**: Edges are ordered based on the shortest path from the root node v_0 to the target node v_t.
- **Longest Path Order**: Edges are ordered according to the longest path from v_0 to v_t.



- ➢ **Shortest path order (maximum answer overlap) achieves 78.57% accuracy - still far from 100%**
- ➢ **Performance drops to random-level with longest path order (minimum overlap)**
- ➢ **Ordered descriptions genuinely improve structural comprehension, not merely exploiting answer patterns**

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

Thanks!

Contact: yuyao.ge.work@gmail.com