nature medicine

Supplementary information

https://doi.org/10.1038/s41591-024-03328-5

An evaluation framework for clinical use of large language models in patient interaction tasks

In the format provided by the authors and unedited

Supplementary Information

Contents

Supplementary Table 1: Mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 2: Adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 3: Adjusted p-values between 4-choice MCQ and FRQ settings for each experimental setup (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 4: Mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 5: Adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 6: Medical specialty wise mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 7: Medical specialty wise adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 8: Medical specialty wise mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 9: Medical specialty wise adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 10: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 11: Adjusted p-values for Dermatology, calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 12: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 13: Adjusted p-values for Dermatology, calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 14: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 15: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 16: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Supplementary Table 17: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 18: P-values between pairs of evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b) for clinical LLM assessment by medical experts. All p-values were calculated using a McNemar test (see Methods).

Supplementary Table 19: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for GPT-4V and GPT-4V-without-image.

Supplementary Table 20: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to GPT-4V and GPT-4V-without-image. All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 21: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated models (GPT-4V and GPT-4V-without-image) corresponding to each of the experimental setups (vignette,

multi-turn conversation, single-turn conversation and summarized conversation). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Table 22: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for Mistral-v1-7b and Mistral-v2-7b.

Supplementary Table 23: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to Mistral-v1-7b. All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Supplementary Tables

Model	Experiment	Mean Accuracy	95% C.I.
GPT-4	Vignette	0.82	(0.804, 0.837)
	Multi-turn conversation	0.627	(0.609, 0.645)
	Single-turn conversation	0.52	(0.5, 0.539)
	Summarized conversation	0.669	(0.652, 0.686)
GPT-3.5	Vignette	0.657	(0.637, 0.676)
	Multi-turn conversation	0.467	(0.448, 0.485)
	Single-turn conversation	0.435	(0.416, 0.454)
	Summarized conversation	0.507	(0.489, 0.526)
Mistral-v2-7b	Vignette	0.637	(0.616, 0.658)
	Multi-turn conversation	0.426	(0.409, 0.443)
	Single-turn conversation	0.448	(0.429, 0.468)
	Summarized conversation	0.513	(0.496, 0.529)
LLaMA-2-7b	Vignette	0.395	(0.376, 0.415)
	Multi-turn conversation	0.319	(0.303, 0.335)
	Single-turn conversation	0.304	(0.286, 0.323)
	Summarized conversation	0.335	(0.318, 0.352)

Supplementary Table 1: Mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0001	0.0001
GPT-3.5	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0001	0.0001
Mistral-v2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0009	0.001
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0001	0.0001
LLaMA-2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0314	0.0314
	Multi-turn conversation	Summarized conversation	0.011	0.0115
	Single-turn conversation	Summarized conversation	0.0001	0.0001

Supplementary Table 2: Adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	p-value	Adjusted p-value
GPT-4	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001
GPT-3.5	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001
Mistral-v2-7b	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001
LLaMA-2-7b	Vignette	0.0001	0.0001
	Multi-turn conversation	0.0001	0.0001
	Single-turn conversation	0.0001	0.0001
	Summarized conversation	0.0001	0.0001

Supplementary Table 3: Adjusted p-values between 4-choice MCQ and FRQ settings for each experimental setup (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Mean Accuracy	95% C.I.
GPT-4	Vignette	0.486	(0.804, 0.837)
	Multi-turn conversation	0.264	(0.609, 0.645)
	Single-turn conversation	0.133	(0.5, 0.539)
	Summarized conversation	0.272	(0.652, 0.686)
GPT-3.5	Vignette	0.375	(0.637, 0.676)
	Multi-turn conversation	0.169	(0.448, 0.485)
	Single-turn conversation	0.123	(0.416, 0.454)
	Summarized conversation	0.174	(0.489, 0.526)
Mistral-v2-7b	Vignette	0.222	(0.616, 0.658)
	Multi-turn conversation	0.066	(0.409, 0.443)
	Single-turn conversation	0.056	(0.429, 0.468)
	Summarized conversation	0.056	(0.496, 0.529)
LLaMA-2-7b	Vignette	0.169	(0.376, 0.415)
	Multi-turn conversation	0.066	(0.303, 0.335)
	Single-turn conversation	0.065	(0.286, 0.323)
	Summarized conversation	0.081	(0.318, 0.352)

Supplementary Table 4: Mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.0936	0.107
	Single-turn conversation	Summarized conversation	0.0001	0.0001
GPT-3.5	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0001	0.0001
	Multi-turn conversation	Summarized conversation	0.2743	0.2992
	Single-turn conversation	Summarized conversation	0.0001	0.0001
Mistral-v2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.0015	0.0018
	Multi-turn conversation	Summarized conversation	0.0013	0.0016
	Single-turn conversation	Summarized conversation	0.8302	0.8302
LLaMA-2-7b	Vignette	Multi-turn conversation	0.0001	0.0001
	Vignette	Single-turn conversation	0.0001	0.0001
	Vignette	Summarized conversation	0.0001	0.0001
	Multi-turn conversation	Single-turn conversation	0.7258	0.7574
	Multi-turn conversation	Summarized conversation	0.0001	0.0001
	Single-turn conversation	Summarized conversation	0.0002	0.0003

Supplementary Table 5: Adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment Name	Specialty	Mean Accuracy	95% C.I.
GPT-4	Vignette	Cardiology	0.802	(0.744, 0.86)
		Dermatology	0.841	(0.802, 0.88)
		Endocrinology	0.798	(0.713, 0.882)
		Gastroenterology	0.758	(0.699, 0.816)
		Hematology and Oncology	0.835	(0.79, 0.881)
		Infectious Disease	0.842	(0.783, 0.902)
		Neurology	0.825	(0.777, 0.873)
		Obstetrics and Gynecology	0.795	(0.72, 0.869)
		Other	0.826	(0.775, 0.877)
		Pediatrics and Neonatology	0.876	(0.829, 0.923)
		Rheumatology	0.841	(0.766, 0.917)
		Urology and Nephrology	0.751	(0.662, 0.839)
	Multi-turn conversation	Cardiology	0.539	(0.473, 0.605)
		Dermatology	0.732	(0.691, 0.774)
		Endocrinology	0.527	(0.439, 0.616)
		Gastroenterology	0.563	(0.506, 0.619)
		Hematology and Oncology	0.609	(0.559, 0.658)
		Infectious Disease	0.616	(0.547, 0.685)
		Neurology	0.636	(0.58, 0.691)
		Obstetrics and Gynecology	0.68	(0.608, 0.752)
		Other	0.589	(0.534, 0.644)
		Pediatrics and Neonatology	0.724	(0.669, 0.778)
		Rheumatology	0.627	(0.543, 0.711)
		Urology and Nephrology	0.528	(0.435, 0.621)
	Single-turn conversation	Cardiology	0.448	(0.377, 0.518)
		Dermatology	0.647	(0.601, 0.694)
		Endocrinology	0.452	(0.357, 0.547)
		Gastroenterology	0.469	(0.406, 0.532)
		Hematology and Oncology	0.456	(0.398, 0.513)
		Infectious Disease	0.542	(0.465, 0.618)
		Neurology	0.529	(0.468, 0.59)
		Obstetrics and Gynecology	0.524	(0.44, 0.607)
		Other	0.465	(0.402, 0.528)

		Pediatrics and Neonatology	0.624	(0.561, 0.686)
		Rheumatology	0.507	(0.409, 0.605)
		Urology and Nephrology	0.42	(0.325, 0.515)
	Summarized conversation	Cardiology	0.599	(0.536, 0.663)
		Dermatology	0.756	(0.716, 0.795)
		Endocrinology	0.605	(0.521, 0.688)
		Gastroenterology	0.601	(0.542, 0.659)
		Hematology and Oncology	0.644	(0.595, 0.694)
		Infectious Disease	0.699	(0.636, 0.761)
		Neurology	0.702	(0.651, 0.753)
		Obstetrics and Gynecology	0.685	(0.611, 0.758)
		Other	0.602	(0.548, 0.657)
		Pediatrics and Neonatology	0.764	(0.713, 0.816)
		Rheumatology	0.663	(0.579, 0.748)
		Urology and Nephrology	0.602	(0.517, 0.688)
GPT-3.5	Vignette	Cardiology	0.601	(0.533, 0.67)
		Dermatology	0.707	(0.661, 0.754)
		Endocrinology	0.576	(0.476, 0.677)
		Gastroenterology	0.634	(0.575, 0.694)
		Hematology and Oncology	0.636	(0.579, 0.693)
		Infectious Disease	0.677	(0.603, 0.751)
		Neurology	0.624	(0.563, 0.686)
		Obstetrics and Gynecology	0.66	(0.575, 0.745)
		Other	0.712	(0.654, 0.771)
		Pediatrics and Neonatology	0.688	(0.622, 0.754)
		Rheumatology	0.646	(0.55, 0.743)
		Urology and Nephrology	0.625	(0.533, 0.716)
	Multi-turn conversation	Cardiology	0.378	(0.315, 0.441)
		Dermatology	0.549	(0.502, 0.595)
		Endocrinology	0.367	(0.28, 0.454)
		Gastroenterology	0.428	(0.371, 0.485)
		Hematology and Oncology	0.375	(0.323, 0.427)
		Infectious Disease	0.469	(0.399, 0.538)
		Neurology	0.504	(0.446, 0.561)
		Obstetrics and Gynecology	0.482	(0.404, 0.56)

		Other	0.482	(0.424, 0.539)
		Pediatrics and Neonatology	0.584	(0.522, 0.645)
		Rheumatology	0.478	(0.386, 0.57)
		Urology and Nephrology	0.373	(0.283, 0.463)
	Single-turn conversation	Cardiology	0.354	(0.288, 0.419)
		Dermatology	0.507	(0.459, 0.556)
		Endocrinology	0.409	(0.319, 0.499)
		Gastroenterology	0.377	(0.318, 0.435)
		Hematology and Oncology	0.401	(0.346, 0.456)
		Infectious Disease	0.46	(0.385, 0.535)
		Neurology	0.459	(0.399, 0.518)
		Obstetrics and Gynecology	0.413	(0.331, 0.494)
		Other	0.443	(0.382, 0.503)
		Pediatrics and Neonatology	0.511	(0.446, 0.576)
		Rheumatology	0.41	(0.315, 0.504)
		Urology and Nephrology	0.362	(0.275, 0.449)
	Summarized conversation	Cardiology	0.417	(0.352, 0.483)
		Dermatology	0.564	(0.518, 0.609)
		Endocrinology	0.428	(0.34, 0.517)
		Gastroenterology	0.427	(0.37, 0.484)
		Hematology and Oncology	0.437	(0.383, 0.49)
		Infectious Disease	0.559	(0.491, 0.628)
		Neurology	0.521	(0.463, 0.578)
		Obstetrics and Gynecology	0.528	(0.448, 0.607)
		Other	0.545	(0.488, 0.603)
		Pediatrics and Neonatology	0.612	(0.551, 0.673)
		Rheumatology	0.529	(0.44, 0.618)
		Urology and Nephrology	0.465	(0.38, 0.551)
Mistral-v2-7b	Vignette	Cardiology	0.646	(0.572, 0.72)
		Dermatology	0.694	(0.643, 0.745)
		Endocrinology	0.576	(0.469, 0.684)
		Gastroenterology	0.621	(0.553, 0.689)
		Hematology and Oncology	0.59	(0.526, 0.653)
		Infectious Disease	0.606	(0.523, 0.689)
		Neurology	0.584	(0.518, 0.651)

		Obstetrics and Gynecology	0.645	(0.555, 0.736)
		Other	0.687	(0.622, 0.751)
		Pediatrics and Neonatology	0.665	(0.594, 0.736)
		Rheumatology	0.671	(0.567, 0.775)
		Urology and Nephrology	0.607	(0.503, 0.71)
	Multi-turn conversation	Cardiology	0.373	(0.316, 0.43)
		Dermatology	0.511	(0.466, 0.556)
		Endocrinology	0.362	(0.281, 0.444)
		Gastroenterology	0.359	(0.306, 0.411)
		Hematology and Oncology	0.379	(0.33, 0.428)
		Infectious Disease	0.394	(0.327, 0.461)
		Neurology	0.424	(0.37, 0.478)
		Obstetrics and Gynecology	0.464	(0.387, 0.541)
		Other	0.448	(0.391, 0.504)
		Pediatrics and Neonatology	0.481	(0.421, 0.541)
		Rheumatology	0.434	(0.347, 0.521)
		Urology and Nephrology	0.402	(0.319, 0.485)
	Single-turn conversation	Cardiology	0.427	(0.361, 0.493)
		Dermatology	0.531	(0.482, 0.58)
		Endocrinology	0.36	(0.266, 0.454)
		Gastroenterology	0.352	(0.292, 0.411)
		Hematology and Oncology	0.405	(0.35, 0.46)
		Infectious Disease	0.429	(0.353, 0.505)
		Neurology	0.431	(0.373, 0.489)
		Obstetrics and Gynecology	0.467	(0.385, 0.55)
		Other	0.474	(0.411, 0.537)
		Pediatrics and Neonatology	0.509	(0.444, 0.574)
		Rheumatology	0.493	(0.396, 0.59)
		Urology and Nephrology	0.438	(0.344, 0.532)
	Summarized conversation	Cardiology	0.493	(0.438, 0.547)
		Dermatology	0.575	(0.533, 0.618)
		Endocrinology	0.48	(0.397, 0.563)
		Gastroenterology	0.507	(0.454, 0.56)
		Hematology and Oncology	0.45	(0.402, 0.497)
		Infectious Disease	0.485	(0.421, 0.549)

		Neurology	0.521	(0.471, 0.572)
		Obstetrics and Gynecology	0.571	(0.498, 0.644)
		Other	0.471	(0.418, 0.523)
		Pediatrics and Neonatology	0.558	(0.5, 0.617)
		Rheumatology	0.512	(0.43, 0.594)
		Urology and Nephrology	0.49	(0.409, 0.571)
LLaMA-2-7b	Vignette	Cardiology	0.372	(0.304, 0.44)
		Dermatology	0.411	(0.362, 0.461)
		Endocrinology	0.353	(0.26, 0.446)
		Gastroenterology	0.392	(0.331, 0.453)
		Hematology and Oncology	0.321	(0.266, 0.376)
		Infectious Disease	0.388	(0.312, 0.465)
		Neurology	0.365	(0.307, 0.424)
		Obstetrics and Gynecology	0.342	(0.26, 0.424)
		Other	0.482	(0.417, 0.546)
		Pediatrics and Neonatology	0.486	(0.418, 0.553)
		Rheumatology	0.422	(0.321, 0.523)
		Urology and Nephrology	0.378	(0.286, 0.469)
	Multi-turn conversation	Cardiology	0.28	(0.229, 0.332)
		Dermatology	0.346	(0.305, 0.388)
		Endocrinology	0.32	(0.242, 0.398)
		Gastroenterology	0.281	(0.231, 0.331)
		Hematology and Oncology	0.291	(0.245, 0.338)
		Infectious Disease	0.318	(0.257, 0.38)
		Neurology	0.296	(0.249, 0.343)
		Obstetrics and Gynecology	0.325	(0.256, 0.395)
		Other	0.364	(0.309, 0.42)
		Pediatrics and Neonatology	0.345	(0.286, 0.403)
		Rheumatology	0.4	(0.319, 0.481)
		Urology and Nephrology	0.272	(0.194, 0.35)
	Single-turn conversation	Cardiology	0.298	(0.234, 0.362)
		Dermatology	0.326	(0.279, 0.373)
		Endocrinology	0.256	(0.168, 0.345)
		Gastroenterology	0.256	(0.199, 0.312)
		Hematology and Oncology	0.279	(0.226, 0.333)

		Infectious Disease	0.318	(0.247, 0.39)
		Neurology	0.289	(0.233, 0.344)
		Obstetrics and Gynecology	0.331	(0.248, 0.414)
		Other	0.343	(0.283, 0.404)
		Pediatrics and Neonatology	0.336	(0.274, 0.399)
		Rheumatology	0.38	(0.281, 0.48)
		Urology and Nephrology	0.225	(0.147, 0.302)
	Summarized conversation	Cardiology	0.291	(0.231, 0.352)
		Dermatology	0.374	(0.33, 0.418)
		Endocrinology	0.294	(0.215, 0.373)
		Gastroenterology	0.309	(0.257, 0.361)
		Hematology and Oncology	0.289	(0.24, 0.337)
		Infectious Disease	0.353	(0.284, 0.423)
		Neurology	0.329	(0.277, 0.381)
		Obstetrics and Gynecology	0.329	(0.254, 0.405)
		Other	0.369	(0.314, 0.425)
		Pediatrics and Neonatology	0.388	(0.328, 0.449)
		Rheumatology	0.385	(0.298, 0.472)
		Urology and Nephrology	0.265	(0.189, 0.341)

Supplementary Table 6: Medical specialty wise mean accuracy and 95% confidence intervals for 4-choice MCQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Medical specialty	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Summarized conversation	0.0001	0.0003
	Cardiology	Multi-turn conversation	Single-turn conversation	0.0003	0.0007
	Cardiology	Multi-turn conversation	Summarized conversation	0.0001	0.0003
	Cardiology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Summarized conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Summarized conversation	0.0351	0.0513
	Dermatology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Endocrinology	Vignette	Multi-turn conversation	0.0001	0.0003
	Endocrinology	Vignette	Single-turn conversation	0.0001	0.0003
	Endocrinology	Vignette	Summarized conversation	0.0002	0.0005
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.0272	0.0417
	Endocrinology	Multi-turn conversation	Summarized conversation	0.0069	0.0117
	Endocrinology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003

	Gastroenterology	Vignette	Summarized conversation	0.0001	0.0003
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.0007	0.0015
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.0182	0.0286
	Gastroenterology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0003
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0094	0.0156
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0003
	Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0003
	Infectious Disease	Vignette	Summarized conversation	0.0001	0.0003
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.0287	0.0433
	Infectious Disease	Multi-turn conversation	Summarized conversation	0.0002	0.0005
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0003
	Neurology	Vignette	Single-turn conversation	0.0001	0.0003
	Neurology	Vignette	Summarized conversation	0.0001	0.0003
	Neurology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Neurology	Multi-turn conversation	Summarized conversation	0.0002	0.0005
	Neurology	Single-turn	Summarized	0.0001	0.0003

		conversation	conversation		
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0011	0.0022
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0003
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0023	0.0043
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0002	0.0005
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.8411	0.8714
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Other	Multi-turn conversation	Summarized conversation	0.3659	0.4157
	Other	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0057	0.0099
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0003

	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0003
	Rheumatology	Multi-turn conversation	Summarized conversation	0.1444	0.1857
	Rheumatology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Summarized conversation	0.0006	0.0013
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.0044	0.0077
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0174	0.0275
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.0001	0.0003
GPT-3.5	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Summarized conversation	0.0001	0.0003
	Cardiology	Multi-turn conversation	Single-turn conversation	0.2938	0.344
	Cardiology	Multi-turn conversation	Summarized conversation	0.0713	0.0978
	Cardiology	Single-turn conversation	Summarized conversation	0.0124	0.0202
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Summarized conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0158	0.0253
	Dermatology	Multi-turn conversation	Summarized conversation	0.3195	0.3695
	Dermatology	Single-turn conversation	Summarized conversation	0.0051	0.0089
	Endocrinology	Vignette	Multi-turn	0.0004	0.0009

			conversation		
	Endocrinology	Vignette	Single-turn conversation	0.0085	0.0142
	Endocrinology	Vignette	Summarized conversation	0.0033	0.006
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.2877	0.3382
	Endocrinology	Multi-turn conversation	Summarized conversation	0.0443	0.0632
	Endocrinology	Single-turn conversation	Summarized conversation	0.6448	0.7008
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Summarized conversation	0.0001	0.0003
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.0183	0.0286
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.9662	0.9696
	Gastroenterology	Single-turn conversation	Summarized conversation	0.0475	0.0669
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0003
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.2599	0.3093
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0036	0.0065
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.119	0.1537
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0003
	Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0003
	Infectious Disease	Vignette	Summarized conversation	0.0008	0.0017
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.7508	0.7921

	Infectious Disease	Multi-turn conversation	Summarized conversation	0.0002	0.0005
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0005	0.0011
	Neurology	Vignette	Multi-turn conversation	0.0002	0.0005
	Neurology	Vignette	Single-turn conversation	0.0001	0.0003
	Neurology	Vignette	Summarized conversation	0.0001	0.0003
	Neurology	Multi-turn conversation	Single-turn conversation	0.0471	0.0668
	Neurology	Multi-turn conversation	Summarized conversation	0.3892	0.4396
	Neurology	Single-turn conversation	Summarized conversation	0.0118	0.0194
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0001	0.0003
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0003
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0003	0.0007
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0284	0.043
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.1076	0.1409
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0022	0.0042
	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.0476	0.0669
	Other	Multi-turn conversation	Summarized conversation	0.0027	0.005
	Other	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0006	0.0013
	Pediatrics and Neonatology	Vignette	Single-turn	0.0001	0.0003

			conversation		
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0128	0.0207
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0022	0.0042
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.1542	0.1956
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0007	0.0015
	Rheumatology	Vignette	Multi-turn conversation	0.002	0.0038
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Summarized conversation	0.0122	0.02
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0582	0.0814
	Rheumatology	Multi-turn conversation	Summarized conversation	0.1087	0.1417
	Rheumatology	Single-turn conversation	Summarized conversation	0.0004	0.0009
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Summarized conversation	0.0002	0.0005
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.6949	0.7486
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0004	0.0009
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.001	0.0021
Mistral-v2-7b	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0003
	Cardiology	Vignette	Summarized conversation	0.0002	0.0005
	Cardiology	Multi-turn conversation	Single-turn conversation	0.013	0.0209
	Cardiology	Multi-turn conversation	Summarized conversation	0.0001	0.0003

	Cardiology	Single-turn conversation	Summarized conversation	0.0268	0.0413
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0003
	Dermatology	Vignette	Summarized conversation	0.0001	0.0003
	Dermatology	Multi-turn conversation	Single-turn conversation	0.2429	0.2903
	Dermatology	Multi-turn conversation	Summarized conversation	0.0002	0.0005
	Dermatology	Single-turn conversation	Summarized conversation	0.0231	0.0358
	Endocrinology	Vignette	Multi-turn conversation	0.0011	0.0022
	Endocrinology	Vignette	Single-turn conversation	0.0028	0.0051
	Endocrinology	Vignette	Summarized conversation	0.0904	0.12
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.9471	0.9571
	Endocrinology	Multi-turn conversation	Summarized conversation	0.0058	0.01
	Endocrinology	Single-turn conversation	Summarized conversation	0.0062	0.0106
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Summarized conversation	0.0018	0.0035
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.735	0.7782
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.0001	0.0003
	Gastroenterology	Single-turn conversation	Summarized conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0003
	Hematology and Oncology	Vignette	Summarized	0.0002	0.0005

			conversation		
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.1834	0.2287
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0014	0.0028
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0674	0.0929
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0003
	Infectious Disease	Vignette	Single-turn conversation	0.0004	0.0009
	Infectious Disease	Vignette	Summarized conversation	0.0087	0.0145
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.1983	0.2441
	Infectious Disease	Multi-turn conversation	Summarized conversation	0.0013	0.0026
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0858	0.1144
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0003
	Neurology	Vignette	Single-turn conversation	0.0001	0.0003
	Neurology	Vignette	Summarized conversation	0.0372	0.0541
	Neurology	Multi-turn conversation	Single-turn conversation	0.6992	0.7486
	Neurology	Multi-turn conversation	Summarized conversation	0.0001	0.0003
	Neurology	Single-turn conversation	Summarized conversation	0.0002	0.0005
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0002	0.0005
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0006	0.0013
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0797	0.1078
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.8467	0.8714
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.0007	0.0015
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0043	0.0076

	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.1483	0.1898
	Other	Multi-turn conversation	Summarized conversation	0.2698	0.3185
	Other	Single-turn conversation	Summarized conversation	0.9117	0.9245
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0004	0.0009
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0026	0.0048
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.1787	0.2238
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0019	0.0037
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0652	0.0907
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Rheumatology	Vignette	Single-turn conversation	0.0025	0.0047
	Rheumatology	Vignette	Summarized conversation	0.0005	0.0011
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.106	0.1394
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0213	0.0332
	Rheumatology	Single-turn conversation	Summarized conversation	0.6076	0.673
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0002	0.0005
	Urology and Nephrology	Vignette	Single-turn conversation	0.0035	0.0063
	Urology and Nephrology	Vignette	Summarized conversation	0.033	0.0487
	Urology and Nephrology	Multi-turn	Single-turn	0.3229	0.372

		conversation	conversation		
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0166	0.0264
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.2055	0.2508
LLaMA-2-7b	Cardiology	Vignette	Multi-turn conversation	0.0018	0.0035
	Cardiology	Vignette	Single-turn conversation	0.0307	0.0461
	Cardiology	Vignette	Summarized conversation	0.0037	0.0066
	Cardiology	Multi-turn conversation	Single-turn conversation	0.4995	0.5598
	Cardiology	Multi-turn conversation	Summarized conversation	0.6265	0.6913
	Cardiology	Single-turn conversation	Summarized conversation	0.8348	0.8711
	Dermatology	Vignette	Multi-turn conversation	0.0042	0.0075
	Dermatology	Vignette	Single-turn conversation	0.0007	0.0015
	Dermatology	Vignette	Summarized conversation	0.0406	0.0582
	Dermatology	Multi-turn conversation	Single-turn conversation	0.1844	0.2289
	Dermatology	Multi-turn conversation	Summarized conversation	0.1012	0.1337
	Dermatology	Single-turn conversation	Summarized conversation	0.0075	0.0126
	Endocrinology	Vignette	Multi-turn conversation	0.4788	0.5387
	Endocrinology	Vignette	Single-turn conversation	0.0387	0.0557
	Endocrinology	Vignette	Summarized conversation	0.188	0.2324
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.0745	0.1012
	Endocrinology	Multi-turn conversation	Summarized conversation	0.3524	0.4027
	Endocrinology	Single-turn conversation	Summarized conversation	0.3062	0.357
	Gastroenterology	Vignette	Multi-turn conversation	0.0005	0.0011

	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0003
	Gastroenterology	Vignette	Summarized conversation	0.0008	0.0017
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.2663	0.3156
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.1747	0.2197
	Gastroenterology	Single-turn conversation	Summarized conversation	0.0311	0.0464
	Hematology and Oncology	Vignette	Multi-turn conversation	0.2238	0.2708
	Hematology and Oncology	Vignette	Single-turn conversation	0.0811	0.1091
	Hematology and Oncology	Vignette	Summarized conversation	0.1145	0.1485
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.5763	0.6433
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.899	0.9149
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.6075	0.673
	Infectious Disease	Vignette	Multi-turn conversation	0.0319	0.0474
	Infectious Disease	Vignette	Single-turn conversation	0.0664	0.0919
	Infectious Disease	Vignette	Summarized conversation	0.222	0.2698
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.9875	0.9875
	Infectious Disease	Multi-turn conversation	Summarized conversation	0.1598	0.2019
	Infectious Disease	Single-turn conversation	Summarized conversation	0.2261	0.2725
	Neurology	Vignette	Multi-turn conversation	0.002	0.0038
	Neurology	Vignette	Single-turn conversation	0.0006	0.0013
	Neurology	Vignette	Summarized conversation	0.0828	0.1109
	Neurology	Multi-turn conversation	Single-turn conversation	0.7036	0.7505
	Neurology	Multi-turn	Summarized	0.0723	0.0987

		conversation	conversation		
	Neurology	Single-turn conversation	Summarized conversation	0.038	0.055
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.6343	0.6954
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.7654	0.8045
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.6576	0.712
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.8472	0.8714
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.8743	0.8961
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.9586	0.9653
	Other	Vignette	Multi-turn conversation	0.0001	0.0003
	Other	Vignette	Single-turn conversation	0.0001	0.0003
	Other	Vignette	Summarized conversation	0.0001	0.0003
	Other	Multi-turn conversation	Single-turn conversation	0.3146	0.3653
	Other	Multi-turn conversation	Summarized conversation	0.8103	0.8486
	Other	Single-turn conversation	Summarized conversation	0.1992	0.2441
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0003
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0003
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.7241	0.7695
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0279	0.0425
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0334	0.0491
	Rheumatology	Vignette	Multi-turn conversation	0.6379	0.6959
	Rheumatology	Vignette	Single-turn conversation	0.3666	0.4157

	Rheumatology	Vignette	Summarized conversation	0.3378	0.3876
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.635	0.6954
	Rheumatology	Multi-turn conversation	Summarized conversation	0.6992	0.7486
	Rheumatology	Single-turn conversation	Summarized conversation	0.8946	0.9136
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0061	0.0105
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0003
	Urology and Nephrology	Vignette	Summarized conversation	0.0015	0.003
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.1506	0.1919
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.846	0.8714
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.232	0.2784

Supplementary Table 7: Medical specialty wise adjusted p-values for 4-choice MCQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment Name	Specialty	Mean Accuracy	95% C.I.
GPT-4	Vignette	Cardiology	0.345	(0.281, 0.409)
		Dermatology	0.645	(0.597, 0.693)
		Endocrinology	0.379	(0.286, 0.472)
		Gastroenterology	0.365	(0.304, 0.425)
		Hematology and Oncology	0.467	(0.409, 0.525)
		Infectious Disease	0.416	(0.339, 0.493)
		Neurology	0.513	(0.453, 0.574)
		Obstetrics and Gynecology	0.453	(0.366, 0.54)
		Other	0.484	(0.422, 0.545)
		Pediatrics and Neonatology	0.557	(0.49, 0.624)
		Rheumatology	0.595	(0.492, 0.698)
		Urology and Nephrology	0.449	(0.354, 0.545)
	Multi-turn conversation	Cardiology	0.134	(0.093, 0.176)
		Dermatology	0.399	(0.354, 0.443)
		Endocrinology	0.108	(0.063, 0.154)
		Gastroenterology	0.152	(0.11, 0.193)
		Hematology and Oncology	0.203	(0.163, 0.243)
		Infectious Disease	0.206	(0.15, 0.261)
		Neurology	0.321	(0.267, 0.376)
		Obstetrics and Gynecology	0.264	(0.194, 0.333)
		Other	0.295	(0.242, 0.347)
		Pediatrics and Neonatology	0.369	(0.31, 0.428)
		Rheumatology	0.329	(0.244, 0.414)
		Urology and Nephrology	0.196	(0.126, 0.265)
	Single-turn conversation	Cardiology	0.07	(0.039, 0.1)
		Dermatology	0.224	(0.184, 0.264)
		Endocrinology	0.024	(-0.003, 0.05)
		Gastroenterology	0.085	(0.051, 0.118)
		Hematology and Oncology	0.078	(0.05, 0.105)
		Infectious Disease	0.117	(0.071, 0.163)
		Neurology	0.185	(0.138, 0.232)
		Obstetrics and Gynecology	0.071	(0.034, 0.107)
		Other	0.142	(0.101, 0.184)

		Pediatrics and Neonatology	0.195	(0.145, 0.246)
		Rheumatology	0.132	(0.068, 0.195)
		Urology and Nephrology	0.115	(0.056, 0.173)
	Summarized conversation	Cardiology	0.122	(0.082, 0.161)
		Dermatology	0.436	(0.391, 0.482)
		Endocrinology	0.129	(0.073, 0.186)
		Gastroenterology	0.165	(0.121, 0.208)
		Hematology and Oncology	0.219	(0.177, 0.261)
		Infectious Disease	0.257	(0.196, 0.318)
		Neurology	0.287	(0.239, 0.336)
		Obstetrics and Gynecology	0.271	(0.2, 0.342)
		Other	0.284	(0.236, 0.333)
		Pediatrics and Neonatology	0.328	(0.272, 0.383)
		Rheumatology	0.371	(0.282, 0.459)
		Urology and Nephrology	0.236	(0.162, 0.31)
GPT-3.5	Vignette	Cardiology	0.254	(0.196, 0.311)
		Dermatology	0.504	(0.454, 0.554)
		Endocrinology	0.212	(0.135, 0.288)
		Gastroenterology	0.311	(0.252, 0.371)
		Hematology and Oncology	0.372	(0.317, 0.426)
		Infectious Disease	0.364	(0.29, 0.437)
		Neurology	0.341	(0.284, 0.398)
		Obstetrics and Gynecology	0.338	(0.256, 0.421)
		Other	0.372	(0.312, 0.432)
		Pediatrics and Neonatology	0.445	(0.378, 0.512)
		Rheumatology	0.463	(0.357, 0.57)
		Urology and Nephrology	0.375	(0.284, 0.467)
	Multi-turn conversation	Cardiology	0.085	(0.051, 0.119)
		Dermatology	0.241	(0.202, 0.28)
		Endocrinology	0.054	(0.016, 0.093)
		Gastroenterology	0.092	(0.061, 0.123)
		Hematology and Oncology	0.13	(0.096, 0.164)
		Infectious Disease	0.137	(0.09, 0.185)
		Neurology	0.206	(0.16, 0.251)
		Obstetrics and Gynecology	0.198	(0.14, 0.256)

		Other	0.214	(0.168, 0.26)
		Pediatrics and Neonatology	0.24	(0.186, 0.295)
		Rheumatology	0.18	(0.111, 0.25)
		Urology and Nephrology	0.121	(0.07, 0.172)
	Single-turn conversation	Cardiology	0.057	(0.029, 0.085)
		Dermatology	0.169	(0.133, 0.205)
		Endocrinology	0.038	(0.005, 0.071)
		Gastroenterology	0.083	(0.05, 0.116)
		Hematology and Oncology	0.092	(0.061, 0.124)
		Infectious Disease	0.092	(0.05, 0.134)
		Neurology	0.153	(0.111, 0.195)
		Obstetrics and Gynecology	0.131	(0.077, 0.184)
		Other	0.166	(0.123, 0.21)
		Pediatrics and Neonatology	0.182	(0.131, 0.232)
		Rheumatology	0.102	(0.044, 0.161)
		Urology and Nephrology	0.099	(0.046, 0.151)
	Summarized conversation	Cardiology	0.079	(0.047, 0.111)
		Dermatology	0.275	(0.234, 0.315)
		Endocrinology	0.045	(0.013, 0.077)
		Gastroenterology	0.133	(0.096, 0.171)
		Hematology and Oncology	0.152	(0.116, 0.189)
		Infectious Disease	0.155	(0.107, 0.203)
		Neurology	0.179	(0.135, 0.222)
		Obstetrics and Gynecology	0.205	(0.144, 0.267)
		Other	0.167	(0.128, 0.206)
		Pediatrics and Neonatology	0.207	(0.158, 0.256)
		Rheumatology	0.238	(0.16, 0.315)
		Urology and Nephrology	0.135	(0.077, 0.193)
Mistral-v2-7b	Vignette	Cardiology	0.155	(0.103, 0.207)
		Dermatology	0.217	(0.173, 0.261)
		Endocrinology	0.167	(0.091, 0.243)
		Gastroenterology	0.197	(0.143, 0.251)
		Hematology and Oncology	0.227	(0.176, 0.278)
		Infectious Disease	0.19	(0.128, 0.252)
		Neurology	0.247	(0.192, 0.302)

		Obstetrics and Gynecology	0.198	(0.128, 0.269)
		Other	0.286	(0.226, 0.345)
		Pediatrics and Neonatology	0.239	(0.178, 0.301)
		Rheumatology	0.276	(0.179, 0.372)
		Urology and Nephrology	0.261	(0.171, 0.35)
	Multi-turn conversation	Cardiology	0.033	(0.015, 0.051)
		Dermatology	0.099	(0.073, 0.125)
		Endocrinology	0.028	(-0.002, 0.059)
		Gastroenterology	0.032	(0.014, 0.05)
		Hematology and Oncology	0.033	(0.015, 0.051)
		Infectious Disease	0.045	(0.02, 0.07)
		Neurology	0.083	(0.053, 0.113)
		Obstetrics and Gynecology	0.042	(0.018, 0.066)
		Other	0.11	(0.076, 0.145)
		Pediatrics and Neonatology	0.108	(0.068, 0.147)
		Rheumatology	0.073	(0.027, 0.12)
		Urology and Nephrology	0.036	(0.013, 0.059)
	Single-turn conversation	Cardiology	0.027	(0.009, 0.045)
		Dermatology	0.086	(0.06, 0.113)
		Endocrinology	0.028	(-0.002, 0.058)
		Gastroenterology	0.028	(0.01, 0.046)
		Hematology and Oncology	0.037	(0.017, 0.057)
		Infectious Disease	0.029	(0.004, 0.055)
		Neurology	0.061	(0.034, 0.087)
		Obstetrics and Gynecology	0.033	(0.01, 0.055)
		Other	0.096	(0.062, 0.129)
		Pediatrics and Neonatology	0.09	(0.052, 0.129)
		Rheumatology	0.061	(0.015, 0.107)
		Urology and Nephrology	0.047	(0.015, 0.08)
	Summarized conversation	Cardiology	0.024	(0.01, 0.039)
		Dermatology	0.09	(0.066, 0.114)
		Endocrinology	0.016	(-0.005, 0.038)
		Gastroenterology	0.021	(0.006, 0.036)
		Hematology and Oncology	0.032	(0.017, 0.046)
		Infectious Disease	0.039	(0.018, 0.061)

		Neurology	0.082	(0.054, 0.111)
		Obstetrics and Gynecology	0.06	(0.028, 0.092)
		Other	0.071	(0.046, 0.096)
		Pediatrics and Neonatology	0.069	(0.04, 0.098)
		Rheumatology	0.085	(0.041, 0.13)
		Urology and Nephrology	0.036	(0.004, 0.067)
LLaMA-2-7b	Vignette	Cardiology	0.101	(0.063, 0.139)
		Dermatology	0.175	(0.136, 0.213)
		Endocrinology	0.082	(0.034, 0.131)
		Gastroenterology	0.143	(0.102, 0.185)
		Hematology and Oncology	0.203	(0.156, 0.251)
		Infectious Disease	0.166	(0.11, 0.223)
		Neurology	0.191	(0.144, 0.237)
		Obstetrics and Gynecology	0.151	(0.091, 0.211)
		Other	0.192	(0.146, 0.239)
		Pediatrics and Neonatology	0.182	(0.13, 0.233)
		Rheumatology	0.249	(0.163, 0.334)
		Urology and Nephrology	0.157	(0.091, 0.224)
	Multi-turn conversation	Cardiology	0.039	(0.022, 0.056)
		Dermatology	0.084	(0.061, 0.107)
		Endocrinology	0.028	(0.007, 0.049)
		Gastroenterology	0.026	(0.01, 0.042)
		Hematology and Oncology	0.022	(0.009, 0.035)
		Infectious Disease	0.031	(0.01, 0.051)
		Neurology	0.076	(0.048, 0.104)
		Obstetrics and Gynecology	0.122	(0.077, 0.166)
		Other	0.132	(0.097, 0.168)
		Pediatrics and Neonatology	0.094	(0.059, 0.128)
		Rheumatology	0.093	(0.045, 0.141)
		Urology and Nephrology	0.029	(0.008, 0.051)
	Single-turn conversation	Cardiology	0.049	(0.021, 0.077)
		Dermatology	0.074	(0.049, 0.1)
		Endocrinology	0.007	(-0.007, 0.021)
		Gastroenterology	0.027	(0.007, 0.048)
		Hematology and Oncology	0.027	(0.011, 0.044)

		Infectious Disease	0.039	(0.011, 0.068)
		Neurology	0.059	(0.031, 0.087)
		Obstetrics and Gynecology	0.105	(0.053, 0.158)
		Other	0.143	(0.1, 0.186)
		Pediatrics and Neonatology	0.079	(0.046, 0.111)
		Rheumatology	0.112	(0.05, 0.174)
		Urology and Nephrology	0.056	(0.015, 0.098)
	Summarized conversation	Cardiology	0.048	(0.023, 0.072)
		Dermatology	0.102	(0.075, 0.128)
		Endocrinology	0.033	(0.006, 0.059)
		Gastroenterology	0.032	(0.011, 0.053)
		Hematology and Oncology	0.051	(0.03, 0.072)
		Infectious Disease	0.079	(0.043, 0.115)
		Neurology	0.085	(0.056, 0.114)
		Obstetrics and Gynecology	0.116	(0.07, 0.163)
		Other	0.124	(0.09, 0.158)
		Pediatrics and Neonatology	0.094	(0.064, 0.124)
		Rheumatology	0.149	(0.08, 0.217)
		Urology and Nephrology	0.067	(0.027, 0.108)

Supplementary Table 8: Medical specialty wise mean accuracy and 95% confidence intervals for FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Medical specialty	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Summarized conversation	0.0001	0.0002
	Cardiology	Multi-turn conversation	Single-turn conversation	0.0003	0.0006
	Cardiology	Multi-turn conversation	Summarized conversation	0.4286	0.486
	Cardiology	Single-turn conversation	Summarized conversation	0.0057	0.0089
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Summarized conversation	0.0013	0.0023
	Dermatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Endocrinology	Vignette	Multi-turn conversation	0.0001	0.0002
	Endocrinology	Vignette	Single-turn conversation	0.0001	0.0002
	Endocrinology	Vignette	Summarized conversation	0.0001	0.0002
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Endocrinology	Multi-turn conversation	Summarized conversation	0.1678	0.2129
	Endocrinology	Single-turn conversation	Summarized conversation	0.0003	0.0006
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002

	Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.0002	0.0004
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.2019	0.2517
	Gastroenterology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0002
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.2584	0.314
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Summarized conversation	0.0001	0.0002
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.0014	0.0025
	Infectious Disease	Multi-turn conversation	Summarized conversation	0.0042	0.0067
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
	Neurology	Vignette	Single-turn conversation	0.0001	0.0002
	Neurology	Vignette	Summarized conversation	0.0001	0.0002
	Neurology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Neurology	Multi-turn conversation	Summarized conversation	0.0553	0.0788
	Neurology	Single-turn	Summarized	0.0001	0.0002

		conversation	conversation		
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0001	0.0002
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.7675	0.7923
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Other	Vignette	Multi-turn conversation	0.0001	0.0002
	Other	Vignette	Single-turn conversation	0.0001	0.0002
	Other	Vignette	Summarized conversation	0.0001	0.0002
	Other	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Other	Multi-turn conversation	Summarized conversation	0.5739	0.6167
	Other	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0402	0.0585
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0002

	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0472	0.068
	Rheumatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.0007	0.0013
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0226	0.0334
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.0001	0.0002
GPT-3.5	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Summarized conversation	0.0001	0.0002
	Cardiology	Multi-turn conversation	Single-turn conversation	0.0211	0.0313
	Cardiology	Multi-turn conversation	Summarized conversation	0.5866	0.628
	Cardiology	Single-turn conversation	Summarized conversation	0.0838	0.1155
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Summarized conversation	0.0038	0.0061
	Dermatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Endocrinology	Vignette	Multi-turn	0.0003	0.0006

			conversation		
	Endocrinology	Vignette	Single-turn conversation	0.0001	0.0002
	Endocrinology	Vignette	Summarized conversation	0.0001	0.0002
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.2519	0.3087
	Endocrinology	Multi-turn conversation	Summarized conversation	0.5935	0.6331
	Endocrinology	Single-turn conversation	Summarized conversation	0.689	0.7216
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.5041	0.5532
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.0024	0.0041
	Gastroenterology	Single-turn conversation	Summarized conversation	0.0034	0.0056
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0002
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.003	0.005
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.056	0.0794
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0003	0.0006
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Summarized conversation	0.0001	0.0002
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.0245	0.036

	Infectious Disease	Multi-turn conversation	Summarized conversation	0.3615	0.4249
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0043	0.0068
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
	Neurology	Vignette	Single-turn conversation	0.0001	0.0002
	Neurology	Vignette	Summarized conversation	0.0001	0.0002
	Neurology	Multi-turn conversation	Single-turn conversation	0.002	0.0035
	Neurology	Multi-turn conversation	Summarized conversation	0.0857	0.1175
	Neurology	Single-turn conversation	Summarized conversation	0.1727	0.2181
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0004	0.0008
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0001	0.0002
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.0002	0.0004
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.7258	0.7546
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.0035	0.0057
	Other	Vignette	Multi-turn conversation	0.0001	0.0002
	Other	Vignette	Single-turn conversation	0.0001	0.0002
	Other	Vignette	Summarized conversation	0.0001	0.0002
	Other	Multi-turn conversation	Single-turn conversation	0.0023	0.0039
	Other	Multi-turn conversation	Summarized conversation	0.0063	0.0097
	Other	Single-turn conversation	Summarized conversation	0.9552	0.9653
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Single-turn	0.0001	0.0002

			conversation		
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0013	0.0023
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.1245	0.1668
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.2319	0.2866
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0002
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.0063	0.0097
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0031	0.0051
	Rheumatology	Single-turn conversation	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.1299	0.1708
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.5503	0.5958
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.0462	0.0669
Mistral-v2-7b	Cardiology	Vignette	Multi-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Single-turn conversation	0.0001	0.0002
	Cardiology	Vignette	Summarized conversation	0.0001	0.0002
	Cardiology	Multi-turn conversation	Single-turn conversation	0.4526	0.5013
	Cardiology	Multi-turn conversation	Summarized conversation	0.2343	0.2884

	Cardiology	Single-turn conversation	Summarized conversation	0.7807	0.803
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.1451	0.1891
	Dermatology	Multi-turn conversation	Summarized conversation	0.2174	0.2699
	Dermatology	Single-turn conversation	Summarized conversation	0.7464	0.7732
	Endocrinology	Vignette	Multi-turn conversation	0.0006	0.0011
	Endocrinology	Vignette	Single-turn conversation	0.0006	0.0011
	Endocrinology	Vignette	Summarized conversation	0.0003	0.0006
	Endocrinology	Multi-turn conversation	Single-turn conversation	1.0	1.0
	Endocrinology	Multi-turn conversation	Summarized conversation	0.1031	0.1401
	Endocrinology	Single-turn conversation	Summarized conversation	0.3447	0.4085
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.4457	0.4975
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.1289	0.1708
	Gastroenterology	Single-turn conversation	Summarized conversation	0.4384	0.4932
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Summarized	0.0001	0.0002

			conversation		
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.6697	0.7039
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.8405	0.8614
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.613	0.6491
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Summarized conversation	0.0001	0.0002
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.1297	0.1708
	Infectious Disease	Multi-turn conversation	Summarized conversation	0.4056	0.4669
	Infectious Disease	Single-turn conversation	Summarized conversation	0.3286	0.3911
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
	Neurology	Vignette	Single-turn conversation	0.0001	0.0002
	Neurology	Vignette	Summarized conversation	0.0001	0.0002
	Neurology	Multi-turn conversation	Single-turn conversation	0.0041	0.0066
	Neurology	Multi-turn conversation	Summarized conversation	0.9391	0.9523
	Neurology	Single-turn conversation	Summarized conversation	0.0745	0.1042
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.0001	0.0002
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.0002	0.0004
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.4919	0.5428
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.1661	0.2118
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.1326	0.1736

	Other	Vignette	Multi-turn conversation	0.0001	0.0002
	Other	Vignette	Single-turn conversation	0.0001	0.0002
	Other	Vignette	Summarized conversation	0.0001	0.0002
	Other	Multi-turn conversation	Single-turn conversation	0.1878	0.2352
	Other	Multi-turn conversation	Summarized conversation	0.0089	0.0136
	Other	Single-turn conversation	Summarized conversation	0.1592	0.2062
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0001	0.0002
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.0802	0.1116
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.0057	0.0089
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.1087	0.147
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Summarized conversation	0.0001	0.0002
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.5254	0.5732
	Rheumatology	Multi-turn conversation	Summarized conversation	0.4354	0.4917
	Rheumatology	Single-turn conversation	Summarized conversation	0.1656	0.2118
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Summarized conversation	0.0001	0.0002
	Urology and Nephrology	Multi-turn	Single-turn	0.4232	0.4829

		conversation	conversation		
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.9699	0.9767
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.6103	0.6486
LLaMA-2-7b	Cardiology	Vignette	Multi-turn conversation	0.0009	0.0017
	Cardiology	Vignette	Single-turn conversation	0.0057	0.0089
	Cardiology	Vignette	Summarized conversation	0.0065	0.01
	Cardiology	Multi-turn conversation	Single-turn conversation	0.3935	0.4551
	Cardiology	Multi-turn conversation	Summarized conversation	0.4069	0.4669
	Cardiology	Single-turn conversation	Summarized conversation	0.8954	0.9145
	Dermatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Single-turn conversation	0.0001	0.0002
	Dermatology	Vignette	Summarized conversation	0.0001	0.0002
	Dermatology	Multi-turn conversation	Single-turn conversation	0.3827	0.4462
	Dermatology	Multi-turn conversation	Summarized conversation	0.0672	0.0944
	Dermatology	Single-turn conversation	Summarized conversation	0.0193	0.0288
	Endocrinology	Vignette	Multi-turn conversation	0.0125	0.0188
	Endocrinology	Vignette	Single-turn conversation	0.0025	0.0042
	Endocrinology	Vignette	Summarized conversation	0.0052	0.0082
	Endocrinology	Multi-turn conversation	Single-turn conversation	0.082	0.1135
	Endocrinology	Multi-turn conversation	Summarized conversation	0.7129	0.7439
	Endocrinology	Single-turn conversation	Summarized conversation	0.0988	0.1349
	Gastroenterology	Vignette	Multi-turn conversation	0.0001	0.0002

	Gastroenterology	Vignette	Single-turn conversation	0.0001	0.0002
	Gastroenterology	Vignette	Summarized conversation	0.0001	0.0002
	Gastroenterology	Multi-turn conversation	Single-turn conversation	0.9041	0.9201
	Gastroenterology	Multi-turn conversation	Summarized conversation	0.3672	0.4299
	Gastroenterology	Single-turn conversation	Summarized conversation	0.4409	0.4941
	Hematology and Oncology	Vignette	Multi-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Single-turn conversation	0.0001	0.0002
	Hematology and Oncology	Vignette	Summarized conversation	0.0001	0.0002
	Hematology and Oncology	Multi-turn conversation	Single-turn conversation	0.4521	0.5013
	Hematology and Oncology	Multi-turn conversation	Summarized conversation	0.0024	0.0041
	Hematology and Oncology	Single-turn conversation	Summarized conversation	0.0012	0.0022
	Infectious Disease	Vignette	Multi-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Single-turn conversation	0.0001	0.0002
	Infectious Disease	Vignette	Summarized conversation	0.0003	0.0006
	Infectious Disease	Multi-turn conversation	Single-turn conversation	0.4242	0.4829
	Infectious Disease	Multi-turn conversation	Summarized conversation	0.0004	0.0008
	Infectious Disease	Single-turn conversation	Summarized conversation	0.0026	0.0044
	Neurology	Vignette	Multi-turn conversation	0.0001	0.0002
	Neurology	Vignette	Single-turn conversation	0.0001	0.0002
	Neurology	Vignette	Summarized conversation	0.0001	0.0002
	Neurology	Multi-turn conversation	Single-turn conversation	0.1278	0.1704
	Neurology	Multi-turn	Summarized	0.3604	0.4249

		conversation	conversation		
	Neurology	Single-turn conversation	Summarized conversation	0.0495	0.0709
	Obstetrics and Gynecology	Vignette	Multi-turn conversation	0.3211	0.3837
	Obstetrics and Gynecology	Vignette	Single-turn conversation	0.1807	0.2273
	Obstetrics and Gynecology	Vignette	Summarized conversation	0.1662	0.2118
	Obstetrics and Gynecology	Multi-turn conversation	Single-turn conversation	0.2929	0.353
	Obstetrics and Gynecology	Multi-turn conversation	Summarized conversation	0.6524	0.6882
	Obstetrics and Gynecology	Single-turn conversation	Summarized conversation	0.5052	0.5532
	Other	Vignette	Multi-turn conversation	0.0037	0.006
	Other	Vignette	Single-turn conversation	0.0325	0.0475
	Other	Vignette	Summarized conversation	0.0012	0.0022
	Other	Multi-turn conversation	Single-turn conversation	0.5344	0.5808
	Other	Multi-turn conversation	Summarized conversation	0.5592	0.6032
	Other	Single-turn conversation	Summarized conversation	0.2583	0.314
	Pediatrics and Neonatology	Vignette	Multi-turn conversation	0.0007	0.0013
	Pediatrics and Neonatology	Vignette	Single-turn conversation	0.0001	0.0002
	Pediatrics and Neonatology	Vignette	Summarized conversation	0.0002	0.0004
	Pediatrics and Neonatology	Multi-turn conversation	Single-turn conversation	0.1597	0.2062
	Pediatrics and Neonatology	Multi-turn conversation	Summarized conversation	0.985	0.9884
	Pediatrics and Neonatology	Single-turn conversation	Summarized conversation	0.2708	0.3277
	Rheumatology	Vignette	Multi-turn conversation	0.0001	0.0002
	Rheumatology	Vignette	Single-turn conversation	0.0013	0.0023

	Rheumatology	Vignette	Summarized conversation	0.0093	0.0141
	Rheumatology	Multi-turn conversation	Single-turn conversation	0.3063	0.3676
	Rheumatology	Multi-turn conversation	Summarized conversation	0.0022	0.0038
	Rheumatology	Single-turn conversation	Summarized conversation	0.1176	0.1583
	Urology and Nephrology	Vignette	Multi-turn conversation	0.0001	0.0002
	Urology and Nephrology	Vignette	Single-turn conversation	0.0022	0.0038
	Urology and Nephrology	Vignette	Summarized conversation	0.002	0.0035
	Urology and Nephrology	Multi-turn conversation	Single-turn conversation	0.0573	0.0809
	Urology and Nephrology	Multi-turn conversation	Summarized conversation	0.0109	0.0164
	Urology and Nephrology	Single-turn conversation	Summarized conversation	0.3912	0.4543

Supplementary Table 9: Medical specialty wise adjusted p-values for FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.791	(0.721, 0.862)
	Vignette	Derm-Private	0.852	(0.782, 0.922)
	Vignette	Derm-Public	0.888	(0.826, 0.95)
	Multi-turn conversation	MedQA-USMLE	0.631	(0.557, 0.705)
	Multi-turn conversation	Derm-Private	0.796	(0.727, 0.865)
	Multi-turn conversation	Derm-Public	0.788	(0.719, 0.857)
	Single-turn conversation	MedQA-USMLE	0.566	(0.487, 0.645)
	Single-turn conversation	Derm-Private	0.678	(0.596, 0.76)
	Single-turn conversation	Derm-Public	0.712	(0.631, 0.793)
	Summarized conversation	MedQA-USMLE	0.679	(0.609, 0.749)
	Summarized conversation	Derm-Private	0.787	(0.717, 0.857)
	Summarized conversation	Derm-Public	0.814	(0.751, 0.877)
GPT-3.5	Vignette	MedQA-USMLE	0.603	(0.522, 0.685)
	Vignette	Derm-Private	0.75	(0.669, 0.831)
	Vignette	Derm-Public	0.786	(0.709, 0.863)
	Multi-turn conversation	MedQA-USMLE	0.46	(0.383, 0.536)
	Multi-turn conversation	Derm-Private	0.588	(0.505, 0.671)
	Multi-turn conversation	Derm-Public	0.614	(0.532, 0.696)
	Single-turn conversation	MedQA-USMLE	0.451	(0.371, 0.531)
	Single-turn conversation	Derm-Private	0.494	(0.407, 0.581)
	Single-turn conversation	Derm-Public	0.586	(0.501, 0.671)
	Summarized conversation	MedQA-USMLE	0.458	(0.381, 0.536)
	Summarized conversation	Derm-Private	0.602	(0.521, 0.683)
	Summarized conversation	Derm-Public	0.65	(0.574, 0.725)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.607	(0.517, 0.697)
	Vignette	Derm-Private	0.77	(0.686, 0.854)
	Vignette	Derm-Public	0.72	(0.63, 0.81)
	Multi-turn conversation	MedQA-USMLE	0.453	(0.379, 0.527)
	Multi-turn conversation	Derm-Private	0.522	(0.441, 0.603)
	Multi-turn conversation	Derm-Public	0.568	(0.487, 0.649)
	Single-turn conversation	MedQA-USMLE	0.515	(0.433, 0.596)
	Single-turn conversation	Derm-Private	0.508	(0.416, 0.6)
	Single-turn conversation	Derm-Public	0.574	(0.489, 0.659)

	Summarized conversation	MedQA-USMLE	0.562	(0.493, 0.631)
	Summarized conversation	Derm-Private	0.548	(0.47, 0.626)
	Summarized conversation	Derm-Public	0.618	(0.542, 0.694)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.385	(0.304, 0.465)
	Vignette	Derm-Private	0.44	(0.349, 0.531)
	Vignette	Derm-Public	0.414	(0.323, 0.505)
	Multi-turn conversation	MedQA-USMLE	0.383	(0.316, 0.45)
	Multi-turn conversation	Derm-Private	0.27	(0.197, 0.343)
	Multi-turn conversation	Derm-Public	0.38	(0.301, 0.459)
	Single-turn conversation	MedQA-USMLE	0.371	(0.291, 0.451)
	Single-turn conversation	Derm-Private	0.244	(0.165, 0.323)
	Single-turn conversation	Derm-Public	0.354	(0.268, 0.44)
	Summarized conversation	MedQA-USMLE	0.405	(0.33, 0.48)
	Summarized conversation	Derm-Private	0.31	(0.236, 0.384)
	Summarized conversation	Derm-Public	0.4	(0.319, 0.481)

Supplementary Table 10: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0003	0.001
		Multi-turn conversation	Single-turn conversation	0.0234	0.0443
		Multi-turn conversation	Summarized conversation	0.0181	0.0372
		Single-turn conversation	Summarized conversation	0.0003	0.001
	Derm-Public	Vignette	Multi-turn conversation	0.0025	0.0062
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0132	0.028
		Multi-turn conversation	Single-turn conversation	0.0219	0.0426
		Multi-turn conversation	Summarized conversation	0.1866	0.2742
		Single-turn conversation	Summarized conversation	0.0015	0.004
	Derm-Private	Vignette	Multi-turn conversation	0.0065	0.0142
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0033	0.0079
		Multi-turn conversation	Single-turn conversation	0.0002	0.0007
		Multi-turn conversation	Summarized conversation	0.5506	0.6293
		Single-turn conversation	Summarized conversation	0.0011	0.0032
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0013	0.0036
		Vignette	Single-turn conversation	0.0004	0.0013
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.7558	0.8003
		Multi-turn conversation	Summarized conversation	0.9599	0.9599
		Single-turn conversation	Summarized conversation	0.8409	0.8649
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.3121	0.4086
		Multi-turn conversation	Summarized conversation	0.1919	0.2763
		Single-turn conversation	Summarized conversation	0.067	0.1149
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004

		Multi-turn conversation	Single-turn conversation	0.0047	0.0106
		Multi-turn conversation	Summarized conversation	0.5018	0.5923
		Single-turn conversation	Summarized conversation	0.0023	0.0059
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.001	0.003
		Vignette	Single-turn conversation	0.0736	0.1232
		Vignette	Summarized conversation	0.307	0.4086
		Multi-turn conversation	Single-turn conversation	0.0584	0.1051
		Multi-turn conversation	Summarized conversation	0.0001	0.0004
		Single-turn conversation	Summarized conversation	0.1977	0.2791
	Derm-Public	Vignette	Multi-turn conversation	0.0002	0.0007
		Vignette	Single-turn conversation	0.004	0.0093
		Vignette	Summarized conversation	0.0189	0.0378
		Multi-turn conversation	Single-turn conversation	0.8129	0.8482
		Multi-turn conversation	Summarized conversation	0.0602	0.1057
		Single-turn conversation	Summarized conversation	0.1336	0.2186
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.6331	0.7122
		Multi-turn conversation	Summarized conversation	0.3858	0.4873
		Single-turn conversation	Summarized conversation	0.1858	0.2742
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.9505	0.9599
		Vignette	Single-turn conversation	0.7234	0.7774
		Vignette	Summarized conversation	0.4754	0.5705
		Multi-turn conversation	Single-turn conversation	0.6705	0.7427
		Multi-turn conversation	Summarized conversation	0.4286	0.523
		Single-turn conversation	Summarized conversation	0.2417	0.3347
	Derm-Public	Vignette	Multi-turn conversation	0.4243	0.523
		Vignette	Single-turn conversation	0.1789	0.2741
		Vignette	Summarized conversation	0.6821	0.7441
		Multi-turn conversation	Single-turn conversation	0.3237	0.4162
		Multi-turn conversation	Summarized conversation	0.5145	0.5975
		Single-turn conversation	Summarized conversation	0.155	0.248
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004

		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.2841	0.3859
		Multi-turn conversation	Summarized conversation	0.1611	0.2522
		Single-turn conversation	Summarized conversation	0.0381	0.0703

Supplementary Table 11: Adjusted p-values for Dermatology, calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.552	(0.467, 0.637)
	Vignette	Derm-Private	0.818	(0.751, 0.885)
	Vignette	Derm-Public	0.582	(0.495, 0.669)
	Multi-turn conversation	MedQA-USMLE	0.345	(0.274, 0.416)
	Multi-turn conversation	Derm-Private	0.572	(0.493, 0.651)
	Multi-turn conversation	Derm-Public	0.288	(0.215, 0.361)
	Single-turn conversation	MedQA-USMLE	0.244	(0.174, 0.314)
	Single-turn conversation	Derm-Private	0.256	(0.18, 0.332)
	Single-turn conversation	Derm-Public	0.168	(0.105, 0.231)
	Summarized conversation	MedQA-USMLE	0.397	(0.321, 0.472)
	Summarized conversation	Derm-Private	0.586	(0.507, 0.664)
	Summarized conversation	Derm-Public	0.334	(0.258, 0.41)
GPT-3.5	Vignette	MedQA-USMLE	0.496	(0.41, 0.582)
	Vignette	Derm-Private	0.608	(0.524, 0.692)
	Vignette	Derm-Public	0.41	(0.324, 0.496)
	Multi-turn conversation	MedQA-USMLE	0.238	(0.17, 0.306)
	Multi-turn conversation	Derm-Private	0.34	(0.263, 0.417)
	Multi-turn conversation	Derm-Public	0.146	(0.091, 0.201)
	Single-turn conversation	MedQA-USMLE	0.173	(0.116, 0.23)
	Single-turn conversation	Derm-Private	0.236	(0.159, 0.313)
	Single-turn conversation	Derm-Public	0.098	(0.05, 0.146)
	Summarized conversation	MedQA-USMLE	0.244	(0.178, 0.31)
	Summarized conversation	Derm-Private	0.401	(0.32, 0.483)
	Summarized conversation	Derm-Public	0.183	(0.127, 0.24)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.253	(0.176, 0.33)
	Vignette	Derm-Private	0.296	(0.207, 0.385)
	Vignette	Derm-Public	0.096	(0.039, 0.153)
	Multi-turn conversation	MedQA-USMLE	0.097	(0.054, 0.141)
	Multi-turn conversation	Derm-Private	0.14	(0.084, 0.196)
	Multi-turn conversation	Derm-Public	0.06	(0.028, 0.092)
	Single-turn conversation	MedQA-USMLE	0.101	(0.052, 0.15)
	Single-turn conversation	Derm-Private	0.098	(0.046, 0.15)
	Single-turn conversation	Derm-Public	0.058	(0.02, 0.096)

	Summarized conversation	MedQA-USMLE	0.089	(0.05, 0.127)
	Summarized conversation	Derm-Private	0.13	(0.077, 0.183)
	Summarized conversation	Derm-Public	0.05	(0.019, 0.081)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.244	(0.173, 0.316)
	Vignette	Derm-Private	0.212	(0.136, 0.288)
	Vignette	Derm-Public	0.056	(0.017, 0.095)
	Multi-turn conversation	MedQA-USMLE	0.111	(0.069, 0.154)
	Multi-turn conversation	Derm-Private	0.108	(0.06, 0.156)
	Multi-turn conversation	Derm-Public	0.028	(0.006, 0.05)
	Single-turn conversation	MedQA-USMLE	0.104	(0.056, 0.153)
	Single-turn conversation	Derm-Private	0.078	(0.028, 0.128)
	Single-turn conversation	Derm-Public	0.036	(0.005, 0.067)
	Summarized conversation	MedQA-USMLE	0.14	(0.09, 0.19)
	Summarized conversation	Derm-Private	0.11	(0.061, 0.159)
	Summarized conversation	Derm-Public	0.048	(0.014, 0.082)

Supplementary Table 12: Mean accuracy and 95% confidence intervals for Dermatology, calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0002	0.0004
		Multi-turn conversation	Summarized conversation	0.0122	0.0201
		Single-turn conversation	Summarized conversation	0.0001	0.0002
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0001	0.0002
		Multi-turn conversation	Summarized conversation	0.0382	0.0598
		Single-turn conversation	Summarized conversation	0.0001	0.0002
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0001	0.0002
		Multi-turn conversation	Summarized conversation	0.4567	0.5391
		Single-turn conversation	Summarized conversation	0.0001	0.0002
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0018	0.0032
		Multi-turn conversation	Summarized conversation	0.7346	0.7665
		Single-turn conversation	Summarized conversation	0.004	0.007
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0143	0.0229
		Multi-turn conversation	Summarized conversation	0.0615	0.0942
		Single-turn conversation	Summarized conversation	0.0005	0.001
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002

		Multi-turn conversation	Single-turn conversation	0.0013	0.0025
		Multi-turn conversation	Summarized conversation	0.0042	0.0072
		Single-turn conversation	Summarized conversation	0.0001	0.0002
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.8394	0.8634
		Multi-turn conversation	Summarized conversation	0.4993	0.5706
		Single-turn conversation	Summarized conversation	0.4866	0.5651
	Derm-Public	Vignette	Multi-turn conversation	0.2204	0.2811
		Vignette	Single-turn conversation	0.2225	0.2811
		Vignette	Summarized conversation	0.1207	0.1704
		Multi-turn conversation	Single-turn conversation	0.869	0.8812
		Multi-turn conversation	Summarized conversation	0.3195	0.3899
		Single-turn conversation	Summarized conversation	0.6318	0.6789
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0123	0.0201
		Multi-turn conversation	Summarized conversation	0.5561	0.6256
		Single-turn conversation	Summarized conversation	0.1073	0.1577
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0003	0.0006
		Multi-turn conversation	Single-turn conversation	0.7042	0.7456
		Multi-turn conversation	Summarized conversation	0.0998	0.1497
		Single-turn conversation	Summarized conversation	0.1191	0.1704
	Derm-Public	Vignette	Multi-turn conversation	0.1541	0.2061
		Vignette	Single-turn conversation	0.2914	0.3617
		Vignette	Summarized conversation	0.6131	0.6688
		Multi-turn conversation	Single-turn conversation	0.5778	0.64
		Multi-turn conversation	Summarized conversation	0.1383	0.1915
		Single-turn conversation	Summarized conversation	0.4087	0.4904
	Derm-Private	Vignette	Multi-turn conversation	0.0014	0.0026
		Vignette	Single-turn conversation	0.0002	0.0004

		Vignette	Summarized conversation	0.0014	0.0026
		Multi-turn conversation	Single-turn conversation	0.1825	0.2389
		Multi-turn conversation	Summarized conversation	0.9422	0.9422
		Single-turn conversation	Summarized conversation	0.1546	0.2061

Supplementary Table 13: Adjusted p-values for Dermatology, calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.867	(0.796, 0.937)
	Vignette	Derm-Private	0.861	(0.792, 0.929)
	Vignette	Derm-Public	0.919	(0.859, 0.979)
	Multi-turn conversation	MedQA-USMLE	0.71	(0.628, 0.791)
	Multi-turn conversation	Derm-Private	0.8	(0.73, 0.87)
	Multi-turn conversation	Derm-Public	0.797	(0.72, 0.875)
	Single-turn conversation	MedQA-USMLE	0.629	(0.54, 0.717)
	Single-turn conversation	Derm-Private	0.675	(0.592, 0.758)
	Single-turn conversation	Derm-Public	0.732	(0.642, 0.821)
	Summarized conversation	MedQA-USMLE	0.755	(0.68, 0.83)
	Summarized conversation	Derm-Private	0.791	(0.721, 0.861)
	Summarized conversation	Derm-Public	0.82	(0.749, 0.891)
GPT-3.5	Vignette	MedQA-USMLE	0.712	(0.624, 0.799)
	Vignette	Derm-Private	0.758	(0.677, 0.838)
	Vignette	Derm-Public	0.83	(0.751, 0.91)
	Multi-turn conversation	MedQA-USMLE	0.493	(0.402, 0.584)
	Multi-turn conversation	Derm-Private	0.594	(0.51, 0.677)
	Multi-turn conversation	Derm-Public	0.618	(0.527, 0.708)
	Single-turn conversation	MedQA-USMLE	0.517	(0.42, 0.613)
	Single-turn conversation	Derm-Private	0.499	(0.411, 0.587)
	Single-turn conversation	Derm-Public	0.587	(0.49, 0.685)
	Summarized conversation	MedQA-USMLE	0.519	(0.425, 0.613)
	Summarized conversation	Derm-Private	0.608	(0.528, 0.689)
	Summarized conversation	Derm-Public	0.653	(0.566, 0.739)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.655	(0.551, 0.759)
	Vignette	Derm-Private	0.768	(0.683, 0.852)
	Vignette	Derm-Public	0.785	(0.692, 0.877)
	Multi-turn conversation	MedQA-USMLE	0.486	(0.397, 0.575)
	Multi-turn conversation	Derm-Private	0.517	(0.436, 0.598)
	Multi-turn conversation	Derm-Public	0.605	(0.517, 0.693)
	Single-turn conversation	MedQA-USMLE	0.536	(0.44, 0.631)
	Single-turn conversation	Derm-Private	0.511	(0.419, 0.603)
	Single-turn conversation	Derm-Public	0.61	(0.515, 0.705)

	Summarized conversation	MedQA-USMLE	0.555	(0.472, 0.637)
	Summarized conversation	Derm-Private	0.543	(0.465, 0.622)
	Summarized conversation	Derm-Public	0.646	(0.56, 0.731)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.388	(0.291, 0.485)
	Vignette	Derm-Private	0.434	(0.343, 0.526)
	Vignette	Derm-Public	0.433	(0.33, 0.536)
	Multi-turn conversation	MedQA-USMLE	0.379	(0.297, 0.46)
	Multi-turn conversation	Derm-Private	0.269	(0.195, 0.342)
	Multi-turn conversation	Derm-Public	0.385	(0.294, 0.475)
	Single-turn conversation	MedQA-USMLE	0.374	(0.276, 0.471)
	Single-turn conversation	Derm-Private	0.236	(0.158, 0.315)
	Single-turn conversation	Derm-Public	0.372	(0.272, 0.473)
	Summarized conversation	MedQA-USMLE	0.414	(0.324, 0.505)
	Summarized conversation	Derm-Private	0.305	(0.231, 0.379)
	Summarized conversation	Derm-Public	0.397	(0.307, 0.488)

Supplementary Table 14: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for 4-choice MCQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0003	0.001
		Multi-turn conversation	Single-turn conversation	0.0234	0.0443
		Multi-turn conversation	Summarized conversation	0.0181	0.0372
		Single-turn conversation	Summarized conversation	0.0003	0.001
	Derm-Public	Vignette	Multi-turn conversation	0.0025	0.0062
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0132	0.028
		Multi-turn conversation	Single-turn conversation	0.0219	0.0426
		Multi-turn conversation	Summarized conversation	0.1866	0.2742
		Single-turn conversation	Summarized conversation	0.0015	0.004
	Derm-Private	Vignette	Multi-turn conversation	0.0065	0.0142
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0033	0.0079
		Multi-turn conversation	Single-turn conversation	0.0002	0.0007
		Multi-turn conversation	Summarized conversation	0.5506	0.6293
		Single-turn conversation	Summarized conversation	0.0011	0.0032
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0013	0.0036
		Vignette	Single-turn conversation	0.0004	0.0013
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.7558	0.8003
		Multi-turn conversation	Summarized conversation	0.9599	0.9599
		Single-turn conversation	Summarized conversation	0.8409	0.8649
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0002	0.0007
		Multi-turn conversation	Single-turn conversation	0.3121	0.4086
		Multi-turn conversation	Summarized conversation	0.1919	0.2763
		Single-turn conversation	Summarized conversation	0.067	0.1149
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004

		Multi-turn conversation	Single-turn conversation	0.0047	0.0106
		Multi-turn conversation	Summarized conversation	0.5018	0.5923
		Single-turn conversation	Summarized conversation	0.0023	0.0059
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.001	0.003
		Vignette	Single-turn conversation	0.0736	0.1232
		Vignette	Summarized conversation	0.307	0.4086
		Multi-turn conversation	Single-turn conversation	0.0584	0.1051
		Multi-turn conversation	Summarized conversation	0.0001	0.0004
		Single-turn conversation	Summarized conversation	0.1977	0.2791
	Derm-Public	Vignette	Multi-turn conversation	0.0002	0.0007
		Vignette	Single-turn conversation	0.004	0.0093
		Vignette	Summarized conversation	0.0189	0.0378
		Multi-turn conversation	Single-turn conversation	0.8129	0.8482
		Multi-turn conversation	Summarized conversation	0.0602	0.1057
		Single-turn conversation	Summarized conversation	0.1336	0.2186
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004
		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.6331	0.7122
		Multi-turn conversation	Summarized conversation	0.3858	0.4873
		Single-turn conversation	Summarized conversation	0.1858	0.2742
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.9505	0.9599
		Vignette	Single-turn conversation	0.7234	0.7774
		Vignette	Summarized conversation	0.4754	0.5705
		Multi-turn conversation	Single-turn conversation	0.6705	0.7427
		Multi-turn conversation	Summarized conversation	0.4286	0.523
		Single-turn conversation	Summarized conversation	0.2417	0.3347
	Derm-Public	Vignette	Multi-turn conversation	0.4243	0.523
		Vignette	Single-turn conversation	0.1789	0.2741
		Vignette	Summarized conversation	0.6821	0.7441
		Multi-turn conversation	Single-turn conversation	0.3237	0.4162
		Multi-turn conversation	Summarized conversation	0.5145	0.5975
		Single-turn conversation	Summarized conversation	0.155	0.248
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0004
		Vignette	Single-turn conversation	0.0001	0.0004

		Vignette	Summarized conversation	0.0001	0.0004
		Multi-turn conversation	Single-turn conversation	0.2841	0.3859
		Multi-turn conversation	Summarized conversation	0.1611	0.2522
		Single-turn conversation	Summarized conversation	0.0381	0.0703

Supplementary Table 15: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the 4-choice MCQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Model	Experiment	Dataset	Mean Accuracy	95% C.I.
GPT-4	Vignette	MedQA-USMLE	0.65	(0.555, 0.745)
	Vignette	Derm-Private	0.826	(0.761, 0.892)
	Vignette	Derm-Public	0.641	(0.546, 0.735)
	Multi-turn conversation	MedQA-USMLE	0.412	(0.327, 0.496)
	Multi-turn conversation	Derm-Private	0.578	(0.499, 0.657)
	Multi-turn conversation	Derm-Public	0.319	(0.233, 0.405)
	Single-turn conversation	MedQA-USMLE	0.29	(0.204, 0.377)
	Single-turn conversation	Derm-Private	0.259	(0.182, 0.335)
	Single-turn conversation	Derm-Public	0.195	(0.119, 0.271)
	Summarized conversation	MedQA-USMLE	0.45	(0.36, 0.54)
	Summarized conversation	Derm-Private	0.591	(0.513, 0.67)
	Summarized conversation	Derm-Public	0.375	(0.287, 0.462)
GPT-3.5	Vignette	MedQA-USMLE	0.579	(0.478, 0.68)
	Vignette	Derm-Private	0.614	(0.53, 0.698)
	Vignette	Derm-Public	0.451	(0.349, 0.552)
	Multi-turn conversation	MedQA-USMLE	0.25	(0.168, 0.332)
	Multi-turn conversation	Derm-Private	0.343	(0.266, 0.421)
	Multi-turn conversation	Derm-Public	0.165	(0.099, 0.23)
	Single-turn conversation	MedQA-USMLE	0.193	(0.124, 0.262)
	Single-turn conversation	Derm-Private	0.238	(0.161, 0.316)
	Single-turn conversation	Derm-Public	0.124	(0.064, 0.184)
	Summarized conversation	MedQA-USMLE	0.259	(0.181, 0.336)
	Summarized conversation	Derm-Private	0.406	(0.324, 0.488)
	Summarized conversation	Derm-Public	0.197	(0.13, 0.263)
Mistral-v2-7b	Vignette	MedQA-USMLE	0.307	(0.211, 0.403)
	Vignette	Derm-Private	0.299	(0.209, 0.389)
	Vignette	Derm-Public	0.122	(0.051, 0.192)
	Multi-turn conversation	MedQA-USMLE	0.124	(0.067, 0.181)
	Multi-turn conversation	Derm-Private	0.141	(0.085, 0.198)
	Multi-turn conversation	Derm-Public	0.071	(0.032, 0.11)
	Single-turn conversation	MedQA-USMLE	0.126	(0.063, 0.19)
	Single-turn conversation	Derm-Private	0.099	(0.047, 0.151)
	Single-turn conversation	Derm-Public	0.068	(0.021, 0.115)

	Summarized conversation	MedQA-USMLE	0.107	(0.059, 0.155)
	Summarized conversation	Derm-Private	0.131	(0.078, 0.185)
	Summarized conversation	Derm-Public	0.063	(0.025, 0.102)
LLaMA-2-7b	Vignette	MedQA-USMLE	0.271	(0.186, 0.357)
	Vignette	Derm-Private	0.214	(0.138, 0.291)
	Vignette	Derm-Public	0.071	(0.022, 0.12)
	Multi-turn conversation	MedQA-USMLE	0.126	(0.072, 0.18)
	Multi-turn conversation	Derm-Private	0.109	(0.061, 0.157)
	Multi-turn conversation	Derm-Public	0.03	(0.004, 0.056)
	Single-turn conversation	MedQA-USMLE	0.129	(0.064, 0.193)
	Single-turn conversation	Derm-Private	0.079	(0.028, 0.129)
	Single-turn conversation	Derm-Public	0.043	(0.005, 0.081)
	Summarized conversation	MedQA-USMLE	0.162	(0.101, 0.223)
	Summarized conversation	Derm-Private	0.111	(0.061, 0.161)
	Summarized conversation	Derm-Public	0.051	(0.01, 0.091)

Supplementary Table 16: Mean accuracy and 95% confidence intervals for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for FRQ setting, reported for all evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) and models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b).

Model	Dataset	Experiment 1	Experiment 2	p-value	Adjusted p-value
GPT-4	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0004	0.0009
		Multi-turn conversation	Summarized conversation	0.1376	0.1981
		Single-turn conversation	Summarized conversation	0.0003	0.0007
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0005	0.0011
		Multi-turn conversation	Summarized conversation	0.0405	0.0663
		Single-turn conversation	Summarized conversation	0.0002	0.0005
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0001	0.0003
		Multi-turn conversation	Summarized conversation	0.4547	0.528
		Single-turn conversation	Summarized conversation	0.0001	0.0003
GPT-3.5	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.0222	0.0372
		Multi-turn conversation	Summarized conversation	0.7198	0.7735
		Single-turn conversation	Summarized conversation	0.0171	0.0293
	Derm-Public	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.048	0.0768
		Multi-turn conversation	Summarized conversation	0.1802	0.2317
		Single-turn conversation	Summarized conversation	0.0038	0.0068
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003

		Multi-turn conversation	Single-turn conversation	0.0007	0.0014
		Multi-turn conversation	Summarized conversation	0.0038	0.0068
		Single-turn conversation	Summarized conversation	0.0001	0.0003
Mistral-v2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0001	0.0003
		Multi-turn conversation	Single-turn conversation	0.9125	0.9254
		Multi-turn conversation	Summarized conversation	0.3342	0.401
		Single-turn conversation	Summarized conversation	0.4193	0.4949
	Derm-Public	Vignette	Multi-turn conversation	0.1719	0.2292
		Vignette	Single-turn conversation	0.1785	0.2317
		Vignette	Summarized conversation	0.1227	0.1803
		Multi-turn conversation	Single-turn conversation	0.8608	0.8982
		Multi-turn conversation	Summarized conversation	0.5251	0.5907
		Single-turn conversation	Summarized conversation	0.8178	0.8659
	Derm-Private	Vignette	Multi-turn conversation	0.0001	0.0003
		Vignette	Single-turn conversation	0.0001	0.0003
		Vignette	Summarized conversation	0.0002	0.0005
		Multi-turn conversation	Single-turn conversation	0.0117	0.0205
		Multi-turn conversation	Summarized conversation	0.5379	0.5958
		Single-turn conversation	Summarized conversation	0.1004	0.1506
LLaMA-2-7b	MedQA-USMLE	Vignette	Multi-turn conversation	0.0003	0.0007
		Vignette	Single-turn conversation	0.0004	0.0009
		Vignette	Summarized conversation	0.0011	0.0021
		Multi-turn conversation	Single-turn conversation	0.9431	0.9431
		Multi-turn conversation	Summarized conversation	0.0912	0.1397
		Single-turn conversation	Summarized conversation	0.2334	0.2948
	Derm-Public	Vignette	Multi-turn conversation	0.0866	0.1355
		Vignette	Single-turn conversation	0.272	0.3377
		Vignette	Summarized conversation	0.3328	0.401
		Multi-turn conversation	Single-turn conversation	0.4691	0.5361
		Multi-turn conversation	Summarized conversation	0.1521	0.2106
		Single-turn conversation	Summarized conversation	0.6356	0.6934
	Derm-Private	Vignette	Multi-turn conversation	0.0015	0.0028
		Vignette	Single-turn conversation	0.0004	0.0009

		Vignette	Summarized conversation	0.0011	0.0021
		Multi-turn conversation	Single-turn conversation	0.1652	0.2244
		Multi-turn conversation	Summarized conversation	0.9063	0.9254
		Single-turn conversation	Summarized conversation	0.1459	0.206

Supplementary Table 17: Adjusted p-values for Dermatology (single most likely diagnosis case vignettes), calculated by dataset source for the FRQ setting, reported for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to each of the models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

Question	Model 1	Model 2	P-value (McNemar's test)
Did the clinical LLM stop asking questions when only a single most likely diagnosis was possible?	GPT-4	GPT-3.5	0.0129
	GPT-4	Mistral-v2-7b	<0.0001
	GPT-4	LLaMA-2-7b	0.0573
	GPT-3.5	Mistral-v2-7b	0.0039
	GPT-3.5	LLaMA-2-7b	0.7744
	Mistral-v2-7b	LLaMA-2-7b	0.0009
Did the clinical LLM elicit the relevant medical history?	GPT-4	GPT-3.5	0.0001
	GPT-4	Mistral-v2-7b	<0.0001
	GPT-4	LLaMA-2-7b	0.0635
	GPT-3.5	Mistral-v2-7b	0.0212
	GPT-3.5	LLaMA-2-7b	0.0635
	Mistral-v2-7b	LLaMA-2-7b	<0.0001

Supplementary Table 18: P-values between pairs of evaluated models (GPT-4, GPT-3.5, Mistral-v2-7b, LLaMA-2-7b) for clinical LLM assessment by medical experts. All p-values were calculated using McNemar's test (see Methods).

	Model	Experiment	Mean Accuracy	95% C.I.
4-choice MCQ	GPT-4V	Vignette	0.842	(0.763, 0.921)
		Multi-turn conversation	0.492	(0.397, 0.587)
		Single-turn conversation	0.474	(0.369, 0.578)
		Summarized conversation	0.547	(0.456, 0.639)
	GPT-4V-without-image	Vignette	0.787	(0.695, 0.878)
		Multi-turn conversation	0.468	(0.37, 0.567)
		Single-turn conversation	0.4	(0.297, 0.503)
		Summarized conversation	0.503	(0.405, 0.601)
FRQ	GPT-4V	Vignette	0.492	(0.398, 0.586)
		Multi-turn conversation	0.145	(0.084, 0.205)
		Single-turn conversation	0.063	(0.021, 0.106)
		Summarized conversation	0.163	(0.1, 0.226)
	GPT-4V-without-image	Vignette	0.471	(0.367, 0.575)
		Multi-turn conversation	0.087	(0.037, 0.137)
		Single-turn conversation	0.039	(0.01, 0.069)
		Summarized conversation	0.108	(0.052, 0.164)

Supplementary Table 19: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across the evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for GPT-4V and GPT-4V-without-image.

	Model	Experiment 1	Experiment 2	p-values	Adjusted p-value
4-choice MCQ	GPT-4V	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.6586	0.6586
		Multi-turn conversation	Summarized conversation	0.1105	0.1357
		Single-turn conversation	Summarized conversation	0.1131	0.1357
	GPT-4V-without-image	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.1097	0.1357
		Multi-turn conversation	Summarized conversation	0.2719	0.2966
		Single-turn conversation	Summarized conversation	0.019	0.0326
FRQ	GPT-4V	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0002	0.0003
		Multi-turn conversation	Summarized conversation	0.3699	0.3699
		Single-turn conversation	Summarized conversation	0.0007	0.0011
	GPT-4V-without-image	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn	0.0304	0.0365

			conversation		
		Multi-turn conversation	Summarized conversation	0.3625	0.3699
		Single-turn conversation	Summarized conversation	0.0052	0.0069

Supplementary Table 20: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to GPT-4V and GPT-4V-without-image.

	Experiment	**Model 1**	**Model 2**	**p-value**	**Adjusted p-value**
4-choice MCQ	Vignette	GPT-4V	GPT-4V-without-image	0.0631	0.2064
	Multi-turn conversation	GPT-4V	GPT-4V-without-image	0.5356	0.5356
	Single-turn conversation	GPT-4V	GPT-4V-without-image	0.1032	0.2064
	Summarized conversation	GPT-4V	GPT-4V-without-image	0.1885	0.2513
FRQ	Vignette	GPT-4V	GPT-4V-without-image	0.5309	0.5309
	Multi-turn conversation	GPT-4V	GPT-4V-without-image	0.0148	0.0506
	Single-turn conversation	GPT-4V	GPT-4V-without-image	0.1312	0.1749
	Summarized conversation	GPT-4V	GPT-4V-without-image	0.0253	0.0506

Supplementary Table 21: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated models (GPT-4V and GPT-4V-without-image) corresponding to each of the experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation). All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).

	Model	Experiment	Accuracy	95% C.I.
4-choice MCQ	Mistral-v1-7b	Vignette	0.441	(0.419, 0.462)
		Multi-turn conversation	0.331	(0.314, 0.349)
		Single-turn conversation	0.324	(0.305, 0.344)
		Summarized conversation	0.361	(0.343, 0.379)
	Mistral-v2-7b	Vignette	0.637	(0.616, 0.658)
		Multi-turn conversation	0.426	(0.409, 0.443)
		Single-turn conversation	0.448	(0.429, 0.468)
		Summarized conversation	0.513	(0.496, 0.529)
FRQ	Mistral-v1-7b	Vignette	0.165	(0.142, 0.189)
		Multi-turn conversation	0.08	(0.065, 0.095)
		Single-turn conversation	0.06	(0.046, 0.074)
		Summarized conversation	0.082	(0.068, 0.097)
	Mistral-v2-7b	Vignette	0.211	(0.186, 0.237)
		Multi-turn conversation	0.065	(0.052, 0.077)
		Single-turn conversation	0.055	(0.043, 0.068)
		Summarized conversation	0.055	(0.044, 0.066)

Supplementary Table 22: Mean accuracy and 95% confidence intervals for 4-choice MCQ and FRQ setting, across experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) for Mistral-v1-7b and Mistral-v2-7b.

	Model	Experiment 1	Experiment 2	p-values	Adjusted p-value
4-choice MCQ	Mistral-v1-7b	Vignette	Multi-turn conversation	0.0001	0.0001
		Vignette	Single-turn conversation	0.0001	0.0001
		Vignette	Summarized conversation	0.0001	0.0001
		Multi-turn conversation	Single-turn conversation	0.2975	0.2975
		Multi-turn conversation	Summarized conversation	0.0001	0.0001
		Single-turn conversation	Summarized conversation	0.0001	0.0001
FRQ	Mistral-v1-7b	Vignette	Multi-turn conversation	0.0001	0.0002
		Vignette	Single-turn conversation	0.0001	0.0002
		Vignette	Summarized conversation	0.0001	0.0002
		Multi-turn conversation	Single-turn conversation	0.0003	0.0004
		Multi-turn conversation	Summarized conversation	0.5212	0.5686
		Single-turn conversation	Summarized conversation	0.0001	0.0002

Supplementary Table 23: Adjusted p-values for 4-choice MCQ and FRQ setting for pairs of evaluated experimental setups (vignette, multi-turn conversation, single-turn conversation and summarized conversation) corresponding to Mistral-v1-7b. All p-values were calculated using a two-sided bootstrapping test, followed by Holm-Bonferroni correction (see Methods).