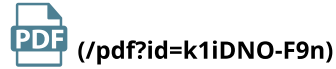



# InstructGraph: Boosting Large Language Models via Graph-centric Instruction Tuning and Preference Alignment



*Anonymous*

16 Feb 2024 ACL ARR 2024 February Blind Submission Readers:  Everyone Show

Revisions (/revisions?id=k1iDNO-F9n)

**Abstract:** Do current large language models (LLMs) better solve graph reasoning and generation tasks with parameter updates? In this paper, we propose `\textbf{InstructGraph}`, a framework that empowers LLMs with the abilities of graph reasoning and generation by instruction tuning and preference alignment. Specifically, we first propose a structured format verbalizer to unify all graph data into a universal code-like format, which can simply represent the graph without any external graph-specific encoders. Furthermore, a graph instruction tuning stage is introduced to guide LLMs in solving graph reasoning and generation tasks. Finally, we identify potential hallucination problems in graph tasks and sample negative instances for preference alignment, the target of which is to enhance the output's reliability of the model. Extensive experiments across multiple graph-centric tasks exhibit that InstructGraph can achieve the best performance and outperform GPT-4 and LLaMA2 by more than 13% and 38%, respectively.

**Paper Type:** long

**Research Area:** NLP Applications

**Contribution Types:** NLP engineering experiment, Publicly available software and/or pre-trained models, Data resources

**Languages Studied:** English

*Revealed to Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, Julian McAuley*


14 Feb 2024 (modified: 15 Feb 2024) ACL ARR 2024 February Submission


**Authors:** *Jianing Wang* (/profile?id=~Jianing\_Wang4), *Junda Wu* (/profile?id=~Junda\_Wu1), *Yupeng Hou* (/profile?id=~Yupeng\_Hou1), *Yao Liu* (/profile?id=~Yao\_Liu12), *Ming Gao* (/profile?id=~Ming\_Gao1), *Julian McAuley* (/profile?id=~Julian\_McAuley1)

**TL;DR:** InstructGraph: Boosting Large Language Models via Graph-centric Instruction Tuning and Preference Alignment

**Reassignment Request Action Editor:** This is not a resubmission

**Reassignment Request Reviewers:** This is not a resubmission

**Software:**  zip (/attachment?id=og80g\_X2RA&name=software)

**Data:**  zip (/attachment?id=og80g\_X2RA&name=data)

**Preprint:** yes

**Preprint Status:** There is a non-anonymous preprint (URL specified in the next question).

**Existing Preprints:** <https://arxiv.org/pdf/2402.08785.pdf> (<https://arxiv.org/pdf/2402.08785.pdf>)

**Consent To Share Data:** yes

**Consent To Review:** yes

**Consent To Share Submission Details:** On behalf of all authors, we agree to the terms above to share our submission details.

**A1:** yes

**A1 Elaboration For Yes Or No:** The "Limitations" section.

**A2:** yes

**A2 Elaboration For Yes Or No:** The "Social Impact and Ethics" section.

A3: yes

**A3 Elaboration For Yes Or No:** The "Abstract" and "Introduction" section.

B: no

B1: n/a

B2: n/a

B3: n/a

B4: n/a

B5: n/a

B6: n/a

C: yes

C1: yes

**C1 Elaboration For Yes Or No:** The "Experiments" section.

C2: yes

**C2 Elaboration For Yes Or No:** The "Experiments" section.

C3: yes

**C3 Elaboration For Yes Or No:** The "Experiments" section.

C4: yes

**C4 Elaboration For Yes Or No:** The "Experiments" section.

D: no

D1: n/a

D2: n/a

D3: n/a

D4: n/a

D5: n/a

E: no

E1: n/a

Add

Author-Editors Confidential Comment

Withdraw

Reply Type:  Author:

15 Replies

Visible To:  Hidden From:

## **[ - ] Meta Review of Paper1200 by Area Chair QrJc**

*ACL ARR 2024 February Paper1200 Area Chair QrJc*

08 Apr 2024, 00:26 ACL ARR 2024 February Paper1200 Meta Review Readers:

Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Authors, Paper1200 Reviewers Submitted, Program Chairs [Show Revisions \(/revisions?id=91AD3f5hS4\)](/revisions?id=91AD3f5hS4)

### **Paper Summary:**

The paper introduces InstructGraph, a framework designed for solving graph reasoning and generation tasks. Specifically, a structured format verbalizer is introduced to transform graph data into a code-like format.

### **Summary Of Strengths:**

1. The proposed method is novel and interesting.
2. The paper conducts a comprehensive array of experiments across various models and tasks.

### **Summary Of Weaknesses:**

1. some important baselines should be included in the main body of this paper instead of appendix.
2. sota results on individual tasks should be reported to show the potential of the proposed method.

**Overall Assessment:** 3 = There are major points that may be revised

**Best Paper Ae:** No

**Needs Ethics Review:** No

**Information Regarding The New ACL Policy On Deanonymized Preprints:** I confirm I have read the information above about changes to the anonymity policy.

Add

**Author-Editors Confidential Comment**

## [−] Official Review of Paper1200 by Reviewer tExi

*ACL ARR 2024 February Paper1200 Reviewer tExi*

22 Mar 2024, 11:49 ACL ARR 2024 February Paper1200 Official Review Readers:

Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200

Reviewers Submitted, Paper1200 Authors [Show Revisions \(/revisions?id=6NSorE9Lzp\)](/revisions?id=6NSorE9Lzp)

**Recommended Process Of Reviewing:** I have read the instructions above

### **Paper Summary:**

The paper introduces InstructGraph, a framework designed for solving graph reasoning and generation tasks. It comprises three key components: structured format verbalizer, graph instruction tuning, and graph performance alignment. Experimental results demonstrate a superior performance of this framework over other large language models (LLMs) in zero-shot and few-shot learning scenarios.

### **Summary Of Strengths:**

1. The paper demonstrates extensive engineering effort in processing benchmark datasets and providing instructions for various graph-related tasks.
2. The analysis of the cause of hallucinations.
3. The paper conducts a comprehensive array of experiments across multiple large language models (LLMs) and various learning settings, including few-shot and zero-shot learning.
4. The paper is easy to understand and follow.

### **Summary Of Weaknesses:**

1. Lack of novelty.
2. The ablation study section is problematic.
3. Experiment results lack sufficient explanation.
4. Some experiment settings are unclear.

### **Comments, Suggestions And Typos:**

1. The paper lacks novelty. This paper combines existing methodologies such as instruction tuning and direct preference optimization (DPO), which are widely used and well-understood. It is more about engineering than innovation.
2. Instead of comparing model performance by removing task clusters, the ablation study should explore scenarios such as degenerating graph input to plain text or removing components such as instruction tuning or graph preference alignment.
3. In addition to simply providing references about LLM's code understanding and generation ability, this paper needs to justify why converting a graph into code format is superior to other methods, such as flattening the graph into triples and plain text.
4. In the method section, this paper needs to justify why choosing instruction tuning instead of directly fine-tuning the LLM on downstream graph-related tasks. If LLM fine-tuning performance is not as good as instruction tuning, the paper should also include performance comparisons in the experiment section.
5. In the experiment section, the paper mainly describes the experiment result but lacks analysis explaining why InstructGraph-INS and InstructGraph-PRE outperform other models, such as a specific case that indicates one of the three components takes effect.
6. In Table 2, the paper should include the state-of-the-art model for each task to provide a comprehensive comparison, rather than only comparing InstructGraph-INS with other LLMs. In addition, an explanation for why LLaMa2 and Vicuna only achieved a score of 0 in certain tasks such as FB15K-237 should be included.

7. In the experiment section, the experiment setting of GPT-3.5, GPT4, LLaMa2, and Vicuna is not clear. In addition to the model size, the paper should mention whether graph input engineering, instruction tuning, or preference aligning is applied to those models.
8. In section 3.5, the dataset information such as AQuA is not mentioned anywhere else except In Table 4.
9. Typos:
  - An extra period in line 96 of the introduction section.
  - In Figure 2, there is an extra hyphen in the word "understand."
  - Inconsistent usage of quotation marks throughout the paper including the appendix.

**Soundness:** 2 = Poor: Some of the main claims/arguments are not sufficiently supported. There are major technical/methodological problems.

**Overall Assessment:** 2 = Revisions Needed: This paper has some merit, but also significant flaws, and needs work before it would be of interest to the community.

**Confidence:** 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

**Best Paper:** No

**Limitations And Societal Impact:**

None

**Ethical Concerns:**

None

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 3 = Potentially useful: Someone might find the new software useful for their work.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** tExi

Add

**Author-Editors Confidential Comment**

## **[–] Response to Reviewer tExi**

*ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~Jianing\_Wang4) (privately revealed to you)*

29 Mar 2024, 22:44 (modified: 29 Mar 2024, 22:50) ACL ARR 2024 February Paper1200 Official Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=mdZjovhhdL8)

**Comment:**

Thank you for your reviews.

Q1: The paper lacks novelty. It is more about engineering than innovation.

A1: Our innovation positioning is higher than engineering. We want to emphasize two main novelties:

- We designed a code-like format with task-specific instruction to enable the LLM to better perform graph reasoning and generation, which is more effective than graph description and GNN embedding fusion.
- We are the first to consider the hallucination on graph reasoning and generation, and propose multiple graph-oriented negative sampling for preference alignment.

Q2: The ablation study should explore scenarios such as degenerating graph input to plain text or removing components such as instruction tuning or graph preference alignment.

A2: Thank you for your suggestions. We have conducted the necessary analysis for these scenarios. The results for the "degenerating graph input to plain text" can be found in Section B.4 (Table 10) in the Appendix. As for the components removing settings, the results are available in Table 5 (two rows in bold). We will also include an analysis of this scenario.

Q3: This paper needs to justify why converting a graph into code format is superior to other methods.

A3: The reasons that we chose a code-like format are threefold:

- Explicitly utilizing code-like format input can activate the LLM to reuse the ability of code understanding and generation, making it easier to capture the structure and semantics of nodes (entities) and edges (triples) in the graph.
- The plain text and triple flattening strategies require different template rules for the graph in different scenarios (e.g., KB, RecSys, Science, etc.). The most critical drawback is that they can not support graph generation because the generated text is hard to be transformed into a graph.
- We have conducted the experiment in Table 10 to compare the effectiveness between code-like format and plain text. The results indicate that converting the graph into a code can achieve the best performance

Q4: This paper needs to justify why choosing instruction tuning instead of directly fine-tuning the LLM on downstream graph-related tasks.

A4: Different graph tasks have various paradigms and objectives. Traditional fine-tuning is limited to optimizing specific graph tasks independently. In contrast, instruction-tuning can standardize all graph tasks and NLP tasks into the same format, allowing for the reuse of the causal language modeling objective in LLM to realize modal alignment. In other words, instruction-tuning can effectively bridge the gap between pre-training and downstream graph tasks, making prior knowledge adaptation easier.

Similar findings are also suggested in other modal domains, such as symbolic with instruction-tuning [1] and visual + instruction-tuning [2].

[1] Fangzhi Xu, Zhiyong Wu, Qiushi Sun, etc.: Symbol-LLM: Towards Foundational Symbol-centric Interface For Large Language Models

[2] Wenliang Dai, Junnan Li, Dongxu Li, etc.: InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. NeurIPS 2023

Q5: Require a specific case that indicates one of the three components takes effect.

A5: Actually, we have completed a thorough analysis, which is detailed in Section B of the Appendix due to space constraints. For instance, the code format analysis is presented in Section B.4 (Table 10), and the specific case study of instruction-tuning and preference alignment is detailed in Section B.6 (Table 11).

Q6.1: The paper should include the state-of-the-art model for each task.

A6.1: Thank you for your review. We want to emphasize that this work mainly focuses on the standardization ability of LLM on graph reasoning and generation, which may not be comparable in some of the tasks due to different paradigms. In other words, SOTA methods of some graph tasks may not be suitable for the instruction-tuning paradigm, such as RecSys, NodeCLS, and Link Pred., so they cannot be directly optimized on the created corpus. In the next version, we will gather relevant SOTA methods to include the original results in the comparison in the Appendix.

Q6.2: Needs an explanation for why LLaMa2 and Vicuna only achieved a score of 0 in certain tasks such as FB15K-237.

A6.2: The results indicate that the tasks of Bipt. Match, Shrt. Path and FB15K-237 require a higher ability of structure understanding, which is a big challenge for current LLMs (e.g., LLaMA2 and Vicuna).

Q7: The experiment setting of GPT-3.5, GPT4, LLaMa2, and Vicuna is not clear.

A7: By default, these four baselines do not undergo any parameter update process. The inference settings remain the same as InstructGraph. It is important to note that the InstructGraph-INS is equivalent to LLaMA2 (or vicuna) instruction-tuning on 29 tasks.

Q8: The dataset information such as AQuA is not mentioned anywhere else except In Table 4.

A8: Thank you for your reminder. We will add the statistics of these datasets in the final version.

Add **Author-Editors Confidential Comment**

## [ - ] Further comments

*ACL ARR 2024 February Paper1200 Reviewer tExi*

02 Apr 2024, 23:35 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs,

Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200

Authors Show Revisions (/revisions?id=VSf6pc9Ojf)

### **Comment:**

Thank you for your detailed response. Please see my further comment.

A1:

- The code-like format seems straightforward as it primarily involves a simple step of data preprocessing or prompt engineering, even though the experiment result shows it is effective. The paper should provide the benefits of the code-like format compared to the flattening and other formats to establish the significance of the code-like format.
- Recent studies such as [1] put a lot of emphasis on hallucination in graph-to-text tasks, which is essentially caption generation. Therefore, the claim that "we are the first to consider the hallucination on graph reasoning and generation" is not valid.

[1] Shi, Xiao, et al. "Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs." EMNLP 2023.

A2:

- Table 10 should be merged into Table 5 to enhance clarity.
- For the components removing setting, the fundamental problem is the clusters are not completely different tasks. For instance, Graph Caption Generation (in GLM) is the reverse problem of Knowledge Graph Generation (in GGM), and Structure Graph Generation (in GGM) is the reverse problem of GSM. Therefore, in the ablation study, even if GSM is taken out during instruction tuning, the model was trained on Structure Graph Generation and thus will help inference on GSM tasks; similarly, if GGM is taken out during instruction tuning, the model being trained on GSM and Graph Caption Generation could be comfortably in inference on GGM tasks.

A3: The provided reasons should be included in the paper to underscore the significance of the code-like format.

A4: Thank you for your explanation. However, in addition to providing the related works, the paper should provide experiment results to show the proposed InstructGraph with instruction-tuning is better than supervised fine-tuning.

A6.1: The paper needs to compare the InstructGraph results with the SOTA method on the same task.

A7:

- I am not asking if there is any parameter update for the baselines in the experiment setting. My concern is whether the paper uses the code-like format or graph preference alignment for the GPT-3.5, GPT4, LLaMa2, and Vicuna. If not, the paper should explain the graph input format for those models.
- InstructGraph is instruct-tuned on the 1.6M examples from 29 tasks, which are the same datasets for evaluation in Table 2, so it is not zero-shot. The paper should clarify the zero-shot setting.

**[-] Response to the further comments from Reviewer tExi**

ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~Jianing\_Wang4) (privately revealed to you)

03 Apr 2024, 02:47 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=QYGRJI8DKY)

**Comment:**

Thank you very much for your comments.

Reply for A1:

- We conducted experiments to compare our proposed code-like format with other formats such as plain text with the template. In Table 10, we present the results, which demonstrate the advantages of our proposed format. The flattening strategy is similar to plain text as it concatenates the head entity, relation, and tailed entity into a sequence, so we can still obtain similar suggestions. We will take the suggestion of the reviewer and add a new baseline named 'triple flattening' in Table 10.
- The reviewer's mentioned reference [1] focuses solely on natural language generation, which is not our only area of focus. We want to clarify that we are considering hallucination in both graph reasoning tasks (such as classification, link prediction, and RecSys, etc.) and graph generation tasks (such as knowledge graph and structure graph generation). Therefore, we assert that we are the first to consider hallucination in both graph reasoning and generation.

[1] Shi, Xiao, et al. "Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs." EMNLP 2023.

Reply for A2:

- Thank you very much for your suggestions, we will merge the results of Table 5 and Table 10 in the final version.
- We would like to state that the ablation experiment aims to explore whether reasoning or generation abilities (i.e. GSM and GLM for reasoning, GGM and GTM for generation) can enhance the LLM on graph-centric tasks. For instance, in Table 5, by using only the corpus from the GSM cluster for instruction tuning, we can see that the 'w/ GSM' can improve the LLaMA2 on multiple graph-centric tasks. Furthermore, we acknowledge that the same corpus can influence the LLM on the corresponding tasks, but the task paradigms of graph reasoning and generation are quite distinct. We believe that even if the corpus is used, the training paradigm and objectives are different, making the improvement challenging.

Reply for A3:

- Thank you for your suggestions, we will add the reason in the final version.

Reply for A4:

- We will add the task-specific baselines for supervised fine-tuning, for example, fine-tuning an LLaMA2 by only using the training set of WebNLG.

Reply for A6:

- InstructGraph is a framework that aims to provide advanced graph reasoning and generation capabilities to LLMs. Our current experiments demonstrate the effectiveness of our method and we will also include a comparison with relevant state-of-the-art methods in the Appendix.

Reply for A7:



- By stating that "the inference settings remain the same as InstructGraph", we mean that we use the same code-like format and graph preference alignment for GPT-3.5, GPT4, LLaMa2, and Vicuna.
- Thank you for your comments. In our paper, we have used a zero-shot setting for both the baselines and InstructGraph. This means that neither of the methods has been provided with any data samples for the zero-shot tasks. As a result, we believe that the setting is fair and can be strictly considered as zero-shot. To avoid any further confusion, we will provide more detailed descriptions of this zero-shot setting.

Add **Author-Editors Confidential Comment**

### **[ - ] Request Reply from Reviewer tExi**

*ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~jianing\_wang4)  
(privately revealed to you)*

02 Apr 2024, 02:51 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs,  
Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200  
Authors Show Revisions (/revisions?id=z2u9\_yYmMh1)

#### **Comment:**

Dear Reviewer tExi:

Thank you very much for your review. I haven't received your reply yet. I believe your response will greatly assist our work. We are looking forward to hearing from you and discussing it further.

Please let me know if you have any questions.

Best regards

Authors.

Add **Author-Editors Confidential Comment**

### **[ - ] Official Review of Paper1200 by Reviewer NpW3**

*ACL ARR 2024 February Paper1200 Reviewer NpW3*

20 Mar 2024, 19:14 (modified: 20 Mar 2024, 19:28) ACL ARR 2024 February Paper1200

Official Review Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area  
Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?  
id=jxGDLG0eIe)

**Recommended Process Of Reviewing:** I have read the instructions above

#### **Paper Summary:**

This paper first constructs a novel code-like structure to use text sequence to represent the graph as prompts for LLMs and then designs several instruction prompts for different graph tasks. This paper also leverages DPO to reduce the hallucination in graph tasks.

#### **Summary Of Strengths:**

This paper has enough novelty in the methods to publish. 1. They constructed a novel code-like graph structure that can be used in various graph tasks by LLMs in-context learning. 2. They designed several instruction prompts for different graph tasks. 3. They used DPO to reduce the hallucination in graph tasks. The paper has solid results on various graph tasks compared to the LLM baselines.

#### **Summary Of Weaknesses:**

Lack of comparison with previous methods for graphs using LLMs. This paper only compares with LLM baselines.



1. It would be interesting to know the results compared to graph description or embedding fusion, as this paper listed them as previous methods.
2. It would also be interesting to know how your approach compares to the SOTA methods for each task.

**Comments, Suggestions And Typos:**

N/A

**Soundness:** 4 = Strong: This study provides sufficient support for all of its claims/arguments. Some extra experiments could be nice, but not essential.

**Overall Assessment:** 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.

**Confidence:** 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

**Best Paper:** No

**Ethical Concerns:**

N/A

**Needs Ethics Review:** No

**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 3 = Potentially useful: Someone might find the new software useful for their work.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** NpW3

Add

Author-Editors Confidential Comment

**[ - ] Response to Reviewer NpW3**

*ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~Jianing\_Wang4) (privately revealed to you)*

29 Mar 2024, 22:45 (modified: 29 Mar 2024, 22:50) ACL ARR 2024 February  
Paper1200 Official Comment Readers: Program Chairs, Paper1200 Senior Area  
Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200  
Authors Show Revisions (/revisions?id=yyGJ2e-c3Lz)

**Comment:**

Thank you for your reviews.

Q1: It would be interesting to know the results compared to graph description or embedding fusion

A1: Thank you for your suggestions. We have conducted the comparison with the naive graph description in Section B.4 (Table 10) in the Appendix due to the space limitation.

Q2: It would also be interesting to know how your approach compares to the SOTA methods for each task.

A2: Thank you for your review. We want to emphasize that this work mainly focuses on the standardization ability of LLM on graph reasoning and generation, which may not be comparable in some of the tasks due to different paradigms. In other words, SOTA methods of some graph tasks may not be suitable for the instruction-tuning paradigm, such as RecSys, NodeCLS, and Link Pred., so they cannot be directly optimized on the created corpus. In the next version, we will gather relevant SOTA methods to include the original results in the comparison in the Appendix.

Add

Author-Editors Confidential Comment

## [-] Comments after Author Discussion

ACL ARR 2024 February Paper1200 Reviewer NpW3

02 Apr 2024, 00:51 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=U5GUMBAUI8g)

### Comment:

Thanks for your reply, I decided to keep the rating.

Add

Author-Editors Confidential Comment

## [-] Response to Reviewer NpW3

ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~jianing\_Wang4) (privately revealed to you)

02 Apr 2024, 02:43 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=bflmvgHHoY)

### Comment:

Thank you very much.

Add

Author-Editors Confidential Comment

## [-] Official Review of Paper1200 by Reviewer mbWq

ACL ARR 2024 February Paper1200 Reviewer mbWq

15 Mar 2024, 13:45 (modified: 23 Mar 2024, 09:18) ACL ARR 2024 February Paper1200

Official Review Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=Z0YV-uX6-d)

**Recommended Process Of Reviewing:** I have read the instructions above

### Paper Summary:

This paper proposes a framework called InstructGraph to empower large language models (LLMs) with the ability to solve graph reasoning and generation tasks. In this paper, a structured format verbalizer is introduced to transform graph data into a code-like format that can be easily understood by LLMs. This bridges the gap between graph data and textual LLMs. Additionally, the authors collect 29 different tasks into an instruction dataset. The LLM is continually tuned on this dataset to improve its performance on graph tasks: graph structure modeling, graph language modeling, graph generation modeling, and graph thought modeling. Besides, to mitigate hallucinations in graph tasks, a preference alignment is applied.

### Summary Of Strengths:

Firstly, the work provides a systematic way to enhance the abilities of LLMs in graph reasoning and generation, which is a relatively underexplored area. Second, the structured format verbalizer to unify graph data into a code-like format understandable by LLMs, bridging the gap between graph and text data. Third, a wide range of graph-centric tasks, including structure modeling, language modeling, generation modeling, and thought modeling for graphs are used. Finally, the results demonstrate better performance, outperforming baselines like GPT-4 and LLaMA2.

### Summary Of Weaknesses:

1.While the paper acknowledges the potential for hallucinations in graph tasks, there is limited analysis or discussion of specific cases; 2. The paper compares InstructGraph primarily to other LLMs like GPT-4 and LLaMA2. However, a more comprehensive comparison to state-of-the-art graph learning methods that use techniques like graph neural networks,

graph embeddings, or graph-specific architectures would provide a better understanding of the relative strengths and weaknesses of the proposed approach

**Comments, Suggestions And Typos:**

None

**Soundness:** 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.

**Overall Assessment:** 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

**Best Paper:** No

**Ethical Concerns:**

None

**Needs Ethics Review:** No

**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

**Datasets:** 3 = Potentially useful: Someone might find the new datasets useful for their work.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** Yes

**Knowledge Of Paper:** After the review process started

**Knowledge Of Paper Source:** Preprint on arxiv

**Impact Of Knowledge Of Paper:** Not at all

**Reviewer Certification:** mbWq

Add

**Author-Editors Confidential Comment**

**[ - ] Response to Reviewer mbWq**

*ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~Jianing\_Wang4) (privately revealed to you)*

29 Mar 2024, 22:47 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200

Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions

(/revisions?id=upwe0nDfFR)

**Comment:**

Thank you for your reviews.

Q1: While the paper acknowledges the potential for hallucinations in graph tasks, there is limited analysis or discussion of specific cases;

A1: Thank you for your suggestions. We showcase some potential hallucination settings to simulate the wrong answer in Table 7.

Because graph reasoning and generation are essentially text generation, so the hallucination consideration in this paper depends on the combination of existing text generation (e.g., dialogue systems, summarization) and the characteristics of the graph. We will extend the case study in the final version.

Q2: A more comprehensive comparison to SOTA graph learning methods that use techniques like graph neural networks, graph embeddings, or graph-specific architectures would provide a better understanding of the relative strengths and weaknesses of the proposed approach

A2: Thank you for your suggestions. We want to emphasize that this work mainly focuses on the standardization ability of LLM on graph reasoning and generation, which may not be comparable in some of the tasks due to different paradigms. In other words, SOTA methods of some graph tasks may not be suitable for the instruction-

tuning paradigm, such as RecSys, NodeCLS, and Link Pred., so they cannot be directly optimized on the created corpus. In the next version, we will gather relevant SOTA methods to include the original results in the comparison in the Appendix.

Add **Author-Editors Confidential Comment**

### **[ - ] Comments after Author Discussion**

*ACL ARR 2024 February Paper1200 Reviewer mbWq*

31 Mar 2024, 05:35 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=feww47LmJB0)

**Comment:**

Thanks for your responses. I appreciate you would consider improving your paper by including a case study and relevant results in your final version.

Add **Author-Editors Confidential Comment**

### **[ - ] Response to Review mbWq**

*ACL ARR 2024 February Paper1200 Authors Jianing Wang (/profile?id=~jianing\_Wang4) (privately revealed to you)*

31 Mar 2024, 05:41 ACL ARR 2024 February Paper1200 Official

Comment Readers: Program Chairs, Paper1200 Senior Area Chairs, Paper1200 Area Chairs, Paper1200 Reviewers Submitted, Paper1200 Authors Show Revisions (/revisions?id=3z2p21KWwQO)

**Comment:**

Thank you so much for your reply.


Add **Author-Editors Confidential Comment**


### **[ - ] Supplementary Materials by Program Chairs**

*ACL ARR 2024 February Program Chairs*

16 Feb 2024, 13:45 ACL ARR 2024 February Paper1200 Supplementary

Materials Readers: Program Chairs, Paper1200 Reviewers, Paper1200 Authors, Paper1200 Area Chairs, Paper1200 Senior Area Chairs Show Revisions (/revisions?id=\_pj4PgPGrE)

**Software:**  zip (/attachment?id=\_pj4PgPGrE&name=software)

**Data:**  zip (/attachment?id=\_pj4PgPGrE&name=data)

**Reassignment Request Action Editor:** This is not a resubmission

**Reassignment Request Reviewers:** This is not a resubmission

**A1:** yes

**A1 Elaboration For Yes Or No:** The "Limitations" section.

**A2:** yes

**A2 Elaboration For Yes Or No:** The "Social Impact and Ethics" section.

**A3:** yes

**A3 Elaboration For Yes Or No:** The "Abstract" and "Introduction" section.

**B:** no

**B1:** n/a

**B2:** n/a

**B3:** n/a

**B4:** n/a

**B5:** n/a

**B6:** n/a

**C:** yes

**C1:** yes

**C1 Elaboration For Yes Or No:** The "Experiments" section.

**C2:** yes

**C2 Elaboration For Yes Or No:** The "Experiments" section.

**C3:** yes

**C3 Elaboration For Yes Or No:** The "Experiments" section.

**C4:** yes

**C4 Elaboration For Yes Or No:** The "Experiments" section.

**D:** no

**D1:** n/a

**D2:** n/a

**D3:** n/a

**D4:** n/a

**D5:** n/a

**E:** no

**E1:** n/a

**Note From EiCs:** These are the confidential supplementary materials of the submission. If you see no entries in this comment, this means there haven't been submitted any.

Add

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Sponsors \(/sponsors\)](/sponsors)

[Frequently Asked Questions](#)

[\(https://docs.openreview.net/getting-started/frequently-asked-questions\)](https://docs.openreview.net/getting-started/frequently-asked-questions)

[Contact \(/contact\)](/contact)

[Feedback](#)

[Terms of Use \(/legal/terms\)](/legal/terms)

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2024 OpenReview