Supplementary Material 3. Statistics

Descriptive statistics

Evaluator 1

		Frequency	Percent	Valid Percent	Cumulative Percent
Score	1	7	2.2	2.2	2.2
	2	47	14.5	14.5	16.7
	3	105	32.4	32.4	49.1
	4	165	50.9	50.9	100.0
	Total	324	100.0	100.0	

All three categories (n=324), Likert Scale 1 to 4

Evaluator 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Score	1	1	.3	.3	.3
	2	25	7.7	7.7	8.0
	3	156	48.1	48.1	56.2
	4	142	43.8	43.8	100.0
	Total	324	100.0	100.0	

All three categories (n=324), Likert Scale 1 to 4

Evaluator 3

		Frequency	Percent	Valid Percent	Cumulative Percent
Score	1	0	0	0	0
	2	31	9.6	9.6	9.6
	3	156	48.1	48.1	57.7
	4	137	42.3	42.3	100.0
	Total	324	100.0	100.0	

 https://doi.org/10.12701/jyms.2024.01151

Statistics

		Evaluator 1	Evaluator 2	Evaluator 3
N	Valid	324	324	324
	Missing	0	0	0
Mean		3.32	3.35	3.33
Median		4.00	3.00	3.00
Mode		4	3	3
Std. Deviation		.800	.635	.642
Variance		.640	.403	.413
Range		3	3	2
Minimum		1	1	2
Maximum		4	4	4

Frequencies compatibility

Compatibility score Evaluators 1-2-3

		Frequency	Percent	Percent	Cumulative Percent
Evaluations	Different score	254	78.4	78.4	78.4
	Identical score	70	21.6	21.6	100.0
	Total	324	100.0	100.0	

Comptability score Evaluator 1 vs 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Different score	197	60.8	60.8	60.8
	Identical score	127	39.2	39.2	100.0
	Total	324	100.0	100.0	

Comptability score Evaluator 1 vs 3

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Different score	169	52.2	52.2	52.2
	Identical score	155	47.8	47.8	100.0
	Total	324	100.0	100.0	

 https://doi.org/10.12701/jyms.2024.01151

Comptability score Evaluator 2 vs 3

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Different score	180	55.6	55.6	55.6
	Identical score	144	44.4	44.4	100.0
	Total	324	100.0	100.0	

Compatibility Scores of 3 or 4 Evaluators 1-2-3

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Mixed scores (1-2 vs 3-4)	93	28.7	28.7	28.7
	Consistent scores (3-4)	231	71.3	71.3	100.0
	Total	324	100.0	100.0	

Compatibility Scores of 3 or 4 Evaluator 1 vs 2

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Mixed scores (1-2 vs 3-4)	77	23.8	23.8	23.8
	Consistent scores (3-4)	247	76.2	76.2	100.0
	Total	324	100.0	100.0	

Compatibility Scores of 3 or 4 of Evaluator 1 vs 3

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Mixed scores (1-2 vs 3-4)	72	22.2	22.2	22.2
	Consistent scores (3-4)	252	77.8	77.8	100.0
	Total	324	100.0	100.0	

Compatibility Scores 3 or 4 of Evaluator 2 vs 3

		Frequency	Percent	Valid Percent	Cumulative Percent
Evaluations	Mixed scores (1-2 vs 3-4)	53	16.4	16.4	16.4
	Consistent scores (3-4)	271	83.6	83.6	100.0
	Total	324	100.0	100.0	

 https://doi.org/10.12701/jyms.2024.01151

Evaluator 1 vs Evaluator 2 Crosstabulation

Count

		Evaluator 2				
		1	2	3	4	Total
Evaluator 1	1	0	0	0	7	7
	2	1	2	23	21	47
	3	0	16	50	39	105
	4	0	7	83	75	165
Total		1	25	156	142	324

Cohen's Kappa Evaluator 1 vs 2

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.002	.040	.059	.953
N of Valid Cases		324			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Evaluator 1 vs Evaluator 3 Crosstabulation

Count

		Evaluator 3				
		1	2	3	4	Total
Evaluator 1	1	0	3	2	2	7
	2	0	10	26	11	47
	3	0	12	57	36	105
	4	0	6	71	88	165
Total		0	31	156	137	324

Cohen's Kappa Evaluator 1 vs 3

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.152	.042	3.713	<.001
N of Valid Cases		324			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

 https://doi.org/10.12701/jyms.2024.01151

Evaluator 2 vs Evaluator 3 Crosstabulation

Count

		Evaluator3				
		1	2	3	4	Total
Evaluator 2	1	0	1	0	0	1
	2	0	3	11	11	25
	3	0	13	79	64	156
	4	0	14	66	62	142
Total		0	31	156	137	324

Cohen's Kappa Evaluator 2 vs 3

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.035	.046	.763	.445
N of Valid Cases		324			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Krippendof's alpha scores

Colonne1	Krippendorf's alpha score	Lower bound	Upper bound
Data	Interval	LL95%CI	UL95%CI
Validity	0.0715	–0.788	0.2085
Safety	0.0149	–0.1745	0.1749
Usefulness	0.0539	–0.0988	0.2008
Total	0.1143	0.0272	0.1971

Multivariate ANOVA test Mean scores Evaluator 1-2-3

Safety score

Descriptive Statistics

	Mean	Std. Deviation	N
ChatGPT Without Clinical Context	3.5926	.36201	27
ChatGPT With Clinical Context	3.6173	.35450	27
Copilot Without Clinical Context	3.5309	.36114	27
Copilot With Clinical Context	3.6173	.35450	27

https://doi.org/10.12701/jyms.2024.01151

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.012	.308[b]	1.000	26.000	.583	.012
	Wilks' Lambda	.988	.308[b]	1.000	26.000	.583	.012
	Hotelling's Trace	.012	.308[b]	1.000	26.000	.583	.012
	Roy's Largest Root	.012	.308[b]	1.000	26.000	.583	.012
Context	Pillai's Trace	.040	1.083[b]	1.000	26.000	.308	.040
	Wilks' Lambda	.960	1.083[b]	1.000	26.000	.308	.040
	Hotelling's Trace	.042	1.083[b]	1.000	26.000	.308	.040
	Roy's Largest Root	.042	1.083[b]	1.000	26.000	.308	.040
LLM * Context	Pillai's Trace	.020	.523[b]	1.000	26.000	.476	.020
	Wilks' Lambda	.980	.523[b]	1.000	26.000	.476	.020
	Hotelling's Trace	.020	.523[b]	1.000	26.000	.476	.020
	Roy's Largest Root	.020	.523[b]	1.000	26.000	.476	.020

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

Validity score

Descriptive Statistics

	Mean	Std. Deviation	N
ChatGPT Without Clinical Context	3.3457	.39744	27
ChatGPT With Clinical Context	3.3086	.36894	27
Copilot Without Clinical Context	3.1111	.48920	27
Copilot With Clinical Context	3.2469	.39864	27

 https://doi.org/10.12701/jyms.2024.01151

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.085	2.426[b]	1.000	26.000	.131	.085
	Wilks' Lambda	.915	2.426[b]	1.000	26.000	.131	.085
	Hotelling's Trace	.093	2.426[b]	1.000	26.000	.131	.085
	Roy's Largest Root	.093	2.426[b]	1.000	26.000	.131	.085
Context	Pillai's Trace	.025	.673[b]	1.000	26.000	.420	.025
	Wilks' Lambda	.975	.673[b]	1.000	26.000	.420	.025
	Hotelling's Trace	.026	.673[b]	1.000	26.000	.420	.025
	Roy's Largest Root	.026	.673[b]	1.000	26.000	.420	.025
LLM * Context	Pillai's Trace	.125	3.720[b]	1.000	26.000	.065	.125
	Wilks' Lambda	.875	3.720[b]	1.000	26.000	.065	.125
	Hotelling's Trace	.143	3.720[b]	1.000	26.000	.065	.125
	Roy's Largest Root	.143	3.720[b]	1.000	26.000	.065	.125

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

Usefulness scores

Descriptive Statistics

	Mean	Std. Deviation	N
ChatGPT Without Clinical Context	3.2099	.42515	27
With Clinical With Clinical Context	3.1728	.44693	27
Copilot Without Clinical Context	3.0988	.46972	27
Copilot With Clinical Context	3.1605	.43726	27

 https://doi.org/10.12701/jyms.2024.01151

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.017	.445[b]	1.000	26.000	.510	.017
	Wilks' Lambda	.983	.445[b]	1.000	26.000	.510	.017
	Hotelling's Trace	.017	.445[b]	1.000	26.000	.510	.017
	Roy's Largest Root	.017	.445[b]	1.000	26.000	.510	.017
Context	Pillai's Trace	.001	.031[b]	1.000	26.000	.861	.001
	Wilks' Lambda	.999	.031[b]	1.000	26.000	.861	.001
	Hotelling's Trace	.001	.031[b]	1.000	26.000	.861	.001
	Roy's Largest Root	.001	.031[b]	1.000	26.000	.861	.001
LLM * Context	Pillai's Trace	.030	.794[b]	1.000	26.000	.381	.030
	Wilks' Lambda	.970	.794[b]	1.000	26.000	.381	.030
	Hotelling's Trace	.031	.794[b]	1.000	26.000	.381	.030
	Roy's Largest Root	.031	.794[b]	1.000	26.000	.381	.030

a. Design: Intercept
Within Subjects Design: LLM + Context+ LLM * Context

b. Exact statistic

Individual multivariate ANOVA tests

Evaluator 1

Safety score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.005	.129[b]	1.000	26.000	.722	.005
	Wilks' Lambda	.995	.129[b]	1.000	26.000	.722	.005
	Hotelling's Trace	.005	.129[b]	1.000	26.000	.722	.005
	Roy's Largest Root	.005	.129[b]	1.000	26.000	.722	.005
Context	Pillai's Trace	.027	.720[b]	1.000	26.000	.404	.027
	Wilks' Lambda	.973	.720[b]	1.000	26.000	.404	.027
	Hotelling's Trace	.028	.720[b]	1.000	26.000	.404	.027
	Roy's Largest Root	.028	.720[b]	1.000	26.000	.404	.027
LLM * Context	Pillai's Trace	.042	1.149[b]	1.000	26.000	.294	.042
	Wilks' Lambda	.958	1.149[b]	1.000	26.000	.294	.042
	Hotelling's Trace	.044	1.149[b]	1.000	26.000	.294	.042
	Roy's Largest Root	.044	1.149[b]	1.000	26.000	.294	.042

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

 https://doi.org/10.12701/jyms.2024.01151

Validity score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.030	.815[b]	1.000	26.000	.375	.030
	Wilks' Lambda	.970	.815[b]	1.000	26.000	.375	.030
	Hotelling's Trace	.031	.815[b]	1.000	26.000	.375	.030
	Roy's Largest Root	.031	.815[b]	1.000	26.000	.375	.030
Context	Pillai's Trace	.029	.776[b]	1.000	26.000	.386	.029
	Wilks' Lambda	.971	.776[b]	1.000	26.000	.386	.029
	Hotelling's Trace	.030	.776[b]	1.000	26.000	.386	.029
	Roy's Largest Root	.030	.776[b]	1.000	26.000	.386	.029
LLM * Context	Pillai's Trace	.091	2.608[b]	1.000	26.000	.118	.091
	Wilks' Lambda	.909	2.608[b]	1.000	26.000	.118	.091
	Hotelling's Trace	.100	2.608[b]	1.000	26.000	.118	.091
	Roy's Largest Root	.100	2.608[b]	1.000	26.000	.118	.091

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

Usefulness score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.016	.419[b]	1.000	26.000	.523	.016
	Wilks' Lambda	.984	.419[b]	1.000	26.000	.523	.016
	Hotelling's Trace	.016	.419[b]	1.000	26.000	.523	.016
	Roy's Largest Root	.016	.419[b]	1.000	26.000	.523	.016
Context	Pillai's Trace	.006	.161[b]	1.000	26.000	.691	.006
	Wilks' Lambda	.994	.161[b]	1.000	26.000	.691	.006
	Hotelling's Trace	.006	.161[b]	1.000	26.000	.691	.006
	Roy's Largest Root	.006	.161[b]	1.000	26.000	.691	.006
LLM * Context	Pillai's Trace	.119	3.496[b]	1.000	26.000	.073	.119
	Wilks' Lambda	.881	3.496[b]	1.000	26.000	.073	.119
	Hotelling's Trace	.134	3.496[b]	1.000	26.000	.073	.119
	Roy's Largest Root	.134	3.496[b]	1.000	26.000	.073	.119

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

 https://doi.org/10.12701/jyms.2024.01151

Evaluator 2

Safety score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.005	.121[b]	1.000	26.000	.731	.005
	Wilks' Lambda	.995	.121[b]	1.000	26.000	.731	.005
	Hotelling's Trace	.005	.121[b]	1.000	26.000	.731	.005
	Roy's Largest Root	.005	.121[b]	1.000	26.000	.731	.005
Context	Pillai's Trace	.021	.562[b]	1.000	26.000	.460	.021
	Wilks' Lambda	.979	.562[b]	1.000	26.000	.460	.021
	Hotelling's Trace	.022	.562[b]	1.000	26.000	.460	.021
	Roy's Largest Root	.022	.562[b]	1.000	26.000	.460	.021
LLM * Context	Pillai's Trace	.009	.243[b]	1.000	26.000	.626	.009
	Wilks' Lambda	.991	.243[b]	1.000	26.000	.626	.009
	Hotelling's Trace	.009	.243[b]	1.000	26.000	.626	.009
	Roy's Largest Root	.009	.243[b]	1.000	26.000	.626	.009

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

Validity score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.003	.080[b]	1.000	26.000	.779	.003
	Wilks' Lambda	.997	.080[b]	1.000	26.000	.779	.003
	Hotelling's Trace	.003	.080[b]	1.000	26.000	.779	.003
	Roy's Largest Root	.003	.080[b]	1.000	26.000	.779	.003
Context	Pillai's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Wilks' Lambda	1.000	.000[b]	1.000	26.000	1.000	.000
	Hotelling's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Roy's Largest Root	.000	.000[b]	1.000	26.000	1.000	.000
LLM * Context	Pillai's Trace	.005	.121[b]	1.000	26.000	.731	.005
	Wilks' Lambda	.995	.121[b]	1.000	26.000	.731	.005
	Hotelling's Trace	.005	.121[b]	1.000	26.000	.731	.005
	Roy's Largest Root	.005	.121[b]	1.000	26.000	.731	.005

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

 https://doi.org/10.12701/jyms.2024.01151

Usefulness score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.005	.121[b]	1.000	26.000	.731	.005
	Wilks' Lambda	.995	.121[b]	1.000	26.000	.731	.005
	Hotelling's Trace	.005	.121[b]	1.000	26.000	.731	.005
	Roy's Largest Root	.005	.121[b]	1.000	26.000	.731	.005
Context	Pillai's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Wilks' Lambda	1.000	.000[b]	1.000	26.000	1.000	.000
	Hotelling's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Roy's Largest Root	.000	.000[b]	1.000	26.000	1.000	.000
LLM * Context	Pillai's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Wilks' Lambda	1.000	.000[b]	1.000	26.000	1.000	.000
	Hotelling's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Roy's Largest Root	.000	.000[b]	1.000	26.000	1.000	.000

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

Evaluator 3

Safety score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.040	1.091[b]	1.000	26.000	.306	.040
	Wilks' Lambda	.960	1.091[b]	1.000	26.000	.306	.040
	Hotelling's Trace	.042	1.091[b]	1.000	26.000	.306	.040
	Roy's Largest Root	.042	1.091[b]	1.000	26.000	.306	.040
Context	Pillai's Trace	.002	.064[b]	1.000	26.000	.802	.002
	Wilks' Lambda	.998	.064[b]	1.000	26.000	.802	.002
	Hotelling's Trace	.002	.064[b]	1.000	26.000	.802	.002
	Roy's Largest Root	.002	.064[b]	1.000	26.000	.802	.002
LLM * Context	Pillai's Trace	.003	.088[b]	1.000	26.000	.769	.003
	Wilks' Lambda	.997	.088[b]	1.000	26.000	.769	.003
	Hotelling's Trace	.003	.088[b]	1.000	26.000	.769	.003
	Roy's Largest Root	.003	.088[b]	1.000	26.000	.769	.003

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

 https://doi.org/10.12701/jyms.2024.01151

Validity score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.169	5.294[b]	1.000	26.000	.030	.169
	Wilks' Lambda	.831	5.294[b]	1.000	26.000	.030	.169
	Hotelling's Trace	.204	5.294[b]	1.000	26.000	.030	.169
	Roy's Largest Root	.204	5.294[b]	1.000	26.000	.030	.169
Context	Pillai's Trace	.003	.074[b]	1.000	26.000	.787	.003
	Wilks' Lambda	.997	.074[b]	1.000	26.000	.787	.003
	Hotelling's Trace	.003	.074[b]	1.000	26.000	.787	.003
	Roy's Largest Root	.003	.074[b]	1.000	26.000	.787	.003
LLM * Context	Pillai's Trace	.103	2.982[b]	1.000	26.000	.096	.103
	Wilks' Lambda	.897	2.982[b]	1.000	26.000	.096	.103
	Hotelling's Trace	.115	2.982[b]	1.000	26.000	.096	.103
	Roy's Largest Root	.115	2.982[b]	1.000	26.000	.096	.103

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

Usefulness score

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
LLM	Pillai's Trace	.042	1.130[b]	1.000	26.000	.297	.042
	Wilks' Lambda	.958	1.130[b]	1.000	26.000	.297	.042
	Hotelling's Trace	.043	1.130[b]	1.000	26.000	.297	.042
	Roy's Largest Root	.043	1.130[b]	1.000	26.000	.297	.042
Context	Pillai's Trace	.016	.435[b]	1.000	26.000	.515	.016
	Wilks' Lambda	.984	.435[b]	1.000	26.000	.515	.016
	Hotelling's Trace	.017	.435[b]	1.000	26.000	.515	.016
	Roy's Largest Root	.017	.435[b]	1.000	26.000	.515	.016
LLM * Context	Pillai's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Wilks' Lambda	1.000	.000[b]	1.000	26.000	1.000	.000
	Hotelling's Trace	.000	.000[b]	1.000	26.000	1.000	.000
	Roy's Largest Root	.000	.000[b]	1.000	26.000	1.000	.000

a. Design: Intercept
Within Subjects Design: LLM + Context + LLM * Context

b. Exact statistic

 https://doi.org/10.12701/jyms.2024.01151