GraphChain: Large Language Models for Large-scale Graph Analysis via Tool Chaining

Chunyu Wei[1], Wenji Hu[1], Xingjia Hao[2], Xin Wang[1], Yifan Yang[3],

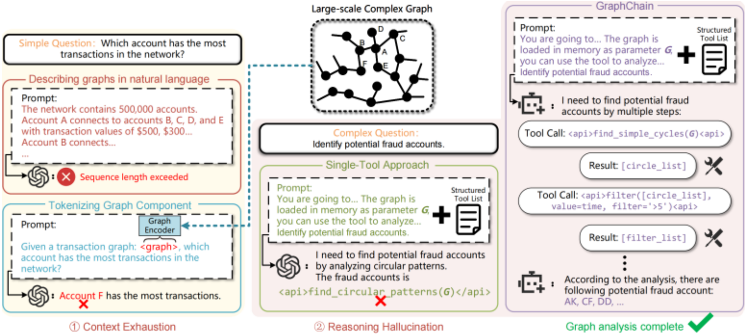
Yueguo Chen[1], Yang Tian[2], Yunhai Wang[1*]

[1]Renmin University of China, China [2]Guangxi University, China [3]Beijing Jiaotong University, China

39th Conference on Neural Information Processing Systems (NeurIPS 2025)

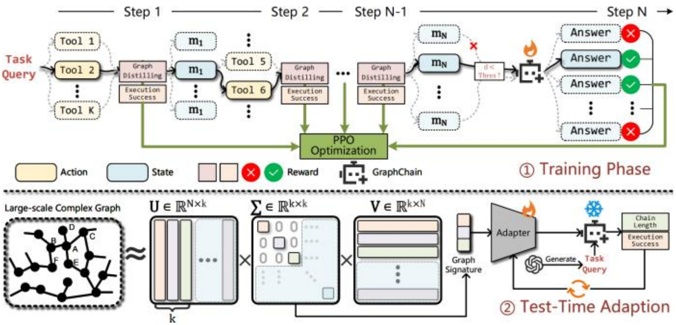
Outline

Motivation and Challenges



- Graphs ubiquitous: Social networks, finance, molecules (e.g., fraud detection on 200k+ nodes)
- LLMs struggle: Context Exhaustion (graphs exceed windows) and Reasoning Hallucination (single tools fail complex tasks)
- Prior work limitations: Text-based (NLGraph) limited by size; Tool-based (GraphForge) lacks adaptive chaining
- Inspiration: Mimic human exploration – broad to focused

GraphChain Overview



- Framework: LLMs chain specialized tools (NetworkX-based) for dynamic graph analysis
- MDP Modeling: States (query + memory), Actions (tool selection), Rewards (success + distillation)
- Innovations:

  Progressive Graph Distillation: RL to reduce complexity while keeping relevance

  Structure-aware Test-Time Adaptation: Adapt to topologies via SVD fingerprint and lightweight adapter

Key Methods

- Distillation: Quantify GDL (volume) and Relevance; Reward for reduction + gain; PPO optimization
- Adaptation: Graph fingerprint zG from Laplacian SVD; Adapter generates prompt for frozen LLM
- Self-supervised: Minimize chain length + KL on auxiliary queries
- Benefits: Scales to large graphs, no retraining

$$R_t = \begin{cases} w_1 \cdot \hat{r}_t^{\text{Succ}} + w_2 \cdot \hat{r}_t^{\Delta\text{GDL}} + w_3 \cdot \hat{r}_t^{\Delta\text{Rel}} & \text{if } t < N \\ w_{\text{solve}} \cdot \text{EvaluateTaskSuccess}(\mathcal{Q}, s_{N+1}) & \text{if } t = N \end{cases}$$

$$L_{\text{STTA}}(\psi) = \mathbb{E}_{\mathcal{Q}_{\text{aux},i}, \tau_i \sim \pi_\psi(\cdot|s; G_{\text{test}})} \left[w_L N_{\tau_i} + w_{KL} \sum_{t=0}^{N_{\tau_i}-1} D_{KL}(\pi_\psi(\cdot|s_t; G_{\text{test}}) || \pi_{\text{orig}}(\cdot|s_t)) \right]$$

Experimental Results

- Datasets: 5 domains (Cora, Facebook, QM9, etc.) – Table 1 stats
- GraphChain: 84.7% avg. accuracy (+20.7% over GraphForge at 70.2%)
- Baselines: Text (GPT-4o: 59.4%); Tools (GraphForge: 70.2%)
- Scalability: Consistent on 200k nodes; Ablations confirm distillation/STTA value

Model	Parameters	Financial Network	Chemical Molecule	Social Network	Citation Graph	Traffic Network	Average
Text-Instruction Methods							
Claude-3-Sonnet	-	21.7 ± 1.8	47.0 ± 2.2	21.5 ± 3.2	17.7 ± 2.1	16.8 ± 2.0	24.9 ± 2.3
GPT-3.5-turbo	~175B	36.6 ± 2.1	23.0 ± 3.7	18.2 ± 3.6	12.2 ± 0.8	19.4 ± 1.9	21.9 ± 2.4
Claude-3-Haiku	~20B	12.2 ± 3.0	52.9 ± 3.2	50.3 ± 3.4	19.8 ± 2.0	13.9 ± 2.4	29.8 ± 2.8
Claude-3-Opus	~137B	23.6 ± 2.1	42.4 ± 1.4	51.9 ± 1.3	36.7 ± 3.1	43.4 ± 3.3	39.6 ± 2.2
GraphWiz	13B	41.1 ± 3.9	52.4 ± 2.6	61.5 ± 3.5	68.0 ± 2.1	38.4 ± 1.9	52.3 ± 2.9
NLGraph	~100B	52.1 ± 3.4	58.4 ± 2.5	65.2 ± 2.3	59.4 ± 0.5	39.8 ± 1.8	55.0 ± 2.1
GPT-4o	~200B	57.5 ± 1.9	62.7 ± 3.6	65.2 ± 3.9	71.5 ± 3.4	43.4 ± 1.6	59.4 ± 2.6
Claude-4-Sonnet	-	58.2 ± 2.1	62.9 ± 1.7	61.7 ± 4.3	77.5 ± 1.4	32.8 ± 1.9	58.6 ± 2.3
GPT-4.1	-	52.2 ± 1.5	63.4 ± 2.6	67.4 ± 2.3	70.0 ± 1.9	55.5 ± 3.1	61.7 ± 2.2
Gemini-2.5-Flash	-	25.1 ± 1.3	67.3 ± 1.6	28.1 ± 2.1	24.1 ± 1.8	24.9 ± 1.8	33.9 ± 1.7
Tool-Instruction Methods							
Graph-ToolFormer	8B	47.5 ± 1.9	68.1 ± 4.8	74.7 ± 4.2	61.4 ± 3.4	65.8 ± 4.5	62.4 ± 3.5
GraphForge	8B	63.5 ± 3.5	70.9 ± 5.4	80.4 ± 4.2	63.4 ± 4.4	73.5 ± 3.1	70.2 ± 3.8
ToolGen	8B	75.8 ± 1.1	57.9 ± 2.9	79.4 ± 2.3	61.2 ± 1.3	62.7 ± 1.5	67.4 ± 1.8
GraphChain	7B	$\mathbf{81.5 \pm 2.2}$	$\mathbf{81.1 \pm 0.7}$	$\mathbf{89.6 \pm 2.0}$	$\mathbf{83.6 \pm 2.6}$	$\mathbf{84.1 \pm 0.3}$	$\mathbf{84.7 \pm 1.8}$
Relative improvement (%)	-	*+7.5%*	*+14.4%*	*+11.4%*	*+7.9%*	*+14.4%*	*+20.7%*

Conclusion and Contributions

- GraphChain enables scalable LLM graph analysis via chaining
- Contributions: RL distillation, topology adaptation, +20.7% SOTA, open-source code
- Future: More tools, online adaptation