논문: https://doi.org/10.48550/arXiv.2402.07140

Can Graph Descriptive Order Affect Solving Graph Problems with LLMs?

그래프 설명 순서가 LLM의 그래프 문제 해결에 미치는 영향

박세현
LAMDA Lab in SKKU

Can Graph Descriptive Order Affect Solving Graph Problems with LLMs?

- 그래프 설명의 순서가 LLM 성능에 미치는 영향 분석
  - 기존 방식은 랜덤으로 배열된 그래프 설명을 사용함
    -> 설명 순서의 역할을 간과
- 4가지의 그래프 설명 순서로 6가지 문제에 대해 실험
  - Orders
    - BFS
    - DFS
    - PageRank
    - Personalized PageRank
  - Problems
    - Connectivity
    - Cycle Detect
    - Shortest Path
    - Hamilton Path
    - Topological Sort
    - Node classification

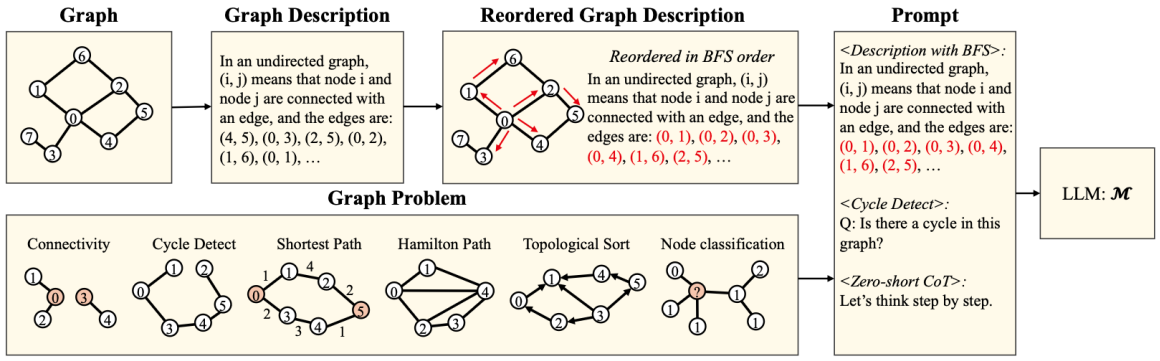


Figure 2: Overview of our framework for solving graph problems with LLMs. *In node classification task, node labels no longer represent identifiers; instead, they indicate the categories the nodes belong to.*

Can Graph Descriptive Order Affect Solving Graph Problems with LLMs?

- 기여점
  - 그래프 설명의 순서가 LLM의 그래프 추론 성능에 영향을 미친다는 점을 증명
  - 다양한 그래프 추론 작업에서 설명 순서가 LLM 성능에 미치는 차별적 영향을 분석
  - 그래프 집합, 대응 프롬프트, 설명 순서로 구성된 데이터셋 GraphDO (Graph Description with Order) 제안
- 한계점
  1. 다양한 그래프 구조와 유형에 대한 이 순서의 영향을 심층적으로 탐구하지는 않음
  2. 관찰된 현상에 대한 엄밀한 수학적·이론적 설명을 제시하지 못함

Prompt Engineering for Graph

- G = (V, E): 그래프
  - V: 노드 집합
  - E: 엣지 집합
- g(G, o): 그래프를 텍스트로 인코딩하기 위한 함수
  - o: 설명 순서 (o ∈ O)
- q(T): 작업 T를 기반으로 질문 Q를 생성하는 함수
  - q: T⟼Q
  - Q: 정답 Y를 갖음
- M(p, g, q): LLM 모델
  - p: 프롬프트 스타일 (p ∈ P)
- S(M, Y): LLM 평가 함수

$$\max_{o \in \mathcal{O}} \mathbb{E}_{\mathcal{G},T,Y \in D}\, \mathcal{S}(\mathcal{M}(p, g(\mathcal{G}, o), q(T)), Y)$$

Graph Problems

1. Connectivity: 두 노드 사이의 경로가 존재하는지 판단
2. Cycle: 시작 노드와 끝 노드가 동일한 경로가 존재하는지 판단
3. Hamilton Path: 각 노드를 한번 씩 지나는 경로가 존재하는지 판단
4. Shortest Path: 두 노드 사이의 최단 경로
5. Topological Sort: 노드들의 선형 정렬 (여러개의 해 존재)
6. Node Classification: 인접 노드들의 라벨을 기반으로 예측

- 1-5: 순수 그래프 구조에 초점
  - 1, 2: 국소적 추론
  - 3-5: 전체 그래프 이해
- 6: 그래프 속성 학습에 초점

Graph Encoder

- 인접 형식 (adjacency format)
  -> 그래프를 에지 리스트로 표현
  -> 순수 그래프와 속성 그래프 모두에 적용 가능

$$g(\mathcal{G}, o) = \mathcal{T}(\mathcal{G}, \mathcal{L}_o), o \in \mathcal{O}$$

Prompt Template for Unweighted Graphs

In an undirected/directed graph, (i, j) means that node i and node j are connected with an edge, and the edges are: [(0, 1), (1, 3), (3, 5), . . .].

Prompt Template for Weighted Graphs

In an undirected/directed graph, (i, j, w) means that node i and node j are connected by an edge with weight w, and the edges are: [(1, 3, 2), (0, 3, 1), (0, 1, 4), . . .].

Prompt Template for Node Classification Task

Adjacency list: [(1758, 2217), (2217, 2645), . . .]
Node to label mapping: node 1758: label 3 | node 2217: label 2 | node 2645: label ? | . . .

Graph Description Ordering

- Random Order: 랜덤 순서
  - 그래프의 엣지 E 를 무작위로 섞어서 나열.
- Breadth-First Search (BFS) Order: 너비 우선 탐색
  - 무작위로 선택한 루트 노드 v_0에서 시작
  - 그래프를 레벨 단위로 탐색
  - 각 레벨에서, 현재 노드 v의 이웃 노드 u와 연결된 엣지 (v, u)를 순서대로 추가한 뒤 다음 레벨로 넘어감
- Depth-First Search (DFS) Order: 깊이 우선 탐색
  - 루트 노드 v_0에서 시작
  - 가능한 한 깊숙이 탐색하다가 더 이상 진행할 수 없을 때 백트래킹

Graph Description Ordering

- PageRank (PR) Order: 노드 중요도에 대한 전역적 확률 분포
  - 모든 노드 v에 대해 PR(v) 계산
  - 점수가 높은 노드부터 순서대로 이웃 엣지를 나열
  - 이미 포함된 엣지는 중복되지 않도록 건너뜀

$$PR(v) = \alpha \sum_{u \in N^{-1}(v)} \frac{PR(u)}{|N(u)|} + (1 - \alpha)$$

$\alpha = 0.85$는 $damping\ factor$, $N^{-1}(v)$는 v로 들어오는 이웃 노드 집합

- Personalized PageRank (PPR) Order: 국소적 확률 분포
  - PageRank에 개인화 벡터를 도입
  - 특정 작업에서 중요도가 높은 노드들에 더 높은 확률 가중치 부여

$$PR_S(v) = \alpha \sum_{u \in N^{-1}(v)} \frac{PR_S(u)}{|N(u)|} + (1 - \alpha) \cdot e_v$$

e_v는 개인화 벡터 값

Experiments

- Datasets
  - GraphDO
    - 8,500 case
    - graph description, question, answer
  - 전통 테스크(connectivity, cycle, shortest, Hamilton, topological)
    - ER 그래프 생성 방법
  - 그래프 학습 테스크(node classification)
    - CORA, Citeseer, Pubmed 데이터셋 사용
      -> LLM input 한계 초과
      -> Ego, FF(forest fire smapling)을 이용한 샘플링
    - 적은 추론단계 -> zero-shot 프롬프팅

Experiments

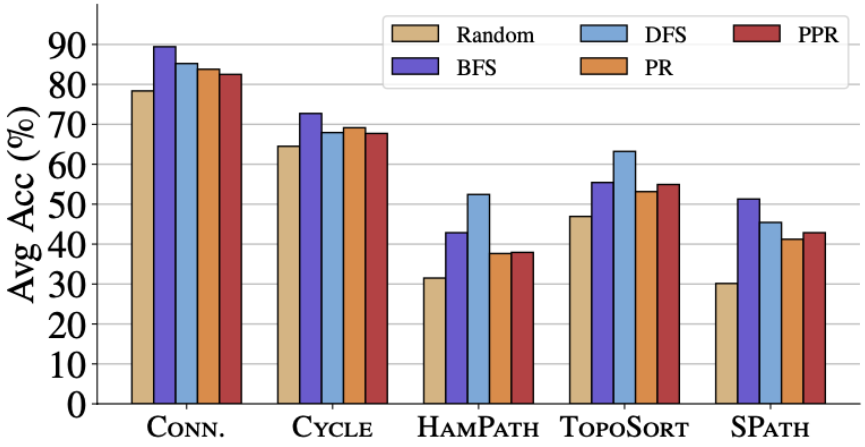
- Models
  - GPT-3.5-TURBO-0613 (default)
  - LLAMA2-7B-CHAT, LLAMA2-13B-CHAT
  - QWEN2-7B
  - MISTRAL-7B
  - VICUNA-7B-v1.5

  * decoding temperature: 0
- Metric

$$\text{Acc} = \frac{\#\textit{correct answers}}{\#\textit{total questions}}$$

- Baseline
  - random order graph

(Q1) Does the order of graph description impact the LLM's performance in solving graph problems?



Task	Order	Zero-shot	Zero-shot CoT	Few-shot	CoT	CoT-BAG	Avg.
Conn.	Random	73.93(-)	70.71(-)	81.07(-)	83.93(-)	82.14(-)	78.36(-)
	BFS	82.14(↑11.11)	87.50(↑23.74)	89.29(↑10.14)	92.50(↑10.21)	95.71(↑16.52)	89.43(↑14.13)
	DFS	79.29(↑7.25)	82.14(↑16.16)	87.14(↑7.49)	88.21(↑5.10)	89.29(↑8.70)	85.21(↑8.75)
	PR	77.86(↑5.32)	83.57(↑18.19)	85.71(↑5.72)	84.29(↑0.43)	87.50(↑6.53)	83.79(↑6.93)
	PPR	76.79(↑3.87)	81.07(↑14.65)	83.93(↑3.53)	84.64(↑0.85)	86.07(↑4.78)	82.50(↑5.29)
Cycle	Random	51.79(-)	53.57(-)	65.36(-)	75.71(-)	76.07(-)	64.50(-)
	BFS	55.71(↑7.57)	56.07(↑4.67)	79.29(↑21.31)	86.07(↑13.68)	86.43(↑13.62)	72.71(↑12.73)
	DFS	52.14(↑0.68)	53.93(↑0.67)	73.21(↑12.01)	79.29(↑4.73)	81.07(↑6.57)	67.93(↑5.31)
	PR	55.36(↑6.89)	56.43(↑5.33)	70.36(↑7.65)	80.36(↑6.14)	83.21(↑9.39)	69.14(↑7.20)
	PPR	54.29(↑4.83)	55.00(↑2.67)	70.00(↑7.10)	79.29(↑4.73)	80.00(↑5.17)	67.72(↑4.99)
HamPath	Random	10.71(-)	15.36(-)	40.00(-)	46.07(-)	45.36(-)	31.50(-)
	BFS	20.00(↑86.74)	20.71(↑34.83)	57.86(↑44.65)	58.57(↑27.13)	57.14(↑25.97)	42.86(↑36.05)
	DFS	33.93(↑216.81)	37.50(↑144.14)	67.50(↑68.75)	63.93(↑38.77)	59.29(↑30.71)	52.43(↑66.44)
	PR	15.00(↑40.06)	19.29(↑25.59)	48.93(↑22.32)	55.00(↑19.38)	50.00(↑10.23)	37.64(↑19.50)
	PPR	16.43(↑53.41)	18.93(↑23.24)	50.00(↑25.00)	53.93(↑17.06)	50.36(↑11.02)	37.93(↑20.41)
TopoSort	Random	28.93(-)	31.07(-)	58.21(-)	56.07(-)	60.36(-)	46.93(-)
	BFS	43.21(↑49.36)	40.36(↑29.90)	67.14(↑15.34)	61.43(↑9.56)	65.00(↑7.69)	55.43(↑18.11)
	DFS	42.14(↑45.66)	48.93(↑57.48)	77.86(↑33.76)	74.29(↑32.50)	72.86(↑20.71)	63.21(↑34.71)
	PR	35.36(↑22.23)	35.71(↑14.93)	71.07(↑22.09)	58.21(↑3.82)	65.36(↑8.28)	53.14(↑13.24)
	PPR	37.14(↑28.38)	39.64(↑27.58)	72.50(↑24.55)	58.93(↑5.10)	66.43(↑10.06)	54.93(↑17.05)
SPath	Random	20.00(-)	25.00(-)	26.07(-)	38.93(-)	40.71(-)	30.14(-)
	BFS	35.36(↑76.80)	42.50(↑70.00)	45.36(↑73.99)	67.50(↑73.39)	65.71(↑61.41)	51.29(↑70.15)
	DFS	32.14(↑60.70)	34.29(↑37.16)	45.00(↑72.61)	58.57(↑50.45)	57.14(↑40.36)	45.43(↑50.71)
	PR	30.36(↑51.80)	43.93(↑75.72)	38.93(↑49.33)	43.93(↑12.84)	48.93(↑20.19)	41.21(↑36.74)
	PPR	32.50(↑62.50)	44.64(↑78.56)	42.14(↑61.64)	45.36(↑16.52)	49.64(↑21.94)	42.86(↑42.18)

Table 1: Results of the performance of various orders on different graph tasks. (↑) indicates the improvement compared to the baseline under the same setting.

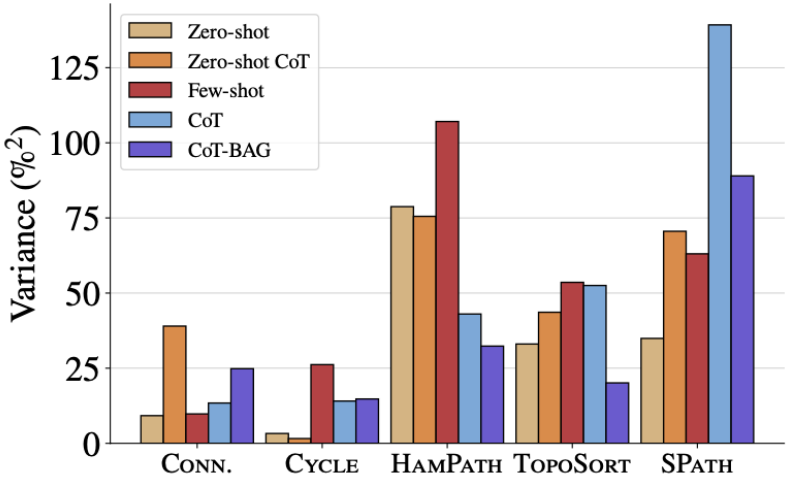
(Q1) Does the order of graph description impact the LLM's performance in solving graph problems?

Sampling	Order	CORA		Citeseer		Pubmed	
		Acc.	Δ	Acc.	Δ	Acc.	Δ
Ego	Random	70.00	-	67.33	-	72.00	-
	BFS	72.00	↑ 2.86	68.67	↑ 1.99	74.00	↑ 2.78
	DFS	71.33	↑ 1.90	68.66	↑ 1.98	77.33	↑ 7.40
	PR	75.33	↑ 7.61	71.33	↑ 5.94	82.67	↑ 14.82
	PPR	73.33	↑ 4.76	69.33	↑ 2.97	77.33	↑ 7.40
Forest Fire	Random	79.33	-	68.67	-	69.99	-
	BFS	82.67	↑ 4.21	71.33	↑ 3.87	74.00	↑ 5.73
	DFS	81.33	↑ 2.52	70.00	↑ 1.94	76.00	↑ 8.59
	PR	83.33	↑ 5.04	71.33	↑ 3.87	76.00	↑ 8.59
	PPR	82.00	↑ 3.36	70.67	↑ 2.91	74.67	↑ 6.69

Table 2: The accuracy of the LLM in solving node classification task across various orders, datasets, and sampling methods. ↑ indicates the improvement compared to the baseline under the same setting.

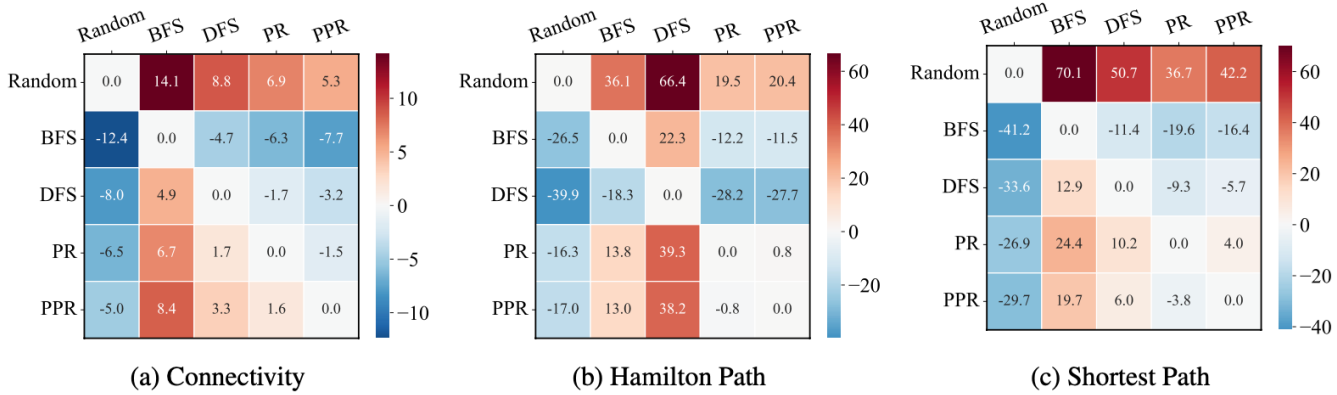
- 정렬된 순서는 항상 랜덤 순서보다 높은 성능을 보임
  -> 단순히 엣지를 나열하는 방식만 달라도 LLM의 추론 능력이 크게 달라짐

(Q2) Is the robustness of LLM to graph description order consistent across different tasks?



- 단순한 과제(Connectivity, Cycle)
  - 순서 변화에 따른 성능 분산이 작음 -> LLM이 비교적 강건함
- 복잡한 과제(Hamilton Path, Topological Sort, Shortest Path)
  - 순서 변화에 따른 성능 분산이 큼 -> LLM이 순서에 매우 민감
  - 특히 Shortest Path는 가중치 정보까지 필요하기 때문에 가장 큰 분산을 보임

(Q3) Does a specific graph description order favor certain graph tasks?



(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).

- 지역적 구조(Cycle, Connectivity, shortest Path)
  - BFS > DFS
- 전역적 구조(Hamilton Path, Topological Sort)
  - DFS > BFS

(Q3) Does a specific graph description order favor certain graph tasks?

Sampling	Order	CORA		Citeseer		Pubmed	
		Acc.	Δ	Acc.	Δ	Acc.	Δ
Ego	Random	70.00	-	67.33	-	72.00	-
	BFS	72.00	↑ 2.86	68.67	↑ 1.99	74.00	↑ 2.78
	DFS	71.33	↑ 1.90	68.66	↑ 1.98	77.33	↑ 7.40
	PR	75.33	↑ 7.61	71.33	↑ 5.94	82.67	↑ 14.82
	PPR	73.33	↑ 4.76	69.33	↑ 2.97	77.33	↑ 7.40
Forest Fire	Random	79.33	-	68.67	-	69.99	-
	BFS	82.67	↑ 4.21	71.33	↑ 3.87	74.00	↑ 5.73
	DFS	81.33	↑ 2.52	70.00	↑ 1.94	76.00	↑ 8.59
	PR	83.33	↑ 5.04	71.33	↑ 3.87	76.00	↑ 8.59
	PPR	82.00	↑ 3.36	70.67	↑ 2.91	74.67	↑ 6.69

Table 2: The accuracy of the LLM in solving node classification task across various orders, datasets, and sampling methods. ↑ indicates the improvement compared to the baseline under the same setting.

- 노드 분류
  - PR이 가장 우수

Better graph understanding or just more overlap with the answer?

- 최단 경로 문제
  - BFS와 DFS의 엣지 리스트에 정답이 부분적으로 중첩되어서 성능이 더 좋은가?
  - 추가 실험 (정답이 중첩 되어있는 순서)
    - 최단 경로: 루트 노드 v_0에서 v_t 까지의 최단 경로
    - 최장 경로: 루트 노드 v_0에서 v_t 까지의 최장 경로

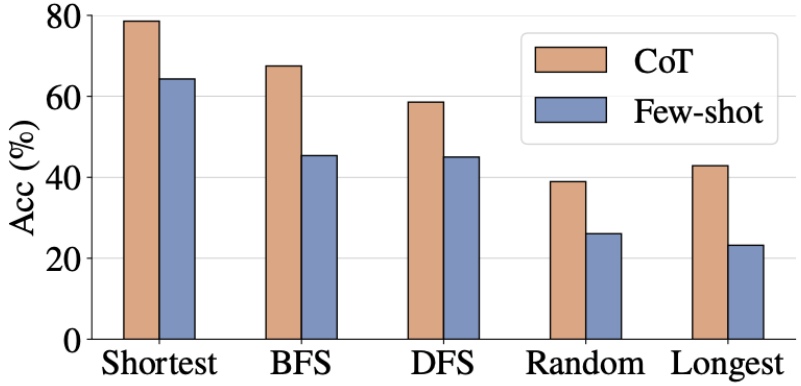


Figure 6: Results of the accuracy of various orders on shortest path task.

Better graph understanding or just more overlap with the answer?

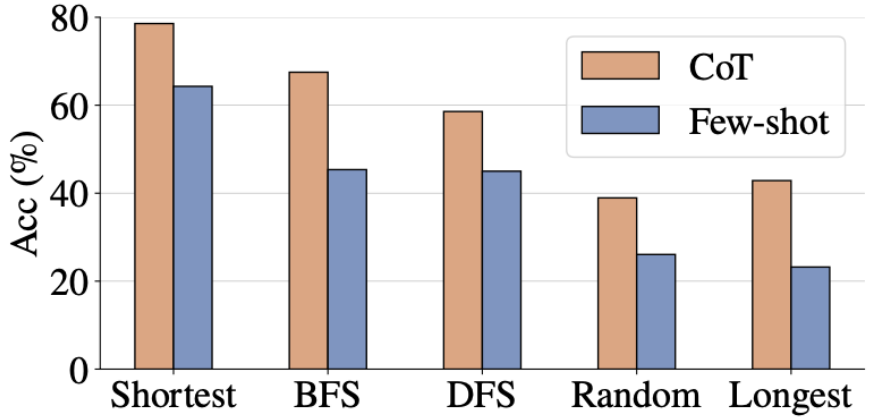


Figure 6: Results of the accuracy of various orders on shortest path task.

- Longest는 Random 정렬과 성능의 큰 차이는 없음
- Shortest에서 가장 높은 성능을 보였지만 100%에 미치지 못함

-> 중첩이 영향은 있지만 유일한 요인은 아님
-> LLM의 그래프 이해 능력 향상에 영향을 미침

Model Comparison Study

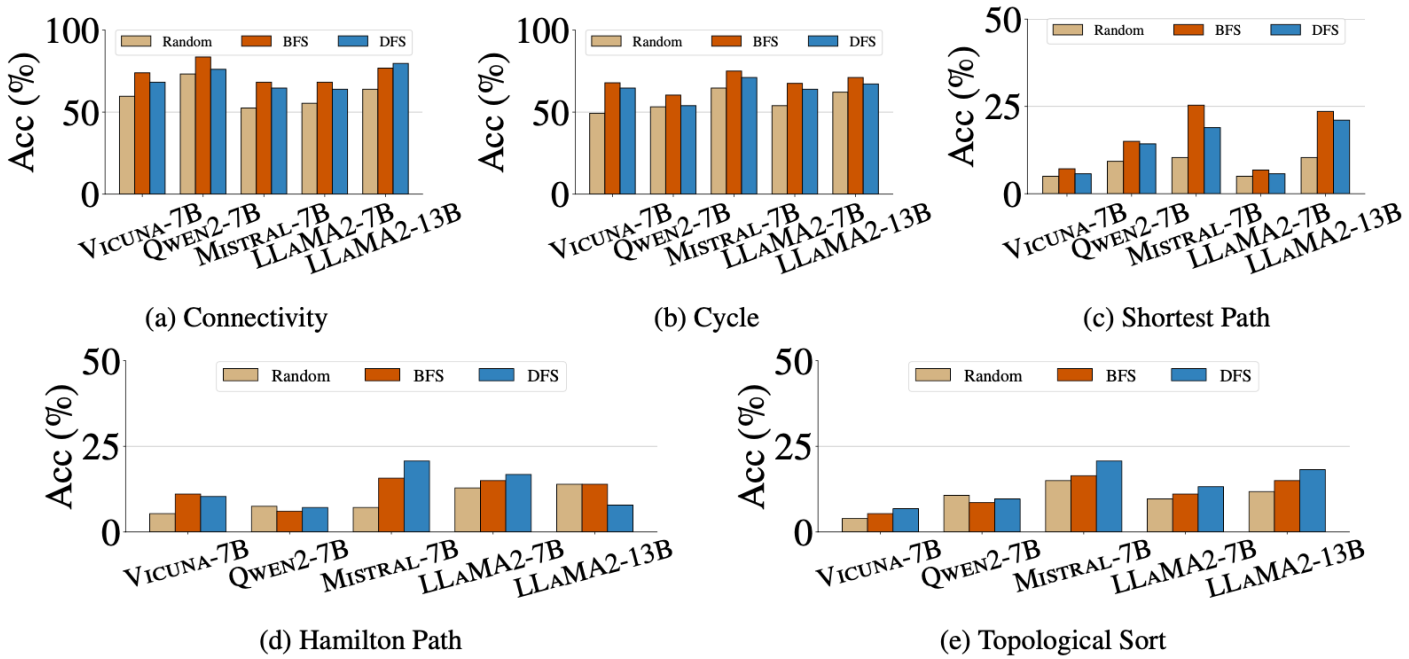


Figure 7: The impact of model differences on solving graph reasoning problems.

- GPT-3.5-TURBO-0613에 비해 효과는 덜 두드러졌으나, 유사한 패턴으로 일관성을 보임

* 일부 모델에서는 특정 작업에서 우수한 성능을 보임
  ex) QWEN2-7B: Connectivity에서 다른 모델 보다 우수함