Supplementary Material
GRAB: A Challenging GRaph Analysis Benchmark for Large Multimodal Models

Appendix

We structure this Appendix into 9 parts. In the first part (§A), we include curation details for the Screenshots and Noise splits of the *Real* task questions. In the next four parts, we report additional results and ablations, including a question format ablation (§B), a comparison of different performance metrics (§C), model performance results on an OCR-based graph analysis question set (§D), and a graph plotting libraries performance ablation (§E) and display example plots with different libraries. Finally, we detail the inference settings used for GRAB-Lite evaluation (§F), include examples of the prompts used for inference and LLM evaluation (§G) along with details of the compute costs of our work (§H) and the specific API model versions used throughout this evaluation (§I).

A. *Real* task curation details

A.1. Screenshots

To construct the Screenshots split, after creating the question graphs, we randomly selected one of the following digital contexts to embed the image:

- Presentations (Google Slides, PowerPoint)
- Video Calling (Teams, Zoom, Meet)
- Image Viewing (Preview)
- IDE (VSCode)
- Email
- Word Processing (Word, Google Docs)

We curated a set of different designs for each digital context (screenshot), including different combinations of background applications/OS. To each, we added an arbitrarily coloured box to the screenshots, denoting where the graph would be embedded. Once created, the coloured box was then replaced with the graph and the composite figure saved and used as part of the question.

A.2. Noise

To construct the Noise split, after initially creating the graphs, we randomly applied one of the following types of noise to the image:

- Gaussian noise
- Salt and pepper noise
- Brightness/contrast
- Blur
- Spatter/smear
- JPEG artifacts
- Rotations
- Flips

B. Question format ablation

Model	Accuracy (%)		
	Single-answer		Multiple-choice
Random chance	-	-	20.0
Closed-source LMMs			
Claude 3 Haiku [2]	14.2	+9.4	23.6
Claude 3 Sonnet [2]	15.3	+10.5	25.8
Claude 3.5 Sonnet [3]	**41.8**	-10.0	31.8
GPT-4 Turbo [10]	18.5	+11.3	29.8
GPT-4o mini [12]	15.8	+12.2	28.0
GPT-4o [11]	24.7	+3.0	27.7
Gemini 1.0 Pro Vision [5]	20.2	-7.5	12.7
Gemini 1.5 Flash [18]	28.5	+2.4	30.9
Gemini 1.5 Pro [18]	34.2	-1.9	**32.3**
Reka Edge [16]	11.8	+11.5	23.3
Reka Flash [16]	13.2	+7.6	20.8
Reka Core [16]	1.7	+5.6	7.3
Open-source LMMs			
CogVLM-Chat [19]	7.0	+14.1	21.1
Qwen-VL-Chat [4]	**10.2**	+11.3	21.5
OmniLMM-3b [15]	6.7	+13.6	20.3
TransCore-M [17]	7.9	+12.9	20.8
Yi-VL-6b [1]	5.6	-1.5	4.1
Yi-VL-34b [1]	7.6	+5.6	13.2
LLaVA-1.5 7b [9]	4.7	+14.2	18.9
LLaVA-1.5 13b [9]	5.0	+17.4	**22.4**

Table 1. **Accuracy on the *Properties* task with different question formats**. Score differences between the formats are shown as coloured text. The highest open and closed-source model scores for each format are highlighted and **bold**.

We carry out an ablation analysing the effect of question format on model performance by re-evaluating the questions from the *Properties* task in a multiple-choice setting with 5 candidate answers. The results of this comparison are displayed in Tab. 1. Multiple-choice options were generated adversarially by sampling values close to the correct answer.

For the majority of models, accuracy scores are higher in the multiple-choice setting, and most models score above the random chance score. However, the highest attained score is only 32.3%, further reflecting the difficulty of the GRAB benchmark. For a few models, including the two leading models, the opposite is true. In these cases, it is possible that the presence of plausible incorrect answers (*i.e.*, close to the true answer) can cause confusion and lead the models to make incorrect selections.

C. Additional metrics

For a broader view, we report some additional metrics in Tab. 2. In addition to the accuracy scores reported in the main paper (pass@1), here we also report pass@5, 5/5 reliability, root mean squared error (RMSE) and mean absolute error (MAE). These additional metrics largely preserve the pass@1 model rankings, though there is some variation. Our core analysis focuses on pass@1 accuracy as each alternative metric has limitations, making use on GRAB unfeasible. While the pass@5 and 5/5 reliability metrics provide insights into model performance outside of near-deterministic settings, running each evaluation on 3284 questions five times is impractical. Although most GRAB answers are numeric, distance-based error metrics, such as RMSE and MAE are problematic due to differing scales, ground-truth equal to 0 and treating non-numeric LMM answers.

	Performance metric				
	pass@1 ↑	*pass@5* ↑	*5/5* ↑	*RMSE* ↓	*MAE* ↓
Claude 3.5 Sonnet	18.6	22.7	14.0	297.5	29.7
Gemini 1.5 Pro	**21.9**	**28.9**	**14.7**	**209.9**	28.3
GPT-4o	18.0	26.0	8.3	226.9	**27.7**

Table 2. **Performance metrics on *Real* Screenshots split.**

D. OCR experiments

We construct a small set of 35 OCR questions based on text in the legend, title, and axis labels. The graphs are created using a similar pipeline to the *Properties* task except rather than using purely randomly selected data generation processes, we use plausible data ranges and functions for the axis label. We leverage GPT-4 Turbo to construct a database of 'realistic' dependent and independent variables along with their ranges, distributions and directions (*e.g.* linear, increasing). We evaluate these questions in a multiple-choice setting with 5 options and present the results in Tab 3. Compared to performance on the Properties task with multiple-choice options (Tab. 1), the models attain much higher scores on these OCR-style questions, with many models achieving either 100% or close to 100% accuracy. These high scores suggest this particular question type is too easy for current frontier LMMs, therefore we refrain from including it in GRAB.

Model	**Accuracy (%)**
Random chance	20.0
Closed-source LMMs	
Claude 3 Haiku [2]	42.9
Claude 3 Sonnet [2]	91.4
Gemini 1.0 Pro Vision [5]	**100.0**
Gemini 1.5 Flash [18]	**100.0**
Gemini 1.5 Pro [18]	**100.0**
GPT-4 Turbo [10]	97.1
GPT-4V [10]	97.1
GPT-4o [11]	82.9
Reka Core [16]	97.1
Reka Edge [16]	88.6
Reka Flash [16]	**100.0**
Open-source LMMs	
TransCore-M [17]	**80.0**
Yi-VL-6b [1]	51.4
OmniLMM-3b [15]	65.7
Qwen-VL-Chat [4]	71.4

Table 3. **Accuracy scores on OCR experiments**.

E. Plotting libraries ablation

To compare different plotting libraries, we sample 100 GRAB questions and synthesise them using both Matplotlib [8] and the Seaborn [20] library to create two sets of questions that differ in the aesthetics and styles of the libraries but are otherwise identical. The near agreement of Claude 3.5 Sonnet's scores on these sets (Tab. 4) suggests that plotting with different libraries does not significantly impact overall performance; therefore, we focus on Matplotlib plots for GRAB. However, to increase the diversity of GRAB-Lite, we re-generate half of the synthetic plots using Seaborn. Fig. 1 displays examples of identical functions and series plotted with both the Matplotlib and Seaborn libraries.

	Accuracy (%)				
Plotting Library	*Properties* (25)	*Functions* (25)	*Series* (25)	*Transforms* (25)	**Overall** (100)
Matplotlib [8]	**48.0**	16.0	**36.0**	**20.0**	**30.0**
Seaborn [20]	40.0	**20.0**	24.0	**20.0**	26.0

Table 4. **Claude 3.5 Sonnet accuracy on different plotting libraries** on a 100-question subset of GRAB.

F. GRAB-Lite inference

For inference on the GRAB-Lite questions we evaluated o1 [13] and Gemini 2.5 Flash [6] via the OpenAI API [14] and Gemini Developer API [7], respectively. Greedy decoding was used. The 'reasoning effort' parameter for o1 was set to 'high'.

G. Prompts

Inference prompt:

```
<question>\n Only provide the answer, no
reasoning steps.  If you are unsure, still
provide an answer.  Answer:\n '
```

LLM evaluation prompt:

```
A generative model has been asked this
question:  "<question>" about a plot.\n The
output from the model answering the question
is:  "<output>"\n Extract just the answer
from the generative model output.  Maintain
the same precision given by the model.
Convert any numbers to digits (e.g., "one"
to "1").  Remove any additional terms like
'approximately'.  Return only the extracted
numeric answer, without any additional text or
explanation.  If no answer is provided, return
"None".
```

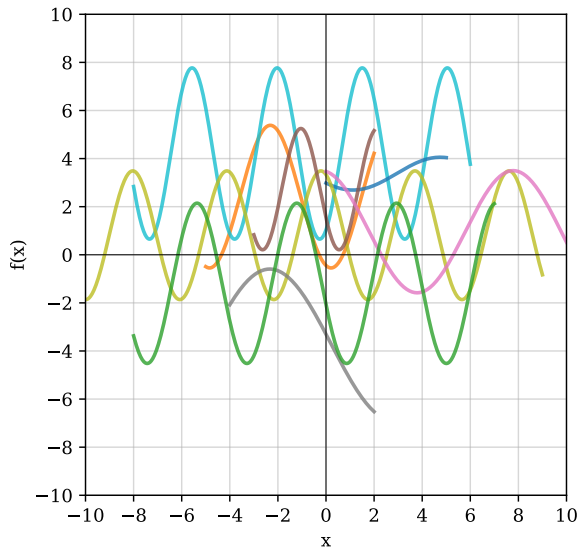
H. Compute

As we restricted the size of our benchmark to 3284, the total compute required for this work is relatively low. Negligible compute was needed to create the synthetic images. Inference with the closed-source models was carried out via API calls. Using a single NVIDIA A100-80GB GPU, inference on the entire GRAB benchmark using LLaVA-1.5 7b can be carried out in approximately 45 minutes, using a single process (inference time is significantly reduced with multiprocessing).

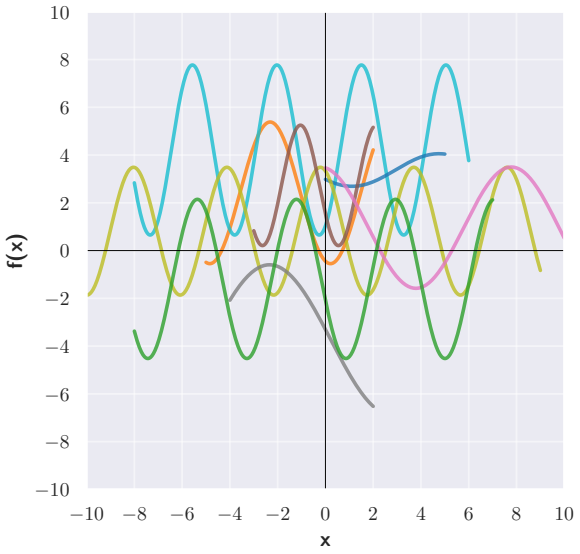
I. API model versions

These are the specific versions of the API models used in this work:

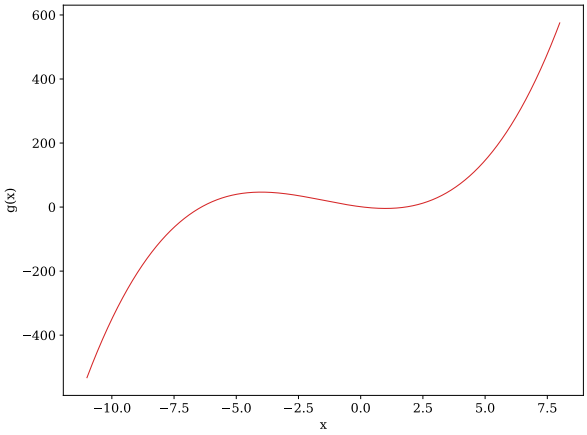
- GPT-4 Turbo:
  gpt-4-turbo-2024-04-09
- GPT-4o mini:
  gpt-4o-mini-2024-07-18
- GPT-4o:
  gpt-4o-2024-05-13
- o1:
  o1-2024-12-17
- Gemini Pro Vision:
  gemini-1.0-pro-vision-001
- Gemini 1.5 Flash:
  gemini-1.5-flash-preview-0514
- Gemini 1.5 Pro:
  gemini-1.5-pro-preview-0514
- Claude 3 Haiku:
  claude-3-haiku@20240307
- Claude 3 Sonnet:
  claude-3-sonnet@20240229
- Claude 3.5 Sonnet:
  claude-3-5-sonnet@20240620
- Reka Edge:
  reka-edge-20240208
- Reka Flash:
  reka-flash-20240226
- Reka Core:
  reka-core-20240501
- Gemini-Pro:
  gemini-1.0-pro-001
- Claude 3.7 Sonnet:
  claude-3-7-sonnet@20250219
- Gemini 2.0 Flash:
  gemini-2.0-flash-001
- Gemini 2.5 Flash:
  gemini-2.5-flash-preview-05-20



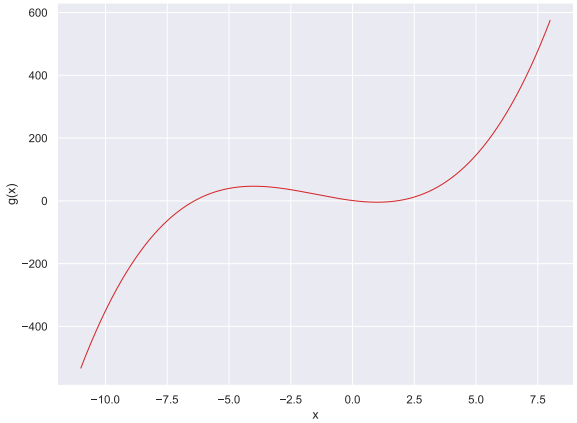
(a) Matplotlib *Functions* example



(b) Seaborn *Functions* example



(c) Matplotlib *Series* example



(d) Seaborn *Series* example

Figure 1. Example equivalent plots from the Matplotlib and Seaborn plotting libraries.

References

[1] 01-ai. Yi. https://github.com/01-ai/Yi, 2023. 1, 2

[2] Anthropic. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family, 2024. 1, 2

[3] Anthropic. Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. 1

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[5] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[6] Google. Gemini 2.5 Flash Model Card. https://storage.googleapis.com/model-cards/documents/gemini-2.5-flash.pdf, 2025. 2

[7] Google. Gemini Developer API. https://ai.google.dev/gemini-api/, 2025. 2

[8] J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. 2

[9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[10] OpenAI. GPT-4V(ision) System Card. https://cdn.

openai.com/papers/GPTV_System_Card.pdf, 2023. 1, 2

[11] OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024. 1, 2

[12] OpenAI. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. 1

[13] OpenAI. OpenAI o1: A Large Language Model for Complex Reasoning. OpenAI website, 2024. https://openai.com/o1/. 2

[14] OpenAI. API Reference. https://platform.openai.com/docs/api-reference, 2024. 2

[15] OpenBMB. OmniLMM. https://https://github.com/OpenBMB/OmniLMM, 2024. 1, 2

[16] Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, et al. Reka Core, Flash, and Edge: A Series of Powerful Multimodal Language Models. *arXiv preprint arXiv:2404.12387*, 2024. 1, 2

[17] PCIResearch. TransCore-M. https://github.com/PCIResearch/TransCore-M, 2023. 1, 2

[18] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 2

[19] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*, 2023. 1

[20] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. 2