

LLM	Project	Total Cost (TC)	Number of Tests (T)	Failed Tests	Passed Tests	Success Rate (SR)	Coverage Branch (C)	Mutation Score (M)	Calculated Score (S)
claude	todo-api	US\$ 0,51	49	0	49	100,0%	85,7%	48,7%	78,1%
gpt	todo-api	US\$ 0,28	32	0	32	100,0%	85,7%	44,8%	76,8%
maritaca	todo-api	US\$ 0,11	45	0	45	100,0%	80,4%	47,2%	75,8%
llama	todo-api	US\$ 0,02	44	3	41	93,2%	85,7%	50,2%	76,4%
mistral	todo-api	US\$ 0,30	40	4	36	90,0%	82,1%	45,2%	72,5%
gemini	todo-api	US\$ 0,21	53	6	47	88,7%	85,7%	21,4%	65,3%
deepseek	todo-api	US\$ 0,95	34	1	33	97,1%	78,6%	43,5%	73,0%
qwen	todo-api	US\$ 0,12	47	0	47	100,0%	73,2%	47,8%	73,7%
claude	restaurants-api	US\$ 0,60	53	0	53	100,0%	75,8%	32,2%	69,3%
gpt	restaurants-api	US\$ 0,34	31	0	31	100,0%	57,8%	23,9%	60,6%
maritaca	restaurants-api	US\$ 0,12	27	4	23	85,2%	45,3%	19,3%	49,9%
llama	restaurants-api	US\$ 0,02	37	0	37	100,0%	64,1%	21,9%	62,0%
mistral	restaurants-api	US\$ 0,44	57	1	56	98,2%	64,1%	24,8%	62,4%
gemini	restaurants-api	US\$ 0,27	59	0	59	100,0%	64,1%	26,2%	63,4%
deepseek	restaurants-api	US\$ 0,97	57	0	57	100,0%	75,0%	31,4%	68,8%
qwen	restaurants-api	US\$ 0,11	44	0	44	100,0%	68,8%	25,4%	64,7%
claude	shortener-api	US\$ 0,25	13	0	13	100,0%	69,0%	43,1%	70,7%
gpt	shortener-api	US\$ 0,16	11	0	11	100,0%	65,5%	54,6%	73,3%
maritaca	shortener-api	US\$ 0,05	8	0	8	100,0%	65,5%	35,7%	67,1%
llama	shortener-api	US\$ 0,01	15	0	15	100,0%	66,7%	9,8%	58,8%
mistral	shortener-api	US\$ 0,17	13	0	13	100,0%	66,7%	9,8%	58,8%
gemini	shortener-api	US\$ 0,09	15	1	14	93,3%	66,7%	9,8%	56,6%
deepseek	shortener-api	US\$ 0,53	20	3	17	85,0%	64,3%	38,7%	62,7%
qwen	shortener-api	US\$ 0,05	15	2	13	86,7%	67,9%	40,7%	65,1%
claude	books-api	US\$ 0,30	25	0	25	100,0%	84,6%	67,0%	83,9%
gpt	books-api	US\$ 0,17	17	0	17	100,0%	84,6%	37,5%	74,0%
maritaca	books-api	US\$ 0,05	11	0	11	100,0%	84,6%	58,0%	80,9%
llama	books-api	US\$ 0,01	32	0	32	100,0%	80,8%	38,4%	73,1%
mistral	books-api	US\$ 0,18	17	0	17	100,0%	73,1%	32,1%	68,4%
gemini	books-api	US\$ 0,12	32	0	32	100,0%	84,6%	39,3%	74,6%
deepseek	books-api	US\$ 0,52	28	0	28	100,0%	84,6%	60,7%	81,8%
qwen	books-api	US\$ 0,05	24	0	24	100,0%	84,6%	39,3%	74,6%
claude	hotels-api	US\$ 0,73	58	0	58	100,0%	44,0%	17,9%	54,0%
gpt	hotels-api	US\$ 0,31	17	2	15	88,2%	17,9%	8,3%	38,1%
maritaca	hotels-api	US\$ 0,11	23	0	23	100,0%	42,9%	16,8%	53,2%
llama	hotels-api	US\$ 0,02	39	0	39	100,0%	38,1%	15,9%	51,3%
mistral	hotels-api	US\$ 0,49	89	0	89	100,0%	44,0%	12,3%	52,1%
gemini	hotels-api	US\$ 0,27	62	2	60	96,8%	44,0%	7,5%	49,4%
deepseek	hotels-api	US\$ 0,99	25	0	25	100,0%	33,3%	10,5%	47,9%
qwen	hotels-api	US\$ 0,11	31	1	30	96,8%	42,9%	10,6%	50,1%
claude	supermarket-ap	US\$ 0,44	32	0	32	100,0%	71,2%	36,2%	69,1%
gpt	supermarket-ap	US\$ 0,24	23	0	23	100,0%	44,2%	28,6%	57,6%
maritaca	supermarket-ap	US\$ 0,07	16	0	16	100,0%	67,3%	31,1%	66,1%
llama	supermarket-ap	US\$ 0,02	50	4	46	92,0%	61,5%	31,6%	61,7%
mistral	supermarket-ap	US\$ 0,31	44	0	44	100,0%	48,1%	29,1%	59,1%
gemini	supermarket-ap	US\$ 0,16	36	0	36	100,0%	75,0%	31,6%	68,9%
deepseek	supermarket-ap	US\$ 0,72	38	0	38	100,0%	69,2%	36,7%	68,7%
qwen	supermarket-ap	US\$ 0,09	47	5	42	89,4%	75,0%	35,7%	66,7%

$S = w \cdot SuccessRate + w \cdot Coverage + w \cdot MutationScore$

LLM	AVG					
	S	SR	C	M	T	TC
claude	70,85%	100,00%	71,73%	40,83%	38,3	US\$ 0,47
deepseek	67,15%	97,01%	67,50%	36,93%	33,7	US\$ 0,78
qwen	65,81%	95,47%	68,71%	33,26%	34,7	US\$ 0,09
maritaca	65,51%	97,53%	64,32%	34,69%	21,7	US\$ 0,08
llama	63,88%	97,53%	66,14%	27,97%	36,2	US\$ 0,02
gpt	63,42%	98,04%	59,28%	32,94%	21,8	US\$ 0,25
gemini	63,04%	96,46%	70,01%	22,64%	42,8	US\$ 0,19
mistral	62,20%	98,04%	63,01%	25,55%	43,3	US\$ 0,31

Aggregated (Min, Max, Variation)					
Min			Max		
Metric	Value	Model	Value	Model	Variation
S	62,20%	mistral	70,85%	claude	8,65%
SR	95,47%	qwen	100,00%	claude	4,53%
C	59,28%	gpt	71,73%	claude	12,45%
M	22,64%	gemini	40,83%	claude	18,20%
T	21,7	maritaca	43,3	mistral	21,7
TC	US\$ 0,02	llama	US\$ 0,78	deepseek	US\$ 0,76

Ranking by Metric				
Position	S	SR	C	M
1	claude	claude	claude	claude
2	deepseek	mistral	gemini	deepseek
3	qwen	gpt	qwen	maritaca
4	maritaca	maritaca	deepseek	qwen
5	llama	llama	llama	gpt
6	gpt	deepseek	maritaca	llama
7	gemini	gemini	mistral	mistral
8	mistral	qwen	gpt	gemini

		Per Project (Min, Max, Variation)					Variation
		Min		Max			
		Metric	Value	Model	Value	Model	
todo-api	S	65,3%	gemin	78,1%	claude	12,9%	
	SR	88,7%	gemin	100,0%	claude, gpt, maritaca, qwen	11,3%	
	C	73,2%	qwen	85,7%	claude, gpt, llama, gemini	12,5%	
	M	21,4%	gemin	50,2%	llama	28,8%	
	T	32	gpt	53	gemin	21	
restaurants-api	TC	US\$ 0,02	llama	US\$ 0,95	deepseek	US\$ 0,93	
	S	49,9%	maritaca	69,3%	claude	19,4%	
	SR	85,2%	maritaca	100,0%	claude, gpt, llama, gemini, deepseek, qwen	14,8%	
	C	45,3%	maritaca	75,8%	claude	30,5%	
	M	19,3%	maritaca	32,2%	claude	12,9%	
shortener-api	T	27	maritaca	59	gemin	32	
	TC	US\$ 0,02	llama	US\$ 0,97	deepseek	US\$ 0,94	
	S	56,6%	gemin	73,3%	gpt	16,8%	
	SR	85,0%	deepseek	100,0%	claude, gpt, maritaca, llama, mistral	15,0%	
	C	64,3%	deepseek	69,0%	claude	4,8%	
books-api	M	9,8%	llama, mistral, gemini	54,6%	gpt	44,8%	
	T	8	maritaca	20	deepseek	12	
	TC	US\$ 0,01	llama	US\$ 0,53	deepseek	US\$ 0,52	
	S	68,4%	mistral	83,9%	claude	15,5%	
	SR	100,0%	all	100,0%	all	0,0%	
hotels-api	C	73,1%	mistral	84,6%	claude, gpt, maritaca, gemini, deepseek, qwen	11,5%	
	M	32,1%	mistral	67,0%	claude	34,8%	
	T	11	maritaca	32	llama, gemini	21	
	TC	US\$ 0,01	llama	US\$ 0,52	deepseek	US\$ 0,50	
	S	38,1%	gpt	54,0%	claude	15,8%	
supermarket-api	SR	88,2%	gpt	100,0%	claude, maritaca, llama, mistral, deepseek	11,8%	
	C	17,9%	gpt	44,0%	claude, mistral, gemini	26,2%	
	M	7,5%	gemin	17,9%	claude	10,4%	
	T	17	gpt	89	mistral	72	
	TC	US\$ 0,02	llama	US\$ 0,99	deepseek	US\$ 0,97	
	S	57,6%	gpt	69,1%	claude	11,5%	
	SR	89,4%	qwen	100,0%	claude, gpt, maritaca, mistral, gemini, deepseek	10,6%	
	C	44,2%	gpt	75,0%	gemin, qwen	30,8%	
	M	28,6%	gpt	36,7%	deepseek	8,2%	
	T	16	maritaca	50	llama	34	
	TC	US\$ 0,02	llama	US\$ 0,72	deepseek	US\$ 0,71	