Dotted Red Line Analysis

Zayne D Muhs-Oretta
Department of Computer Science
Rice University
zdm5@rice.edu
zdm5

***Abstract*—Extracting precise numerical data from visual graphs is a common challenge. This paper evaluates the accuracy of Multimodal Large Language Models (MLLMs), specifically Gemini 2.5 Pro, GPT 4.1 (Base and mini), and LLaMa 4, in reading exact values from a line graph at one-minute intervals. A baseline dataset and smoothed curve were generated using WebPlotDigitizer and spline smoothing. MLLM performance was assessed by comparing their visually extracted points against the baseline curve and quantifying the error using Mean Squared Error (MSE). While visual alignment varied, the results provide insights into the current capabilities and limitations of MLLMs for quantitative graph analysis tasks.**

***Index Terms*—Visual Graph Analysis, Data Extraction, Multimodal Large Language Models, WebPlotDigitizer.**

I. Introduction

Visual graphs are essential for representing data trends, but extracting the underlying numerical data from static images can be difficult. Manual estimation is error-prone, and specialized tools like WebPlotDigitizer [1] require user interaction. The emergence of Multimodal Large Language Models (MLLMs) capable of processing visual information offers potential for automating this task.

This study investigates the accuracy of four current MLLMs - Gemini 2.5 Pro, GPT 4.1, GPT 4.1-mini and LLaMa 4 - for quantitative data extraction from a specific line graph, shown in Fig.1. The models were tasked with identifying the y-value of a target line at regular one-minute intervals along the x-axis. A high-resolution dataset was first extracted from the source graph image using WebPlotDigitizer. This dataset was then processed using smoothing splines to generate a canonical baseline curve and derive reference values at the target intervals.

The performance of each MLLM was evaluated both visually, by plotting their extracted points against the baseline curve (see Section III), and numerically, by calculating the Mean Squared Error (MSE) between their reported values and the baseline values. Grok was initially included but could not participate due to input processing issues. This analysis aims to quantify the precision of current MLLMs for this task and discuss their suitability for extracting exact numerical data from visual graphs.

The paper proceeds as follows: Section II describes the baseline generation and MLLM querying procedures. Section III presents the baseline data table, the visual comparison graphs, and the MSE results. Section IV offers a discussion of these results. Section V provides concluding remarks.

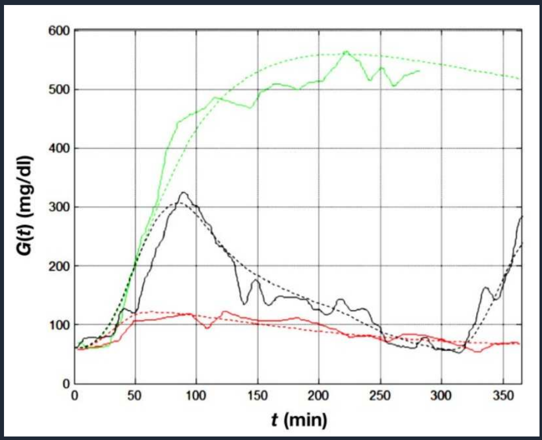


Fig. 1: The source line graph image used for data extraction by MLLMs and WebPlotDigitizer.

II. Methodology

The analysis compares MLLM-extracted data points against a baseline derived from meticulous digitization and smoothing.

A. Source Graph and Initial Extraction

The source image for this analysis is the 2D line graph shown in Fig.1, featuring a red line plotted against time (x-axis, 't [min]', approx. 0-350 min) and a measured quantity (y-axis, 'G(t) [mg/dl]'). The goal was to extract the y-value of the dotted red line at each interval of 1 minute, and plot those values.

WebPlotDigitizer (WPD) version 5 was employed to digitize the dotted red line from the source image, yielding a set of raw (x, y) coordinates after axis calibration, automatic extraction, and manual outlier removal. A truncated example of the raw data is:

```
3.390249058694458, 59.33243935793098
...
362.9665066538365, 68.4993162415858
```

(The full raw dataset consists of 114 extracted points). See Appendix for full list of estimated points at 1 minute intervals from the extracted points.

B. Baseline Curve Generation and Data Derivation

To create a smooth, representative baseline from the potentially noisy WPD points, a cubic smoothing spline ($k = 3$)

was fitted using ‘scipy.interpolate.make_splrep‘. This method balances fidelity to the data points with overall curve smoothness by minimizing $\sum w_i * (g(x_i) - y_i))^2$, where $g(x)$ is the estimation output at a given x (weighted least squares) [2]. The smoothing factor s was left to scipy for automatic evaluation.

The resulting smoothed curve represents the canonical baseline trend. This smoothed spline function was evaluated at the target 1-minute intervals to obtain the precise baseline y-values used for numerical comparison, listed in Table I. The visual representation of the baseline curve derived from the raw points is shown in the top-left panel of Fig.2.

TABLE I: Smoothed Baseline Y-Values at Select Intervals

X-Value (minutes)	**Baseline Y-Value**
0	*55.2353*
60	*121.1254*
120	*110.5229*
180	*94.5300*
240	*81.6698*
300	*72.4751*

C. Multimodal LLM Data Extraction

Four MLLMs were tested: Gemini 2.5 Pro, GPT 4.1 (Base and mini), and LLaMa 4. Each model received the original source graph image (Fig.1) and was prompted to provide the y-value of the dotted red line at each 1 minute interval. Each LLM was given one shot and supplied with the prompt: ”The provided graph has 6 lines, and measures glucose levels over time [G(t)]. The only line we are interested in is the ’dotted red line’. Reply with a markdown format of an estimation of points on the dotted red line in 1 minute intervals. Your response should include one X, Y on each line. You should attempt to get results with y values accurate to e-4.” Additions to the prompt were added when necessary to get the correctly *formatted* output. For access to Gemini 2.5 Pro (3-25) Google’s AiStudio was used [3]. For access to LLaMa 4 Maverick [4] and GPT-4.1 (Base [5] and mini [6]) OpenRouter chat was used (Note: Grok 3 & 2 were also attempted but refused to admit they had access to the graph. Further testing through Grok.com or X.ai might show results).

D. Accuracy Assessment

MLLM accuracy was assessed numerically using the Mean Squared Error (MSE) between the MLLM-reported y-values (y_{mllm}) and the baseline y-values ($y_{baseline}$) from Table I at the corresponding x-intervals:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{mllm,i} - y_{baseline,i})^2 \tag{1}$$

where N is the number of points for which both the baseline and the MLLM provided a value.

III. Results

This section presents the visual comparisons and the quantitative error summary.

A. Visual Comparison of MLLM Extractions

Figure 2 displays the comparison plots in a grid. The top-left panel shows the baseline smoothed curve derived from the raw WPD points. The subsequent panels overlay the points extracted by each MLLM onto this baseline curve, facilitating a direct visual assessment of their alignment.

B. Quantitative Accuracy Summary

The calculated Mean Squared Error (MSE), based on comparing MLLM-reported y-values to the baseline values in Table I, is summarized in Table II. Lower MSE values indicate better quantitative agreement with the baseline. Equation (1) was used for this calculation.

TABLE II: Mean Squared Error (MSE) for MLLM Data Extraction

MLLM Model	**Mean Squared Error (MSE)**
Gemini 2.5 Pro	*31.38*
LLaMa 4	*108.46*
GPT 4.1	*315.88*
GPT 4.1-mini	*596.06*

IV. Comparison and Discussion

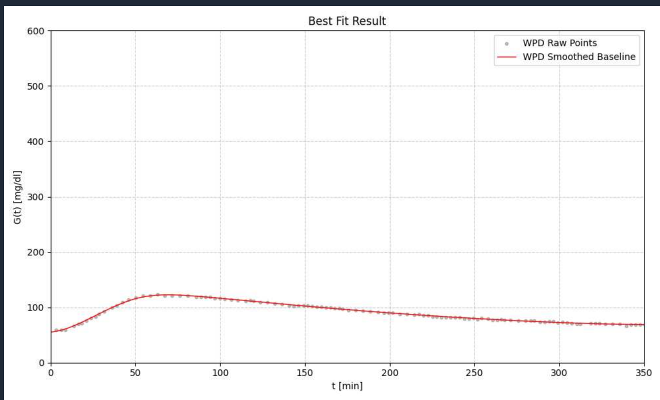
For each MLLM tested, follow up calibration questions were asked to ensure the image was properly available. These included the colors and types of all the lines of the graph to ensure visual accessibility. All models tested passed, providing correct information regarding the visual details of the graph.

From the results (Figure 2), it can be seen that the WPD extracted graph quite closely resembles the original graph, making it an accurate extraction of the dotted red line for the glucose levels. With it as a baseline, interesting patterns appear between the MLLMs evaluated. The first appearing pattern is that each MLLM struggled most on the noisiest areas of the graph, which in this case represents the beginning of the graph. Every MLLM tested had a result that overestimated the Y value of the start of the graph. In this test GPT-4.1-mini came the closest to the correct starting value, but underestimated the following bump and remaining values.

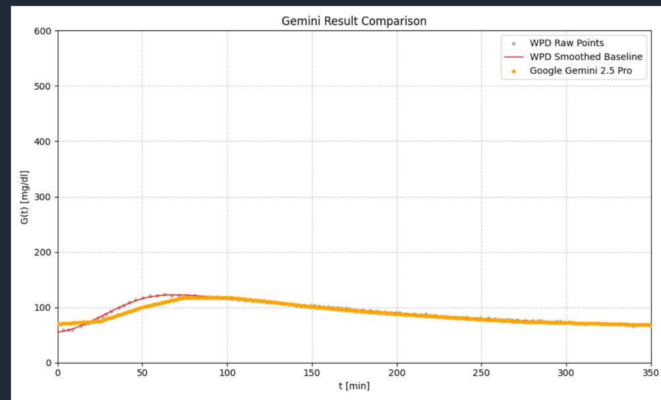
Of the four MLLMs tested, only three of them appeared to correctly estimate the overall shape of the glucose graph curve. These were Gemini 2.5 Pro, LLaMa 4, and GPT-4.1-mini. GPT-4.1 base was the only MLLM in this test which failed to identify the curvature in one shot, incorrectly indicating a positive slope on the trailing arc of the curve.

Quantitatively, Gemini 2.5 Pro gave the best performance. Excluding the first arc of the curve, the coordinates produced find themselves consistently placed in close proximity with the same curvature as the baseline graph (WPD), providing a MSE of only 31.38. Given Google’s multimodal first approach to creating their models [7], it is unsurprising that it features a top ranking.

The remaining models, especially the GPT models, all struggled greatly, showing a critical lack of understanding for the given images.



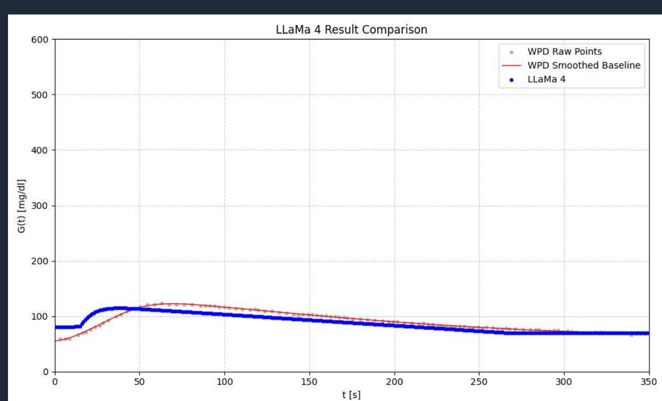
(a) Baseline: WPD Smoothed Curve



(b) Gemini 2.5 Pro vs Baseline

(c) GPT 4.1 vs Baseline

(d) GPT 4.1-mini vs Baseline



(e) LLaMa 4 vs Baseline

Fig. 2: Visual comparison of MLLM extracted points against the WPD smoothed baseline curve. (a) The baseline curve derived from WPD data. (b)-(e) Points extracted by Gemini, GPT, and LLaMa (markers) overlaid on the baseline (red line).

V. Conclusion

This study evaluated the capability of four state of the art Multimodal Large Language Models (MLLMs) - Gemini 2.5 Pro, GPT 4.1 Base, GPT 4.1-mini, and LLaMa 4 - to extract precise numerical data from a specific visual line graph at one-minute intervals. Performance was assessed against a baseline derived from WebPlotDigitizer and smoothing spline techniques, using both visual comparison (Fig.2) and quantitative Mean Squared Error (MSE) analysis (Table II).

The findings indicate significant variability in MLLM accuracy for this task. Google's Gemini 2.5 Pro demonstrated the strongest performance, achieving the lowest MSE (31.38) and providing visually aligned points that captured the overall curve shape reasonably well, particularly after the initial noisy segment. LLaMa 4 also replicated the general curve shape but with considerably higher error (MSE 108.46) and sharper value changes. The GPT models tested, especially GPT 4.1 Base (MSE 315.88), struggled significantly, with GPT 4.1 Base failing to correctly interpret the curve's overall trend in this single-shot test. While GPT 4.1-mini (MSE 596.06) captured the shape somewhat better than its base counterpart, its numerical accuracy was the lowest among the tested models. Notably, all models exhibited difficulty accurately estimating values in the initial, potentially noisier, section of the graph.

While MLLMs offer a potentially automated and user-friendly approach to graph data extraction, this analysis highlights that their current ability to deliver highly precise numerical data (such as the requested e-4 accuracy) directly from complex visual graphs remains limited and model-dependent. Gemini 2.5 Pro shows promise, aligning with its multimodal design focus, but even its output deviates from the baseline. For applications requiring high fidelity, methods combining specialized digitization tools like WebPlotDigitizer with appropriate data processing currently offer more reliable results. Future work should explore the performance of these and newer MLLMs across a wider variety of graph types and complexities, investigate the impact of different prompting strategies, and assess robustness to variations in image quality and style.

References

[1] A. Rohatgi, "Webplotdigitizer," accessed: May 5, 2025. [Online]. Available: https://automeris.io

[2] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. [Online]. Available: https://scipy.github.io/devdocs/reference/generated/scipy.interpolate.make_splrep.html#scipy.interpolate.make_splrep

[3] Google, "Gemini 2.5 pro," accessed: May 5, 2025. [Online]. Available: https://aistudio.google.com

[4] Meta, "Llama 4 maverick," accessed: May 5, 2025. [Online]. Available: https://openrouter.com

[5] OpenAI, "Gpt-4.1," accessed: May 5, 2025. [Online]. Available: https://openrouter.com

[6] ——, "Gpt-4.1-mini," accessed: May 5, 2025. [Online]. Available: https://openrouter.com

[7] S. Pichai and D. Hassabis, "Introducing gemini: Our largest and most capable ai model," Blog post on Google website, Dec. 06, 2023. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/

Appendix

Smoothed Baseline Data

0	55.24	44	109.46	88	119.60	132	107.14	176	95.51	220	85.55	264	77.55	308	71.58
1	55.55	45	110.58	89	119.35	133	106.87	177	95.27	221	85.35	265	77.39	309	71.47
2	55.96	46	111.65	90	119.09	134	106.59	178	95.02	222	85.14	266	77.23	310	71.37
3	56.46	47	112.67	91	118.82	135	106.31	179	94.77	223	84.94	267	77.07	311	71.27
4	57.04	48	113.65	92	118.55	136	106.03	180	94.53	224	84.74	268	76.92	312	71.17
5	57.71	49	114.57	93	118.28	137	105.76	181	94.29	225	84.54	269	76.76	313	71.07
6	58.46	50	115.43	94	118.00	138	105.48	182	94.04	226	84.34	270	76.61	314	70.97
7	59.28	51	116.24	95	117.72	139	105.21	183	93.80	227	84.15	271	76.45	315	70.88
8	60.18	52	116.99	96	117.44	140	104.93	184	93.56	228	83.95	272	76.30	316	70.78
9	61.14	53	117.68	97	117.15	141	104.66	185	93.32	229	83.75	273	76.15	317	70.69
10	62.16	54	118.32	98	116.87	142	104.39	186	93.08	230	83.56	274	76.00	318	70.61
11	63.25	55	118.91	99	116.58	143	104.11	187	92.84	231	83.37	275	75.85	319	70.52
12	64.39	56	119.45	100	116.29	144	103.84	188	92.61	232	83.17	276	75.70	320	70.43
13	65.59	57	119.94	101	116.00	145	103.57	189	92.37	233	82.98	277	75.55	321	70.35
14	66.83	58	120.38	102	115.71	146	103.30	190	92.14	234	82.79	278	75.41	322	70.27
15	68.12	59	120.77	103	115.41	147	103.03	191	91.90	235	82.60	279	75.26	323	70.19
16	69.45	60	121.13	104	115.12	148	102.76	192	91.67	236	82.41	280	75.12	324	70.12
17	70.82	61	121.43	105	114.83	149	102.49	193	91.44	237	82.23	281	74.98	325	70.04
18	72.22	62	121.70	106	114.54	150	102.22	194	91.20	238	82.04	282	74.83	326	69.97
19	73.65	63	121.93	107	114.25	151	101.96	195	90.97	239	81.85	283	74.69	327	69.90
20	75.11	64	122.12	108	113.96	152	101.69	196	90.74	240	81.67	284	74.55	328	69.83
21	76.59	65	122.28	109	113.68	153	101.42	197	90.52	241	81.49	285	74.41	329	69.77
22	78.09	66	122.40	110	113.39	154	101.16	198	90.29	242	81.30	286	74.27	330	69.71
23	79.61	67	122.49	111	113.10	155	100.89	199	90.06	243	81.12	287	74.14	331	69.64
24	81.13	68	122.54	112	112.81	156	100.63	200	89.84	244	80.94	288	74.00	332	69.59
25	82.67	69	122.57	113	112.52	157	100.37	201	89.61	245	80.76	289	73.87	333	69.53
26	84.20	70	122.57	114	112.24	158	100.11	202	89.39	246	80.58	290	73.73	334	69.48
27	85.74	71	122.55	115	111.95	159	99.84	203	89.17	247	80.41	291	73.60	335	69.42
28	87.27	72	122.50	116	111.66	160	99.58	204	88.95	248	80.23	292	73.47	336	69.37
29	88.80	73	122.42	117	111.38	161	99.32	205	88.73	249	80.06	293	73.34	337	69.33
30	90.32	74	122.33	118	111.09	162	99.06	206	88.51	250	79.88	294	73.21	338	69.28
31	91.83	75	122.21	119	110.81	163	98.80	207	88.29	251	79.71	295	73.09	339	69.24
32	93.33	76	122.08	120	110.52	164	98.55	208	88.07	252	79.54	296	72.96	340	69.20
33	94.81	77	121.93	121	110.24	165	98.29	209	87.86	253	79.36	297	72.84	341	69.17
34	96.28	78	121.77	122	109.96	166	98.03	210	87.64	254	79.19	298	72.72	342	69.13
35	97.73	79	121.59	123	109.67	167	97.78	211	87.43	255	79.03	299	72.59	343	69.10
36	99.15	80	121.40	124	109.39	168	97.52	212	87.22	256	78.86	300	72.48	344	69.07
37	100.55	81	121.20	125	109.11	169	97.27	213	87.00	257	78.69	301	72.36	345	69.05
38	101.92	82	121.00	126	108.82	170	97.02	214	86.79	258	78.52	302	72.24	346	69.02
39	103.26	83	120.78	127	108.54	171	96.76	215	86.58	259	78.36	303	72.13	347	69.00
40	104.58	84	120.56	128	108.26	172	96.51	216	86.37	260	78.19	304	72.01	348	68.98
41	105.85	85	120.33	129	107.98	173	96.26	217	86.17	261	78.03	305	71.90	349	68.97
42	107.09	86	120.09	130	107.70	174	96.01	218	85.96	262	77.87	306	71.79	350	68.95
43	108.30	87	119.85	131	107.42	175	95.76	219	85.75	263	77.71	307	71.68		