

---

# Supplementary Material

---

Anonymous Authors

## 1 Download URL

You can download the datasets from Dropbox Link.

## 2 Motivation

### 1. For what purpose was the dataset created?

These datasets are intended to study the ability of large language models to solve graph-based tasks using Python API calls. Previous work has not explored using APIs for graph-based tasks, instead relying on direct responses from large language models.

### 2. Any other comments?

None.

## 3 Composition

### 1. What do the instances that comprise the dataset represent?

The benchmark dataset's instance is a test case for LLM to evaluate the ability to use API to solve graph foundation tasks.

The document dataset's instance is API info.

The Code(QA) dataset's instance is a QA pair for finetuning open-source models.

The Doc+Code(QA) dataset's instance is a QA pair for finetuning open-source models.

### 2. How many instances are there in total (of each type, if appropriate)?

The benchmark dataset contains 512 instances.

The document dataset contains 1413 instances.

The Code(QA) dataset contains 29,260 instances.

The Doc+Code(QA) dataset contains 29,260 instances.

Table 1: Statistics of datasets.

	Benchmark	Document	Code (QA)	Doc+Code (QA)
Category 1	312	1,115	23,324	23,324
Category 2	154	253	5,136	5,136
Category 3	46	45	800	800
Total	512	1,413	29,260	29,260

### 3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

It contains all possible instances.

### 4. What data does each instance consist of?

For the benchmark dataset, each instance includes a human-expert-annotated question, a

new role-play rephrased question from GPT-4-turbo-0409, the corresponding answers, the reference code for those answers, the problem type (True/False, calculations, draw, multi), the key APIs needed to solve the question, the number of lines in the reference code (excluding empty lines), the category of the question (basic graph theory, graph statistical learning, graph embedding), and the difficulty level of the question.

For the document dataset, each instance contains the API info for an API.

For the Code(QA) dataset, each instance contains a question, a reference code corresponding to the question, and an API name.

For the Doc+Code(QA) dataset, each instance contains a question, a reference code corresponding to the question, and an API name.

**5. Is there a label or target associated with each instance? If so, please provide a description.**

For the benchmark dataset, the targets are the third column in the data.

For the document dataset, there is no target or label with each instance.

For the Code(QA) dataset, the targets are the second column in the data.

For the Doc+Code(QA) dataset, the targets are the second column in the data.

**6. Is any information missing from individual instances?**

None.

**7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

None.

**8. Are there recommended data splits (e.g., training, development/validation, testing)?**

We expect to use the benchmark dataset for testing, and other 3 datasets for enhancing or finetuning.

**9. Are there any errors, sources of noise, or redundancies in the dataset?**

There are some instances that the code can't run properly in Code(QA) dataset and Doc+Code(QA) dataset, but these instances can be used for finetuning open-source models to enhance the ability to solve graph foundation tasks.

**10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

The dataset is self-contained.

**11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**

No.

**12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

No.

**13. Does the dataset relate to people?**

Yes. The role-play rephrased questions contain profession information.

**14. Does the dataset identify any subpopulations (e.g., by age, gender)?**

No.

**15. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

No.

**16. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

No.

78 17. **Any other comments?**  
79 No.

## 80 4 Collection Process

81 You can see the details in the paper.

## 82 5 Uses

- 83 1. **Has the dataset been used for any tasks already?**  
84 The dataset has been used for evaluating LLMs' ability to solve graph foundation tasks  
85 using APIs.
- 86 2. **What (other) tasks could the dataset be used for?**  
87 The dataset could possibly be used for other LLM's evaluations and enhancements.
- 88 3. **Are there tasks for which the dataset should not be used?**  
89 None.
- 90 4. **Any other comments?**  
91 None.

## 92 6 Distribution

- 93 1. **Will the dataset be distributed to third parties outside of the entity (e.g., company,  
94 institution, organization) on behalf of which the dataset was created?**  
95 The dataset will be open-sourced later, but commercial use is prohibited.
- 96 2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**  
97 The dataset is free for download at dropbox now, and will be possible at  
98 github.com/huggingface.com later.
- 99 3. **Will the dataset be distributed under a copyright or other intellectual property (IP)  
100 license, and/or under applicable terms of use (ToU)?**  
101 The dataset is licensed under a CC BY-NC 4.0 license.
- 102 4. **Have any third parties imposed IP-based or other restrictions on the data associated  
103 with the instances?**  
104 Not to our knowledge.
- 105 5. **Do any export controls or other regulatory restrictions apply to the dataset or to  
106 individual instances?**  
107 Not to our knowledge.
- 108 6. **Any other comments?**  
109 None.

## 110 7 Maintenance

- 111 1. **Who is supporting/hosting/maintaining the dataset?**  
112 All the authors.
- 113 2. **Is there an erratum?**  
114 None.
- 115 3. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete  
116 instances')?**  
117 Yes. We will update our datasets when we open-source all content on Github.

- 118 4. **If the dataset relates to people, are there applicable limits on the retention of the data**  
119 **associated with the instances (e.g., were individuals in question told that their data**  
120 **would be retained for a fixed period of time and then deleted)?**  
121 No.
- 122 5. **Any other comments?**  
123 None.