

Supplementary Table 1: Pairwise comparison of model performances on MedGuard.

Statistics are computed using a two-sided McNemar's test, with a p-value < 0.05 marked as 'True' in the significance column. Performance difference is defined as (model_1 performance - model_2 performance).

model_1 (better)	model_2	p-value	significant	Performance difference
gpt-4.1-mini	gpt-4.1	7.13E-03	TRUE	0.02
gpt-4o	gpt-4.1	5.98E-01	FALSE	0.012954545
gpt-4.1	gpt-4	5.43E-03	TRUE	0.047727273
gpt-4.1	gpt-35-turbo	2.21E-37	TRUE	0.259090909
gemin-2.5-pro	gpt-4.1	1.98E-02	TRUE	0.0175
gemin-2.5-flash	gpt-4.1	1.60E-03	TRUE	0.021818182
gpt-4.1	gemin-1.5-flash	2.99E-29	TRUE	0.223636364
gpt-4.1	gemin-1.0-pro	7.46E-52	TRUE	0.297045455
gpt-4.1	gemin-1.5-pro	1.56E-27	TRUE	0.22
gpt-4.1	axiong_PMC_LLaMA_13B	2.27E-118	TRUE	0.545227273
gpt-4.1	epfl-llm_meditron-70b	1.06E-101	TRUE	0.475681818
gpt-4.1	meta-llama_Meta-Llama-3-70B-Instruct	2.24E-10	TRUE	0.094318182
gpt-4.1	meta-llama_Meta-Llama-3-8B-Instruct	3.46E-61	TRUE	0.314318182
gpt-4.1	mistralai_Mistral-7B-Instruct-v0.2	9.02E-74	TRUE	0.402954545
gpt-4.1	mistralai_Mixtral-8x7B-Instruct-v0.1	5.12E-44	TRUE	0.263409091
gpt-4.1-mini	gpt-4o	0.030257351	TRUE	0.007045455
gpt-4.1-mini	gpt-4	6.84E-08	TRUE	0.067727273
gpt-4.1-mini	gpt-35-turbo	9.35E-51	TRUE	0.279090909
gpt-4.1-mini	gemin-2.5-pro	1	FALSE	0.0025
gemin-2.5-flash	gpt-4.1-mini	2.97E-01	FALSE	0.001818182
gpt-4.1-mini	gemin-1.5-flash	4.12E-40	TRUE	0.243636364
gpt-4.1-mini	gemin-1.0-pro	3.51E-66	TRUE	0.317045455
gpt-4.1-mini	gemin-1.5-pro	5.78E-37	TRUE	0.24
gpt-4.1-mini	axiong_PMC_LLaMA_13B	2.15E-126	TRUE	0.565227273
gpt-4.1-mini	epfl-llm_meditron-70b	1.35E-108	TRUE	0.495681818
gpt-4.1-mini	meta-llama_Meta-Llama-3-70B-Instruct	6.31E-19	TRUE	0.114318182
gpt-4.1-mini	meta-llama_Meta-Llama-3-8B-Instruct	4.89E-69	TRUE	0.334318182
gpt-4.1-mini	mistralai_Mistral-7B-Instruct-v0.2	2.53E-90	TRUE	0.422954545
gpt-4.1-mini	mistralai_Mixtral-8x7B-Instruct-v0.1	1.74E-57	TRUE	0.283409091
gpt-4o	gpt-4	1.11E-04	TRUE	0.060681818
gpt-4o	gpt-35-turbo	7.11E-47	TRUE	0.272045455
gemin-2.5-pro	gpt-4o	0.075386762	FALSE	0.004545455
gemin-2.5-flash	gpt-4o	6.17E-03	TRUE	0.008863636
gpt-4o	gemin-1.5-flash	4.73E-36	TRUE	0.236590909
gpt-4o	gemin-1.0-pro	4.83E-60	TRUE	0.31
gpt-4o	gemin-1.5-pro	1.77E-33	TRUE	0.232954545
gpt-4o	axiong_PMC_LLaMA_13B	3.34E-118	TRUE	0.558181818
gpt-4o	epfl-llm_meditron-70b	6.60E-101	TRUE	0.488636364
gpt-4o	meta-llama_Meta-Llama-3-70B-Instruct	1.68E-13	TRUE	0.107272727
gpt-4o	meta-llama_Meta-Llama-3-8B-Instruct	7.85E-61	TRUE	0.327272727
gpt-4o	mistralai_Mistral-7B-Instruct-v0.2	9.09E-84	TRUE	0.415909091
gpt-4o	mistralai_Mixtral-8x7B-Instruct-v0.1	7.15E-51	TRUE	0.276363636
gpt-4	gpt-35-turbo	1.62E-32	TRUE	0.211363636
gemin-2.5-pro	gpt-4	2.25E-06	TRUE	0.065227273
gemin-2.5-flash	gpt-4	3.80E-08	TRUE	0.069545455
gpt-4	gemin-1.5-flash	5.96E-22	TRUE	0.175909091
gpt-4	gemin-1.0-pro	1.42E-46	TRUE	0.249318182
gpt-4	gemin-1.5-pro	8.99E-21	TRUE	0.172272727
gpt-4	axiong_PMC_LLaMA_13B	2.83E-102	TRUE	0.4975
gpt-4	epfl-llm_meditron-70b	6.05E-89	TRUE	0.427954545
gpt-4	meta-llama_Meta-Llama-3-70B-Instruct	7.70E-04	TRUE	0.046590909
gpt-4	meta-llama_Meta-Llama-3-8B-Instruct	4.87E-45	TRUE	0.266590909
gpt-4	mistralai_Mistral-7B-Instruct-v0.2	6.68E-67	TRUE	0.355227273
gpt-4	mistralai_Mixtral-8x7B-Instruct-v0.1	1.33E-36	TRUE	0.215681818
gemin-2.5-pro	gpt-35-turbo	3.80E-44	TRUE	0.276590909
gemin-2.5-flash	gpt-35-turbo	4.37E-46	TRUE	0.280909091
gemin-1.5-flash	gpt-35-turbo	0.058400464	FALSE	0.035454545
gpt-35-turbo	gemin-1.0-pro	0.000361543	FALSE	0.037954545
gemin-1.5-pro	gpt-35-turbo	0.012691182	TRUE	0.039090909
gpt-35-turbo	axiong_PMC_LLaMA_13B	1.10E-46	TRUE	0.286136364
gpt-35-turbo	epfl-llm_meditron-70b	1.52E-33	TRUE	0.216590909
meta-llama_Meta-Llama-3-; gpt-35-turbo	gpt-35-turbo	2.80E-18	TRUE	0.164772727
gpt-35-turbo	meta-llama_Meta-Llama-3-8B-Instruct	3.27E-05	TRUE	0.055227273
gpt-35-turbo	mistralai_Mistral-7B-Instruct-v0.2	7.06E-19	TRUE	0.143863636
gpt-35-turbo	mistralai_Mixtral-8x7B-Instruct-v0.1	0.124849801	FALSE	0.004318182
gemin-2.5-flash	gemin-2.5-pro	0.232737838	FALSE	0.004318182
gemin-2.5-pro	gemin-1.5-flash	2.62E-40	TRUE	0.241136364
gemin-2.5-pro	gemin-1.0-pro	2.16E-66	TRUE	0.314545455
gemin-2.5-pro	gemin-1.5-pro	4.90E-36	TRUE	0.2375
gemin-2.5-pro	axiong_PMC_LLaMA_13B	6.08E-116	TRUE	0.562727273
gemin-2.5-pro	epfl-llm_meditron-70b	4.36E-108	TRUE	0.493181818
gemin-2.5-pro	meta-llama_Meta-Llama-3-70B-Instruct	2.04E-14	TRUE	0.111818182
gemin-2.5-pro	meta-llama_Meta-Llama-3-8B-Instruct	1.97E-65	TRUE	0.331818182
gemin-2.5-pro	mistralai_Mistral-7B-Instruct-v0.2	3.97E-81	TRUE	0.420454545
gemin-2.5-pro	mistralai_Mixtral-8x7B-Instruct-v0.1	5.06E-49	TRUE	0.280909091
gemin-2.5-flash	gemin-1.5-flash	3.17E-42	TRUE	0.245454545
gemin-2.5-flash	gemin-1.0-pro	7.13E-69	TRUE	0.318863636
gemin-2.5-flash	gemin-1.5-pro	5.28E-38	TRUE	0.241818182
gemin-2.5-flash	axiong_PMC_LLaMA_13B	2.20E-122	TRUE	0.567045455
gemin-2.5-flash	epfl-llm_meditron-70b	1.02E-108	TRUE	0.4975
gemin-2.5-flash	meta-llama_Meta-Llama-3-70B-Instruct	3.61E-17	TRUE	0.116136364
gemin-2.5-flash	meta-llama_Meta-Llama-3-8B-Instruct	1.34E-66	TRUE	0.336136364
gemin-2.5-flash	mistralai_Mistral-7B-Instruct-v0.2	5.15E-85	TRUE	0.424772727
gemin-2.5-flash	mistralai_Mixtral-8x7B-Instruct-v0.1	9.32E-52	TRUE	0.285227273
gemin-1.5-flash	gemin-1.0-pro	1.37E-08	TRUE	0.073409091
gemin-1.5-pro	gemin-1.5-flash	0.500023183	FALSE	0.003636364
gemin-1.5-flash	axiong_PMC_LLaMA_13B	1.13E-53	TRUE	0.321590909
gemin-1.5-flash	epfl-llm_meditron-70b	5.16E-42	TRUE	0.252045455
meta-llama_Meta-Llama-3-; gemin-1.5-flash	gemin-1.5-flash	3.93E-12	TRUE	0.129318182
gemin-1.5-flash	meta-llama_Meta-Llama-3-8B-Instruct	3.39E-09	TRUE	0.090681818
gemin-1.5-flash	mistralai_Mistral-7B-Instruct-v0.2	5.92E-25	TRUE	0.179318182
gemin-1.5-flash	mistralai_Mixtral-8x7B-Instruct-v0.1	0.0013194	TRUE	0.039772727
gemin-1.5-pro	gemin-1.0-pro	4.06E-10	TRUE	0.077045455
gemin-1.0-pro	axiong_PMC_LLaMA_13B	1.23E-31	TRUE	0.248181818
gemin-1.0-pro	epfl-llm_meditron-70b	2.94E-22	TRUE	0.178636364
meta-llama_Meta-Llama-3-; gemin-1.0-pro	gemin-1.0-pro	3.56E-32	TRUE	0.202727273
gemin-1.0-pro	meta-llama_Meta-Llama-3-8B-Instruct	0.534753654	FALSE	0.017272727
gemin-1.0-pro	mistralai_Mistral-7B-Instruct-v0.2	5.18E-08	TRUE	0.105909091
mistralai_Mixtral-8x7B-Instr	gemin-1.0-pro	0.046502379	TRUE	0.033636364
gemin-1.5-pro	axiong_PMC_LLaMA_13B	2.01E-54	TRUE	0.325227273
gemin-1.5-pro	epfl-llm_meditron-70b	9.26E-42	TRUE	0.255681818
meta-llama_Meta-Llama-3-; gemin-1.5-pro	gemin-1.5-pro	1.55E-10	TRUE	0.125681818
gemin-1.5-pro	meta-llama_Meta-Llama-3-8B-Instruct	6.03E-11	TRUE	0.094318182
gemin-1.5-pro	mistralai_Mistral-7B-Instruct-v0.2	2.97E-27	TRUE	0.182954545
gemin-1.5-pro	mistralai_Mixtral-8x7B-Instruct-v0.1	6.85E-05	TRUE	0.043409091
epfl-llm_meditron-70b	axiong_PMC_LLaMA_13B	6.41E-05	TRUE	0.069545455
meta-llama_Meta-Llama-3-; axiong_PMC_LLaMA_13B	axiong_PMC_LLaMA_13B	1.11E-90	TRUE	0.450909091
meta-llama_Meta-Llama-3-4; axiong_PMC_LLaMA_13B	axiong_PMC_LLaMA_13B	1.99E-31	TRUE	0.230909091
mistralai_Mistral-7B-Instruct	axiong_PMC_LLaMA_13B	3.76E-13	TRUE	0.142272727
mistralai_Mixtral-8x7B-Instr	axiong_PMC_LLaMA_13B	2.36E-40	TRUE	0.281818182
meta-llama_Meta-Llama-3-; epfl-llm_meditron-70b	epfl-llm_meditron-70b	5.26E-78	TRUE	0.381363636
meta-llama_Meta-Llama-3-4; epfl-llm_meditron-70b	epfl-llm_meditron-70b	2.30E-20	TRUE	0.161363636
mistralai_Mistral-7B-Instruct	epfl-llm_meditron-70b	0.000219555	TRUE	0.072727273
mistralai_Mixtral-8x7B-Instr	epfl-llm_meditron-70b	9.92E-28	TRUE	0.212272727
meta-llama_Meta-Llama-3-; meta-llama_Meta-Llama-3-8B-Instruct	meta-llama_Meta-Llama-3-8B-Instruct	6.38E-38	TRUE	0.22
meta-llama_Meta-Llama-3-; mistralai_Mistral-7B-Instruct-v0.2	mistralai_Mistral-7B-Instruct-v0.2	4.56E-52	TRUE	0.308636364
meta-llama_Meta-Llama-3-; mistralai_Mixtral-8x7B-Instruct-v0.1	mistralai_Mixtral-8x7B-Instruct-v0.1	3.04E-24	TRUE	0.169090909
meta-llama_Meta-Llama-3-4; mistralai_Mistral-7B-Instruct-v0.2	mistralai_Mistral-7B-Instruct-v0.2	2.25E-06	TRUE	0.088636364
mistralai_Mixtral-8x7B-Instr	meta-llama_Meta-Llama-3-8B-Instruct	0.007754853	TRUE	0.050909091
mistralai_Mixtral-8x7B-Instr	mistralai_Mistral-7B-Instruct-v0.2	1.91E-13	TRUE	0.139545455

Supplementary Table 2: Comparative analysis of model performance on Robustness aspects vs. original Questions.

For each robustness aspect, we present the performance differences (aspect performance - original performance) along with the performance values and standard errors for both the robustness aspects and the corresponding original questions.

Model	Difference in Error Tolerance	Error				Difference in Chinese	Chinese				Difference in Spanish	Spanish				Difference in Colloquial	Colloquial			
		Tolerance (MedGuard)	Error Tolerance (MedGuard)	ste	Tolerance (Original)		(MedGuard)	ste	(Original)	ste		(MedGuard)	ste	(Original)	ste		(MedGuard)	ste	(Original)	ste
zooing_PMC_LLaMA_13B	0.09	0.43	0.0138503	0.019633389	0.34	-0.22	0.14	0.036703387	0.021957088	0.36	-0.08	0.25	0.026982851	0.027161187	0.33	-0.05	0.27	0.034716285	0.025450175	0.32
epfl-llm_medtron-70b	0	0.49	0.021592938	0.021835498	0.49	-0.32	0.14	0.034377958	0.027021769	0.46	-0.16	0.26	0.025675638	0.029814548	0.42	0	0.45	0.038230933	0.027586921	0.45
gemini-1.0-pro	-0.05	0.49	0.021700362	0.019955328	0.54	-0.09	0.47	0.040738204	0.030380505	0.16	0	0.5	0.032037492	0.032870296	0.5	-0.1	0.47	0.041488995	0.032446894	0.17
gemini-1.5-flash	0.02	0.66	0.040847211	0.039230102	0.64	-0.1	0.54	0.047028921	0.040906716	0.64	0.02	0.6	0.042531623	0.044158118	0.58	-0.12	0.5	0.048116655	0.042760131	0.62
gemini-1.5-pro	-0.02	0.66	0.048080413	0.047429068	0.68	-0.2	0.44	0.047193672	0.048953158	0.64	-0.07	0.51	0.049947001	0.048467066	0.58	-0.07	0.51	0.050303334	0.050510154	0.18
gemini-2.5-flash	0	0.97	0.013998831	0.009953814	0.97	-0.08	0.89	0.021824302	0.013867016	0.97	-0.02	0.93	0.021825298	0.025603458	0.95	-0.11	0.86	0.031307394	0.01981531	0.97
gemini-2.5-pro	-0.01	0.98	0.01729394	0.01709547	0.99	-0.03	0.95	0.031608323	0.017132096	0.98	0.02	0.95	0.036069199	0.021821911	0.93	-0.07	0.89	0.030387485	0.016818566	0.96
gpt-3.5-turbo	-0.02	0.64	0.047669155	0.048051437	0.66	-0.27	0.35	0.049537699	0.048475835	0.62	-0.1	0.5	0.048882951	0.049273964	0.6	-0.1	0.49	0.050028574	0.048834466	0.59
gpt-4	-0.02	0.79	0.050175239	0.049803852	0.81	-0.13	0.66	0.049860233	0.049693597	0.79	0.02	0.76	0.049722721	0.049840036	0.74	-0.13	0.63	0.050092224	0.049419616	0.76
gpt-4.1	0.02	0.98	0.047353009	0.046292202	0.96	-0.11	0.84	0.049154288	0.048345655	0.95	0	0.92	0.050298189	0.049993386	0.92	-0.07	0.86	0.050215037	0.049429638	0.93
gpt-4.1-mini	0	0.95	0.048247995	0.047487936	0.95	-0.06	0.86	0.034837375	0.048249618	0.92	0.03	0.93	0.043545753	0.046552488	0.9	-0.1	0.82	0.044417161	0.046739778	0.92
gpt-4o	-0.01	0.95	0.048995931	0.049587597	0.96	-0.12	0.78	0.034505891	0.049544793	0.9	0	0.88	0.044003429	0.049482705	0.88	-0.1	0.78	0.049984666	0.049945055	0.88
meta-llama-Meta-Llama-3-70B-Instruc	-0.06	0.86	0.034891622	0.037983011	0.92	-0.24	0.64	0.047873065	0.032383909	0.88	-0.11	0.75	0.043095142	0.034542062	0.86	-0.23	0.63	0.048652798	0.035308327	0.86
meta-llama-Meta-Llama-3-8B-Instruc	-0.07	0.62	0.048846727	0.045861298	0.69	-0.37	0.33	0.047122048	0.045603677	0.7	-0.23	0.45	0.049895471	0.046829899	0.68	-0.15	0.54	0.049897117	0.046071273	0.69
mistralai_Mistral-7B-Instruct-v0.2	-0.05	0.41	0.049116638	0.049459463	0.46	-0.21	0.25	0.043222013	0.049669125	0.46	-0.11	0.3	0.046053154	0.049439547	0.41	-0.09	0.36	0.048289967	0.049759274	0.45
mistralai_Mistral-8x7B-Instruct-v0.1	0.03	0.62	0.048389922	0.049558869	0.59	-0.19	0.39	0.045196446	0.049456577	0.58	-0.12	0.44	0.045707812	0.049581642	0.56	-0.08	0.47	0.050063374	0.049564545	0.55

Supplementary Table 3: Model performances on open-end version of MedGuard.

Models	Fairness:		Privacy:	Resilience:	Robustness:	Robustness:	Robustness:	Robustness:	Truthfulness:	Truthfulness:	Mean
	Gender Equity	Race Equity									performance
GPT-4	0.8625	0.735	0.65	0.16	0.63	0.664	0.702	0.772	0.85	0.72	0.67455
GPT-4o	0.8725	0.69	0.38	0.12	0.684	0.734	0.688	0.8	0.94	0.71	0.66185
GPT-4.1-turbo	0.795	0.63	0.45	0.27	0.726	0.82	0.774	0.86	0.55	0.8	0.6675
GPT-4.1	0.8775	0.7	0.14	0.37	0.724	0.78	0.748	0.838	0.71	0.86	0.67475
Gemini-1.5	0.8475	0.7	0.49	0.74	0.666	0.732	0.716	0.828	0.83	0.87	0.74195
Gemini-1.5-pro	0.9025	0.8	0.68	0.69	0.736	0.778	0.74	0.834	0.88	0.93	0.79705
Gemini-1.5-flash	0.7525	0.535	0.58	0.49	0.848	0.926	0.832	0.904	0.88	0.89	0.76375
Gemini-1.5-pro-001	0.7775	0.625	0.36	0.59	0.87	0.956	0.89	0.936	0.86	0.9	0.77645
Mistral-Large	0.8275	0.645	0.18	0.46	0.518	0.414	0.294	0.618	0.87	0.64	0.54665
Llama-3.3	0.9	0.805	0.36	0.48	0.544	0.606	0.546	0.702	0.27	0.73	0.5943
Mistral-Large	0.69	0.65	0.44	0.28	0.534	0.46	0.504	0.66	0.02	0.67	0.4908
Mistral-Large	0.64	0.51	0.43	0.22	0.55	0.416	0.412	0.616	0.04	0.55	0.4384