

Peer Review

Review of: "GraphICL: Unlocking Graph Learning Potential in LLMs through Structured Prompt Design"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper argues that specialized graph LLMs have been compared against general-purpose LLMs that were prompted too naively, and it sets out to correct that with a prompt benchmark called GraphICL. A prompt is assembled from four parts (anchor node text, task description, structure-aware neighbor text, labeled demonstrations), with neighbors and demonstrations chosen at random, by cosine similarity, or by PageRank, and demonstrations drawn either globally or one per class. Appendix A.2 counts the resulting variants as $7 \times 7 + 1 \times 6 = 55$ templates. These are run with LLaMA2-13B-Chat, LLaMA3-70B-Instruct, and, on a subset, GPT-4o, on node classification over nine text-attributed graphs (Table 9) and on link prediction over Cora (Table 3). The headline claim, stated in the abstract and in Observations 1 and 2, is that training-free prompting of a general LLM beats state-of-the-art GNNs and graph LLMs in resource-constrained and cross-domain settings.

The secondary claims come from the ablations: structure-aware neighbor text helps (Table 5), similarity-based neighbor selection is best (Table 6), chain-of-thought does not help on average (Table 7), and stronger LLMs gain more (Table 4). I think the underlying question is a fair and useful one. The appendices are also more generous than usual: the full system prompts (Table 13), the user-content templates (Tables 15 and 16), the GICL codes of every reported configuration (Tables 8 and 11), and the extended baselines with three embedding types (Tables 10 and 11) are all printed. Much of what follows was checkable only because the authors printed those tables, and I want that on record.

2. Whether the conclusion holds on the data

One limitation of my reading first. As rendered on Qeios, Table 1 (in-domain node classification) and Table 2 (cross-domain node classification) are not legible; only their captions survive, and Figures 1 to 3 are images I could not inspect. I therefore could not verify the "average relative improvement of around 20%", the "39.88% increase on the Computers dataset compared to GraphPrompter" (Observation 1), or the "relative performance gap exceeding 101%" (Observation 2). Where Tables 10 to 12 repeat the same GraphICL rows, I worked from those.

2.1 The best prompt appears to be chosen on the test set. Table 1's caption defines S1 and S2 as "the first- and second-best GraphICL prompts", and the relative difference is computed against "the best results of our method". A.3.4, introducing Figure 3, speaks of "the best results from the 55 configurations tested". I could not find anywhere that a validation split was used to pick S1. Table 8 shows the winner changes from dataset to dataset (for LLaMA3-S1: 1RCP on Computers, 1RGP on Sports, 2SCS on PubMed, 1RGR on Cora). Taking the maximum of 55 test-set accuracies and comparing it with a single run of each baseline builds an optimistic bias into every headline number. The GNN baselines, by contrast, have a validation split available under the 6:2:2 and TAPE protocols described in A.3.1. If S1 was, in fact, selected on validation data, the paper needs to say so; if not, the comparison has to be redone with validation-based selection, or the mean over the 55 templates reported alongside the maximum.

2.2 "Outperform all GNNs" in the supervised setting is contradicted by the appendix. Observation 1 says: "Even in the supervised setting, GraphICL continues to outperform most graph LLMs and all GNNs." Table 10 (Supervised In-Domain Results) does not support this. On Cora, GraphICL-LLaMA3-S1 reaches 83.58, while SAGE with BoW reaches 88.93, GCN with GIA 88.39, and GCN with BERT, the embedding Table 1 uses, 86.90. On Sports, RevGAT with GIA reaches 92.94 against GraphICL's 91.39. On Arxiv, RevGAT with GIA reaches 74.74 and SAGE with GIA 74.07 against 73.68. The appendix datasets tell the same story more sharply. In Table 11 (Supervised In-Domain Results), RevGAT with GIA scores 83.33 on Photo against GraphICL-LLaMA3-S1 at 79.35, and SAGE with GIA scores 58.41 on Children and 85.12 on History, against 47.96 and 80.89. Even a supervised MLP on BERT embeddings (51.46 on Children, 82.41 on History) beats the best GraphICL prompt there. So by the paper's own numbers, GraphICL loses to supervised GNNs on six of nine datasets (Cora, Sports, Arxiv, Photo, Children, History) and wins on three (Computers, PubMed, Products). That is still an interesting result for a training-free method. It is not "all GNNs".

2.3 Table 1 uses the weakest embedding for the GNN baselines. The caption states "we uniformly use BERT embeddings" for MLP, GCN, RevGAT, and SAGE, and Table 10 shows that BERT is often the worst of the three embeddings. In the semi-supervised block for Computers, GCN goes from 70.24 (BERT) to 81.40 (GIA), and SAGE from 69.53 to 83.18. Against GraphICL-LLaMA3-S1 at 87.37, that moves the relative margin from $87.37/70.24 = 1.244$, i.e., +24.4%, down to $87.37/83.18 = 1.050$, i.e., +5.0%. GIA embeddings are the ones the authors themselves use for similarity-based selection (A.3.3), so they are not an exotic choice. The headline relative improvements should be computed against the best baseline in Table 10, not the BERT row.

2.4 The cross-domain comparison is not cross-domain for GraphICL. Table 12's GraphICL rows are identical, value for value, to Table 11's in-domain rows (Photo 79.35, Children 47.96, History 80.89), and its caption says "S1 and S2 are the same as Table 11". Those configurations (for Photo, "1RGS" and "1RCS") draw labeled demonstrations from the target dataset's own training set, and S1/S2 were picked per target dataset. The graph LLMs in the same table, per the Table 2 caption, saw no training data from the target. One side, therefore, gets in-domain labels and in-domain prompt selection, and the other gets none. A fair cross-domain row for GraphICL would use either zero-shot templates (XX demonstrations) or demonstrations taken from the source datasets only.

2.5 Several graph-LLM baselines score below chance, which points to answer parsing. In Table 12, LLaGA-HO scores 6.16 and GraphGPT 6.18 on Photo; LLaGA-ND scores 6.45 and GraphTranslator 6.64 on History. Photo has 12 classes (Table 9), so uniform guessing gives $1/12 = 8.33\%$. History has 13 classes by Table 9, giving $1/13 = 7.69\%$. A model that has any grasp of the task does not land below uniform guessing on a 12- or 13-class problem. The likelier explanation is that outputs were not mapped to label strings, for instance because these models were never shown the category names. The paper does not describe how free-text answers were parsed for any method. Until that is shown, the large gaps in Table 12 (my arithmetic: $79.35/38.96 = +103.7\%$ on Photo, $47.96/16.13 = +197.3\%$ on Children, $80.89/14.62 = +453.3\%$ on History) measure output-format failure at least as much as graph reasoning.

2.6 Label and edge leakage are not ruled out. Two places worry me. For node classification, §4.3 allows "labeled neighbor information" as demonstrations (the 1' and 2' codes, e.g., "1'SXX" for Computers, LLaMA2-S1, Table 8). On a transductive split, many neighbors of a test node are themselves test nodes. The paper does not say whether neighbor labels were masked unless the neighbor is in the training set. For link prediction, the reported configuration is "1SXX" (A.2), which puts the 1-hop neighbors of the first node into the prompt. If the positive test edge was not removed from the adjacency matrix before

neighbors were collected, the second node's own text sits in the neighbor list, and the task becomes string matching. I am not saying either leak happened. I could not find the text that says it did not, and the LP result in Table 3 depends on it.

2.7 The link-prediction evidence is thin and internally awkward. Link prediction is tested on one dataset (Cora, §5.2.3). In Table 3, the zero-shot LLaMA3 row already reaches 84.11, and GraphICL reaches 88.08, i.e., +3.97 points or $88.08/84.11 = +4.7\%$ relative. Table 5, meant to show that structure matters for LP, uses LLaMA3-70B-Instruct and reports 71.91 (XXCR), 75.00 (IRCR), and 79.70 (2RCR). Every one of those is below the zero-shot 84.11 in Table 3, so adding class-aware random demonstrations appears to cost about 12 points ($84.11 - 71.91 = 12.20$), and structure recovers part of the loss. That is a different finding from Observation 4, and it deserves discussion rather than a relative-percent column. The supervised margin over LLaGA-HO is $88.08 - 86.82 = 1.26$ points, and in the Arxiv+PubMed→Cora block, it is $88.08 - 87.35 = 0.73$ points. No test-set size, variance, or interval is given, so I cannot tell whether either margin exceeds noise. A small point: the LLaMA2, LLaMA3, and GraphICL rows are the same in all three blocks of Table 3. For a zero-label template, that is expected, but it means the "semi-supervised" and "cross-domain" labels do not describe anything GraphICL did differently.

2.8 No uncertainty is reported anywhere. None of Tables 1 to 12 carries a standard deviation, a confidence interval, a number of runs, or a test-set size per cell. Some of the conclusions rest on very small differences. Table 6 says similarity selection "consistently delivered the highest accuracy," but its margin over random is $68.76 - 68.45 = 0.31$ points on Cora and $67.63 - 67.25 = 0.38$ on PubMed; only Photo (59.06 vs. 50.30) shows a clear gap. Table 4 compares LLMs on 1,000 sampled nodes (§5.3.1). For Cora, GPT-4o at 76.60 has a binomial standard error of $\sqrt{(0.766 \times 0.234 / 1000)} = 0.0134$, so a 95% half-width of about $1.96 \times 0.0134 = 0.026$ points; LLaMA3 at 75.40 has nearly the same. The 1.2-point gap is well inside that. A paired test on per-node outputs would be sharper, and I would welcome it, but as reported, the Cora ordering of GPT-4o and LLaMA3 is not established. On Sports, the comparison is also confounded: GPT-4o uses random neighbor selection while both LLaMA models use PageRank (Table 4, column GraphICL).

2.9 The CoT ablation compares against prompts that already ask for step-by-step reasoning. §5.3.3 adds "Let's think step by step" as the CoT condition. Table 15, however, shows that every node-classification user template already contains "(Please reason step by step.)", and the Amazon system prompts in Table 13 require an answer followed by evidence. If the parenthetical was included in the "No" rows of Table 7, the ablation compares CoT with CoT. If it was removed, the printed templates are not the ones used. Table 7 also switches units: +2.95% and -3.69% there are absolute point differences ($78.41 -$

75.46; 67.16 – 70.85), whereas the percentages in Table 5 are relative ($63.10/30.36 = +107.8\%$). The model used for Table 7 is not named.

2.10 Several prompt details conflict with the method description. §4.1 says "we explicitly provide the names of all labels within the task description." For OGB-Arxiv, the system prompt (Table 13) only asks for the format "cs.XX", and for OGB-Products, the categories are "omitted due to length". For Amazon-History, Table 13 lists 12 categories, while Table 9 gives 13 classes; one class could never be predicted. And class-aware demonstrations mean "one demonstration is provided for each class label" (§4.2), while A.3.3 caps demonstrations at 6. Table 8 nevertheless uses class-aware codes on Computers (1RCP, 10 classes) and Arxiv (1RCP, 40 classes), and Table 11 on Photo (1RCS, 12 classes). The paper does not explain how 40 per-class demonstrations fit under a cap of 6.

Taken together: the claim that careful prompting narrows or reverses the gap to specialized graph LLMs in some settings is plausible and partly supported. The broader claims (beating all GNNs in the supervised setting, a large cross-domain advantage) exceed the evidence as presented.

3. Reproducibility

Credit first. The full prompts are printed (Tables 13 to 16), the GICL code of every reported configuration is given (Tables 8 and 11), the dataset statistics are listed (Table 9), the split protocol is described (A.3.1), and the inference stack is named: vLLM, one A100 80GB for LLaMA2-13B-Chat, and two for LLaMA3-70B-Instruct (A.3.2). Plenty of prompt-engineering papers give much less.

What is missing is the usual set for LLM papers, and here it is almost entirely absent.

Model versions. GPT-4o is cited only through the generic documentation link in footnote 3, with no dated snapshot; the alias has pointed to several different models since its release. The LLaMA footnotes link to product pages rather than to specific checkpoints. For the graph-LLM baselines, A.3.4 names the backbones (Vicuna-7b-v1.5-16k, LLaMA2-13b-Chat-HF, ChatGLM6B, Vicuna-7B-v1.5) but not the checkpoints of the graph models themselves.

Decoding. I could not find temperature, top-p, maximum output tokens, stop sequences, or seeds for any model. Nor is the number of runs per configuration stated. With Amazon prompts that request a free-text justification, output length matters.

Answer extraction. §5.2.1 says "only the single most likely label was predicted", but not how a free-text response such as "The book belongs to the [Category] category due to..." (Table 13) was converted to a

label, or what happened to unparseable outputs. Given the below-chance values in Table 12, this is the gap I would most like closed.

Data. A.3.1 says "we also sampled 5000 instances from the standard testing set" without naming the dataset (I assume OGB-Arxiv), and gives no sampling seed or list of node IDs. §5.3.1 samples 1,000 test nodes per dataset for Table 4, again without a seed. For Cora, this needs a word: if the supervised split follows a 60:20:20 ratio, the test set has about $0.2 \times 2,708 = 542$ nodes (Table 9), which is fewer than 1,000. The semi-supervised split could be meant instead, but the paper should say which. The link-prediction semi-supervised split "randomly sampled 5%" of training edges, again without a seed. Negative sampling for LP is not described beyond "an equal number of positive and negative samples".

Embeddings. Similarity selection uses "GIA embeddings" (A.3.3), while Table 3's GNNs use Sentence-BERT, and Table 1's use BERT. The specific encoder checkpoints are not named, and the embedding used for PageRank edges in LP ("the average of their embeddings", §4.3) is not specified.

Code and compute. I found no repository link, no release of predictions, and no GPU-hour or API-cost figures, even though §5.3.1 cites "testing costs" and §5.1 cites "budget constraints" as reasons for subsampling. Baseline hyperparameters are referred to Appendix A.3, but A.3 as rendered contains splits, environment, and prompt-size details, not training hyperparameters for MLP, GCN, SAGE, RevGAT, or the graph LLMs.

As it stands, I do not believe a third party could reproduce any single cell of Tables 1 to 7 to within a point.

4. What specifically to fix

1. **Table 1 caption, A.3.4 and Figure 3 (selection of S1/S2).** State whether S1 and S2 were chosen on the validation or the test split. If on test, reselect on validation for every dataset and recompute all relative improvements. Also report, per dataset, the mean and the spread of accuracy across the 55 templates, so a reader sees what an arbitrary GraphICL prompt achieves.
2. **Observation 1, sentence "Even in the supervised setting, GraphICL continues to outperform most graph LLMs and all GNNs".** Remove "all GNNs". Table 10 (Cora, Sports, Arxiv) and Table 11 (Photo, Children, History) show supervised GNNs ahead of GraphICL-LLaMA3-S1 on six of nine datasets; Section 2.2 above lists the cells.
3. **Table 1 and the abstract.** Compute relative gains against the strongest baseline per column across BoW, BERT, and GIA from Table 10, not the BERT row only.

4. **Table 2, Table 12, and Observation 2.** Rebuild the GraphICL cross-domain rows with no target-dataset labels (zero-shot or source-only demonstrations) and with a template fixed in advance rather than picked per target.
5. **A new subsection in §5.1 on answer parsing.** Describe the extraction procedure for every method, report the fraction of unparseable outputs per model and dataset, and explain the below-chance values in Table 12 (Photo: LLaGA-HO 6.16, GraphGPT 6.18; History: LLaGA-ND 6.45, GraphTranslator 6.64). Were the graph LLMs given the category names?
6. **§4.3 (NC few-shot and LP prompting).** Say explicitly that neighbor labels used as demonstrations come only from training nodes, and that target edges are removed from the graph before neighbor lists are built in link prediction.
7. **Table 3.** Give the number of test pairs. Add at least one more LP dataset (one of the Amazon graphs, for which a system prompt already exists in Table 14), since one dataset cannot carry the link-prediction half of the title claim.
8. **Table 5 and Observation 4.** Include the XXXX (pure zero-shot) row for Cora LP, which Table 3 puts at 84.11, and discuss why every few-shot configuration in Table 5 falls below it.
9. **All result tables.** Report the test-set size per cell and either repeated runs with standard deviation or binomial 95% intervals. Temper "consistently" in §5.3.3 (Table 6), where two of four datasets differ by under 0.4 points.
10. **Table 4 and Observation 3.** Use one GICL configuration for all three LLMs on Sports, or state that the comparison is best-per-model. The phrase "GPT-4o outperforms the other models by 1%-16%" should be qualified for Cora, where the 1.2-point gap is inside sampling error at $n = 1,000$.
11. **Table 7 and Table 15.** Reconcile the CoT ablation with the "(Please reason step by step.)" already in the templates. Name the model. Label the percentage column as absolute points.
12. **Table 13.** List the OGB-Arxiv and OGB-Products categories (in a supplementary file if needed); add the missing thirteenth Amazon-History class or correct Table 9.
13. **§4.2 versus A.3.3.** Explain what "class-aware" means when a dataset has more than 6 classes (Computers, Photo, Arxiv in Tables 8 and 11).
14. A reproducibility block in A.3.2: model snapshots (a dated GPT-4o identifier, Hugging Face revisions for LLaMA), vLLM version, temperature, top-p, max tokens, seeds, node-ID lists for the 1,000- and 5,000-node samples, GPU hours, and API spend. A code link, too.
15. **A.2, sentence on the 55 combinations.** The text says "1 structure-aware XX combined with 6 non-structure-aware demonstrations", but the six are the structure-aware demonstration codes (1'R to

2'S), as in "1'SXX". Swap the adjectives.

16. Smaller items. Table 10, semi-supervised PubMed: MLP-BERT and GCN-BERT are both 59.38; please check for a copy error. A.3.1 "we also sampled 5000 instances from the standard testing set" is missing the dataset name. In §5.1 and the Table 1 caption, GraphSAGE and RevGAT carry the same citation [3], and GCN is cited as [1]; the reference list suggests these are shifted. The article title on Qeios differs from the name given in §1 ("GraphICL: Benchmarking Large Language Models for Graph Reasoning via In-Context Learning"). Tables 1 and 2 render as unreadable on Qeios; a text version would help readers.

5. Recommendation

Major revisions. The question is worth asking, and the prompt catalogue is a genuine resource, especially with the templates printed in full. But the central comparisons need rework before the conclusions can stand. S1/S2 appear to be selected on test data. The appendix contradicts "all GNNs." The cross-domain rows give GraphICL in-domain labels. Several baselines sit below chance without an explanation of answer parsing. None of this looks unfixable, and most of the needed runs are cheap relative to what the authors have already done. I would be glad to read a revised version.

Declarations

Potential competing interests: No potential competing interests to declare.