CLASSIFICATIONS OF LLM RESPONSES

Due to the free-form nature of the LLM responses, they are *manually* classified into three categories: "correct", "partially correct", and "incorrect". A "correct" response contains entirely accurate statements, while an "incorrect" response is composed of entirely inaccurate statements. A "partially correct" response includes both correct and incorrect sentences. Below, we present an example of a response classified as "partially correct".

The response in the following box is used to evaluate whether the correct method is used in a student's solution that truly uses the correct method. The second sentence ② in this response is correct, while the third sentence ③ is incorrect. In this case, we consider the method evaluation of the LLM as partially correct.

> ① The student used some correct methods in their solution. ② They correctly calculated power P using $P = VI$ and recognized that power is the rate of energy transfer. ③ However, the method to calculate the final time Δt was not clear or fully executed.

Tables 1–6 provide the classifications of responses from different LLMs and circuit topics.

Table 1. Test summary on the electric circuit variables and elements

LLM	Metric	Response Classification (Total Number of Data: 40)		
		Correct	Partially Correct	Incorrect
GPT-3.5 Turbo `gpt-3.5-turbo-0125`	Completeness	22 (55.00%)	5 (12.50%)	13 (32.50%)
	Method	35 (87.50%)	2 (5.00%)	3 (7.50%)
	Final Answer	19 (47.50%)	14 (35.00%)	7 (17.50%)
	Arithmetic	19 (47.50%)	5 (12.50%)	16 (40.00%)
	Unit	29 (72.50%)	2 (5.00%)	9 (22.50%)
GPT-4o `gpt-4o-2024-05-13`	Completeness	38 (95.00%)	1 (2.50%)	1 (2.50%)
	Method	**40 (100.00%)**	0 (0.00%)	0 (0.00%)
	Final Answer	**36 (90.00%)**	4 (10.00%)	0 (0.00%)
	Arithmetic	**38 (95.00%)**	2 (5.00%)	0 (0.00%)
	Unit	**40 (100.00%)**	0 (0.00%)	0 (0.00%)
Llama 3 70B `llama3-70b-instruct`	Completeness	**39 (97.50%)**	1 (2.50%)	0 (0.00%)
	Method	**40 (100.00%)**	0 (0.00%)	0 (0.00%)
	Final Answer	33 (82.50%)	6 (15.00%)	1 (2.50%)
	Arithmetic	30 (75.00%)	9 (22.50%)	1 (2.50%)
	Unit	34 (85.00%)	4 (10.00%)	2 (5.00%)

*The integers are the numbers of LLM responses that can categorized in the corresponding classes, while the percentages in parentheses represent the ratios of responses in the corresponding classes to the total number of data.

Table 2. Test summary on the analysis of resistive circuits

LLM	Metric	Response Classification (Total Number of Data: 63)		
		Correct	Partially Correct	Incorrect
GPT-3.5 Turbo `gpt-3.5-turbo-0125`	Completeness	27 (42.86%)	12 (19.05%)	24 (38.10%)
	Method	46 (73.02%)	5 (7.94%)	12 (19.05%)
	Final Answer	27 (42.86%)	19 (30.16%)	17 (26.98%)
	Arithmetic	15 (23.81%)	10 (15.87%)	38 (60.32%)
	Unit	35 (55.56%)	7 (11.11%)	21 (33.33%)
GPT-4o `gpt-4o-2024-05-13`	Completeness	**60 (95.24%)**	1 (1.59%)	2 (3.17%)
	Method	58 (92.06%)	3 (4.76%)	2 (3.17%)
	Final Answer	51 (80.95%)	8 (12.70%)	4 (6.35%)
	Arithmetic	51 (80.95%)	9 (14.29%)	3 (4.76%)
	Unit	**62 (98.41%)**	1 (1.59%)	0 (0.00%)
Llama 3 70B `llama3-70b-instruct`	Completeness	**60 (95.24%)**	1 (1.59%)	2 (3.17%)
	Method	**62 (98.41%)**	1 (1.59%)	0 (0.00%)
	Final Answer	**57 (90.48%)**	6 (9.52%)	0 (0.00%)
	Arithmetic	**52 (82.54%)**	11 (17.46%)	0 (0.00%)
	Unit	58 (92.00%)	3 (4.76%)	2 (3.17%)

Table 3. Test summary on the operational amplifier

LLM	Metric	Response Classification (Total Number of Data: 28)		
		Correct	Partially Correct	Incorrect
GPT-3.5 Turbo `gpt-3.5-turbo-0125`	Completeness	14 (50.00%)	2 (7.14%)	12 (42.86%)
	Method	22 (78.57%)	1 (3.57%)	5 (17.86%)
	Final Answer	13 (46.43%)	11 (39.29%)	4 (14.29%)
	Arithmetic	8 (28.57%)	3 (10.71%)	17 (60.71%)
	Unit	15 (53.57%)	7 (25.00%)	6 (21.43%)
GPT-4o `gpt-4o-2024-05-13`	Completeness	22 (78.57%)	5 (17.85%)	1 (3.57%)
	Method	27 (96.43%)	1 (3.57%)	0 (0.00%)
	Final Answer	25 (89.29%)	1 (3.57%)	2 (7.14%)
	Arithmetic	23 (82.14%)	3 (10.71%)	2 (7.14%)
	Unit	**26 (92.86%)**	1 (3.57%)	1 (3.57%)
Llama 3 70B `llama3-70b-instruct`	Completeness	**26 (92.86%)**	1 (3.57%)	1 (3.57%)
	Method	**26 (92.86%)**	1 (3.57%)	1 (3.57%)
	Final Answer	**26 (92.86%)**	2 (7.14%)	0 (0.00%)
	Arithmetic	**25 (89.29%)**	1 (3.57%)	2 (7.14%)
	Unit	25 (89.29%)	3 (10.71%)	0 (0.00%)

Table 4. Test summary on the complete response of circuits with energy storage elements

LLM	Metric	Response Classification (Total Number of Data: 95)		
		Correct	Partially Correct	Incorrect
GPT-3.5 Turbo `gpt-3.5-turbo-0125`	Completeness	48 (50.53%)	0 (0.00%)	47 (49.47%)
	Method	66 (69.47%)	9 (9.47%)	20 (21.05%)
	Final Answer	42 (44.21%)	28 (29.47%)	25 (26.32%)
	Arithmetic	34 (35.79%)	14 (14.74%)	47 (49.47%)
	Unit	47 (49.47%)	26 (27.37%)	22 (23.16%)
GPT-4o `gpt-4o-2024-05-13`	Completeness	85 (89.47%)	7 (7.37%)	3 (3.16%)
	Method	92 (96.84%)	3 (3.16%)	0 (0.00%)
	Final Answer	**76 (80.00%)**	12 (12.63%)	7 (7.37%)
	Arithmetic	**76 (80.00%)**	11 (11.58%)	8 (8.42%)
	Unit	**82 (86.32%)**	10 (10.53%)	3 (3.16%)
Llama 3 70B `llama3-70b-instruct`	Completeness	**87 (91.58%)**	3 (3.16%)	5 (5.26%)
	Method	**93 (97.89%)**	2 (2.11%)	0 (0.00%)
	Final Answer	75 (78.95%)	16 (16.84%)	4 (4.21%)
	Arithmetic	70 (73.68%)	16 (16.84%)	9 (9.47%)
	Unit	56 (58.95%)	30 (31.58%)	9 (9.47%)

Table 5. Test summary on the sinusoidal steady-state analysis

LLM	Metric	Response Classification (Total Number of Data: 29)		
		Correct	Partially Correct	Incorrect
GPT-3.5 Turbo `gpt-3.5-turbo-0125`	Completeness	13 (44.83%)	1 (3.45%)	15 (51.72%)
	Method	25 (86.21%)	0 (0.00%)	4 (13.79%)
	Final Answer	18 (62.07%)	8 (27.59%)	3 (10.34%)
	Arithmetic	17 (58.62%)	3 (10.34%)	9 (31.03%)
	Unit	17 (58.62%)	3 (10.34%)	9 (31.03%)
GPT-4o `gpt-4o-2024-05-13`	Completeness	20 (68.97%)	6 (20.69%)	3 (10.34%)
	Method	27 (93.10%)	1 (3.45%)	1 (3.45%)
	Final Answer	**24 (82.76%)**	2 (6.90%)	3 (10.34%)
	Arithmetic	**22 (75.86%)**	4 (13.79%)	3 (10.34%)
	Unit	**20 (68.97%)**	4 (13.79%)	5 (17.24%)
Llama 3 70B `llama3-70b-instruct`	Completeness	**24 (82.76%)**	3 (10.34%)	2 (6.90%)
	Method	**29 (100.00%)**	0 (0.00%)	0 (0.00%)
	Final Answer	**24 (82.76%)**	3 (10.34%)	2 (6.90%)
	Arithmetic	19 (65.52%)	5 (17.24%)	5 (17.24%)
	Unit	16 (55.17%)	6 (20.69%)	7 (24.14%)

Table 6. Test summary on the frequency response

LLM	Metric	Response Classification (Total Number of Data: 28)		
		Correct	Partially Correct	Incorrect
GPT-3.5 Turbo `gpt-3.5-turbo-0125`	Completeness	12 (42.66%)	3 (10.71%)	13 (46.43%)
	Method	18 (64.29%)	0 (0.00%)	10 (35.71%)
	Final Answer	15 (53.57%)	2 (7.14%)	11 (39.29%)
	Arithmetic	7 (25.00%)	4 (14.29%)	17 (60.71%)
	Unit	6 (21.43%)	19 (67.86%)	3 (10.71%)
GPT-4o `gpt-4o-2024-05-13`	Completeness	**27 (96.43%)**	0 (0.00%)	1 (3.57%)
	Method	27 (96.43%)	1 (3.57%)	0 (0.00%)
	Final Answer	**26 (92.86%)**	1 (3.57%)	1 (3.57%)
	Arithmetic	**26 (92.86%)**	1 (3.57%)	1 (3.57%)
	Unit	**26 (92.86%)**	2 (7.14%)	0 (0.00%)
Llama 3 70B `llama3-70b-instruct`	Completeness	**27 (96.43%)**	0 (0.00%)	1 (3.57%)
	Method	**28 (100.00%)**	0 (0.00%)	0 (0.00%)
	Final Answer	23 (82.14%)	1 (3.57%)	4 (14.29%)
	Arithmetic	25 (89.29%)	1 (3.57%)	2 (7.14%)
	Unit	15 (53.57%)	7 (25.00%)	6 (21.43%)