NeurIPS 2026
40th Conference on Neural Information Processing Systems
Evaluations & Datasets Track

GraphInstruct: A Progressive Benchmark for Diagnosing Capability Gaps in LLM Graph Generation

Zihe Wei[1], Sheng Xiang[1,*], Ying Zhang[2], Changjun Jiang[1]

[1] School of Computer Science & Technology, Tongji University
[2] School of Computer Science & Information Technology and School of Statistics & Mathematics, Zhejiang Gongshang University

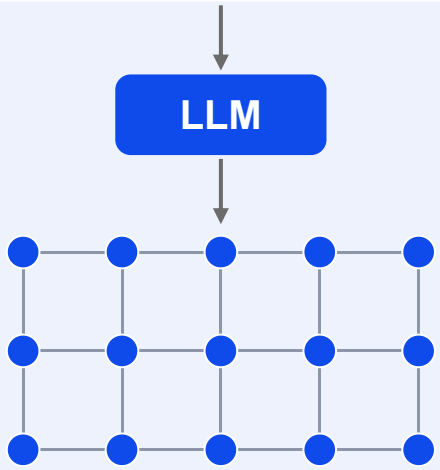
{weizihe, xiangsheng, cjjiang}@tongji.edu.cn · ying.zhang@zjgsu.edu.cn · *Corresponding author

Background

Why a progressive benchmark?

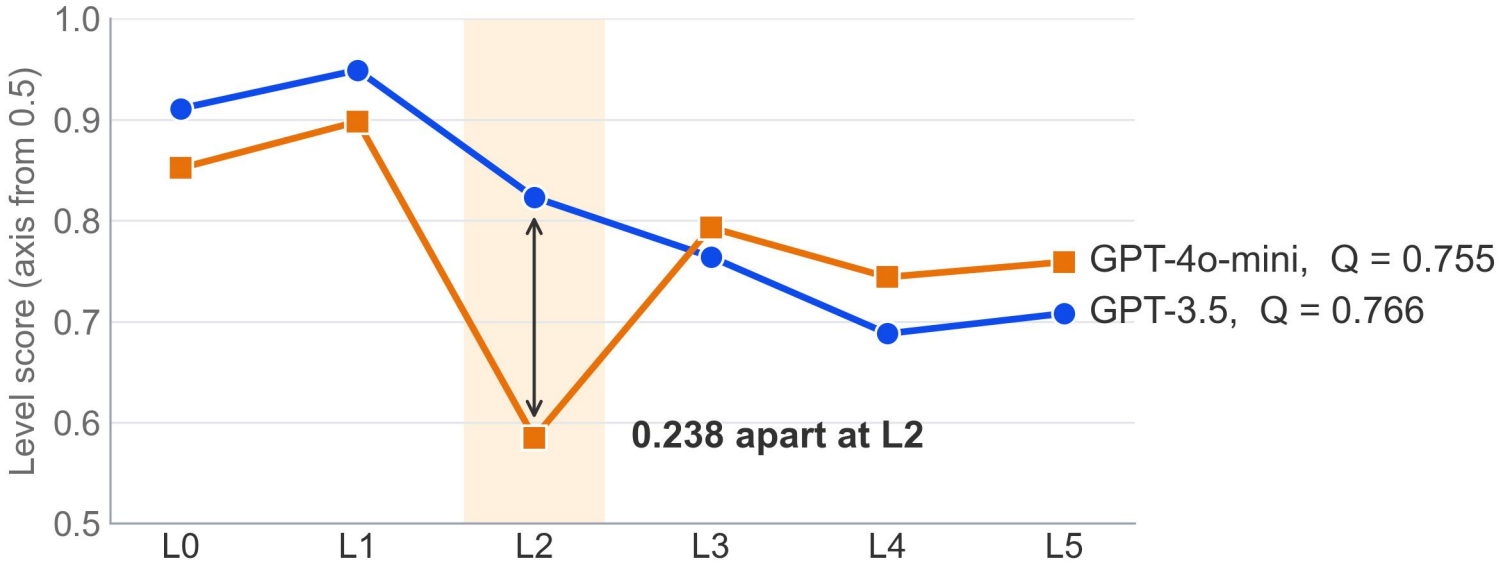
Instruction

"Generate a 3x5 grid graph with 15 nodes that is connected and planar."



LLMs as graph synthesizers

Close overall scores, very different level scores (zero-shot)



Recent benchmarks organize by graph type or task domain and average over structural complexity.

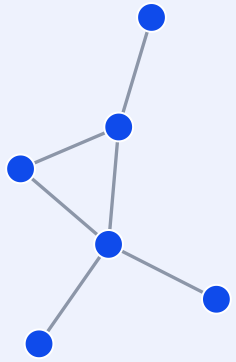
GraphInstruct organizes by constraint type and localizes where models diverge.

Benchmark

Six constraint-type levels, not a difficulty ladder

L0 N = 100

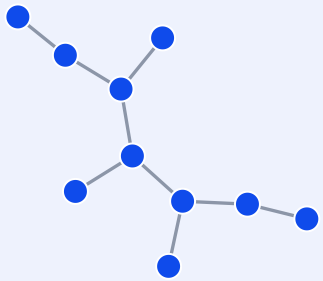
Format generation



6 nodes and 6 edges

L1 N = 200

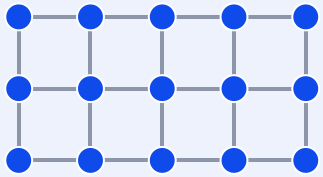
Single explicit constraint



a tree with 10 nodes

L2 N = 200

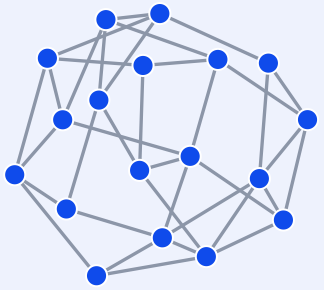
Multi-constraint composition



3x5 grid with 15 nodes, connected and planar

L3 N = 150

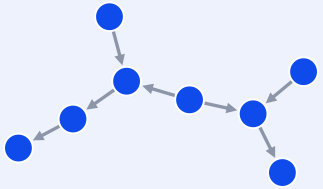
Numerical attribute control



18 nodes, density ≤ 0.2595, clustering ≥ 0.1621, path length ≤ 2.3563

L4 N = 100

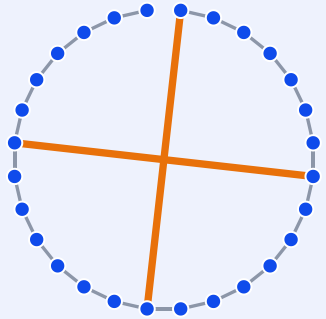
Domain semantics



ecological network: 8 species, prey → predator, density ≤ 0.315

L5 N = 50

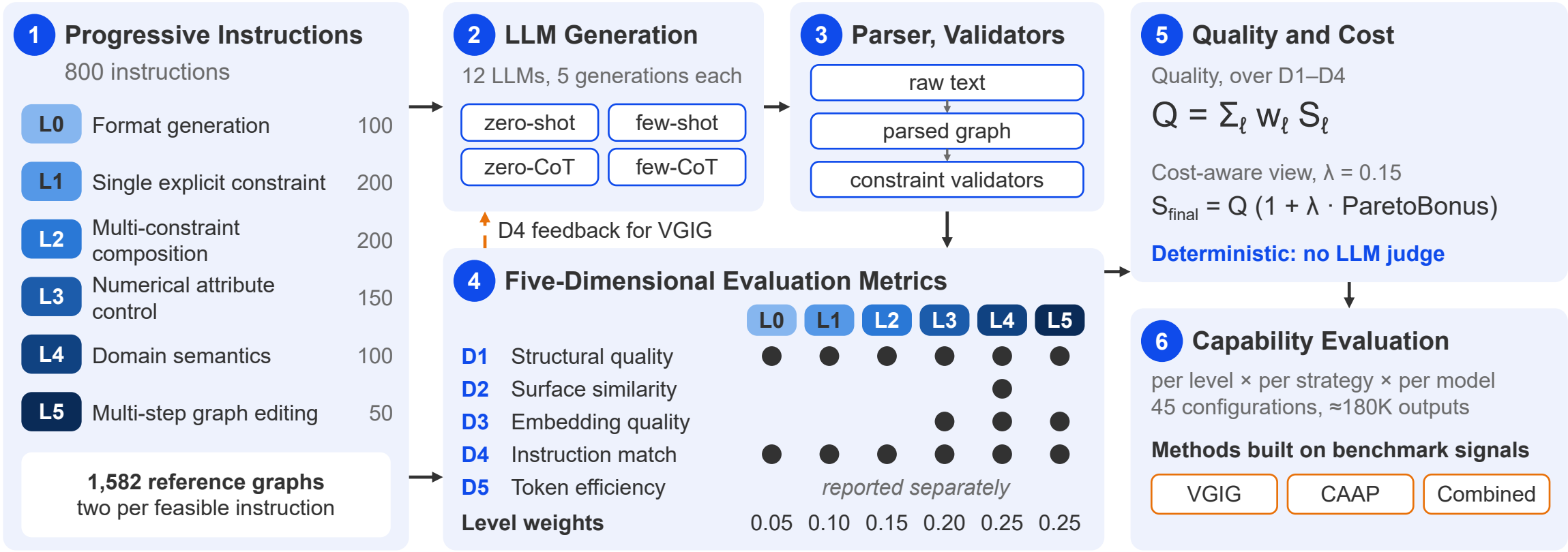
Multi-step graph editing



given a path with diameter 27, add 2 edges to reach diameter ≤ 16

Benchmark

From instruction to score: six levels, five dimensions

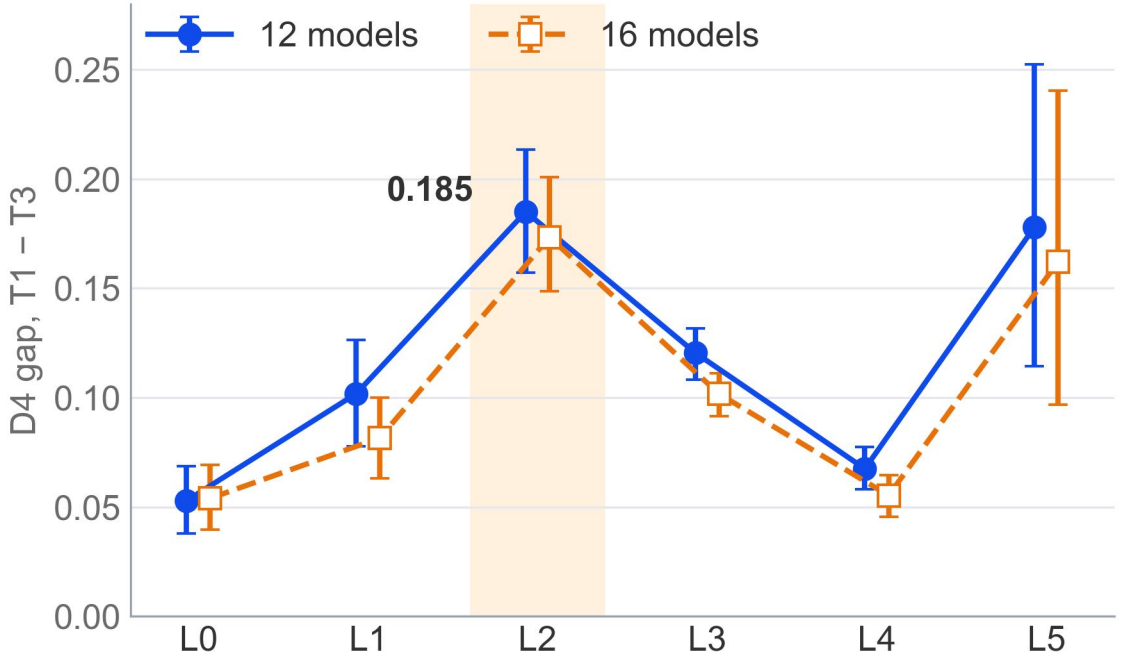


12 LLMs: Sonnet-4.6, -4 (zero-shot only); GPT-4.1, -4o, -4o-mini, -3.5; Qwen3.5-397B, -122B, -35B; DeepSeek-V3; Llama-3.3-70B, -3.1-8B; later 16 and 20.

Evaluation

Dispersion peaks at multi-constraint composition (L2)

(a) D4 tier gap, T1 − T3, best strategy per model, 95% CI



(b) Spread across models, zero-shot level scores (SD)

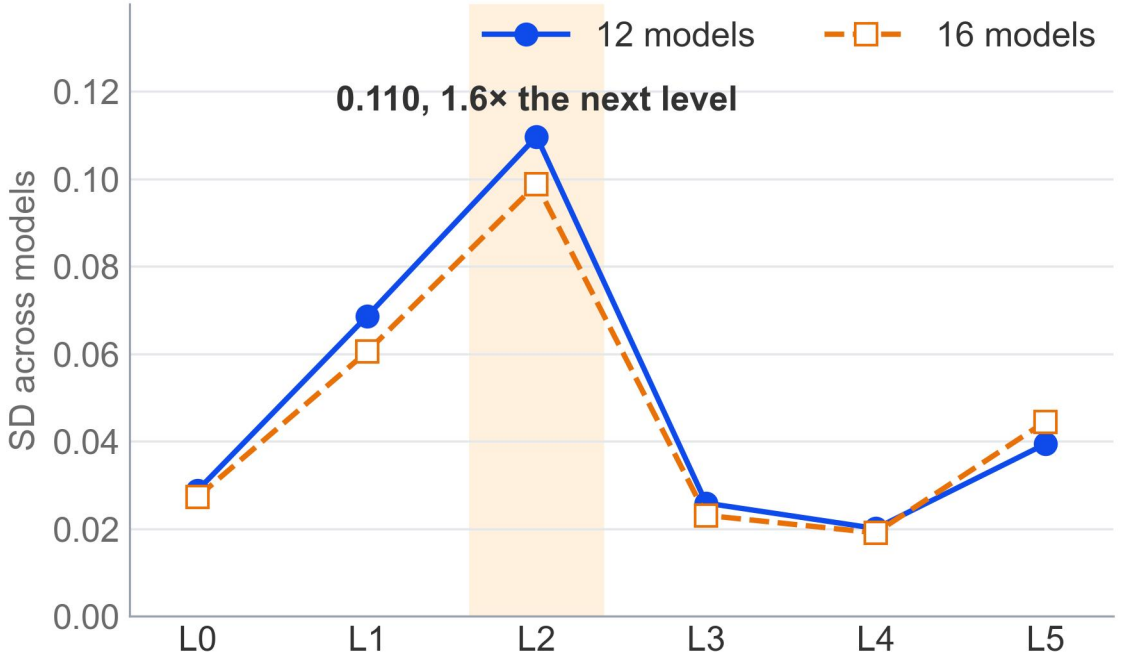


Figure 1 of the paper. The L2 and L5 gap intervals overlap ([0.157, 0.214] vs. [0.114, 0.252]), so their order is not established; L5 statements are tentative (N = 50).

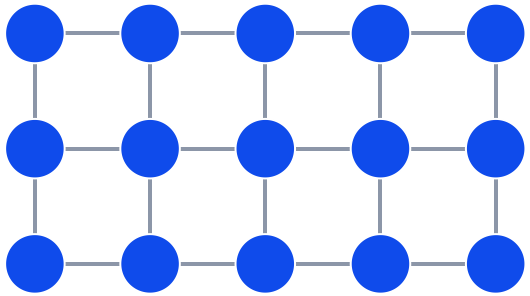
Evaluation

What an L2 failure looks like: one instruction, two models

Instruction L2-143: *"Generate a 3x5 grid graph with 15 nodes that is connected and planar."*

Sonnet-4.6 (T1)

instruction score 1.000

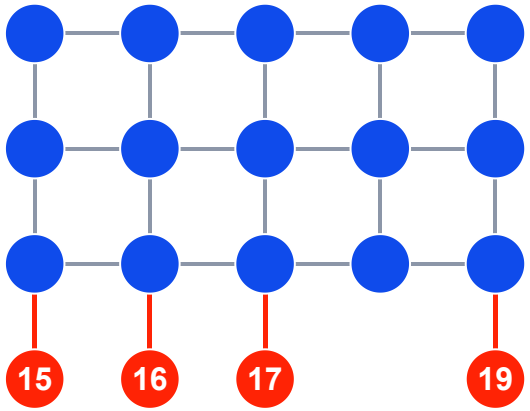


✓ all 7 constraints satisfied

Constraint	Sonnet-4.6	GPT-4o-mini
Graph type: grid	✓ yes	✗ no
Nodes = 15	✓ 15	✗ 19
Edges = 22	✓ 22	✗ 26
Connected	✓ yes	✓ yes
Planar	✓ yes	✓ yes
Bipartite	✓ yes	✓ yes
Min. degree ≥ 2	✓ 2	✗ 1

GPT-4o-mini (T3)

instruction score 0.000



✗ 4 of 7 constraints violated

Graphs and table: the first of five zero-shot samples. Score: over all five; no GPT-4o-mini sample passes the validity check, so its D1 and D4 are zero on this instruction. Full version: Figure 6 of the paper.

Evaluation

No single prompting strategy dominates across levels

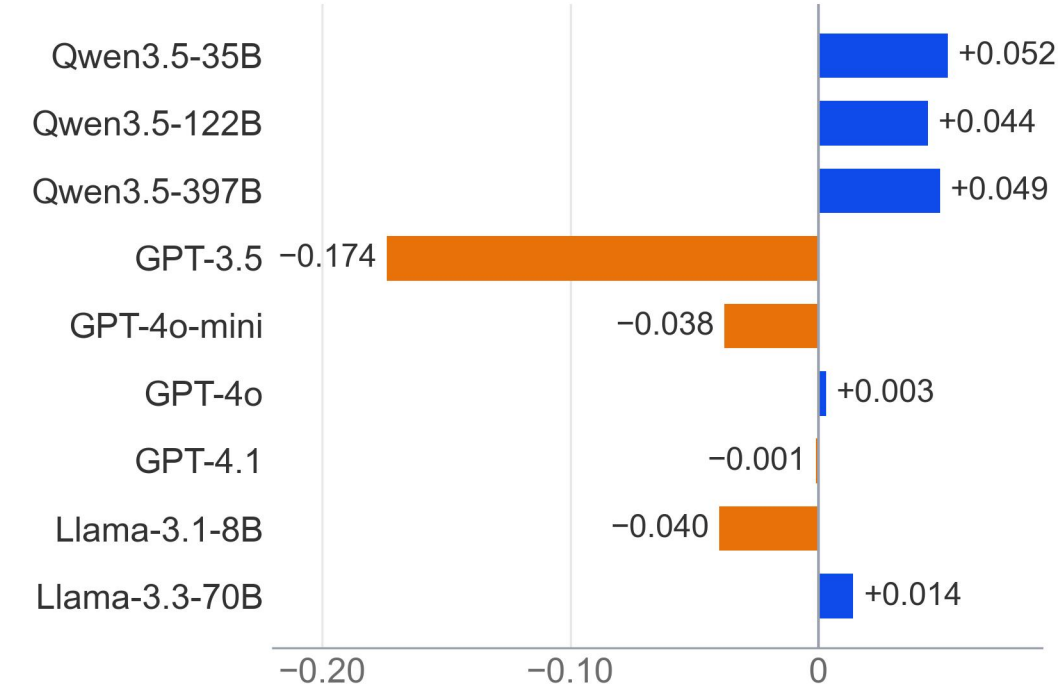
Mean change in level score vs. zero-shot (11 models)

	L0	L1	L2	L3	L4	L5
Few-shot	−0.019	−0.032	**−0.029**	−0.016	**+0.058**	+0.002
Zero-CoT	+0.004	+0.008	+0.028	+0.004	−0.003	+0.036
Few-CoT	−0.025	−0.020	−0.025	**−0.078**	+0.033	**+0.047**

Blue = gain, orange = loss; bold = the cells discussed. L5 is tentative (N = 50).

Zero-CoT never loses much
Its only negative mean is −0.003, at L4.

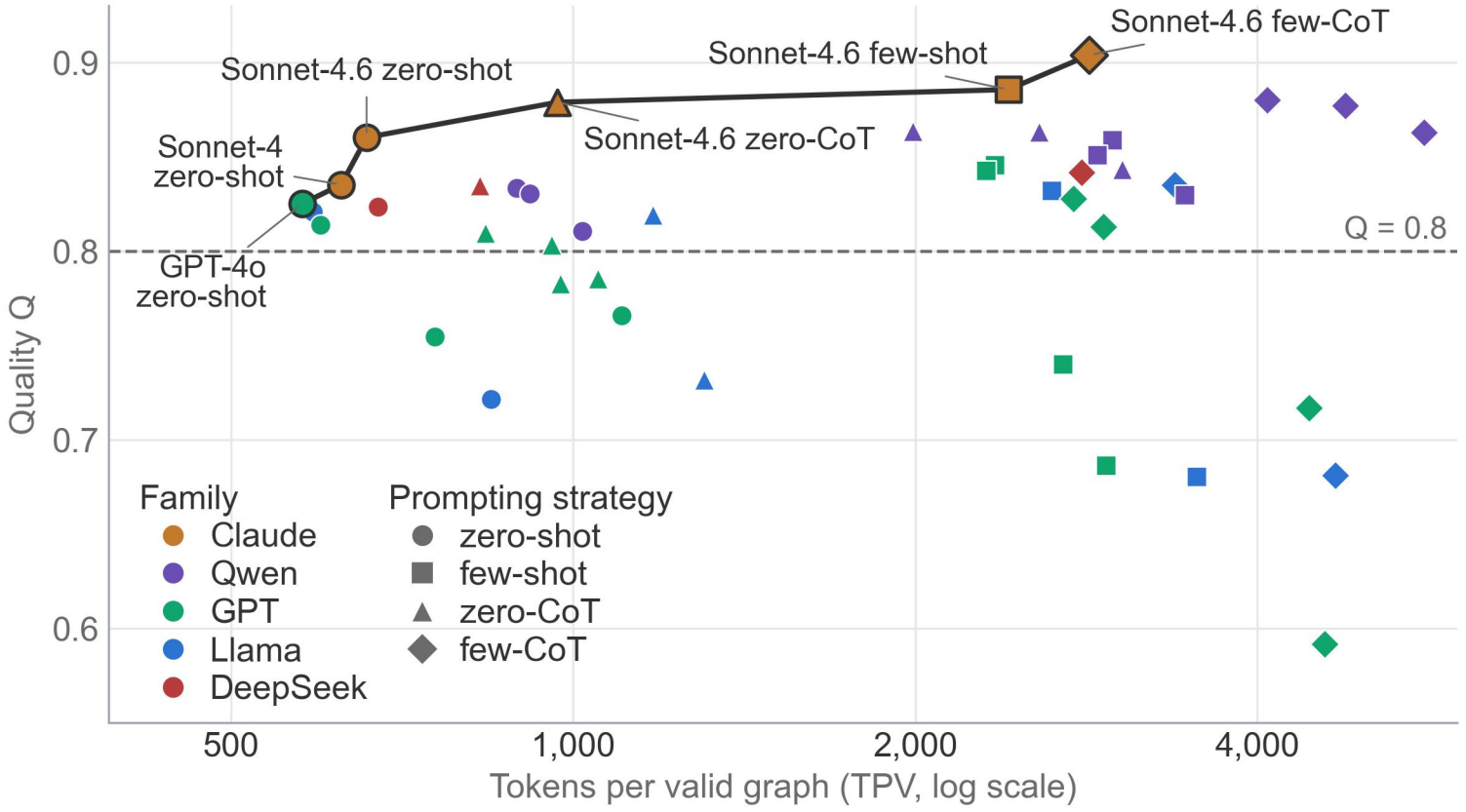
Few-CoT minus zero-shot, total Q



Helps Qwen3.5 at every scale; hurts the weaker GPT models. GPT-4o/4.1 stay within ±0.005; within Llama the sign flips with scale.

Evaluation

Only 6 of 45 configurations are non-dominated



6 / 45

configurations are non-dominated, all from Anthropic or OpenAI

≈600

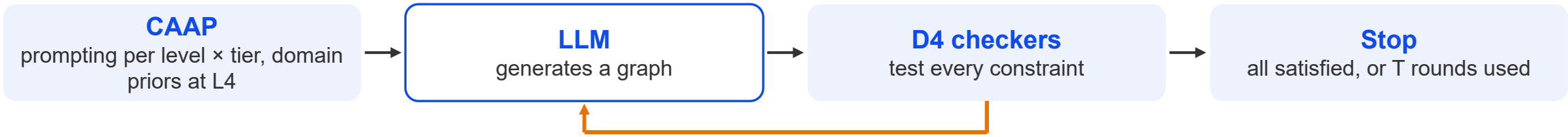
tokens per valid graph on the cheapest paths to Q ≥ 0.8, all zero-shot

3 / 12

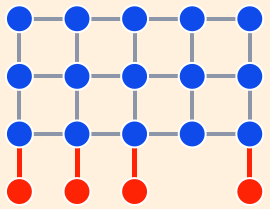
models never reach Q ≥ 0.8 under any of the four strategies

Method

VGIG turns the checker's output into the next prompt

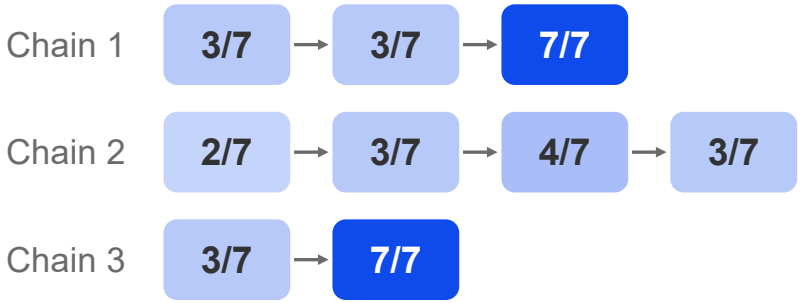


Violated constraints, sent back as text



```
Your previous graph did not satisfy all constraints.
Issues found:
- graph_type should be grid (num_nodes=19;
num_edges=26)
- Graph has 19 nodes but requires 15 (remove 4)
- Graph has 26 edges but requires 22 (remove 4)
- Minimum degree must be >= 2 (current min 1; nodes
below threshold ['15', '16', '17', '19'])
Please regenerate the graph in the same format,
fixing these issues.
```

VGIG on this instruction: constraints met

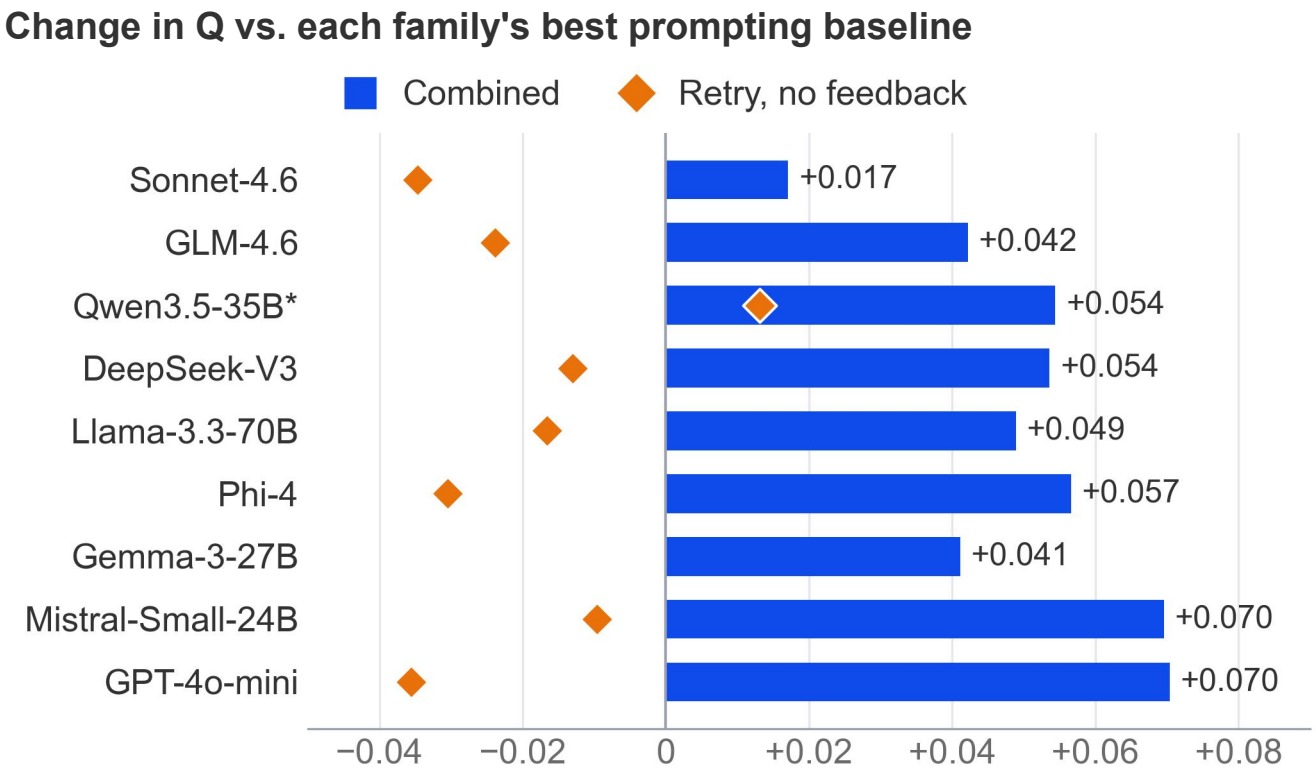


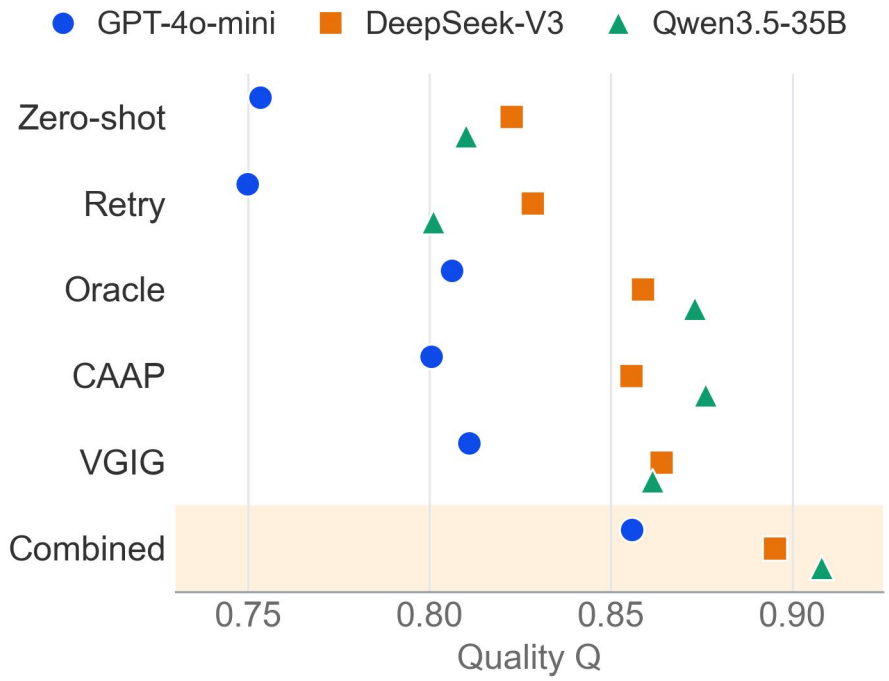
Each chain keeps its best round; chain 2 ends at 4/7.

The text is the checker's output for the case study's zero-shot sample. The chains are a separate VGIG-only run on GPT-4o-mini (no CAAP, fine feedback, T = 3), which drew its own first graphs.

Method

Combined beats the best prompting baseline in all 9 families





*Left chart only: Qwen3.5-35B Combined and Retry use thinking mode, its baseline does not; thinking is off on the right. Retry keeps all its draws and differs from Combined in more than feedback. Gemma-3-27B has no Retry run. Combined uses more tokens and calls.

Conclusion

Three takeaways

1

Report L2 and L5, not just aggregates

Cross-model dispersion peaks at multi-constraint composition (L2); averaging over constraint count hides most of the separation.

2

Tune prompting per level and per model

No strategy wins everywhere, and the CoT sign varies with family and capability. Zero-shot is the cheapest route to Q ≥ 0.8.

3

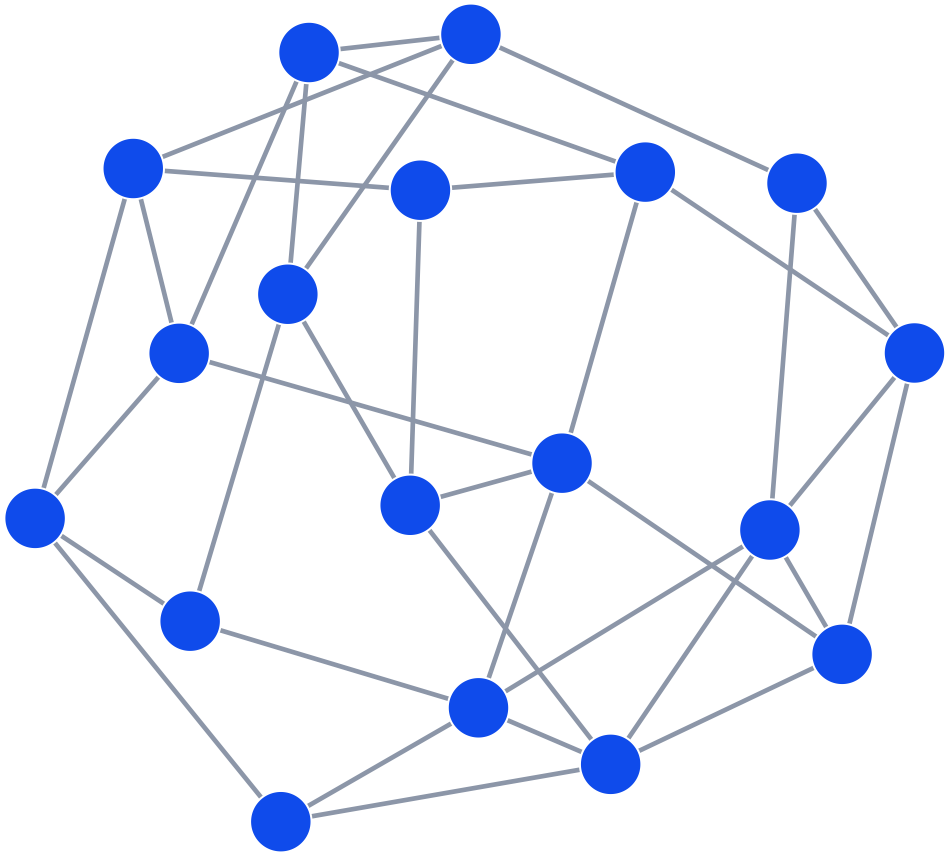
Use the benchmark's signals

D4 verification feedback, prompting chosen per level and L4 domain priors beat the best prompting baseline in all nine tested families.

Scope static, serialized, instruction-following graph generation with programmatically checkable constraints.

Caveats L5 (N = 50) statements are tentative; the L4 conclusion depends partly on D2, which is serialization-sensitive; verifier-guided methods may optimize toward the checker.

Next an L6 level for multi-graph reasoning; a serialization-robust semantic metric.



Thank you!

Zihe Wei, Tongji University
weizihe@tongji.edu.cn

Code & data

github.com/AI4DataSynth/GraphInstruct_formal

Archive

doi.org/10.5281/zenodo.21596997

NeurIPS 2026 · Evaluations & Datasets Track