

Supplementary File 1. Statistical Analysis Results

Table 1 Statistical Analysis Results for Diagnostic Accuracy Between Different Models in Static Question Answering Scenario

Model1	Model2	Chi-square	Original P value	Corrected P value	Significance
ClinDiag-GPT	4omini	2.702819	0.100171	1	No
ClinDiag-GPT	4o	0.714042	0.398105	1	No
ClinDiag-GPT	Claude-3-haiku	7.873123	0.005017	0.105367	No
ClinDiag-GPT	Qwen2.5 72b	0.958402	0.32759	1	No
ClinDiag-GPT	Qwen2.5 32b	0.030181	0.862079	1	No
ClinDiag-GPT	Qwen2.5 14b	5.049382	0.024635	0.517329	No
4omini	4o	0.604438	0.43689	1	No
4omini	Claude-3-haiku	1.301083	0.254015	1	No
4omini	Qwen2.5 72b	6.993037	0.008183	0.171837	No
4omini	Qwen2.5 32b	3.383475	0.065853	1	No
4omini	Qwen2.5 14b	0.338295	0.560815	1	No
4o	Claude-3-haiku	3.761902	0.052433	1	No
4o	Qwen2.5 72b	3.406224	0.064951	1	No
4o	Qwen2.5 32b	1.082502	0.298139	1	No
4o	Qwen2.5 14b	1.906089	0.167398	1	No
Claude-3-haiku	Qwen2.5 72b	14.48506	0.000141	0.002967	Yes
Claude-3-haiku	Qwen2.5 32b	9.007445	0.002689	0.056465	No
Claude-3-haiku	Qwen2.5 14b	0.28893	0.590907	1	No
Qwen2.5 72b	Qwen2.5 32b	0.613929	0.433312	1	No
Qwen2.5 72b	Qwen2.5 14b	10.54501	0.001165	0.024465	Yes
Qwen2.5 32b	Qwen2.5 14b	5.965564	0.014588	0.306347	No

Table 1 Statistical Analysis Results for Diagnostic Accuracy Between Different Models in Clinical Diagnostic Procedure

Model1	Model2	Chi-square	Original P value	Corrected P value	Significance
ClinDiag-GPT	4omini	113.6007	1.59E-26	3.35E-25	Yes
ClinDiag-GPT	4o	53.44201	2.66E-13	5.59E-12	Yes
ClinDiag-GPT	Claude-3-haiku	104.8776	1.3E-24	2.73E-23	Yes
ClinDiag-GPT	Qwen2.5 72b	36.56849	1.47E-09	3.1E-08	Yes
ClinDiag-GPT	Qwen2.5 32b	118.1043	1.65E-27	3.45E-26	Yes
ClinDiag-GPT	Qwen2.5 14b	88.55298	4.95E-21	1.04E-19	Yes
4omini	4o	11.19165	0.000822	0.017255	Yes
4omini	Claude-3-haiku	0.158167	0.69085	1	No
4omini	Qwen2.5 72b	21.26395	4E-06	8.4E-05	Yes
4omini	Qwen2.5 32b	0.035245	0.851083	1	No
4omini	Qwen2.5 14b	1.524949	0.216872	1	No
4o	Claude-3-haiku	8.553909	0.003448	0.072404	No
4o	Qwen2.5 72b	1.549465	0.213214	1	No
4o	Qwen2.5 32b	12.64761	0.000376	0.007897	Yes
4o	Qwen2.5 14b	4.360196	0.036788	0.772549	No
Claude-3-haiku	Qwen2.5 72b	17.56398	2.78E-05	0.000583	Yes
Claude-3-haiku	Qwen2.5 32b	0.370731	0.542606	1	No
Claude-3-haiku	Qwen2.5 14b	0.662398	0.415715	1	No
Qwen2.5 72b	Qwen2.5 32b	23.25221	1.42E-06	2.98E-05	Yes
Qwen2.5 72b	Qwen2.5 14b	11.25536	0.000794	0.016673	Yes
Qwen2.5 32b	Qwen2.5 14b	2.091011	0.148168	1	No

Table 3 Statistical Analysis Results for Diagnostic Accuracy in Challenging Case Subset in Clinical Diagnostic Procedure

Model1	Model2	Chi-square	Original P value	Corrected P value	Significance
ClinDiag-GPT	4omini	62.25636	3.02E-15	6.33E-14	Yes
ClinDiag-GPT	4o	4.701521	0.030136	0.632855	No
ClinDiag-GPT	Claude-3-haiku	63.63399	1.5E-15	3.15E-14	Yes
ClinDiag-GPT	Qwen2.5 72b	22.51574	2.08E-06	4.38E-05	Yes
ClinDiag-GPT	Qwen2.5 32b	44.21541	2.94E-11	6.18E-10	Yes
ClinDiag-GPT	Qwen2.5 14b	35.74415	2.25E-09	4.73E-08	Yes
4omini	4o	32.63708	1.11E-08	2.33E-07	Yes
4omini	Claude-3-haiku	0.001973	0.964573	1	No
4omini	Qwen2.5 72b	9.875491	0.001675	0.035174	Yes
4omini	Qwen2.5 32b	1.489117	0.222354	1	No
4omini	Qwen2.5 14b	3.609439	0.057452	1	No
4o	Claude-3-haiku	33.64652	6.61E-09	1.39E-07	Yes
4o	Qwen2.5 72b	6.478958	0.010916	0.229234	No
4o	Qwen2.5 32b	19.90363	8.14E-06	0.000171	Yes
4o	Qwen2.5 14b	14.33227	0.000153	0.003218	Yes
Claude-3-haiku	Qwen2.5 72b	10.43897	0.001234	0.025911	Yes
Claude-3-haiku	Qwen2.5 32b	1.713501	0.190532	1	No
Claude-3-haiku	Qwen2.5 14b	3.954076	0.046758	0.981918	No
Qwen2.5 72b	Qwen2.5 32b	3.543092	0.059794	1	No
Qwen2.5 72b	Qwen2.5 14b	1.448199	0.228817	1	No
Qwen2.5 32b	Qwen2.5 14b	0.406205	0.523902	1	No

Table 4 Statistical Analysis Results for Diagnostic Accuracy in Emergency Case Subset in Clinical Diagnostic Procedure

Model1	Model2	Chi-square	Original P value	Corrected P value	Significance
ClinDiag-GPT	4omini	43.58603	4.06E-11	8.52E-10	Yes
ClinDiag-GPT	4o	63.83698	1.35E-15	2.84E-14	Yes
ClinDiag-GPT	Claude-3-haiku	39.79019	2.83E-10	5.94E-09	Yes
ClinDiag-GPT	Qwen2.5 72b	12.89329	0.00033	0.006925	Yes
ClinDiag-GPT	Qwen2.5 32b	63.36427	1.72E-15	3.61E-14	Yes
ClinDiag-GPT	Qwen2.5 14b	45.94925	1.21E-11	2.55E-10	Yes
4omini	4o	1.86757	0.171754	1	No
4omini	Claude-3-haiku	0.070692	0.790332	1	No
4omini	Qwen2.5 72b	8.939887	0.00279	0.058592	No
4omini	Qwen2.5 32b	1.786882	0.181307	1	No
4omini	Qwen2.5 14b	0.021885	0.882395	1	No
4o	Claude-3-haiku	2.762108	0.096521	1	No
4o	Qwen2.5 72b	19.2192	1.17E-05	0.000245	Yes
4o	Qwen2.5 32b	3.43E-29	1	1	No
4o	Qwen2.5 14b	1.41393	0.234405	1	No
Claude-3-haiku	Qwen2.5 72b	7.262183	0.007042	0.147885	No
Claude-3-haiku	Qwen2.5 32b	2.663798	0.102655	1	No
Claude-3-haiku	Qwen2.5 14b	0.196587	0.65749	1	No
Qwen2.5 72b	Qwen2.5 32b	18.95892	1.34E-05	0.00028	Yes
Qwen2.5 72b	Qwen2.5 14b	10.03162	0.001539	0.032314	Yes
Qwen2.5 32b	Qwen2.5 14b	1.343833	0.246359	1	No

Table 5 Statistical Analysis Results for Diagnostic Accuracy in Rare Disease Subset in Clinical Diagnostic Procedure

Model1	Model2	Chi-square	Original P value	Corrected P value	Significance
ClinDiag-GPT	4omini	8.065481	0.004512	0.094745	No
ClinDiag-GPT	4o	3.748907	0.052842	1	No
ClinDiag-GPT	Claude-3-haiku	3.291727	0.06963	1	No
ClinDiag-GPT	Qwen2.5 72b	0.214282	0.643432	1	No
ClinDiag-GPT	Qwen2.5 32b	4.060179	0.043906	0.922023	No
ClinDiag-GPT	Qwen2.5 14b	1.746205	0.186354	1	No
4omini	4o	23.27215	1.41E-06	2.95E-05	Yes
4omini	Claude-3-haiku	0.875832	0.349346	1	No
4omini	Qwen2.5 72b	5.244157	0.022021	0.462432	No
4omini	Qwen2.5 32b	0.537093	0.46364	1	No
4omini	Qwen2.5 14b	2.051183	0.152088	1	No
4o	Claude-3-haiku	14.59949	0.000133	0.002792	Yes
4o	Qwen2.5 72b	6.18492	0.012884	0.270571	No
4o	Qwen2.5 32b	16.14336	5.87E-05	0.001233	Yes
4o	Qwen2.5 14b	11.12828	0.00085	0.017854	Yes
Claude-3-haiku	Qwen2.5 72b	1.587299	0.207713	1	No
Claude-3-haiku	Qwen2.5 32b	0.010271	0.919274	1	No
Claude-3-haiku	Qwen2.5 14b	0.157715	0.691269	1	No
Qwen2.5 72b	Qwen2.5 32b	2.135181	0.143953	1	No
Qwen2.5 72b	Qwen2.5 14b	0.586381	0.443822	1	No
Qwen2.5 32b	Qwen2.5 14b	0.358936	0.549098	1	No

Table 7 Statistical Analysis Results for Diagnostic Accuracy Across Specialties in Clinical Diagnostic Procedure

Specialty	Model 1	Model 2	Chi-Square Value	P-Value	Adjusted P- Value (Bonferroni)	Signif icance
Addiction Medicine	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Addiction Medicine	ClinDiag-GPT	GPT-4o	NA	1	1	no
Addiction Medicine	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Addiction Medicine	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Addiction Medicine	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Addiction Medicine	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Addiction Medicine	GPT-4o-mini	GPT-4o	NA	1	1	no
Addiction Medicine	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Addiction Medicine	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Addiction Medicine	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Addiction Medicine	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Addiction Medicine	GPT-4o	Claude-3-Haiku	NA	1	1	no
Addiction Medicine	GPT-4o	Qwen2.5-72b	NA	1	1	no
Addiction Medicine	GPT-4o	Qwen2.5-32b	NA	1	1	no
Addiction Medicine	GPT-4o	Qwen2.5-14b	NA	1	1	no
Addiction Medicine	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Addiction Medicine	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Addiction Medicine	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Addiction Medicine	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Addiction Medicine	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Addiction Medicine	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Allergy and Immunology	ClinDiag-GPT	GPT-4o-mini	NA	0.00829 9	1	no
Allergy and Immunology	ClinDiag-GPT	GPT-4o	8.086544	0.00446	1	no
Allergy and Immunology	ClinDiag-GPT	Claude-3-Haiku	NA	0.06221 8	1	no
Allergy and Immunology	ClinDiag-GPT	Qwen2.5-72b	NA	0.02547 7	1	no
Allergy and Immunology	ClinDiag-GPT	Qwen2.5-32b	NA	0.02547 7	1	no
Allergy and Immunology	ClinDiag-GPT	Qwen2.5-14b	NA	0.02547 7	1	no
Allergy and Immunology	GPT-4o-mini	GPT-4o	NA	0.41742 7	1	no
Allergy and Immunology	GPT-4o-mini	Claude-3-Haiku	NA	0.66710 1	1	no
Allergy and Immunology	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Allergy and Immunology	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Allergy and Immunology	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no

Allergy and Immunology	GPT-4o	Claude-3-Haiku	NA	1	1	no
Allergy and Immunology	GPT-4o	Qwen2.5-72b	NA	0.701939	1	no
Allergy and Immunology	GPT-4o	Qwen2.5-32b	NA	0.701939	1	no
Allergy and Immunology	GPT-4o	Qwen2.5-14b	NA	0.701939	1	no
Allergy and Immunology	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Allergy and Immunology	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Allergy and Immunology	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Allergy and Immunology	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Allergy and Immunology	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Allergy and Immunology	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Anesthesiology	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Anesthesiology	ClinDiag-GPT	GPT-4o	NA	1	1	no
Anesthesiology	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Anesthesiology	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Anesthesiology	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Anesthesiology	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Anesthesiology	GPT-4o-mini	GPT-4o	NA	1	1	no
Anesthesiology	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Anesthesiology	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Anesthesiology	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Anesthesiology	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Anesthesiology	GPT-4o	Claude-3-Haiku	NA	1	1	no
Anesthesiology	GPT-4o	Qwen2.5-72b	NA	1	1	no
Anesthesiology	GPT-4o	Qwen2.5-32b	NA	1	1	no
Anesthesiology	GPT-4o	Qwen2.5-14b	NA	1	1	no
Anesthesiology	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Anesthesiology	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Anesthesiology	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Anesthesiology	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Anesthesiology	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Anesthesiology	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Cardiology	ClinDiag-GPT	GPT-4o-mini	28.54033	9.18E-08	6.2E-05	yes
Cardiology	ClinDiag-GPT	GPT-4o	9.93241	0.001624	1	no
Cardiology	ClinDiag-GPT	Claude-3-Haiku	39.07148	4.09E-10	2.7E-07	yes
Cardiology	ClinDiag-GPT	Qwen2.5-72b	11.91557	0.000557	0.37408	no
Cardiology	ClinDiag-GPT	Qwen2.5-32b	35.43049	2.64E-09	1.8E-06	yes

Cardiology	ClinDiag-GPT	Qwen2.5-14b	32.71656	1.07E-08	7.2E-06	yes
Cardiology	GPT-4o-mini	GPT-4o	3.148277	0.076007	1	no
Cardiology	GPT-4o-mini	Claude-3-Haiku	0.489745	0.484041	1	no
Cardiology	GPT-4o-mini	Qwen2.5-72b	2.320365	0.12769	1	no
Cardiology	GPT-4o-mini	Qwen2.5-32b	0.222869	0.636863	1	no
Cardiology	GPT-4o-mini	Qwen2.5-14b	0.085966	0.769369	1	no
Cardiology	GPT-4o	Claude-3-Haiku	7.286117	0.006949	1	no
Cardiology	GPT-4o	Qwen2.5-72b	0.072257	0.788078	1	no
Cardiology	GPT-4o	Qwen2.5-32b	5.983446	0.014441	1	no
Cardiology	GPT-4o	Qwen2.5-14b	5.053069	0.024582	1	no
Cardiology	Claude-3-Haiku	Qwen2.5-72b	4.624685	0.031515	1	no
Cardiology	Claude-3-Haiku	Qwen2.5-32b	0.047883	0.826788	1	no
Cardiology	Claude-3-Haiku	Qwen2.5-14b	0.152844	0.695833	1	no
Cardiology	Qwen2.5-72b	Qwen2.5-32b	4.622549	0.031554	1	no
Cardiology	Qwen2.5-72b	Qwen2.5-14b	3.820865	0.050618	1	no
Cardiology	Qwen2.5-32b	Qwen2.5-14b	0.03037	0.861653	1	no
Dermatology	ClinDiag-GPT	GPT-4o-mini	81.91916	1.42E-19	9.5E-17	yes
Dermatology	ClinDiag-GPT	GPT-4o	26.91934	2.12E-07	0.00014	yes
Dermatology	ClinDiag-GPT	Claude-3-Haiku	81.91916	1.42E-19	9.5E-17	yes
Dermatology	ClinDiag-GPT	Qwen2.5-72b	64.52701	9.52E-16	6.4E-13	yes
Dermatology	ClinDiag-GPT	Qwen2.5-32b	81.91916	1.42E-19	9.5E-17	yes
Dermatology	ClinDiag-GPT	Qwen2.5-14b	59.59002	1.17E-14	7.9E-12	yes

Dermatology	GPT-4o-mini	GPT-4o	6.959936	0.00833 6	1	no
Dermatology	GPT-4o-mini	Claude-3-Haiku	0	1	1	no
Dermatology	GPT-4o-mini	Qwen2.5-72b	0.42309	0.51539 9	1	no
Dermatology	GPT-4o-mini	Qwen2.5-32b	0	1	1	no
Dermatology	GPT-4o-mini	Qwen2.5-14b	0.733683	0.39169 2	1	no
Dermatology	GPT-4o	Claude-3-Haiku	9.307854	0.00228 2	1	no
Dermatology	GPT-4o	Qwen2.5-72b	5.18252	0.02281 5	1	no
Dermatology	GPT-4o	Qwen2.5-32b	9.307854	0.00228 2	1	no
Dermatology	GPT-4o	Qwen2.5-14b	4.138141	0.04192 7	1	no
Dermatology	Claude-3-Haiku	Qwen2.5-72b	0.42309	0.51539 9	1	no
Dermatology	Claude-3-Haiku	Qwen2.5-32b	0	1	1	no
Dermatology	Claude-3-Haiku	Qwen2.5-14b	0.733683	0.39169 2	1	no
Dermatology	Qwen2.5-72b	Qwen2.5-32b	0.45967	0.49777 8	1	no
Dermatology	Qwen2.5-72b	Qwen2.5-14b	0.045855	0.83043 9	1	no
Dermatology	Qwen2.5-32b	Qwen2.5-14b	0.733683	0.39169 2	1	no
Emergency Medicine	ClinDiag-GPT	GPT-4o-mini	NA	0.23511 6	1	no
Emergency Medicine	ClinDiag-GPT	GPT-4o	NA	0.23511 6	1	no
Emergency Medicine	ClinDiag-GPT	Claude-3-Haiku	NA	0.23511 6	1	no
Emergency Medicine	ClinDiag-GPT	Qwen2.5-72b	NA	0.23511 6	1	no
Emergency Medicine	ClinDiag-GPT	Qwen2.5-32b	NA	0.23511 6	1	no
Emergency Medicine	ClinDiag-GPT	Qwen2.5-14b	NA	0.23511 6	1	no
Emergency Medicine	GPT-4o-mini	GPT-4o	NA	1	1	no
Emergency Medicine	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Emergency Medicine	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Emergency Medicine	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Emergency Medicine	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no

Emergency Medicine	GPT-4o	Claude-3-Haiku	NA	1	1	no
Emergency Medicine	GPT-4o	Qwen2.5-72b	NA	1	1	no
Emergency Medicine	GPT-4o	Qwen2.5-32b	NA	1	1	no
Emergency Medicine	GPT-4o	Qwen2.5-14b	NA	1	1	no
Emergency Medicine	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Emergency Medicine	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Emergency Medicine	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Emergency Medicine	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Emergency Medicine	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Emergency Medicine	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Endocrinology	ClinDiag-GPT	GPT-4o-mini	275.3399	7.78E-62	5.2E-59	yes
Endocrinology	ClinDiag-GPT	GPT-4o	133.083	8.67E-31	5.8E-28	yes
Endocrinology	ClinDiag-GPT	Claude-3-Haiku	133.083	8.67E-31	5.8E-28	yes
Endocrinology	ClinDiag-GPT	Qwen2.5-72b	30.76293	2.92E-08	2E-05	yes
Endocrinology	ClinDiag-GPT	Qwen2.5-32b	133.083	8.67E-31	5.8E-28	yes
Endocrinology	ClinDiag-GPT	Qwen2.5-14b	275.3399	7.78E-62	5.2E-59	yes
Endocrinology	GPT-4o-mini	GPT-4o	5.784314	0.01617	1	no
Endocrinology	GPT-4o-mini	Claude-3-Haiku	5.784314	0.01617	1	no
Endocrinology	GPT-4o-mini	Qwen2.5-72b	35.4	2.68E-09	1.8E-06	yes
Endocrinology	GPT-4o-mini	Qwen2.5-32b	5.784314	0.01617	1	no
Endocrinology	GPT-4o-mini	Qwen2.5-14b	0	1	1	no
Endocrinology	GPT-4o	Claude-3-Haiku	0	1	1	no
Endocrinology	GPT-4o	Qwen2.5-72b	15.73333	7.29E-05	0.04901	yes
Endocrinology	GPT-4o	Qwen2.5-32b	0	1	1	no
Endocrinology	GPT-4o	Qwen2.5-14b	8.194444	0.004202	1	no
Endocrinology	Claude-3-Haiku	Qwen2.5-72b	15.73333	7.29E-05	0.04901	yes
Endocrinology	Claude-3-Haiku	Qwen2.5-32b	0	1	1	no
Endocrinology	Claude-3-Haiku	Qwen2.5-14b	8.194444	0.004202	1	no
Endocrinology	Qwen2.5-72b	Qwen2.5-32b	23.13725	1.51E-06	0.00101	yes
Endocrinology	Qwen2.5-72b	Qwen2.5-14b	73.75	8.87E-18	6E-15	yes

Endocrinology	Qwen2.5-32b	Qwen2.5-14b	8.194444	0.004202	1	no
Gastroenterology	ClinDiag-GPT	GPT-4o-mini	36.14067	1.84E-09	1.2E-06	yes
Gastroenterology	ClinDiag-GPT	GPT-4o	3.998925	0.045529	1	no
Gastroenterology	ClinDiag-GPT	Claude-3-Haiku	71.17234	3.27E-17	2.2E-14	yes
Gastroenterology	ClinDiag-GPT	Qwen2.5-72b	7.302664	0.006885	1	no
Gastroenterology	ClinDiag-GPT	Qwen2.5-32b	15.53465	8.1E-05	0.054436	no
Gastroenterology	ClinDiag-GPT	Qwen2.5-14b	4.070725	0.043633	1	no
Gastroenterology	GPT-4o-mini	GPT-4o	46.72639	8.16E-12	5.5E-09	yes
Gastroenterology	GPT-4o-mini	Claude-3-Haiku	3.846656	0.049845	1	no
Gastroenterology	GPT-4o-mini	Qwen2.5-72b	7.590677	0.005867	1	no
Gastroenterology	GPT-4o-mini	Qwen2.5-32b	2.932932	0.08679	1	no
Gastroenterology	GPT-4o-mini	Qwen2.5-14b	11.13588	0.000847	0.569009	no
Gastroenterology	GPT-4o	Claude-3-Haiku	123.519	1.07E-28	7.2E-26	yes
Gastroenterology	GPT-4o	Qwen2.5-72b	24.59252	7.08E-07	0.00048	yes
Gastroenterology	GPT-4o	Qwen2.5-32b	39.43095	3.4E-10	2.3E-07	yes
Gastroenterology	GPT-4o	Qwen2.5-14b	17.90838	2.32E-05	0.01558	yes
Gastroenterology	Claude-3-Haiku	Qwen2.5-72b	19.43199	1.04E-05	0.00701	yes
Gastroenterology	Claude-3-Haiku	Qwen2.5-32b	11.73173	0.000614	0.4129	no
Gastroenterology	Claude-3-Haiku	Qwen2.5-14b	24.59128	7.09E-07	0.00048	yes
Gastroenterology	Qwen2.5-72b	Qwen2.5-32b	1.303589	0.253559	1	no
Gastroenterology	Qwen2.5-72b	Qwen2.5-14b	0.401742	0.526191	1	no
Gastroenterology	Qwen2.5-32b	Qwen2.5-14b	2.941681	0.086321	1	no
General Surgery	ClinDiag-GPT	GPT-4o-mini	NA	0	0	yes
General Surgery	ClinDiag-GPT	GPT-4o	4513.213	0	0	yes
General Surgery	ClinDiag-GPT	Claude-3-Haiku	NA	0	0	yes

General Surgery	ClinDiag-GPT	Qwen2.5-72b	765.522	1.7E-168	1E-165	yes
General Surgery	ClinDiag-GPT	Qwen2.5-32b	1679.749	0	0	yes
General Surgery	ClinDiag-GPT	Qwen2.5-14b	NA	0	0	yes
General Surgery	GPT-4o-mini	GPT-4o	NA	8.91E-41	6E-38	yes
General Surgery	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
General Surgery	GPT-4o-mini	Qwen2.5-72b	NA	6.3E-130	4E-127	yes
General Surgery	GPT-4o-mini	Qwen2.5-32b	NA	8.28E-84	5.6E-81	yes
General Surgery	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
General Surgery	GPT-4o	Claude-3-Haiku	NA	8.91E-41	6E-38	yes
General Surgery	GPT-4o	Qwen2.5-72b	231.4083	2.94E-52	2E-49	yes
General Surgery	GPT-4o	Qwen2.5-32b	78.10112	9.79E-19	6.6E-16	yes
General Surgery	GPT-4o	Qwen2.5-14b	NA	8.91E-41	6E-38	yes
General Surgery	Claude-3-Haiku	Qwen2.5-72b	NA	6.3E-130	4E-127	yes
General Surgery	Claude-3-Haiku	Qwen2.5-32b	NA	8.28E-84	5.6E-81	yes
General Surgery	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
General Surgery	Qwen2.5-72b	Qwen2.5-32b	78.09925	9.8E-19	6.6E-16	yes
General Surgery	Qwen2.5-72b	Qwen2.5-14b	NA	6.3E-130	4E-127	yes
General Surgery	Qwen2.5-32b	Qwen2.5-14b	NA	8.28E-84	5.6E-81	yes
Gynecology	ClinDiag-GPT	GPT-4o-mini	NA	0.595556	1	no
Gynecology	ClinDiag-GPT	GPT-4o	NA	0.164734	1	no
Gynecology	ClinDiag-GPT	Claude-3-Haiku	NA	0.595556	1	no
Gynecology	ClinDiag-GPT	Qwen2.5-72b	NA	0.325926	1	no
Gynecology	ClinDiag-GPT	Qwen2.5-32b	NA	0.595556	1	no
Gynecology	ClinDiag-GPT	Qwen2.5-14b	NA	0.595556	1	no
Gynecology	GPT-4o-mini	GPT-4o	NA	0.677617	1	no

Gynecology	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Gynecology	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Gynecology	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Gynecology	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Gynecology	GPT-4o	Claude-3-Haiku	NA	0.677617	1	no
Gynecology	GPT-4o	Qwen2.5-72b	NA	1	1	no
Gynecology	GPT-4o	Qwen2.5-32b	NA	0.677617	1	no
Gynecology	GPT-4o	Qwen2.5-14b	NA	0.677617	1	no
Gynecology	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Gynecology	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Gynecology	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Gynecology	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Gynecology	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Gynecology	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Hematology	ClinDiag-GPT	GPT-4o-mini	30.09225	4.12E-08	2.8E-05	yes
Hematology	ClinDiag-GPT	GPT-4o	30.09225	4.12E-08	2.8E-05	yes
Hematology	ClinDiag-GPT	Claude-3-Haiku	37.12609	1.11E-09	7.4E-07	yes
Hematology	ClinDiag-GPT	Qwen2.5-72b	12.99629	0.000312	0.209737	no
Hematology	ClinDiag-GPT	Qwen2.5-32b	16.11432	5.96E-05	0.04007	yes
Hematology	ClinDiag-GPT	Qwen2.5-14b	7.448181	0.00635	1	no
Hematology	GPT-4o-mini	GPT-4o	0	1	1	no
Hematology	GPT-4o-mini	Claude-3-Haiku	0.123631	0.72513	1	no
Hematology	GPT-4o-mini	Qwen2.5-72b	1.367965	0.242162	1	no
Hematology	GPT-4o-mini	Qwen2.5-32b	0.816303	0.366263	1	no
Hematology	GPT-4o-mini	Qwen2.5-14b	3.11081	0.077774	1	no
Hematology	GPT-4o	Claude-3-Haiku	0.123631	0.72513	1	no
Hematology	GPT-4o	Qwen2.5-72b	1.367965	0.242162	1	no
Hematology	GPT-4o	Qwen2.5-32b	0.816303	0.366263	1	no
Hematology	GPT-4o	Qwen2.5-14b	3.11081	0.077774	1	no

Hematology	Claude-3-Haiku	Qwen2.5-72b	2.137446	0.14374 1	1	no
Hematology	Claude-3-Haiku	Qwen2.5-32b	1.451206	0.22833 5	1	no
Hematology	Claude-3-Haiku	Qwen2.5-14b	4.157091	0.04146 1	1	no
Hematology	Qwen2.5-72b	Qwen2.5-32b	0.0907	0.76328 9	1	no
Hematology	Qwen2.5-72b	Qwen2.5-14b	0.43979	0.50722 4	1	no
Hematology	Qwen2.5-32b	Qwen2.5-14b	0.88043	0.34808 4	1	no
Hospice and Palliative Medicine	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Hospice and Palliative Medicine	ClinDiag-GPT	GPT-4o	NA	1	1	no
Hospice and Palliative Medicine	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Hospice and Palliative Medicine	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Hospice and Palliative Medicine	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Hospice and Palliative Medicine	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o-mini	GPT-4o	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o	Claude-3-Haiku	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o	Qwen2.5-72b	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o	Qwen2.5-32b	NA	1	1	no
Hospice and Palliative Medicine	GPT-4o	Qwen2.5-14b	NA	1	1	no
Hospice and Palliative Medicine	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no

Hospice and Palliative Medicine	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Hospice and Palliative Medicine	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Hospice and Palliative Medicine	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Hospice and Palliative Medicine	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Hospice and Palliative Medicine	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Immunology	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Immunology	ClinDiag-GPT	GPT-4o	NA	1	1	no
Immunology	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Immunology	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Immunology	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Immunology	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Immunology	GPT-4o-mini	GPT-4o	NA	1	1	no
Immunology	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Immunology	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Immunology	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Immunology	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Immunology	GPT-4o	Claude-3-Haiku	NA	1	1	no
Immunology	GPT-4o	Qwen2.5-72b	NA	1	1	no
Immunology	GPT-4o	Qwen2.5-32b	NA	1	1	no
Immunology	GPT-4o	Qwen2.5-14b	NA	1	1	no
Immunology	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Immunology	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Immunology	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Immunology	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Immunology	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Immunology	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Infectious Disease	ClinDiag-GPT	GPT-4o-mini	124.2012	7.61E-29	5.1E-26	yes
Infectious Disease	ClinDiag-GPT	GPT-4o	37.77847	7.93E-10	5.3E-07	yes
Infectious Disease	ClinDiag-GPT	Claude-3-Haiku	11.91718	0.000556	0.373757	no
Infectious Disease	ClinDiag-GPT	Qwen2.5-72b	37.77847	7.93E-10	5.3E-07	yes
Infectious Disease	ClinDiag-GPT	Qwen2.5-32b	124.2012	7.61E-29	5.1E-26	yes
Infectious Disease	ClinDiag-GPT	Qwen2.5-14b	11.91718	0.000556	0.373757	no

Infectious Disease	GPT-4o-mini	GPT-4o	4.666661	0.030754	1	no
Infectious Disease	GPT-4o-mini	Claude-3-Haiku	13.99998	0.000183	0.12285	no
Infectious Disease	GPT-4o-mini	Qwen2.5-72b	4.666661	0.030754	1	no
Infectious Disease	GPT-4o-mini	Qwen2.5-32b	0	1	1	no
Infectious Disease	GPT-4o-mini	Qwen2.5-14b	13.99998	0.000183	0.12285	no
Infectious Disease	GPT-4o	Claude-3-Haiku	3.499995	0.061369	1	no
Infectious Disease	GPT-4o	Qwen2.5-72b	0	1	1	no
Infectious Disease	GPT-4o	Qwen2.5-32b	8.399991	0.003752	1	no
Infectious Disease	GPT-4o	Qwen2.5-14b	3.499995	0.061369	1	no
Infectious Disease	Claude-3-Haiku	Qwen2.5-72b	4.666661	0.030754	1	no
Infectious Disease	Claude-3-Haiku	Qwen2.5-32b	33.59996	6.77E-09	4.5E-06	yes
Infectious Disease	Claude-3-Haiku	Qwen2.5-14b	0	1	1	no
Infectious Disease	Qwen2.5-72b	Qwen2.5-32b	8.399991	0.003752	1	no
Infectious Disease	Qwen2.5-72b	Qwen2.5-14b	3.499995	0.061369	1	no
Infectious Disease	Qwen2.5-32b	Qwen2.5-14b	13.99998	0.000183	0.12285	no
Medical Genetics and Genomics	ClinDiag-GPT	GPT-4o-mini	5.979636	0.014472	1	no
Medical Genetics and Genomics	ClinDiag-GPT	GPT-4o	3.322803	0.068325	1	no
Medical Genetics and Genomics	ClinDiag-GPT	Claude-3-Haiku	5.979636	0.014472	1	no
Medical Genetics and Genomics	ClinDiag-GPT	Qwen2.5-72b	50.71337	1.07E-12	7.2E-10	yes
Medical Genetics and Genomics	ClinDiag-GPT	Qwen2.5-32b	50.71337	1.07E-12	7.2E-10	yes
Medical Genetics and Genomics	ClinDiag-GPT	Qwen2.5-14b	3.322803	0.068325	1	no
Medical Genetics and Genomics	GPT-4o-mini	GPT-4o	0.31053	0.577355	1	no
Medical Genetics and Genomics	GPT-4o-mini	Claude-3-Haiku	0	1	1	no
Medical Genetics and Genomics	GPT-4o-mini	Qwen2.5-72b	14.75004	0.000123	0.082485	no

Medical Genetics and Genomics	GPT-4o-mini	Qwen2.5-32b	14.75004	0.000123	0.082485	no
Medical Genetics and Genomics	GPT-4o-mini	Qwen2.5-14b	0.31053	0.577355	1	no
Medical Genetics and Genomics	GPT-4o	Claude-3-Haiku	0.327782	0.566968	1	no
Medical Genetics and Genomics	GPT-4o	Qwen2.5-72b	21.2401	4.05E-06	0.00272	yes
Medical Genetics and Genomics	GPT-4o	Qwen2.5-32b	21.2401	4.05E-06	0.00272	yes
Medical Genetics and Genomics	GPT-4o	Qwen2.5-14b	0	1	1	no
Medical Genetics and Genomics	Claude-3-Haiku	Qwen2.5-72b	14.75004	0.000123	0.082485	no
Medical Genetics and Genomics	Claude-3-Haiku	Qwen2.5-32b	14.75004	0.000123	0.082485	no
Medical Genetics and Genomics	Claude-3-Haiku	Qwen2.5-14b	0.31053	0.577355	1	no
Medical Genetics and Genomics	Qwen2.5-72b	Qwen2.5-32b	0	1	1	no
Medical Genetics and Genomics	Qwen2.5-72b	Qwen2.5-14b	11.17899	0.000827	0.555938	no
Medical Genetics and Genomics	Qwen2.5-32b	Qwen2.5-14b	11.17899	0.000827	0.555938	no
Nephrology	ClinDiag-GPT	GPT-4o-mini	27.82627	1.33E-07	8.9E-05	yes
Nephrology	ClinDiag-GPT	GPT-4o	4.628044	0.031453	1	no
Nephrology	ClinDiag-GPT	Claude-3-Haiku	20.13159	7.23E-06	0.00486	yes
Nephrology	ClinDiag-GPT	Qwen2.5-72b	8.534157	0.003485	1	no
Nephrology	ClinDiag-GPT	Qwen2.5-32b	20.13159	7.23E-06	0.00486	yes
Nephrology	ClinDiag-GPT	Qwen2.5-14b	17.09686	3.55E-05	0.02387	yes
Nephrology	GPT-4o-mini	GPT-4o	4.568884	0.032558	1	no
Nephrology	GPT-4o-mini	Claude-3-Haiku	0.246761	0.619365	1	no
Nephrology	GPT-4o-mini	Qwen2.5-72b	2.449954	0.117528	1	no
Nephrology	GPT-4o-mini	Qwen2.5-32b	0.246761	0.619365	1	no
Nephrology	GPT-4o-mini	Qwen2.5-14b	0.528398	0.467281	1	no

Nephrology	GPT-4o	Claude-3-Haiku	3.948258	0.04692	1	no
Nephrology	GPT-4o	Qwen2.5-72b	0.450016	0.502327	1	no
Nephrology	GPT-4o	Qwen2.5-32b	3.948258	0.04692	1	no
Nephrology	GPT-4o	Qwen2.5-14b	2.876903	0.089859	1	no
Nephrology	Claude-3-Haiku	Qwen2.5-72b	1.249977	0.263557	1	no
Nephrology	Claude-3-Haiku	Qwen2.5-32b	0	1	1	no
Nephrology	Claude-3-Haiku	Qwen2.5-14b	0.058711	0.808545	1	no
Nephrology	Qwen2.5-72b	Qwen2.5-32b	1.542257	0.214282	1	no
Nephrology	Qwen2.5-72b	Qwen2.5-14b	0.939375	0.332439	1	no
Nephrology	Qwen2.5-32b	Qwen2.5-14b	0.058711	0.808545	1	no
Neurological Surgery	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Neurological Surgery	ClinDiag-GPT	GPT-4o	NA	1	1	no
Neurological Surgery	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Neurological Surgery	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Neurological Surgery	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Neurological Surgery	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Neurological Surgery	GPT-4o-mini	GPT-4o	NA	1	1	no
Neurological Surgery	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Neurological Surgery	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Neurological Surgery	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Neurological Surgery	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Neurological Surgery	GPT-4o	Claude-3-Haiku	NA	1	1	no
Neurological Surgery	GPT-4o	Qwen2.5-72b	NA	1	1	no
Neurological Surgery	GPT-4o	Qwen2.5-32b	NA	1	1	no
Neurological Surgery	GPT-4o	Qwen2.5-14b	NA	1	1	no
Neurological Surgery	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Neurological Surgery	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Neurological Surgery	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Neurological Surgery	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Neurological Surgery	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Neurological Surgery	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Neurology	ClinDiag-GPT	GPT-4o-mini	0.202332	0.652845	1	no
Neurology	ClinDiag-GPT	GPT-4o	15.4767	8.35E-05	0.056131	no
Neurology	ClinDiag-GPT	Claude-3-Haiku	0.202332	0.652845	1	no

Neurology	ClinDiag-GPT	Qwen2.5-72b	1.097369	0.294844	1	no
Neurology	ClinDiag-GPT	Qwen2.5-32b	13.66091	0.000219	0.147145	no
Neurology	ClinDiag-GPT	Qwen2.5-14b	3.119279	0.077371	1	no
Neurology	GPT-4o-mini	GPT-4o	18.96287	1.33E-05	0.00896	yes
Neurology	GPT-4o-mini	Claude-3-Haiku	0	1	1	no
Neurology	GPT-4o-mini	Qwen2.5-72b	0.34922	0.554555	1	no
Neurology	GPT-4o-mini	Qwen2.5-32b	10.24862	0.001368	0.919207	no
Neurology	GPT-4o-mini	Qwen2.5-14b	1.691257	0.193435	1	no
Neurology	GPT-4o	Claude-3-Haiku	21.68922	3.21E-06	0.00215	yes
Neurology	GPT-4o	Qwen2.5-72b	28.28571	1.05E-07	7E-05	yes
Neurology	GPT-4o	Qwen2.5-32b	69.27979	8.54E-17	5.7E-14	yes
Neurology	GPT-4o	Qwen2.5-14b	37.476	9.25E-10	6.2E-07	yes
Neurology	Claude-3-Haiku	Qwen2.5-72b	0.34922	0.554555	1	no
Neurology	Claude-3-Haiku	Qwen2.5-32b	10.24862	0.001368	0.919207	no
Neurology	Claude-3-Haiku	Qwen2.5-14b	1.691257	0.193435	1	no
Neurology	Qwen2.5-72b	Qwen2.5-32b	6.559078	0.010435	1	no
Neurology	Qwen2.5-72b	Qwen2.5-14b	0.486939	0.485296	1	no
Neurology	Qwen2.5-32b	Qwen2.5-14b	2.933851	0.08674	1	no
Obstetrics	ClinDiag-GPT	GPT-4o-mini	NA	0.005415	1	no
Obstetrics	ClinDiag-GPT	GPT-4o	2.793998	0.094618	1	no
Obstetrics	ClinDiag-GPT	Claude-3-Haiku	NA	0.005415	1	no
Obstetrics	ClinDiag-GPT	Qwen2.5-72b	0.379851	0.537683	1	no
Obstetrics	ClinDiag-GPT	Qwen2.5-32b	2.793998	0.094618	1	no

Obstetrics	ClinDiag-GPT	Qwen2.5-14b	2.793998	0.094618	1	no
Obstetrics	GPT-4o-mini	GPT-4o	NA	0.055109	1	no
Obstetrics	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Obstetrics	GPT-4o-mini	Qwen2.5-72b	NA	0.001066	0.716263	no
Obstetrics	GPT-4o-mini	Qwen2.5-32b	NA	0.055109	1	no
Obstetrics	GPT-4o-mini	Qwen2.5-14b	NA	0.055109	1	no
Obstetrics	GPT-4o	Claude-3-Haiku	NA	0.055109	1	no
Obstetrics	GPT-4o	Qwen2.5-72b	3.583333	0.058362	1	no
Obstetrics	GPT-4o	Qwen2.5-32b	0	1	1	no
Obstetrics	GPT-4o	Qwen2.5-14b	0	1	1	no
Obstetrics	Claude-3-Haiku	Qwen2.5-72b	NA	0.001066	0.716263	no
Obstetrics	Claude-3-Haiku	Qwen2.5-32b	NA	0.055109	1	no
Obstetrics	Claude-3-Haiku	Qwen2.5-14b	NA	0.055109	1	no
Obstetrics	Qwen2.5-72b	Qwen2.5-32b	6.142857	0.013194	1	no
Obstetrics	Qwen2.5-72b	Qwen2.5-14b	6.142857	0.013194	1	no
Obstetrics	Qwen2.5-32b	Qwen2.5-14b	0	1	1	no
Oncology	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Oncology	ClinDiag-GPT	GPT-4o	NA	0.606004	1	no
Oncology	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Oncology	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Oncology	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Oncology	ClinDiag-GPT	Qwen2.5-14b	NA	0.606004	1	no
Oncology	GPT-4o-mini	GPT-4o	NA	0.606004	1	no
Oncology	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Oncology	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Oncology	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Oncology	GPT-4o-mini	Qwen2.5-14b	NA	0.606004	1	no
Oncology	GPT-4o	Claude-3-Haiku	NA	1	1	no

Oncology	GPT-4o	Qwen2.5-72b	NA	1	1	no
Oncology	GPT-4o	Qwen2.5-32b	NA	1	1	no
Oncology	GPT-4o	Qwen2.5-14b	NA	1	1	no
Oncology	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Oncology	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Oncology	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Oncology	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Oncology	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Oncology	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Ophthalmology	ClinDiag-GPT	GPT-4o-mini	24.06548	9.31E-07	0.00063	yes
Ophthalmology	ClinDiag-GPT	GPT-4o	10.74458	0.001046	0.702812	no
Ophthalmology	ClinDiag-GPT	Claude-3-Haiku	63.97868	1.26E-15	8.5E-13	yes
Ophthalmology	ClinDiag-GPT	Qwen2.5-72b	30.7567	2.92E-08	2E-05	yes
Ophthalmology	ClinDiag-GPT	Qwen2.5-32b	30.7567	2.92E-08	2E-05	yes
Ophthalmology	ClinDiag-GPT	Qwen2.5-14b	63.97868	1.26E-15	8.5E-13	yes
Ophthalmology	GPT-4o-mini	GPT-4o	1.597364	0.206277	1	no
Ophthalmology	GPT-4o-mini	Claude-3-Haiku	4.685787	0.030413	1	no
Ophthalmology	GPT-4o-mini	Qwen2.5-72b	0.22163	0.637801	1	no
Ophthalmology	GPT-4o-mini	Qwen2.5-32b	0.22163	0.637801	1	no
Ophthalmology	GPT-4o-mini	Qwen2.5-14b	4.685787	0.030413	1	no
Ophthalmology	GPT-4o	Claude-3-Haiku	14.34998	0.000152	0.101997	no
Ophthalmology	GPT-4o	Qwen2.5-72b	3.545914	0.059692	1	no
Ophthalmology	GPT-4o	Qwen2.5-32b	3.545914	0.059692	1	no
Ophthalmology	GPT-4o	Qwen2.5-14b	14.34998	0.000152	0.101997	no
Ophthalmology	Claude-3-Haiku	Qwen2.5-72b	1.994608	0.15786	1	no
Ophthalmology	Claude-3-Haiku	Qwen2.5-32b	1.994608	0.15786	1	no
Ophthalmology	Claude-3-Haiku	Qwen2.5-14b	0	1	1	no
Ophthalmology	Qwen2.5-72b	Qwen2.5-32b	0	1	1	no

Ophthalmology	Qwen2.5-72b	Qwen2.5-14b	2.635734	0.10448 4	1	no
Ophthalmology	Qwen2.5-32b	Qwen2.5-14b	2.635734	0.10448 4	1	no
Orthopaedic Surgery	ClinDiag-GPT	GPT-4o-mini	0.592397	0.44149 4	1	no
Orthopaedic Surgery	ClinDiag-GPT	GPT-4o	4.249839	0.03925 4	1	no
Orthopaedic Surgery	ClinDiag-GPT	Claude-3-Haiku	0.592397	0.44149 4	1	no
Orthopaedic Surgery	ClinDiag-GPT	Qwen2.5-72b	0.714744	0.39787 3	1	no
Orthopaedic Surgery	ClinDiag-GPT	Qwen2.5-32b	0.198284	0.65610 9	1	no
Orthopaedic Surgery	ClinDiag-GPT	Qwen2.5-14b	0.198284	0.65610 9	1	no
Orthopaedic Surgery	GPT-4o-mini	GPT-4o	7.201613	0.00728 4	1	no
Orthopaedic Surgery	GPT-4o-mini	Claude-3-Haiku	0	1	1	no
Orthopaedic Surgery	GPT-4o-mini	Qwen2.5-72b	2.282661	0.13082 7	1	no
Orthopaedic Surgery	GPT-4o-mini	Qwen2.5-32b	1.280137	0.25787 3	1	no
Orthopaedic Surgery	GPT-4o-mini	Qwen2.5-14b	1.280137	0.25787 3	1	no
Orthopaedic Surgery	GPT-4o	Claude-3-Haiku	11.02469	0.00089 9	0.60417	no
Orthopaedic Surgery	GPT-4o	Qwen2.5-72b	1.848956	0.17390 5	1	no
Orthopaedic Surgery	GPT-4o	Qwen2.5-32b	3.343614	0.06746 6	1	no
Orthopaedic Surgery	GPT-4o	Qwen2.5-14b	3.343614	0.06746 6	1	no
Orthopaedic Surgery	Claude-3-Haiku	Qwen2.5-72b	2.282661	0.13082 7	1	no
Orthopaedic Surgery	Claude-3-Haiku	Qwen2.5-32b	1.280137	0.25787 3	1	no
Orthopaedic Surgery	Claude-3-Haiku	Qwen2.5-14b	1.280137	0.25787 3	1	no
Orthopaedic Surgery	Qwen2.5-72b	Qwen2.5-32b	0.181865	0.66977 5	1	no
Orthopaedic Surgery	Qwen2.5-72b	Qwen2.5-14b	0.181865	0.66977 5	1	no
Orthopaedic Surgery	Qwen2.5-32b	Qwen2.5-14b	0	1	1	no

Otolaryngology-Head and Neck Surgery	ClinDiag-GPT	GPT-4o-mini	0	1	1	no
Otolaryngology-Head and Neck Surgery	ClinDiag-GPT	GPT-4o	0.601618	0.437961	1	no
Otolaryngology-Head and Neck Surgery	ClinDiag-GPT	Claude-3-Haiku	0.137929	0.710349	1	no
Otolaryngology-Head and Neck Surgery	ClinDiag-GPT	Qwen2.5-72b	3.379289	0.066021	1	no
Otolaryngology-Head and Neck Surgery	ClinDiag-GPT	Qwen2.5-32b	1.317401	0.251059	1	no
Otolaryngology-Head and Neck Surgery	ClinDiag-GPT	Qwen2.5-14b	NA	0.704873	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o-mini	GPT-4o	0.601618	0.437961	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o-mini	Claude-3-Haiku	0.137929	0.710349	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o-mini	Qwen2.5-72b	3.379289	0.066021	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o-mini	Qwen2.5-32b	1.317401	0.251059	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o-mini	Qwen2.5-14b	NA	0.704873	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o	Claude-3-Haiku	0.201385	0.653605	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o	Qwen2.5-72b	1.305001	0.253302	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o	Qwen2.5-32b	0.164779	0.684794	1	no
Otolaryngology-Head and Neck Surgery	GPT-4o	Qwen2.5-14b	NA	0.468772	1	no
Otolaryngology-Head and Neck Surgery	Claude-3-Haiku	Qwen2.5-72b	2.31999	0.127721	1	no
Otolaryngology-Head and Neck Surgery	Claude-3-Haiku	Qwen2.5-32b	0.659097	0.416879	1	no
Otolaryngology-Head and Neck Surgery	Claude-3-Haiku	Qwen2.5-14b	NA	0.704873	1	no
Otolaryngology-Head and Neck Surgery	Qwen2.5-72b	Qwen2.5-32b	0.659077	0.416886	1	no
Otolaryngology-Head and Neck Surgery	Qwen2.5-72b	Qwen2.5-14b	NA	0.101075	1	no
Otolaryngology-Head and Neck Surgery	Qwen2.5-32b	Qwen2.5-14b	NA	0.295464	1	no
Pediatrics	ClinDiag-GPT	GPT-4o-mini	46.59363	8.73E-12	5.9E-09	yes

Pediatrics	ClinDiag-GPT	GPT-4o	46.59363	8.73E-12	5.9E-09	yes
Pediatrics	ClinDiag-GPT	Claude-3-Haiku	115.8163	5.21E-27	3.5E-24	yes
Pediatrics	ClinDiag-GPT	Qwen2.5-72b	46.59363	8.73E-12	5.9E-09	yes
Pediatrics	ClinDiag-GPT	Qwen2.5-32b	72.1068	2.04E-17	1.4E-14	yes
Pediatrics	ClinDiag-GPT	Qwen2.5-14b	67.49404	2.11E-16	1.4E-13	yes
Pediatrics	GPT-4o-mini	GPT-4o	0	1	1	no
Pediatrics	GPT-4o-mini	Claude-3-Haiku	5.333242	0.020922	1	no
Pediatrics	GPT-4o-mini	Qwen2.5-72b	0	1	1	no
Pediatrics	GPT-4o-mini	Qwen2.5-32b	1.041651	0.307438	1	no
Pediatrics	GPT-4o-mini	Qwen2.5-14b	0.733405	0.391782	1	no
Pediatrics	GPT-4o	Claude-3-Haiku	5.333242	0.020922	1	no
Pediatrics	GPT-4o	Qwen2.5-72b	0	1	1	no
Pediatrics	GPT-4o	Qwen2.5-32b	1.041651	0.307438	1	no
Pediatrics	GPT-4o	Qwen2.5-14b	0.733405	0.391782	1	no
Pediatrics	Claude-3-Haiku	Qwen2.5-72b	3.478212	0.062182	1	no
Pediatrics	Claude-3-Haiku	Qwen2.5-32b	1.041651	0.307438	1	no
Pediatrics	Claude-3-Haiku	Qwen2.5-14b	1.332336	0.24839	1	no
Pediatrics	Qwen2.5-72b	Qwen2.5-32b	1.041651	0.307438	1	no
Pediatrics	Qwen2.5-72b	Qwen2.5-14b	0.733405	0.391782	1	no
Pediatrics	Qwen2.5-32b	Qwen2.5-14b	0.022183	0.881602	1	no
Physical Medicine and Rehabilitation	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Physical Medicine and Rehabilitation	ClinDiag-GPT	GPT-4o	NA	1	1	no
Physical Medicine and Rehabilitation	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Physical Medicine and Rehabilitation	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no

Physical Medicine and Rehabilitation	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Physical Medicine and Rehabilitation	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o-mini	GPT-4o	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o	Claude-3-Haiku	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o	Qwen2.5-72b	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o	Qwen2.5-32b	NA	1	1	no
Physical Medicine and Rehabilitation	GPT-4o	Qwen2.5-14b	NA	1	1	no
Physical Medicine and Rehabilitation	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Physical Medicine and Rehabilitation	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Physical Medicine and Rehabilitation	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Physical Medicine and Rehabilitation	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Physical Medicine and Rehabilitation	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Physical Medicine and Rehabilitation	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
				0.050356	1	no
Plastic Surgery	ClinDiag-GPT	GPT-4o-mini	NA	0.844113	1	no
Plastic Surgery	ClinDiag-GPT	GPT-4o	0.038666	0.29349	1	no
Plastic Surgery	ClinDiag-GPT	Claude-3-Haiku	NA	0.844113	1	no
Plastic Surgery	ClinDiag-GPT	Qwen2.5-72b	0.038666	0.490153	1	no
Plastic Surgery	ClinDiag-GPT	Qwen2.5-32b	0.47619			

Plastic Surgery	ClinDiag-GPT	Qwen2.5-14b	0.038666	0.844113	1	no
Plastic Surgery	GPT-4o-mini	GPT-4o	NA	0.050356	1	no
Plastic Surgery	GPT-4o-mini	Claude-3-Haiku	NA	0.610433	1	no
Plastic Surgery	GPT-4o-mini	Qwen2.5-72b	NA	0.050356	1	no
Plastic Surgery	GPT-4o-mini	Qwen2.5-32b	NA	0.100285	1	no
Plastic Surgery	GPT-4o-mini	Qwen2.5-14b	NA	0.050356	1	no
Plastic Surgery	GPT-4o	Claude-3-Haiku	NA	0.29349	1	no
Plastic Surgery	GPT-4o	Qwen2.5-72b	0	1	1	no
Plastic Surgery	GPT-4o	Qwen2.5-32b	0.233339	0.629059	1	no
Plastic Surgery	GPT-4o	Qwen2.5-14b	0	1	1	no
Plastic Surgery	Claude-3-Haiku	Qwen2.5-72b	NA	0.29349	1	no
Plastic Surgery	Claude-3-Haiku	Qwen2.5-32b	NA	0.467157	1	no
Plastic Surgery	Claude-3-Haiku	Qwen2.5-14b	NA	0.29349	1	no
Plastic Surgery	Qwen2.5-72b	Qwen2.5-32b	0.233339	0.629059	1	no
Plastic Surgery	Qwen2.5-72b	Qwen2.5-14b	0	1	1	no
Plastic Surgery	Qwen2.5-32b	Qwen2.5-14b	0.210531	0.646351	1	no
Psychiatry	ClinDiag-GPT	GPT-4o-mini	7.190198	0.00733	1	no
Psychiatry	ClinDiag-GPT	GPT-4o	14.50655	0.00014	0.09386	no
Psychiatry	ClinDiag-GPT	Claude-3-Haiku	14.50655	0.00014	0.09386	no
Psychiatry	ClinDiag-GPT	Qwen2.5-72b	119.7872	7.04E-28	4.7E-25	yes
Psychiatry	ClinDiag-GPT	Qwen2.5-32b	7.190198	0.00733	1	no
Psychiatry	ClinDiag-GPT	Qwen2.5-14b	0.471193	0.492439	1	no
Psychiatry	GPT-4o-mini	GPT-4o	1.037564	0.308389	1	no
Psychiatry	GPT-4o-mini	Claude-3-Haiku	1.037564	0.308389	1	no
Psychiatry	GPT-4o-mini	Qwen2.5-72b	49.9995	1.54E-12	1E-09	yes
Psychiatry	GPT-4o-mini	Qwen2.5-32b	0	1	1	no
Psychiatry	GPT-4o-mini	Qwen2.5-14b	3.449994	0.063252	1	no
Psychiatry	GPT-4o	Claude-3-Haiku	0	1	1	no

Psychiatry	GPT-4o	Qwen2.5-72b	31.99984	1.54E-08	1E-05	yes
Psychiatry	GPT-4o	Qwen2.5-32b	0.958307	0.327614	1	no
Psychiatry	GPT-4o	Qwen2.5-14b	7.76242	0.005334	1	no
Psychiatry	Claude-3-Haiku	Qwen2.5-72b	31.99984	1.54E-08	1E-05	yes
Psychiatry	Claude-3-Haiku	Qwen2.5-32b	0.958307	0.327614	1	no
Psychiatry	Claude-3-Haiku	Qwen2.5-14b	7.76242	0.005334	1	no
Psychiatry	Qwen2.5-72b	Qwen2.5-32b	23.95817	9.85E-07	0.00066	yes
Psychiatry	Qwen2.5-72b	Qwen2.5-14b	42.26227	7.98E-11	5.4E-08	yes
Psychiatry	Qwen2.5-32b	Qwen2.5-14b	3.449994	0.063252	1	no
Pulmonology	ClinDiag-GPT	GPT-4o-mini	0	1	1	no
Pulmonology	ClinDiag-GPT	GPT-4o	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	ClinDiag-GPT	Claude-3-Haiku	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	ClinDiag-GPT	Qwen2.5-72b	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	ClinDiag-GPT	Qwen2.5-32b	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	ClinDiag-GPT	Qwen2.5-14b	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	GPT-4o-mini	GPT-4o	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	GPT-4o-mini	Claude-3-Haiku	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	GPT-4o-mini	Qwen2.5-72b	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	GPT-4o-mini	Qwen2.5-32b	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	GPT-4o-mini	Qwen2.5-14b	32.66667	1.09E-08	7.4E-06	yes
Pulmonology	GPT-4o	Claude-3-Haiku	0	1	1	no
Pulmonology	GPT-4o	Qwen2.5-72b	0	1	1	no
Pulmonology	GPT-4o	Qwen2.5-32b	0	1	1	no
Pulmonology	GPT-4o	Qwen2.5-14b	0	1	1	no
Pulmonology	Claude-3-Haiku	Qwen2.5-72b	0	1	1	no

Pulmonology	Claude-3-Haiku	Qwen2.5-32b	0	1	1	no
Pulmonology	Claude-3-Haiku	Qwen2.5-14b	0	1	1	no
Pulmonology	Qwen2.5-72b	Qwen2.5-32b	0	1	1	no
Pulmonology	Qwen2.5-72b	Qwen2.5-14b	0	1	1	no
Pulmonology	Qwen2.5-32b	Qwen2.5-14b	0	1	1	no
Rheumatology	ClinDiag-GPT	GPT-4o-mini	3.529834	0.06027 4	1	no
Rheumatology	ClinDiag-GPT	GPT-4o	2.724785	0.09880 1	1	no
Rheumatology	ClinDiag-GPT	Claude-3-Haiku	3.891516	0.04853 1	1	no
Rheumatology	ClinDiag-GPT	Qwen2.5-72b	6.621956	0.01007 3	1	no
Rheumatology	ClinDiag-GPT	Qwen2.5-32b	2.80095	0.09420 8	1	no
Rheumatology	ClinDiag-GPT	Qwen2.5-14b	3.345205	0.06740 1	1	no
Rheumatology	GPT-4o-mini	GPT-4o	0.050596	0.82202 9	1	no
Rheumatology	GPT-4o-mini	Claude-3-Haiku	0.008385	0.92703 8	1	no
Rheumatology	GPT-4o-mini	Qwen2.5-72b	0.440747	0.50676 2	1	no
Rheumatology	GPT-4o-mini	Qwen2.5-32b	0.040876	0.83977 8	1	no
Rheumatology	GPT-4o-mini	Qwen2.5-14b	0.002381	0.96108	1	no
Rheumatology	GPT-4o	Claude-3-Haiku	0.100595	0.75111 7	1	no
Rheumatology	GPT-4o	Qwen2.5-72b	0.79723	0.37192 3	1	no
Rheumatology	GPT-4o	Qwen2.5-32b	0.000519	0.98182 8	1	no
Rheumatology	GPT-4o	Qwen2.5-14b	0.031116	0.85998 2	1	no
Rheumatology	Claude-3-Haiku	Qwen2.5-72b	0.32597	0.56804 2	1	no
Rheumatology	Claude-3-Haiku	Qwen2.5-32b	0.086136	0.76914 8	1	no
Rheumatology	Claude-3-Haiku	Qwen2.5-14b	0.019658	0.88849 5	1	no
Rheumatology	Qwen2.5-72b	Qwen2.5-32b	0.729942	0.39290 2	1	no
Rheumatology	Qwen2.5-72b	Qwen2.5-14b	0.492363	0.48287 5	1	no

Rheumatology	Qwen2.5-32b	Qwen2.5-14b	0.023591	0.87792 9	1	no
Thoracic Surgery	ClinDiag-GPT	GPT-4o-mini	NA	1	1	no
Thoracic Surgery	ClinDiag-GPT	GPT-4o	NA	1	1	no
Thoracic Surgery	ClinDiag-GPT	Claude-3-Haiku	NA	1	1	no
Thoracic Surgery	ClinDiag-GPT	Qwen2.5-72b	NA	1	1	no
Thoracic Surgery	ClinDiag-GPT	Qwen2.5-32b	NA	1	1	no
Thoracic Surgery	ClinDiag-GPT	Qwen2.5-14b	NA	1	1	no
Thoracic Surgery	GPT-4o-mini	GPT-4o	NA	1	1	no
Thoracic Surgery	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Thoracic Surgery	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Thoracic Surgery	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no
Thoracic Surgery	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Thoracic Surgery	GPT-4o	Claude-3-Haiku	NA	0.58204 3	1	no
Thoracic Surgery	GPT-4o	Qwen2.5-72b	NA	1	1	no
Thoracic Surgery	GPT-4o	Qwen2.5-32b	NA	0.58204 3	1	no
Thoracic Surgery	GPT-4o	Qwen2.5-14b	NA	1	1	no
Thoracic Surgery	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Thoracic Surgery	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Thoracic Surgery	Claude-3-Haiku	Qwen2.5-14b	NA	0.58204 3	1	no
Thoracic Surgery	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Thoracic Surgery	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Thoracic Surgery	Qwen2.5-32b	Qwen2.5-14b	NA	0.58204 3	1	no
Urology	ClinDiag-GPT	GPT-4o-mini	NA	0.02261 8	1	no
Urology	ClinDiag-GPT	GPT-4o	0.666666	0.41421 7	1	no
Urology	ClinDiag-GPT	Claude-3-Haiku	NA	0.02261 8	1	no
Urology	ClinDiag-GPT	Qwen2.5-72b	NA	0.02261 8	1	no
Urology	ClinDiag-GPT	Qwen2.5-32b	NA	0.02261 8	1	no
Urology	ClinDiag-GPT	Qwen2.5-14b	NA	0.02261 8	1	no
Urology	GPT-4o-mini	GPT-4o	NA	0.00415 2	1	no
Urology	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Urology	GPT-4o-mini	Qwen2.5-72b	NA	1	1	no
Urology	GPT-4o-mini	Qwen2.5-32b	NA	1	1	no

Urology	GPT-4o-mini	Qwen2.5-14b	NA	1	1	no
Urology	GPT-4o	Claude-3-Haiku	NA	0.00415 2	1	no
Urology	GPT-4o	Qwen2.5-72b	NA	0.00415 2	1	no
Urology	GPT-4o	Qwen2.5-32b	NA	0.00415 2	1	no
Urology	GPT-4o	Qwen2.5-14b	NA	0.00415 2	1	no
Urology	Claude-3-Haiku	Qwen2.5-72b	NA	1	1	no
Urology	Claude-3-Haiku	Qwen2.5-32b	NA	1	1	no
Urology	Claude-3-Haiku	Qwen2.5-14b	NA	1	1	no
Urology	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Urology	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Urology	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no
Vascular Surgery	ClinDiag-GPT	GPT-4o-mini	NA	0.03250 8	1	no
Vascular Surgery	ClinDiag-GPT	GPT-4o	NA	0.03250 8	1	no
Vascular Surgery	ClinDiag-GPT	Claude-3-Haiku	NA	0.03250 8	1	no
Vascular Surgery	ClinDiag-GPT	Qwen2.5-72b	NA	0.34984 5	1	no
Vascular Surgery	ClinDiag-GPT	Qwen2.5-32b	NA	0.34984 5	1	no
Vascular Surgery	ClinDiag-GPT	Qwen2.5-14b	NA	0.34984 5	1	no
Vascular Surgery	GPT-4o-mini	GPT-4o	NA	1	1	no
Vascular Surgery	GPT-4o-mini	Claude-3-Haiku	NA	1	1	no
Vascular Surgery	GPT-4o-mini	Qwen2.5-72b	NA	0.47058 8	1	no
Vascular Surgery	GPT-4o-mini	Qwen2.5-32b	NA	0.47058 8	1	no
Vascular Surgery	GPT-4o-mini	Qwen2.5-14b	NA	0.47058 8	1	no
Vascular Surgery	GPT-4o	Claude-3-Haiku	NA	1	1	no
Vascular Surgery	GPT-4o	Qwen2.5-72b	NA	0.47058 8	1	no
Vascular Surgery	GPT-4o	Qwen2.5-32b	NA	0.47058 8	1	no
Vascular Surgery	GPT-4o	Qwen2.5-14b	NA	0.47058 8	1	no
Vascular Surgery	Claude-3-Haiku	Qwen2.5-72b	NA	0.47058 8	1	no

Vascular Surgery	Claude-3-Haiku	Qwen2.5-32b	NA	0.47058 8	1	no
Vascular Surgery	Claude-3-Haiku	Qwen2.5-14b	NA	0.47058 8	1	no
Vascular Surgery	Qwen2.5-72b	Qwen2.5-32b	NA	1	1	no
Vascular Surgery	Qwen2.5-72b	Qwen2.5-14b	NA	1	1	no
Vascular Surgery	Qwen2.5-32b	Qwen2.5-14b	NA	1	1	no

Table 8 Statistical Analysis Results for Error Analysis in Clinical Diagnostic Procedure

Category	Model 1	Model 2	Chi-Square Value	P-Value	Degree of Freedom	Corrected P-Value
Medical History Step	ClinDiag-	GPT-4o-	36.41705099	1.59303	1	3.345E-08
	GPT	mini		E-09		
Medical History Step	ClinDiag-	GPT-4o	38.85669486	4.56087	1	9.578E-09
	GPT			E-10		
Medical History Step	ClinDiag-	Claude-3-	23.30050534	1.38564	1	2.91E-05
	GPT	Haiku		E-06		
Medical History Step	ClinDiag-	Qwen2.5-	21.37349975	3.77959	1	7.937E-05
	GPT	72b		E-06		
Medical History Step	ClinDiag-	Qwen2.5-	37.61663611	8.6107E	1	1.808E-08
	GPT	32b		-10		
Medical History Step	ClinDiag-	Qwen2.5-	15.20329373	9.65349	1	0.0020272
	GPT	14b		E-05		
Medical History Step	GPT-4o-	GPT-4o	0.011088669	0.91613	1	1
	mini			558		
Medical History Step	GPT-4o-	Claude-3-	1.347405636	0.24573	1	1
	mini	Haiku		218		
Medical History Step	GPT-4o-	Qwen2.5-	1.882644882	0.17003	1	1
	mini	72b		3844		
Medical History Step	GPT-4o-	Qwen2.5-	0	1	1	1
	mini	32b				
Medical History Step	GPT-4o-	Qwen2.5-	4.478185145	0.03433	1	0.720933304
	mini	14b		0157		
Medical History Step	GPT-4o	Claude-3-	1.879859981	0.17035	1	1
		Haiku		0063		
Medical History Step	GPT-4o	Qwen2.5-	2.503442754	0.11359	1	1
		72b		7725		
Medical History Step	GPT-4o	Qwen2.5-	7.12893E-07	0.99932	1	1
		32b		6322		
Medical History Step	GPT-4o	Qwen2.5-	5.407924154	0.02004	1	0.420956331
		14b		554		
Medical History Step	Claude-3-	Qwen2.5-	0.011205017	0.91569	1	1
	Haiku	72b		8388		
Medical History Step	Claude-3-	Qwen2.5-	1.600170187	0.20587	1	1
	Haiku	32b		9094		
Medical History Step	Claude-3-	Qwen2.5-	0.727479398	0.39370	1	1
	Haiku	14b		1436		
Medical History Step	Qwen2.5-	Qwen2.5-	2.179209112	0.13988	1	1
	72b	32b		6331		
Medical History Step	Qwen2.5-	Qwen2.5-	0.410931708	0.52149	1	1
	72b	14b		6986		

Medical History Step	Qwen2.5-32b	Qwen2.5-14b	4.927961675	0.02642547	1	0.554934868
History of Present Illness	ClinDiag-GPT	GPT-4o-mini	46.27355989	1.02842E-11	1	2.16E-10
History of Present Illness	ClinDiag-GPT	GPT-4o	38.65399082	5.05997E-10	1	1.063E-08
History of Present Illness	ClinDiag-GPT	Claude-3-Haiku	21.99744325	2.73014E-06	1	5.733E-05
History of Present Illness	ClinDiag-GPT	Qwen2.5-72b	22.98826195	1.62994E-06	1	3.423E-05
History of Present Illness	ClinDiag-GPT	Qwen2.5-32b	37.44773935	9.38955E-10	1	1.972E-08
History of Present Illness	ClinDiag-GPT	Qwen2.5-14b	21.99744325	2.73014E-06	1	5.733E-05
History of Present Illness	GPT-4o-mini	GPT-4o	0.283021062	0.594727908	1	1
History of Present Illness	GPT-4o-mini	Claude-3-Haiku	4.682079092	0.030478839	1	0.640055618
History of Present Illness	GPT-4o-mini	Qwen2.5-72b	4.21874569	0.039979704	1	0.839573794
History of Present Illness	GPT-4o-mini	Qwen2.5-32b	0.407027633	0.52348203	1	1
History of Present Illness	GPT-4o-mini	Qwen2.5-14b	4.682079092	0.030478839	1	0.640055618
History of Present Illness	GPT-4o	Claude-3-Haiku	2.335102977	0.126486626	1	1
History of Present Illness	GPT-4o	Qwen2.5-72b	2.010266987	0.156237821	1	1
History of Present Illness	GPT-4o	Qwen2.5-32b	0	1	1	1
History of Present Illness	GPT-4o	Qwen2.5-14b	2.335102977	0.126486626	1	1
History of Present Illness	Claude-3-Haiku	Qwen2.5-72b	0	1	1	1
History of Present Illness	Claude-3-Haiku	Qwen2.5-32b	2.022697897	0.154963602	1	1
History of Present Illness	Claude-3-Haiku	Qwen2.5-14b	0	1	1	1
History of Present Illness	Qwen2.5-72b	Qwen2.5-32b	1.721066113	0.189555871	1	1
History of Present Illness	Qwen2.5-72b	Qwen2.5-14b	0	1	1	1
History of Present Illness	Qwen2.5-32b	Qwen2.5-14b	2.022697897	0.154963602	1	1

Past Medical History	ClinDiag- GPT	GPT-4o- mini	8.547117028	0.00346 0699	1	0.072674675
Past Medical History	ClinDiag- GPT	GPT-4o	7.787826283	0.00525 9944	1	0.110458827
Past Medical History	ClinDiag- GPT	Claude-3- Haiku	5.630641914	0.01764 9167	1	0.370632514
Past Medical History	ClinDiag- GPT	Qwen2.5- 72b	0	1	1	1
Past Medical History	ClinDiag- GPT	Qwen2.5- 32b	13.60618522	0.00022 5442	1	0.0047343
Past Medical History	ClinDiag- GPT	Qwen2.5- 14b	4.333991287	0.03735 8546	1	0.784529468
Past Medical History	GPT-4o- mini	GPT-4o	0	1	1	1
Past Medical History	GPT-4o- mini	Claude-3- Haiku	0.184491084	0.66754 1748	1	1
Past Medical History	GPT-4o- mini	Qwen2.5- 72b	7.483910059	0.00622 5274	1	0.130730754
Past Medical History	GPT-4o- mini	Qwen2.5- 32b	0.455421222	0.49977 1682	1	1
Past Medical History	GPT-4o- mini	Qwen2.5- 14b	0.527412744	0.46769 6631	1	1
Past Medical History	GPT-4o	Claude-3- Haiku	0.083870602	0.77211 9041	1	1
Past Medical History	GPT-4o	Qwen2.5- 72b	6.769241177	0.00927 4226	1	0.194758744
Past Medical History	GPT-4o	Qwen2.5- 32b	0.66051407	0.41637 8516	1	1
Past Medical History	GPT-4o	Qwen2.5- 14b	0.344014173	0.55752 1279	1	1
Past Medical History	Claude-3- Haiku	Qwen2.5- 72b	4.756466647	0.02918 8408	1	0.612956571
Past Medical History	Claude-3- Haiku	Qwen2.5- 32b	1.541860296	0.21434 1116	1	1
Past Medical History	Claude-3- Haiku	Qwen2.5- 14b	0.022386987	0.88106 2207	1	1
Past Medical History	Qwen2.5- 72b	Qwen2.5- 32b	12.29845043	0.00045 3335	1	0.00952
Past Medical History	Qwen2.5- 72b	Qwen2.5- 14b	3.56403705	0.05904 3977	1	1
Past Medical History	Qwen2.5- 32b	Qwen2.5- 14b	2.35992443	0.12448 822	1	1
Family History	ClinDiag- GPT	GPT-4o- mini	22.53866498	2.05956 E-06	1	4.325E-05

Family History	ClinDiag- GPT	GPT-4o	31.69857248	1.80054 E-08	1	3.781E-07
Family History	ClinDiag- GPT	Claude-3- Haiku	23.6356222	1.16412 E-06	1	2.445E-05
Family History	ClinDiag- GPT	Qwen2.5- 72b	23.6356222	1.16412 E-06	1	2.445E-05
Family History	ClinDiag- GPT	Qwen2.5- 32b	22.53866498	2.05956 E-06	1	4.325E-05
Family History	ClinDiag- GPT	Qwen2.5- 14b	7.545783164	0.00601 5067	1	0.126316413
Family History	GPT-4o- mini	GPT-4o	0.815436504	0.36651 7812	1	1
Family History	GPT-4o- mini	Claude-3- Haiku	0	1	1	1
Family History	GPT-4o- mini	Qwen2.5- 72b	0	1	1	1
Family History	GPT-4o- mini	Qwen2.5- 32b	0	1	1	1
Family History	GPT-4o- mini	Qwen2.5- 14b	4.332157195	0.03739 8819	1	0.785375209
Family History	GPT-4o	Claude-3- Haiku	0.595216289	0.44040 9071	1	1
Family History	GPT-4o	Qwen2.5- 72b	0.595216289	0.44040 9071	1	1
Family History	GPT-4o	Qwen2.5- 32b	0.815436504	0.36651 7812	1	1
Family History	GPT-4o	Qwen2.5- 14b	9.540751642	0.00200 9594	1	0.0422015
Family History	Claude-3- Haiku	Qwen2.5- 72b	0	1	1	1
Family History	Claude-3- Haiku	Qwen2.5- 32b	0	1	1	1
Family History	Claude-3- Haiku	Qwen2.5- 14b	4.891508845	0.02698 9089	1	0.566770877
Family History	Qwen2.5- 72b	Qwen2.5- 32b	0	1	1	1
Family History	Qwen2.5- 72b	Qwen2.5- 14b	4.891508845	0.02698 9089	1	0.566770877
Family History	Qwen2.5- 32b	Qwen2.5- 14b	4.332157195	0.03739 8819	1	0.785375209
Social History	ClinDiag- GPT	GPT-4o- mini	30.7901879	2.87488 E-08	1	6.037E-07
Social History	ClinDiag- GPT	GPT-4o	33.12207133	8.65507 E-09	1	1.818E-07

Social History	ClinDiag-GPT	Claude-3-Haiku	20.94222257	4.73346E-06	1	9.94E-05
Social History	ClinDiag-GPT	Qwen2.5-72b	24.10799916	9.10812E-07	1	1.913E-05
Social History	ClinDiag-GPT	Qwen2.5-32b	21.97059324	2.7686E-06	1	5.814E-05
Social History	ClinDiag-GPT	Qwen2.5-14b	17.90896663	2.31727E-05	1	0.0004866
Social History	GPT-4o-mini	GPT-4o	0.013317393	0.908127278	1	1
Social History	GPT-4o-mini	Claude-3-Haiku	0.915168926	0.33874666	1	1
Social History	GPT-4o-mini	Qwen2.5-72b	0.349658808	0.554306334	1	1
Social History	GPT-4o-mini	Qwen2.5-32b	0.698040548	0.403442883	1	1
Social History	GPT-4o-mini	Qwen2.5-14b	1.768663025	0.183547627	1	1
Social History	GPT-4o	Claude-3-Haiku	1.410573209	0.234961261	1	1
Social History	GPT-4o	Qwen2.5-72b	0.676878334	0.410663736	1	1
Social History	GPT-4o	Qwen2.5-32b	1.137768962	0.286124251	1	1
Social History	GPT-4o	Qwen2.5-14b	2.434775273	0.118670901	1	1
Social History	Claude-3-Haiku	Qwen2.5-72b	0.060058197	0.806403981	1	1
Social History	Claude-3-Haiku	Qwen2.5-32b	0	1	1	1
Social History	Claude-3-Haiku	Qwen2.5-14b	0.063010631	0.801799277	1	1
Social History	Qwen2.5-72b	Qwen2.5-32b	0.015292681	0.901581648	1	1
Social History	Qwen2.5-72b	Qwen2.5-14b	0.383761809	0.535596824	1	1
Social History	Qwen2.5-32b	Qwen2.5-14b	0.139392984	0.708885215	1	1
Physical Examination Step	ClinDiag-GPT	GPT-4o-mini	23.74727017	1.09851E-06	1	2.307E-05
Physical Examination Step	ClinDiag-GPT	GPT-4o	22.75601649	1.83927E-06	1	3.862E-05
Physical Examination Step	ClinDiag-GPT	Claude-3-Haiku	23.74727017	1.09851E-06	1	2.307E-05

Physical Examination Step	ClinDiag-GPT	Qwen2.5-72b	3.438747789	0.063683943	1	1
Physical Examination Step	ClinDiag-GPT	Qwen2.5-32b	27.85675858	1.30637E-07	1	2.743E-06
Physical Examination Step	ClinDiag-GPT	Qwen2.5-14b	15.55252414	8.02444E-05	1	0.0016851
Physical Examination Step	GPT-4o-mini	GPT-4o	7.22918E-07	0.999321602	1	1
Physical Examination Step	GPT-4o-mini	Claude-3-Haiku	0	1	1	1
Physical Examination Step	GPT-4o-mini	Qwen2.5-72b	8.782577131	0.003041212	1	0.0638655
Physical Examination Step	GPT-4o-mini	Qwen2.5-32b	0.10074092	0.750942306	1	1
Physical Examination Step	GPT-4o-mini	Qwen2.5-14b	0.732841956	0.391963548	1	1
Physical Examination Step	GPT-4o	Claude-3-Haiku	7.22918E-07	0.999321602	1	1
Physical Examination Step	GPT-4o	Qwen2.5-72b	8.167052507	0.004265817	1	0.0895822
Physical Examination Step	GPT-4o	Qwen2.5-32b	0.180142343	0.671250942	1	1
Physical Examination Step	GPT-4o	Qwen2.5-14b	0.560933409	0.453884402	1	1
Physical Examination Step	Claude-3-Haiku	Qwen2.5-72b	8.782577131	0.003041212	1	0.0638655
Physical Examination Step	Claude-3-Haiku	Qwen2.5-32b	0.10074092	0.750942306	1	1
Physical Examination Step	Claude-3-Haiku	Qwen2.5-14b	0.732841956	0.391963548	1	1
Physical Examination Step	Qwen2.5-72b	Qwen2.5-32b	11.4270802	0.000723813	1	0.0152001
Physical Examination Step	Qwen2.5-72b	Qwen2.5-14b	4.036986361	0.044513244	1	0.934778131
Physical Examination Step	Qwen2.5-32b	Qwen2.5-14b	1.635737607	0.20091129	1	1
General Examination	ClinDiag-GPT	GPT-4o-mini	26.95593631	2.08147E-07	1	4.371E-06
General Examination	ClinDiag-GPT	GPT-4o	26.95593631	2.08147E-07	1	4.371E-06
General Examination	ClinDiag-GPT	Claude-3-Haiku	25.89133286	3.61187E-07	1	7.585E-06
General Examination	ClinDiag-GPT	Qwen2.5-72b	5.228427996	0.022220607	1	0.466632737

General Examination	ClinDiag-GPT	Qwen2.5-32b	22.74999076	1.84505E-06	1	3.875E-05
General Examination	ClinDiag-GPT	Qwen2.5-14b	10.09398209	0.001487543	1	0.0312384
General Examination	GPT-4o-mini	GPT-4o	0	1	1	1
General Examination	GPT-4o-mini	Claude-3-Haiku	0	1	1	1
General Examination	GPT-4o-mini	Qwen2.5-72b	8.472962301	0.003604639	1	0.075697428
General Examination	GPT-4o-mini	Qwen2.5-32b	0.116338891	0.733039688	1	1
General Examination	GPT-4o-mini	Qwen2.5-14b	4.0522401	0.044112772	1	0.926368202
General Examination	GPT-4o	Claude-3-Haiku	0	1	1	1
General Examination	GPT-4o	Qwen2.5-72b	8.472962301	0.003604639	1	0.075697428
General Examination	GPT-4o	Qwen2.5-32b	0.116338891	0.733039688	1	1
General Examination	GPT-4o	Qwen2.5-14b	4.0522401	0.044112772	1	0.926368202
General Examination	Claude-3-Haiku	Qwen2.5-72b	7.845216948	0.005095534	1	0.10700622
General Examination	Claude-3-Haiku	Qwen2.5-32b	0.052411794	0.818918487	1	1
General Examination	Claude-3-Haiku	Qwen2.5-14b	3.617029327	0.057190908	1	1
General Examination	Qwen2.5-72b	Qwen2.5-32b	6.066825976	0.013774427	1	0.289262971
General Examination	Qwen2.5-72b	Qwen2.5-14b	0.61906192	0.431396096	1	1
General Examination	Qwen2.5-32b	Qwen2.5-14b	2.434775273	0.118670901	1	1
Specialty Examination	ClinDiag-GPT	GPT-4o-mini	38.80231828	4.68971E-10	1	9.848E-09
Specialty Examination	ClinDiag-GPT	GPT-4o	35.16875111	3.02336E-09	1	6.349E-08
Specialty Examination	ClinDiag-GPT	Claude-3-Haiku	42.57526619	6.8014E-11	1	1.428E-09
Specialty Examination	ClinDiag-GPT	Qwen2.5-72b	40.05441307	2.46986E-10	1	5.187E-09
Specialty Examination	ClinDiag-GPT	Qwen2.5-32b	47.7721723	4.78739E-12	1	1.005E-10

Specialty Examination	ClinDiag-GPT	Qwen2.5-14b	13.16168134	0.000285732	1	0.0060004
Specialty Examination	GPT-4o-mini	GPT-4o	0.049199318	0.824462424	1	1
Specialty Examination	GPT-4o-mini	Claude-3-Haiku	0.048957618	0.824887126	1	1
Specialty Examination	GPT-4o-mini	Qwen2.5-72b	7.84364E-07	0.999293359	1	1
Specialty Examination	GPT-4o-mini	Qwen2.5-32b	0.432322233	0.510851658	1	1
Specialty Examination	GPT-4o-mini	Qwen2.5-14b	7.321244911	0.006814419	1	0.143102805
Specialty Examination	GPT-4o	Claude-3-Haiku	0.306805093	0.579647358	1	1
Specialty Examination	GPT-4o	Qwen2.5-72b	0.111286476	0.738684228	1	1
Specialty Examination	GPT-4o	Qwen2.5-32b	0.979776394	0.322254017	1	1
Specialty Examination	GPT-4o	Qwen2.5-14b	5.647937669	0.017475913	1	0.366994181
Specialty Examination	Claude-3-Haiku	Qwen2.5-72b	0.012074487	0.912501397	1	1
Specialty Examination	Claude-3-Haiku	Qwen2.5-32b	0.106704672	0.743927469	1	1
Specialty Examination	Claude-3-Haiku	Qwen2.5-14b	9.19367094	0.002428534	1	0.0509992
Specialty Examination	Qwen2.5-72b	Qwen2.5-32b	0.2983315	0.584930316	1	1
Specialty Examination	Qwen2.5-72b	Qwen2.5-14b	7.928325503	0.004866681	1	0.102200311
Specialty Examination	Qwen2.5-32b	Qwen2.5-14b	11.96585212	0.000541844	1	0.0113787
Diagnostic Test Step	ClinDiag-GPT	GPT-4o-mini	2.843798512	0.091727139	1	1
Diagnostic Test Step	ClinDiag-GPT	GPT-4o	3.601450863	0.057729168	1	1
Diagnostic Test Step	ClinDiag-GPT	Claude-3-Haiku	3.214697299	0.072979689	1	1
Diagnostic Test Step	ClinDiag-GPT	Qwen2.5-72b	2.843798512	0.091727139	1	1
Diagnostic Test Step	ClinDiag-GPT	Qwen2.5-32b	8.158479286	0.004286029	1	0.0900066
Diagnostic Test Step	ClinDiag-GPT	Qwen2.5-14b	2.501910886	0.113708254	1	1

Diagnostic Test Step	GPT-4o-mini	GPT-4o	0.011266912	0.915466745	1	1
Diagnostic Test Step	GPT-4o-mini	Claude-3-Haiku	7.22918E-07	0.999321602	1	1
Diagnostic Test Step	GPT-4o-mini	Qwen2.5-72b	0	1	1	1
Diagnostic Test Step	GPT-4o-mini	Qwen2.5-32b	1.149281287	0.283699652	1	1
Diagnostic Test Step	GPT-4o-mini	Qwen2.5-14b	0	1	1	1
Diagnostic Test Step	GPT-4o	Claude-3-Haiku	0	1	1	1
Diagnostic Test Step	GPT-4o	Qwen2.5-72b	0.011266912	0.915466745	1	1
Diagnostic Test Step	GPT-4o	Qwen2.5-32b	0.739040455	0.389968428	1	1
Diagnostic Test Step	GPT-4o	Qwen2.5-14b	0.044641839	0.832663986	1	1
Diagnostic Test Step	Claude-3-Haiku	Qwen2.5-72b	7.22918E-07	0.999321602	1	1
Diagnostic Test Step	Claude-3-Haiku	Qwen2.5-32b	0.931230401	0.33454341	1	1
Diagnostic Test Step	Claude-3-Haiku	Qwen2.5-14b	0.011234753	0.915587021	1	1
Diagnostic Test Step	Qwen2.5-72b	Qwen2.5-32b	1.149281287	0.283699652	1	1
Diagnostic Test Step	Qwen2.5-72b	Qwen2.5-14b	0	1	1	1
Diagnostic Test Step	Qwen2.5-32b	Qwen2.5-14b	1.385306174	0.239199377	1	1
Failure to recommend Key Test	ClinDiag-GPT	GPT-4o-mini	11.4270802	0.000723813	1	0.0152001
Failure to recommend Key Test	ClinDiag-GPT	GPT-4o	10.06416435	0.001511811	1	0.031748
Failure to recommend Key Test	ClinDiag-GPT	Claude-3-Haiku	12.14540023	0.000492091	1	0.0103339
Failure to recommend Key Test	ClinDiag-GPT	Qwen2.5-72b	4.475131214	0.034391562	1	0.722222794
Failure to recommend Key Test	ClinDiag-GPT	Qwen2.5-32b	16.84031396	4.06603E-05	1	0.0008539
Failure to recommend Key Test	ClinDiag-GPT	Qwen2.5-14b	10.06416435	0.001511811	1	0.031748
Failure to recommend Key Test	GPT-4o-mini	GPT-4o	0.01115487	0.915886538	1	1

Failure to recommend	GPT-4o-	Claude-3-	7.14837E-07	0.99932	1	1
Key Test	mini	Haiku		5404		
Failure to recommend	GPT-4o-	Qwen2.5-	1.374613791	0.24102	1	1
Key Test	mini	72b		075		
Failure to recommend	GPT-4o-	Qwen2.5-	0.400873317	0.52663	1	1
Key Test	mini	32b		8584		
Failure to recommend	GPT-4o-	Qwen2.5-	0.01115487	0.91588	1	1
Key Test	mini	14b		6538		
Failure to recommend	GPT-4o	Claude-3-	0.045023927	0.83196	1	1
Key Test		Haiku		0038		
Failure to recommend	GPT-4o	Qwen2.5-	0.923971313	0.33643	1	1
Key Test		72b		4391		
Failure to recommend	GPT-4o	Qwen2.5-	0.712893344	0.39848	1	1
Key Test		32b		4962		
Failure to recommend	GPT-4o	Qwen2.5-	0	1	1	1
Key Test		14b				
Failure to recommend	Claude-3-	Qwen2.5-	1.635285416	0.20097	1	1
Key Test	Haiku	72b		3557		
Failure to recommend	Claude-3-	Qwen2.5-	0.277419138	0.59839	1	1
Key Test	Haiku	32b		7816		
Failure to recommend	Claude-3-	Qwen2.5-	0.045023927	0.83196	1	1
Key Test	Haiku	14b		0038		
Failure to recommend	Qwen2.5-	Qwen2.5-	3.643834421	0.05627	1	1
Key Test	72b	32b		7183		
Failure to recommend	Qwen2.5-	Qwen2.5-	0.923971313	0.33643	1	1
Key Test	72b	14b		4391		
Failure to recommend	Qwen2.5-	Qwen2.5-	0.712893344	0.39848	1	1
Key Test	32b	14b		4962		
Misinterpretation	ClinDiag-	GPT-4o-	0.111751146	0.73815	1	1
	GPT	mini		9222		
Misinterpretation	ClinDiag-	GPT-4o	0.11417053	0.73544	1	1
	GPT			5105		
Misinterpretation	ClinDiag-	Claude-3-	0.012556976	0.91077	1	1
	GPT	Haiku		7487		
Misinterpretation	ClinDiag-	Qwen2.5-	0.976998457	0.32294	1	1
	GPT	72b		0964		
Misinterpretation	ClinDiag-	Qwen2.5-	7.96721E-07	0.99928	1	1
	GPT	32b		7814		
Misinterpretation	ClinDiag-	Qwen2.5-	1.315379561	0.25142	1	1
	GPT	14b		2962		
Misinterpretation	GPT-4o-	GPT-4o	0.614235895	0.43319	1	1
	mini			7364		
Misinterpretation	GPT-4o-	Claude-3-	0.311444666	0.57679	1	1
	mini	Haiku		503		

Misinterpretation	GPT-4o-mini	Qwen2.5-72b	0.296464808	0.586107204	1	1
Misinterpretation	GPT-4o-mini	Qwen2.5-32b	0.049313999	0.824261295	1	1
Misinterpretation	GPT-4o-mini	Qwen2.5-14b	2.532322545	0.111535938	1	1
Misinterpretation	GPT-4o	Claude-3-Haiku	0.012786035	0.909970817	1	1
Misinterpretation	GPT-4o	Qwen2.5-72b	2.0647621	0.150738451	1	1
Misinterpretation	GPT-4o	Qwen2.5-32b	0.203001152	0.65230923	1	1
Misinterpretation	GPT-4o	Qwen2.5-14b	0.484372638	0.486448339	1	1
Misinterpretation	Claude-3-Haiku	Qwen2.5-72b	1.468475162	0.225586433	1	1
Misinterpretation	Claude-3-Haiku	Qwen2.5-32b	0.050513103	0.822172829	1	1
Misinterpretation	Claude-3-Haiku	Qwen2.5-14b	0.851131993	0.3562323	1	1
Misinterpretation	Qwen2.5-72b	Qwen2.5-32b	0.768015566	0.380831654	1	1
Misinterpretation	Qwen2.5-72b	Qwen2.5-14b	5.02415143	0.024996171	1	0.524919593
Misinterpretation	Qwen2.5-32b	Qwen2.5-14b	1.585744617	0.207935164	1	1
Diagnostic Reasoning	ClinDiag-GPT	GPT-4o-mini	10.84244196	0.000991998	1	0.020832
Diagnostic Reasoning	ClinDiag-GPT	GPT-4o	5.544132794	0.018542742	1	0.38939759
Diagnostic Reasoning	ClinDiag-GPT	Claude-3-Haiku	9.5174256	0.002035299	1	0.0427413
Diagnostic Reasoning	ClinDiag-GPT	Qwen2.5-72b	6.571621381	0.010361751	1	0.217596774
Diagnostic Reasoning	ClinDiag-GPT	Qwen2.5-32b	7.684001137	0.005571251	1	0.116996266
Diagnostic Reasoning	ClinDiag-GPT	Qwen2.5-14b	2.967020712	0.084978223	1	1
Diagnostic Reasoning	GPT-4o-mini	GPT-4o	0.711133412	0.399067803	1	1
Diagnostic Reasoning	GPT-4o-mini	Claude-3-Haiku	0.011088669	0.91613558	1	1
Diagnostic Reasoning	GPT-4o-mini	Qwen2.5-72b	0.400378373	0.526893914	1	1

Diagnostic Reasoning	GPT-4o-mini	Qwen2.5-32b	0.178271286	0.672863098	1	1
Diagnostic Reasoning	GPT-4o-mini	Qwen2.5-14b	2.183199975	0.139524099	1	1
Diagnostic Reasoning	GPT-4o	Claude-3-Haiku	0.40027882	0.526945298	1	1
Diagnostic Reasoning	GPT-4o	Qwen2.5-72b	0.011078996	0.916172031	1	1
Diagnostic Reasoning	GPT-4o	Qwen2.5-32b	0.099782688	0.752090573	1	1
Diagnostic Reasoning	GPT-4o	Qwen2.5-14b	0.280812424	0.596169219	1	1
Diagnostic Reasoning	Claude-3-Haiku	Qwen2.5-72b	0.178139098	0.672977372	1	1
Diagnostic Reasoning	Claude-3-Haiku	Qwen2.5-32b	0.044723984	0.83251238	1	1
Diagnostic Reasoning	Claude-3-Haiku	Qwen2.5-14b	1.606165978	0.205031586	1	1
Diagnostic Reasoning	Qwen2.5-72b	Qwen2.5-32b	0.011068075	0.916213208	1	1
Diagnostic Reasoning	Qwen2.5-72b	Qwen2.5-14b	0.548441065	0.458955375	1	1
Diagnostic Reasoning	Qwen2.5-32b	Qwen2.5-14b	0.904396526	0.34160557	1	1
Anchoring Bias	ClinDiag-GPT	GPT-4o-mini	10.98544057	0.000918304	1	0.0192844
Anchoring Bias	ClinDiag-GPT	GPT-4o	3.813658864	0.050836287	1	1
Anchoring Bias	ClinDiag-GPT	Claude-3-Haiku	6.148718128	0.013150739	1	0.27616551
Anchoring Bias	ClinDiag-GPT	Qwen2.5-72b	4.682079092	0.030478839	1	0.640055618
Anchoring Bias	ClinDiag-GPT	Qwen2.5-32b	5.152528475	0.023212549	1	0.487463524
Anchoring Bias	ClinDiag-GPT	Qwen2.5-14b	1.726883191	0.188809433	1	1
Anchoring Bias	GPT-4o-mini	GPT-4o	1.609745027	0.204527647	1	1
Anchoring Bias	GPT-4o-mini	Claude-3-Haiku	0.546607238	0.459707298	1	1
Anchoring Bias	GPT-4o-mini	Qwen2.5-72b	1.117463169	0.290465356	1	1
Anchoring Bias	GPT-4o-mini	Qwen2.5-32b	0.903381601	0.341876595	1	1

Anchoring Bias	GPT-4o-mini	Qwen2.5-14b	3.635598841	0.056556255	1	1
Anchoring Bias	GPT-4o	Claude-3-Haiku	0.180142343	0.671250942	1	1
Anchoring Bias	GPT-4o	Qwen2.5-72b	0.011234753	0.915587021	1	1
Anchoring Bias	GPT-4o	Qwen2.5-32b	0.04532884	0.831400512	1	1
Anchoring Bias	GPT-4o	Qwen2.5-14b	0.285032241	0.593421718	1	1
Anchoring Bias	Claude-3-Haiku	Qwen2.5-72b	0.045114394	0.831793822	1	1
Anchoring Bias	Claude-3-Haiku	Qwen2.5-32b	0.01115487	0.915886538	1	1
Anchoring Bias	Claude-3-Haiku	Qwen2.5-14b	1.132516592	0.287239181	1	1
Anchoring Bias	Qwen2.5-72b	Qwen2.5-32b	7.1913E-07	0.999323382	1	1
Anchoring Bias	Qwen2.5-72b	Qwen2.5-14b	0.556678659	0.455601565	1	1
Anchoring Bias	Qwen2.5-32b	Qwen2.5-14b	0.727479398	0.393701436	1	1
Confirmation Bias	ClinDiag-GPT	GPT-4o-mini	12.22660425	0.000471129	1	0.0098937
Confirmation Bias	ClinDiag-GPT	GPT-4o	4.410351701	0.035721476	1	0.750150989
Confirmation Bias	ClinDiag-GPT	Claude-3-Haiku	4.410351701	0.035721476	1	0.750150989
Confirmation Bias	ClinDiag-GPT	Qwen2.5-72b	9.454229777	0.002106627	1	0.0442392
Confirmation Bias	ClinDiag-GPT	Qwen2.5-32b	3.116407231	0.077507501	1	1
Confirmation Bias	ClinDiag-GPT	Qwen2.5-14b	3.116407231	0.077507501	1	1
Confirmation Bias	GPT-4o-mini	GPT-4o	1.705529079	0.191566479	1	1
Confirmation Bias	GPT-4o-mini	Claude-3-Haiku	1.705529079	0.191566479	1	1
Confirmation Bias	GPT-4o-mini	Qwen2.5-72b	0.103889074	0.747211586	1	1
Confirmation Bias	GPT-4o-mini	Qwen2.5-32b	2.692156839	0.100843229	1	1
Confirmation Bias	GPT-4o-mini	Qwen2.5-14b	2.692156839	0.100843229	1	1

Confirmation Bias	GPT-4o	Claude-3- Haiku	0	1	1	1
Confirmation Bias	GPT-4o	Qwen2.5- 72b	0.768015566	0.38083 1654	1	1
Confirmation Bias	GPT-4o	Qwen2.5- 32b	0.050513103	0.82217 2829	1	1
Confirmation Bias	GPT-4o	Qwen2.5- 14b	0.050513103	0.82217 2829	1	1
Confirmation Bias	Claude-3- Haiku	Qwen2.5- 72b	0.768015566	0.38083 1654	1	1
Confirmation Bias	Claude-3- Haiku	Qwen2.5- 32b	0.050513103	0.82217 2829	1	1
Confirmation Bias	Claude-3- Haiku	Qwen2.5- 14b	0.050513103	0.82217 2829	1	1
Confirmation Bias	Qwen2.5- 72b	Qwen2.5- 32b	1.468475162	0.22558 6433	1	1
Confirmation Bias	Qwen2.5- 72b	Qwen2.5- 14b	1.468475162	0.22558 6433	1	1
Confirmation Bias	Qwen2.5- 32b	Qwen2.5- 14b	0	1	1	1
Total Prevalence	ClinDiag- GPT	GPT-4o- mini	8.393508435	0.00376 5634	1	0.07907831
Total Prevalence	ClinDiag- GPT	GPT-4o	7.166555896	0.00742 7524	1	0.155978003
Total Prevalence	ClinDiag- GPT	Claude-3- Haiku	7.772545015	0.00530 4626	1	0.111397144
Total Prevalence	ClinDiag- GPT	Qwen2.5- 72b	5.514042617	0.01886 4396	1	0.396152308
Total Prevalence	ClinDiag- GPT	Qwen2.5- 32b	14.39530332	0.00014 8171	1	0.0031116
Total Prevalence	ClinDiag- GPT	Qwen2.5- 14b	4.541291776	0.03308 6614	1	0.694818903
Total Prevalence	GPT-4o- mini	GPT-4o	0.012449739	0.91115 7701	1	1
Total Prevalence	GPT-4o- mini	Claude-3- Haiku	0	1	1	1
Total Prevalence	GPT-4o- mini	Qwen2.5- 72b	0.197977941	0.65635 7968	1	1
Total Prevalence	GPT-4o- mini	Qwen2.5- 32b	0.643295132	0.42252 0095	1	1
Total Prevalence	GPT-4o- mini	Qwen2.5- 14b	0.441469136	0.50641 4217	1	1
Total Prevalence	GPT-4o	Claude-3- Haiku	7.96721E-07	0.99928 7814	1	1

Total Prevalence	GPT-4o	Qwen2.5-72b	0.049313999	0.824261295	1	1
Total Prevalence	GPT-4o	Qwen2.5-32b	1.051714977	0.305112443	1	1
Total Prevalence	GPT-4o	Qwen2.5-14b	0.195037693	0.65875645	1	1
Total Prevalence	Claude-3-Haiku	Qwen2.5-72b	0.111751146	0.738159222	1	1
Total Prevalence	Claude-3-Haiku	Qwen2.5-32b	0.833654642	0.361217876	1	1
Total Prevalence	Claude-3-Haiku	Qwen2.5-14b	0.306717265	0.579701624	1	1
Total Prevalence	Qwen2.5-72b	Qwen2.5-32b	1.844207669	0.174458582	1	1
Total Prevalence	Qwen2.5-72b	Qwen2.5-14b	0.01199146	0.91280154	1	1
Total Prevalence	Qwen2.5-32b	Qwen2.5-14b	2.486054372	0.114859352	1	1

Table 9. Statistical Analysis Results of Diagnostic Accuracy in Three-Arm Comparative Study on Physicians-LLM Collaboration

Test Data	Group 1	Group 2	Chi-square Value	P Value	Adjusted P Value
Total (n=60)	ClinDiag-GPT	ClinDiag-GPT+Physician	7.950635209	0.00480705	0.01442115
Total (n=60)	ClinDiag-GPT	Physician	3.556653288	0.0593072	0.177921599
Total (n=60)	ClinDiag-GPT+Physician	Physician	1.648907408	0.199107704	0.597323112
Test Data	Group 1	Group 2	Chi-square Value	P Value	Adjusted P Value
Challenging Case (n=20)	ClinDiag-GPT	ClinDiag-GPT+Physician	NA	0.5643	NA
Challenging Case (n=20)	ClinDiag-GPT	Physician	NA	1	NA
Challenging Case (n=20)	ClinDiag-GPT+Physician	Physician	NA	0.5361	NA
Test Data	Group 1	Group 2	Chi-square Value	P Value	Adjusted P Value
Emergency Case (n=20)	ClinDiag-GPT	ClinDiag-GPT+Physician	4.862298196	0.027449748	0.082349243
Emergency Case (n=20)	ClinDiag-GPT	Physician	5.604848485	0.017910844	0.053732531
Emergency Case (n=20)	ClinDiag-GPT+Physician	Physician	0	1	1
Test Data	Group 1	Group 2	Chi-square Value	P Value	Adjusted P Value
Rare Disease (n=20)	ClinDiag-GPT	ClinDiag-GPT+Physician	NA	0.5643	NA
Rare Disease (n=20)	ClinDiag-GPT	Physician	NA	0.4825	NA
Rare Disease (n=20)	ClinDiag-GPT+Physician	Physician	NA	0.0448	NA

Table 10. Statistical Analysis Results of Procedure Error Prevalence in Three-Arm Comparative Study on Physicians-LLM Collaboration

Category	Model 1	Model 2	Chi-Square Value	P-Value	Corrected P-Value
Total Prevalence	ClinDiag-GPT	ClinDiag-GPT + Physician	3.112191586	0.077708333	0.233125
Total Prevalence	ClinDiag-GPT	Physician	1.201404043	0.273041236	0.819124
Total Prevalence	ClinDiag-GPT + Physician	Physician	0.24084608	0.623595698	1
Medical History Step	ClinDiag-GPT	ClinDiag-GPT + Physician	0.457663469	0.498717983	1
Medical History Step	ClinDiag-GPT	Physician	0.025949403	0.872023946	1
Medical History Step	ClinDiag-GPT + Physician	Physician	0.070231241	0.79100039	1
History of Present Illness	ClinDiag-GPT	ClinDiag-GPT + Physician	N/A	0.717007411	1
History of Present Illness	ClinDiag-GPT	Physician	N/A	1	1
History of Present Illness	ClinDiag-GPT + Physician	Physician	N/A	0.717007411	1
Past Medical History	ClinDiag-GPT	ClinDiag-GPT + Physician	N/A	0.679275066	1
Past Medical History	ClinDiag-GPT	Physician	N/A	0.717007411	1
Past Medical History	ClinDiag-GPT + Physician	Physician	N/A	1	1
Family History	ClinDiag-GPT	ClinDiag-GPT + Physician	N/A	0.364176886	1
Family History	ClinDiag-GPT	Physician	N/A	0.364176886	1
Family History	ClinDiag-GPT + Physician	Physician	N/A	1	1
Social History	ClinDiag-GPT	ClinDiag-GPT + Physician	N/A	0.061092584	0.183278
Social History	ClinDiag-GPT	Physician	1.933522616	0.164373568	0.493121
Social History	ClinDiag-GPT + Physician	Physician	N/A	1	1
Physical Examination Step	ClinDiag-GPT	ClinDiag-GPT + Physician	0.086575319	0.768576931	1
Physical Examination Step	ClinDiag-GPT	Physician	0.258512123	0.611144194	1
Physical Examination Step	ClinDiag-GPT + Physician	Physician	1.79098E-07	0.999662335	1
General Examination	ClinDiag-GPT	ClinDiag-GPT + Physician	N/A	0.364176886	1
General Examination	ClinDiag-GPT	Physician	N/A	0.679275066	1
General Examination	ClinDiag-GPT + Physician	Physician	N/A	1	1

Specialty Examination	ClinDiag-GPT	ClinDiag-GPT + Physician	0.467500833	0.494139154	1
Specialty Examination	ClinDiag-GPT	Physician	0.374060386	0.540799286	1
Specialty Examination	ClinDiag-GPT + Physician	Physician	0	1	1
Diagnostic Test Step	ClinDiag-GPT	ClinDiag-GPT + Physician	1.985470412	0.158815295	0.476446
Diagnostic Test Step	ClinDiag-GPT	Physician	1.349674897	0.24533496	0.736005
Diagnostic Test Step	ClinDiag-GPT + Physician	Physician	0.003885815	0.950295015	1
Failure to recommend Key Test	ClinDiag-GPT	ClinDiag-GPT + Physician	1.566762864	0.210677748	0.632033
Failure to recommend Key Test	ClinDiag-GPT	Physician	1.386518238	0.238993967	0.716982
Failure to recommend Key Test	ClinDiag-GPT + Physician	Physician	0	1	1
Misinterpretation	ClinDiag-GPT	ClinDiag-GPT + Physician	2.725119093	0.098780753	0.296342
Misinterpretation	ClinDiag-GPT	Physician	0.812390164	0.36741454	1
Misinterpretation	ClinDiag-GPT + Physician	Physician	0.309123183	0.578218734	1
Diagnostic Reasoning	ClinDiag-GPT	ClinDiag-GPT + Physician	1.757402201	0.184948872	0.554847
Diagnostic Reasoning	ClinDiag-GPT	Physician	1.429890847	0.231782253	0.695347
Diagnostic Reasoning	ClinDiag-GPT + Physician	Physician	0	1	1
Anchoring Bias	ClinDiag-GPT	ClinDiag-GPT + Physician	0.250173038	0.616953262	1
Anchoring Bias	ClinDiag-GPT	Physician	0.006443715	0.936020326	1
Anchoring Bias	ClinDiag-GPT + Physician	Physician	0.029622157	0.863350355	1
Confirmation Bias	ClinDiag-GPT	ClinDiag-GPT + Physician	2.283881414	0.130724359	0.392173
Confirmation Bias	ClinDiag-GPT	Physician	4.032135701	0.044641394	0.133924
Confirmation Bias	ClinDiag-GPT + Physician	Physician	0.086575319	0.768576931	1

Supplementary file 2. Example of LLM performing clinical diagnostic procedure

Case No. 24703566

Correct Diagnosis: Type 2 Diabetes

Stage One: Medical History

Initial Information:

A 5-year-old female presented with non-healing sores on thighs and persistent bedwetting.

Doctor Agent:

Doctor Reasoning: The patient is presenting with non-healing sores on her thighs and persistent bedwetting. It is important to gather more details about the sores, including their characteristics, duration, any associated symptoms, and previous treatments.

Doctor Action: Please ask the patient's guardian to describe the sores in detail, including when they first appeared, how they have changed over time, and if there are any other symptoms such as itching or pain.

Provider:

The patient has non-healing sores on her thighs that have been present for roughly 5 weeks. There are no additional details provided regarding itching, pain, or changes over time.

Doctor Agent:

Doctor Reasoning: Since the sores have persisted for several weeks without healing, it may be necessary to consider potential causes such as infections, dermatological conditions, or systemic issues. Additionally, understanding the context of the bedwetting is crucial, especially since it started after a traumatic event.

Doctor Action: Please inquire about the nature of the bedwetting, including when it began, its frequency, and whether there were any changes in the child's behavior or health around that time.

Provider:

The bedwetting has been present for the past 12 months. There are no details provided regarding its frequency or any changes in the child's behavior or health around that time.",

Doctor Agent:

Doctor Reasoning: The onset of bedwetting following a traumatic event raises concerns about possible psychological factors affecting the child's health. It would be beneficial to explore any relevant medical history, including any previous illnesses or treatments.

Doctor Action: Please ask if the patient has had any significant medical history, including prior illnesses, hospitalizations, or treatments related to skin conditions or urinary issues.

Provider:

The patient has a relevant medical history of being born macrosomic (4.5 kg) at 38 weeks

by caesarean section after a pregnancy complicated by poorly controlled gestational diabetes. There is no mention of prior illnesses, hospitalizations, or treatments specifically related to skin conditions or urinary issues.

Doctor Agent:

Doctor Reasoning: The patient's birth history and maternal complications could potentially influence her current health status. Understanding the family history can also provide insights into any genetic predispositions or familial patterns of illness.

Doctor Action: Please inquire about the family history, particularly any significant medical conditions that might relate to the patient's current symptoms.

Provider:

The family history indicates a strong history of type 2 diabetes. There are no other specific medical conditions mentioned that relate to the patient's current symptoms.

Doctor Agent:

Doctor Reasoning: Given the family history of type 2 diabetes, it is essential to assess any potential metabolic or endocrine issues that could contribute to the patient's symptoms. Additionally, exploring the social history will help understand environmental factors that may impact her health.

Doctor Action: Please ask about the patient's social history, including lifestyle factors such as diet, physical activity, and any relevant social circumstances.

Provider:

The patient's diet consists of high portions of refined carbohydrates and simple sugars. There are no specified details regarding physical activity or relevant social circumstances.

Doctor Agent:

Doctor Reasoning: The dietary habits reported could contribute to various health issues, including skin problems and urinary difficulties. Summarizing the collected information will help clarify the situation and guide further evaluation.

Doctor Action: Final History: The patient is a 5-year-old female who presents with non-healing sores on her thighs that have been present for approximately 5 weeks. She also experiences persistent bedwetting for the past 12 months, which began after a traumatic event. Her medical history includes being born macrosomic due to poorly controlled gestational diabetes. Family history reveals a strong presence of type 2 diabetes. The patient's diet is characterized by high portions of refined carbohydrates and simple sugars.

Stage Two: Physical Examination

Initial Information and Final History Report Passed to Doctor Agent

Doctor Agent:

Doctor Reasoning: The patient presents with non-healing sores and a history of bedwetting

following a traumatic event. It is important to assess the characteristics of the sores, including their size, appearance, and any associated symptoms such as pain or itching. This will help determine potential causes, including infections or dermatological conditions.

Doctor Action: Please ask the patient's caregiver to describe the sores on her thighs in detail, including their size, color, any discharge, and whether they are painful or itchy.

Provider:

The relevant findings from the physical examination indicate that there are crusted sores on both upper thighs. Specific details regarding their size, color, discharge, or associated symptoms such as pain or itching are not specified in the records provided.

Doctor Agent:

Doctor Reasoning: The presence of crusted sores could suggest an infectious process, possibly bacterial or fungal, especially given the patient's diet high in refined carbohydrates. Additionally, understanding the duration and progression of these sores will be crucial in determining the appropriate management.

Doctor Action: Please inquire about the duration of the sores, any changes over time, and if there have been any treatments attempted prior to this visit.

Provider:

The sores on the patient's thighs had been present for roughly 5 weeks at the time of the examination. There are no specified changes over time or treatments attempted prior to this visit in the records provided.

Doctor Agent:

Doctor Reasoning: Given the patient's age and the chronicity of the sores, it is essential to consider other factors that may contribute to delayed healing, such as nutritional deficiencies or underlying health issues. Additionally, assessing for any signs of systemic illness or other skin conditions would be prudent.

Doctor Action: Please perform a general inspection of the patient to check for any other skin lesions, signs of systemic illness, or abnormalities in growth and development.

Provider:

The physical examination findings include acanthosis nigricans in the axillae and around the neck. There are no other skin lesions, signs of systemic illness, or abnormalities in growth and development specified in the records provided.

Doctor Agent:

Doctor Reasoning: Acanthosis nigricans can be indicative of insulin resistance, which aligns with the patient's family history of type 2 diabetes and her dietary habits. It is important to evaluate the patient's hydration status and overall nutrition, particularly since she has a high intake of refined carbohydrates.

Doctor Action: Please assess the patient's vital status, hydration status and any signs of malnutrition or obesity during the physical examination.

Provider:

The patient's weight is recorded as 36 kg, and her body mass index (BMI) is 24.5 kg/m², which indicates obesity. There are no specific details regarding her vital status, hydration status or signs of malnutrition provided in the records.

Doctor Agent:

Doctor Reasoning: The patient's obesity and dietary habits raise concerns about metabolic syndrome and its implications for her current condition. Additionally, the history of bedwetting warrants further exploration, as it may relate to psychological factors or underlying health issues.

Doctor Action: Please ask the caregiver about the frequency and nature of the bedwetting episodes, including any associated symptoms such as urgency or daytime wetting.

Provider:

The records indicate that the patient has a history of bedwetting for the past 12 months. However, there are no specific details regarding the frequency, nature of the bedwetting episodes, or any associated symptoms such as urgency or daytime wetting provided.

Doctor Agent:

Doctor Reasoning: Understanding the context of the bedwetting is critical, especially considering the traumatic event mentioned in the history. Psychological factors could significantly influence the patient's urinary control. Summarizing the findings will help clarify the situation and guide further management.

Doctor Action: Final Physical Examination: The patient is a 5-year-old female presenting with non-healing crusted sores on both upper thighs for approximately 5 weeks. Acanthosis nigricans is noted in the axillae and around the neck. The patient weighs 36 kg with a BMI of 24.5 kg/m², indicating obesity. There is a history of bedwetting for the past 12 months following a traumatic event. No additional skin lesions or systemic signs were reported.

Stage Three: Diagnostic Test

Initial Information, Final History Report, and Final Physical Examination Report Passed to Doctor Agent

Doctor Agent:

Doctor Reasoning: The patient's non-healing sores and acanthosis nigricans could suggest an underlying metabolic issue, possibly related to insulin resistance given her family history of type 2 diabetes. I need to assess her blood glucose levels and HbA1c to evaluate for potential diabetes.

Doctor Action: Please provide the patient's fasting blood glucose level and HbA1c results.

Provider:

The patient's HbA1c result is 11.9% (normal range 4.3–6.0). The fasting blood glucose level

is not specified in the records provided. However, the laboratory examinations show random blood glucose levels of 19.2 mmol/L and 18.7 mmol/L.",

Doctor Agent:

Doctor Reasoning: The elevated HbA1c indicates significant hyperglycemia over the past few months, consistent with possible undiagnosed diabetes. The random blood glucose levels further support this concern. Next, I will review the lipid profile to assess for any dyslipidemia associated with insulin resistance.

Doctor Action: Please provide the lipid profile results, focusing on total cholesterol, HDL, LDL, and triglycerides.

Provider:

The lipid profile results are not provided in the patient's records.

Doctor Agent:

Doctor Reasoning: Since the lipid profile information is missing, it would be prudent to consider other tests that may help clarify the patient's metabolic status. Given her age and symptoms, I will check for any signs of infection or inflammation that might be contributing to the non-healing sores.

Doctor Action: Please arrange for a complete blood count (CBC) and inflammatory markers such as C-reactive protein (CRP) to rule out any underlying infections or inflammatory conditions.

Provider:

The results for a complete blood count (CBC) and inflammatory markers such as C-reactive protein (CRP) are not provided in the patient's records.

Doctor Agent:

Doctor Reasoning: Without the CBC and CRP results, we cannot fully assess for infection or inflammation. Additionally, the patient's history of bedwetting following trauma raises concerns about possible psychological factors affecting her health. I will now inquire about any previous evaluations or treatments related to her bedwetting.

Doctor Action: Please gather information regarding any prior assessments or interventions related to the patient's bedwetting and any psychological evaluations conducted since the traumatic event.

Provider:

There is no information provided in the patient's records regarding any prior assessments or interventions related to the bedwetting or any psychological evaluations conducted since the traumatic event.

Doctor Agent:

Doctor Reasoning: The lack of information on prior assessments for bedwetting limits our understanding of its impact on the patient's overall health. Given the patient's obesity and

family history of diabetes, it is crucial to consider lifestyle factors and their implications on her condition. I will now summarize the findings collected thus far.

Doctor Action: Final Test: The patient is a 5-year-old girl with non-healing crusted sores on her thighs for 5 weeks, acanthosis nigricans, and a history of bedwetting for 12 months following a traumatic event. The HbA1c is significantly elevated at 11.9%, and random blood glucose levels are markedly high at 19.2 mmol/L and 18.7 mmol/L. The lipid profile and CBC results are currently unavailable. This suggests a potential diagnosis of type 2 diabetes mellitus, and further evaluation is necessary.

Stage For: Final Diagnosis

Initial Information, Final History Report, Final Physical Examination Report, and Final Diagnostic Test Report passed to Doctor Agent

Doctor Agent:

Doctor Reasoning: The patient's presentation of non-healing sores, acanthosis nigricans, and significantly elevated blood glucose levels strongly indicates a diagnosis of type 2 diabetes mellitus. The history of bedwetting may also be related to metabolic disturbances associated with diabetes.

Doctor Action: Final Diagnosis: Type 2 Diabetes Mellitus (T2DM)

Supplementary file 3. Examples of Errors Identified in Clinical Diagnostic Procedure

#1

Error Type: Medical History-History of Present Illness

Error Content: The Doctor failed to gather key information such as the onset and progression of abdominal pain, and the development of substernal and bilateral shoulder pain.

#2

Error Type: Medical History-Past Medical History

Error Content: Doctor omitted several chronic conditions, including coronary artery disease, chronic renal insufficiency, type 2 diabetes mellitus, and hypertension, which can influence the differential diagnosis.

#3

Error Type: Physical Examination-General Examination

Error Content: The patient is in critical condition but vital signs such as blood pressure and signs of systemic distress (e.g., hypotension) were not gathered.

#4

Error Type: Physical Examination-Specialty Examination

Error Content: Skin examination was not conducted, potentially missing characteristic skin lesions of Langerhans cell histiocytosis.

#5

Error Type: Diagnostic Test-Failure to Recommend Key Tests

Error Content: Doctor did not gather information on genetic testing (e.g., Whole Exome Sequencing) which is critical for diagnosing Alagille syndrome

#6

Error Type: Diagnostic Test-Misinterpretation of Test Results

Error Content: Doctor incorrectly interpreted the inconclusive ultrasound results as sufficient to rule out appendicitis without further testing.

#7

Error Type: Iterative Diagnostic Procedure-Anchoring Bias

Error Content: Doctor prematurely focused on patient's past medical history of COPD, which prevented consideration of alternative diagnoses like acute cholecystitis.

#8

Error Type: Iterative Diagnostic Procedure-Anchoring Bias

Error Content: Doctor prematurely focused on patient's past medical history of COPD, which prevented consideration of alternative diagnoses like acute cholecystitis.

#9

Error Type: Iterative Diagnostic-Confirmation Bias

Error Content: Doctor continued to support the diagnosis of acute pancreatitis, despite of contradictory evidence such as normal amylase and lipase levels.

Supplementary File 4. Results for Ablation Studies

Table 1. Effect of Different Temperature Settings on Diagnostic Accuracy

	Temperature = 0	Temperature = 0.3	Temperature = 0.7	Temperature = 1
GPT-4o-mini	28.21%	29.45%	28.82%	28.30%
Qwen2.5-72b	32.80%	33.54%	31.83%	32.79%

The Chi-Square test results for pairwise comparisons of temperature settings within each model are as follows: For GPT-4o-mini, none of the comparisons (Temperature = 0 vs Temperature = 0.3, Chi-Square = 1.5487, $P = 0.2133$; Temperature = 0 vs Temperature = 0.7, Chi-Square = 0.3751, $P = 0.5402$; Temperature = 0 vs Temperature = 1, Chi-Square = 0.0050, $P = 0.9435$; Temperature = 0.3 vs Temperature = 0.7, Chi-Square = 0.3704, $P = 0.5428$; Temperature = 0.3 vs Temperature = 1, Chi-Square = 1.3225, $P = 0.2501$; Temperature = 0.7 vs Temperature = 1, Chi-Square = 0.2683, $P = 0.6045$) showed significant differences ($P > 0.05$). For Qwen2.5-72b, none of the comparisons (Temperature = 0 vs Temperature = 0.3, Chi-Square = 0.4904, $P = 0.4838$; Temperature = 0 vs Temperature = 0.7, Chi-Square = 0.9122, $P = 0.3395$; Temperature = 0 vs Temperature = 1, Chi-Square = 0.0, $P = 1.0$; Temperature = 0.3 vs Temperature = 0.7, Chi-Square = 2.8153, $P = 0.0934$; Temperature = 0.3 vs Temperature = 1, Chi-Square = 0.5226, $P = 0.4697$; Temperature = 0.7 vs Temperature = 1, Chi-Square = 0.8694, $P = 0.3511$) demonstrated significant differences ($P > 0.05$).

Table 2. Effect of Prompt Engineering on Diagnostic Accuracy

Model	Baseline	Prompt Engineering
GPT-4o-mini	28.98%	33.54%
Qwen2.5-72b	35.92%	36.58%

The Chi-Square test results for the comparison between Baseline and Prompt Engineering are as follows: For GPT-4o-mini, the comparison yields a significant difference (Chi-Square = 21.0571, $P < 0.001$). For Qwen2.5-72b, no significant difference is observed (Chi-Square = 0.3837, $P = 0.5356$).

Table 3. Effect of Including Multiple Agents on Diagnostic Accuracy

	Baseline	Involving Critic Agent	Involving 2 Doctor Agents	Involving 3 Doctor Agents
GPT-4o-mini	28.98%	29.22%	29.97%	30.45%
Qwen2.5-72b	33.54%	33.36%	33.91%	33.15%

The Chi-Square test results for each model and setting are as follows: For GPT-4o-mini, the comparisons between Baseline and Involving Critic Agent (Chi-Square = 0.0444, $P = 0.8331$), Baseline and Involving 2 Doctor Agents (Chi-Square = 0.9600, $P = 0.3272$), and Baseline and Involving 3 Doctor Agents (Chi-Square = 2.2182, $P = 0.1364$) all show no significant differences. Similarly, for Qwen2.5-72b, the comparisons between Baseline and Involving Critic Agent (Chi-Square = 0.0249, $P = 0.8746$), Baseline and Involving 2 Doctor Agents (Chi-Square = 0.1296, $P = 0.7189$), and Baseline and Involving 3 Doctor Agents (Chi-Square = 0.1303, $P = 0.7181$) also show no significant differences.

Supplementary File 5. Reliability Analysis of Each Large Language Models in Clinical Diagnostic Procedure.

Table 1. Reliability Analysis of Each Large Language Models in Clinical Diagnostic Procedure.

Model	ClinDiag- GPT	GPT-4o- mini	GPT-4o	Claude-3- haiku	Qwen2.5- 72b	Qwen2.5- 32b	Qwen2.5- 14b
1st run diagnosis accuracy	39.76%	28.98%	32.28%	29.38%	33.54%	28.77%	30.20%
2nd run diagnosis accuracy	38.18%	29.35%	32.09%	30.27%	34.01%	29.22%	30.92%
3rd run diagnosis accuracy	37.39%	29.16%	33.24%	31.07%	34.10%	29.13%	31.44%
Kappa	0.4941	0.4977	0.4816	0.3738	0.5107	0.4599	0.4702

Supplementary File 6 ClinDiag-GPT Interacting with Human Examiner

Table 1. Diagnostic Accuracy

No.	Specialty	Disease	Correctness
1	Cardiology	Aortic Dissection	Yes
2	Cardiology	Stable Angina Pectoris	No
3	Cardiology	Atrial Fibrillation	Yes
4	Cardiology	Acute Coronary Syndrome	Yes
5	Cardiology	Chronic Heart Failure	Yes
6	Cardiology	Cardiogenic Shock	No
7	Pulmonology	Upper Respiratory Infection	Yes
8	Pulmonology	Bronchitis	No
9	Pulmonology	Bronchial Asthma	No
10	Pulmonology	Community-acquired Pneumonia	Yes
11	Pulmonology	Chronic Obstructive Pulmonary Disease	Yes
12	Pulmonology	Acute Respiratory Distress Syndrome	No
13	Gastroenterology	Gastroesophageal Reflux Disease	Yes
14	Gastroenterology	Peptic Ulcer Disease	No
15	Gastroenterology	Functional Dyspepsia	Yes
16	Gastroenterology	Acute Pancreatitis	Yes
17	Gastroenterology	Inflammatory Bowel Disease	No
18	Gastroenterology	Autoimmune Hepatitis	No
19	Endocrinology	Type 2 Diabetes	Yes
20	Endocrinology	Hyperthyroidism	Yes
21	Endocrinology	Cushing's Syndrome	Yes
22	General Surgery	Acute Appendicitis	Yes
23	General Surgery	Gallstones	No
24	General Surgery	Inguinal Hernia	Yes
25	General Surgery	Acute Cholecystitis	Yes
26	Neurology	Migraine	Yes
27	Neurology	Facial Nerve Paralysis	No
28	Neurology	Transient Ischemic Attack	Yes
29	Neurology	Epilepsy	No
30	Neurology	Subarachnoid Hemorrhage	Yes
31	Orthopedics	Lumbar Disc Herniation	Yes
32	Orthopedics	Knee Osteoarthritis	Yes
33	Multidisciplinary	Diabetes with Diabetic Nephropathy and Coronary Heart Disease	No
34	Multidisciplinary	COPD with Heart Failure	No
35	Multidisciplinary	Multisystem Diseases in the Elderly	Yes

Supplementary File 7. Procedure Performance in Three Arm Comparative Study

Table 1. Error Prevalence of Each Group in Three-Arm Comparative Study on Physicians-LLM Collaboration

Error Prevalence	ClinDiag-GPT	ClinDiag-GPT + Physician	Physician
Total Prevalence	58.33%	40.56%	46.67%
Medical History Step	18.33%	12.22%	15.56%
History of Present Illness	6.67%	5.00%	7.78%
Past Medical History	6.67%	4.44%	5.00%
Family History	8.33%	1.67%	2.78%
Social History	13.33%	2.78%	4.44%
Physical Examination Step	20.00%	23.89%	25.56%
General Examination	6.67%	2.78%	3.89%
Specialty Examination	16.67%	23.33%	22.78%
Diagnostic Test Step	51.67%	37.22%	39.44%
Failure to recommend Key Test	31.67%	20.00%	20.56%
Misinterpretation	38.33%	22.78%	28.89%
Diagnostic Reasoning	43.33%	30.00%	31.11%
Anchoring Bias	18.33%	13.33%	16.11%
Confirmation Bias	38.33%	23.89%	20.00%

Table 2. Procedure Performance of Three-Arm Comparative Study on Physicians-LLM Collaboration

Medical History Performance						
	Gather	Logic	Reason	Reason	Summarization	Summarization
	Completeness	Consistency	Accuracy	Explainability	Accuracy	Completeness
ClinDiag-GPT	3.43	4.05	4.00	3.90	4.43	3.60
ClinDiag-GPT + Physician	3.54	4.15	4.05	3.86	4.68	4.30
Physician	3.69	4.12	4.00	3.67	4.82	4.30
Physical Examination Performance						
	Gather	Logic	Reason	Reason	Summarization	Summarization
	Completeness	Consistency	Accuracy	Explainability	Accuracy	Completeness
ClinDiag-GPT	3.55	4.40	4.05	3.78	4.20	3.47
ClinDiag-GPT + Physician	3.88	4.34	4.15	3.76	4.45	3.78
Physician	3.97	4.03	3.94	3.63	4.17	3.68
Diagnostic Tests Performance						
	Gather	Logic	Reason	Reason	Summarization	Summarization
	Completeness	Consistency	Accuracy	Explainability	Accuracy	Completeness
ClinDiag-GPT	4.05	4.33	4.33	3.93	4.45	4.03
ClinDiag-GPT + Physician	4.14	4.06	4.10	3.85	4.59	4.31
Physician	4.04	4.04	4.04	3.85	4.51	4.14

Supplementary File 8. Quality Assessment Criteria for Case Inclusion

Quality Evaluation Metrics:

1. Completeness

Does the case report include all essential sections: History, Physical Examination, Diagnostics, Diagnosis, Treatment Plan, and Follow-up?

2. Detail and Depth

Is each section sufficiently detailed to support the final diagnosis?

3. Clinical Plausibility

Do the symptoms, signs, diagnostic results, and diagnosis align with established clinical knowledge and current guidelines?

4. Diagnostic Evidence

Is there adequate supporting evidence for the final diagnosis, such as pathological findings, imaging results, or laboratory data?

5. Ethical Compliance

Are patient details appropriately anonymized, and does the report comply with relevant privacy and ethical guidelines?

Adherence to CARE Guidelines:

6. De-identified Patient Specific Information

Is all patient-specific information anonymized to protect patient privacy?

7. Primary Concerns and Symptoms of the Patient

Are the main concerns and symptoms that led the patient to seek medical attention clearly described?

8. Relevant Past Interventions with Outcomes

Are past medical interventions, treatments, and their outcomes relevantly detailed?

9. Describe Significant Physical Examination (PE) and Important Clinical Findings

Are significant findings from the physical examination and other clinical assessments thoroughly described?

10. Diagnostic Testing

Are all relevant diagnostic tests and their results comprehensively reported?

11. Diagnosis

Is the final diagnosis clearly stated, along with any differential diagnoses that were

considered?

Supplementary File 9. Case Example from ClinDiag-Benchmark

Case ID: rare18

Initial Information

```
{  
  "initial_information": "A 31-year-old male smoker presented with atypical chest pain  
radiating to the left arm for 5 days."  
}
```

Medical History

```
{  
  "medical_history": {  
    "chief_complaint": "Episodes of atypical chest pain radiating to the left arm for 5 days.",  
    "history_of_present_illness": "The patient presented with chest pain episodes that  
started 5 days prior to the visit. There was no history of fever, cough, or contact with COVID-  
infected patients. There was also a long-standing history of decreased vision in the left eye,  
starting at age 10, which progressed to blindness.",  
    "past_medical_history": "History of decreased vision in the left eye starting at age 10,  
which progressed to blindness. Diagnosed with IGFBP7 mutation causing eye manifestations  
in the form of Familial Retinal Arterial Macroaneurysms (FRAM).",  
    "family_history": "Younger brother diagnosed with FRAM and IGFBP7 mutation at age  
23. The rest of the siblings and parents are asymptomatic.",  
    "social_history": "31-year-old male smoker."  
  }  
}
```

Physical Examination

```
{  
  "physical_examinations": [  
    {  
      "date": "Not specified",  
      "vital_signs": {  
        "pulse_rate": "74 bpm",  
        "blood_pressure": "125/74 mmHg",  
        "temperature": "Not specified",  
        "respiratory_rate": "Not specified",  
        "oxygen_saturation": "95%"  
      },  
      "relevant_findings": {  
        "cardiac_examination": "Normal character of first and second heart sound  
without additional sounds",  
        "fundoscopy": "Bleeding retinal artery macroaneurysms in the right retina and  
sub-retinal gliosis suggesting prior laser treatment"      }  
    }  
  ]  
}
```

```

    },
    "changes_over_time": "Not specified"
  }
]
}

```

Diagnostic Test

```

{
  "laboratory_examinations": [
    {
      "procedure_name": "White Blood Cell Count",
      "date": "Not specified",
      "findings": "Mildly raised white blood cell count of  $13.08 \times 10^9$  /L (normal range 3.9 -  $11 \times 10^9$  /L).",
      "changes_over_time": "Not specified"
    },
    {
      "procedure_name": "Troponin Level",
      "date": "Not specified",
      "findings": "Troponin was raised at 637 ng/L (normal = 34 ng/L), repeated level showed 1938 ng/L, peaked at 5484 ng/L the next day.",
      "changes_over_time": "Troponin levels increased from 637 ng/L to 1938 ng/L and peaked at 5484 ng/L the next day."
    },
    {
      "procedure_name": "C-reactive Protein",
      "date": "Not specified",
      "findings": "C-reactive proteins were high.",
      "changes_over_time": "Not specified."
    },
    {
      "procedure_name": "PCR for COVID",
      "date": "Not specified",
      "findings": "PCR for COVID was negative.",
      "changes_over_time": "Not specified."
    }
  ]
}

```

```

{
  "radiographic_examinations": [
    {
      "procedure_name": "Echocardiography",
      "date": "Not specified",

```

```

        "findings": "Normal biventricular systolic function with normal valves. No regional
wall motion abnormalities.",
        "changes_over_time": "Not specified"
    },
    {
        "procedure_name": "Chest X-ray",
        "date": "Not specified",
        "findings": "Did not show any significant findings.",
        "changes_over_time": "Not specified"
    },
    {
        "procedure_name": "Coronary angiogram",
        "date": "Not specified",
        "findings": "Large aneurysm in the proximal segment of the left anterior
descending (LAD) artery. No evidence of plaque rupture or thrombus in aneurysm.",
        "changes_over_time": "Not specified"
    },
    {
        "procedure_name": "Cardiac CT scan",
        "date": "Not specified",
        "findings": "2.28 x 1.64 cm coronary aneurysm at the proximal segment of the LAD
artery with peripheral calcification and a narrow neck of about 0.6 cm. Myocardial bridging
noted in the mid-segment of LAD. Non-obstructive calcified plaques in the right coronary
artery.",
        "changes_over_time": "Not specified"
    },
    {
        "procedure_name": "CT scan of the brain",
        "date": "Not specified",
        "findings": "Did not show any significant abnormality.",
        "changes_over_time": "Not specified"
    }
]
}

{
    "other_tests": [
        {
            "procedure_name": "Electrocardiogram",
            "date": "Not specified",
            "findings": "Normal sinus rhythm with T-wave inversions in lateral leads.",
            "changes_over_time": "Not specified."
        }
    ]
}

```

```
}
```

Diagnosis

```
{
```

```
  "diagnosis": {
```

```
    "final_diagnosis": "Familial retinal arterial macroaneurysm",
```

```
    "differential_diagnosis": "Not specified",
```

```
    "diagnostic_reasoning": "The diagnosis was supported by the patient's familial history of IGFBP7 mutation causing eye manifestations in the form of FRAM, alongside clinical findings of bleeding retinal artery macroaneurysms and the patient's history of decreased vision leading to blindness."
```

```
  }
```

```
}
```