Supplementary Table 1. Repeated-measures ANOVA results-all raters

Criterion	F	df	p-value	Partial η^2	Observed power
A/PR	12.84	4.144	<0.001	0.263	1.000
C	10.47	4.144	<0.001	0.225	1.000
SA/PU	11.89	4.144	<0.001	0.248	1.000

Evaluation criteria are expressed as follows: A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons), ANOVA: One way analysis of variance, df: Degrees of freedom

Supplementary Table 2. Repeated-measures ANOVA results-experts (E1-E3)

Criterion	F	df	p-value	Partial η^2	Observed power
A	5.16	4.144	0.001	0.125	0.98
C	2.42	4.144	0.051	0.062	0.62
SA	6.73	4.144	<0.001	0.157	0.99

Evaluation criteria are expressed as follows: A: Accuracy, C: Clarity, SA: Scientific adequacy, ANOVA: One way analysis of variance, df: Degrees of freedom

Supplementary Table 3. Repeated-measures ANOVA results-laypersons (L1-L3)

Criterion	F	df	p-value	Partial η^2	Observed power
PR	29.37	4.144	<0.001	0.449	1.000
C	18.56	4.144	<0.001	0.340	1.000
PU	24.48	4.144	<0.001	0.405	1.000

Evaluation criteria are expressed as follows: PR: Perceived reliability, C: Clarity, PU: Practical usefulness, ANOVA: One way analysis of variance, df: Degrees of freedom

Supplementary Table 4. Detailed Bonferroni-adjusted pairwise comparisons-all raters (criteria A/PR, C, SA/PU)

Criterion	Comparison	Mean diff.	95% CI	t	df	Cohen's d	Bonf. p
A/PR	Claude Sonnet 4 vs. DeepSeek v3.2	0.00	(-0.12, 0.13)	0.07	36	0.01	1.000
	Claude Sonnet 4 vs. GPT-4	-0.13	(-0.25, -0.01)	-2.19	36	-0.36	0.349
	Claude Sonnet 4 vs. Gemini 2.5	-0.31	(-0.42, -0.19)	-5.39	36	-0.89	**<0.001**
	Claude Sonnet 4 vs. Professor	0.14	(-0.03, 0.31)	1.69	36	0.28	1.000
	DeepSeek v3.2 vs. GPT-4	-0.14	(-0.26, -0.01)	-2.17	36	-0.36	0.365
	DeepSeek v3.2 vs. Gemini 2.5	-0.31	(-0.44, -0.18)	-4.83	36	-0.79	**<0.001**
	DeepSeek v3.2 vs. Professor	0.14	(-0.03, 0.30)	1.63	36	0.27	1.000
	GPT-4 vs. Gemini 2.5	-0.18	(-0.27, -0.08)	-3.60	36	-0.59	**0.009**
	GPT-4 vs. Professor	0.27	(0.15, 0.39)	4.45	36	0.73	**<0.001**
	Gemini 2.5 vs. Professor	0.45	(0.31, 0.59)	6.44	36	1.06	**<0.001**
C	Claude Sonnet 4 vs. DeepSeek v3.2	0.09	(-0.04, 0.21)	1.34	36	0.22	1.000
	Claude Sonnet 4 vs. GPT-4	-0.09	(-0.21, 0.03)	-1.49	36	-0.25	1.000
	Claude Sonnet 4 vs. Gemini 2.5	-0.14	(-0.26, -0.01)	-2.17	36	-0.36	0.365
	Claude Sonnet 4 vs. Professor	0.31	(0.13, 0.48)	3.59	36	0.59	**0.010**
	DeepSeek v3.2 vs. GPT-4	-0.18	(-0.29, -0.07)	-3.25	36	-0.53	**0.025**
	DeepSeek v3.2 vs. Gemini 2.5	-0.22	(-0.36, -0.09)	-3.32	36	-0.55	**0.021**
	DeepSeek v3.2 vs. Professor	0.22	(0.03, 0.41)	2.40	36	0.39	0.218
	GPT-4 vs. Gemini 2.5	-0.05	(-0.15, 0.06)	-0.84	36	-0.14	1.000
	GPT-4 vs. Professor	0.40	(0.24, 0.55)	5.16	36	0.85	**<0.001**
	Gemini 2.5 vs. Professor	0.44	(0.28, 0.61)	5.43	36	0.89	**<0.001**
SA/PU	Claude Sonnet 4 vs. DeepSeek v3.2	-0.08	(-0.20, 0.05)	-1.25	36	-0.20	1.000
	Claude Sonnet 4 vs. GPT-4	-0.12	(-0.26, 0.01)	-1.83	36	-0.30	0.752
	Claude Sonnet 4 vs. Gemini 2.5	-0.33	(-0.46, -0.20)	-5.18	36	-0.85	**<0.001**
	Claude Sonnet 4 vs. Professor	0.17	(-0.00, 0.34)	1.99	36	0.33	0.541
	DeepSeek v3.2 vs. GPT-4	-0.05	(-0.18, 0.09)	-0.67	36	-0.11	1.000
	DeepSeek v3.2 vs. Gemini 2.5	-0.25	(-0.41, -0.10)	-3.32	36	-0.55	**0.021**
	DeepSeek v3.2 vs. Professor	0.24	(0.05, 0.43)	2.60	36	0.43	0.136
	GPT-4 vs. Gemini 2.5	-0.21	(-0.33, -0.09)	-3.53	36	-0.58	**0.011**
	GPT-4 vs. Professor	0.29	(0.12, 0.46)	3.43	36	0.56	**0.015**
	Gemini 2.5 vs. Professor	0.50	(0.34, 0.65)	6.62	36	1.09	**<0.001**

Evaluation criteria are expressed as follows: A/PR: Accuracy (experts)/perceived reliability (laypersons), C: Clarity, SA/PU: Scientific adequacy (experts)/practical usefulness (laypersons), CI: Confidence interval, df: Degrees of freedom

Supplementary Table 5. Detailed Bonferroni-adjusted pairwise comparisons-experts only (criteria A, C, SA)

Criterion	Comparison	Mean diff.	95% CI	t	df	Cohen's d	Bonf. p
A	Claude Sonnet 4 vs. DeepSeek v3.2	-0.14	(-0.29, 0.02)	-1.81	36	-0.30	0.789
	Claude Sonnet 4 vs. GPT-4	-0.09	(-0.24, 0.06)	-1.20	36	-0.20	1.000
	Claude Sonnet 4 vs. Gemini 2.5	-0.28	(-0.44, -0.12)	-3.58	36	-0.59	**0.010**
	Claude Sonnet 4 vs. Professor	-0.27	(-0.45, -0.09)	-3.03	36	-0.50	**0.046**
	DeepSeek v3.2 vs. GPT-4	0.05	(-0.11, 0.20)	0.60	36	0.10	1.000
	DeepSeek v3.2 vs. Gemini 2.5	-0.14	(-0.28, -0.01)	-2.13	36	-0.35	0.404
	DeepSeek v3.2 vs. Professor	-0.14	(-0.30, 0.03)	-1.67	36	-0.27	1.000
	GPT-4 vs. Gemini 2.5	-0.19	(-0.33, -0.05)	-2.74	36	-0.45	0.095
	GPT-4 vs. Professor	-0.18	(-0.32, -0.04)	-2.61	36	-0.43	0.132
	Gemini 2.5 vs. Professor	0.01	(-0.14, 0.15)	0.13	36	0.02	1.000
C	Claude Sonnet 4 vs. DeepSeek v3.2	0.01	(-0.14, 0.16)	0.12	36	0.02	1.000
	Claude Sonnet 4 vs. GPT-4	-0.05	(-0.20, 0.10)	-0.73	36	-0.12	1.000
	Claude Sonnet 4 vs. Gemini 2.5	-0.10	(-0.25, 0.05)	-1.32	36	-0.22	1.000
	Claude Sonnet 4 vs. Professor	0.03	(-0.14, 0.20)	0.33	36	0.05	1.000
	DeepSeek v3.2 vs. GPT-4	-0.06	(-0.22, 0.09)	-0.84	36	-0.14	1.000
	DeepSeek v3.2 vs. Gemini 2.5	-0.11	(-0.27, 0.05)	-1.37	36	-0.23	1.000
	DeepSeek v3.2 vs. Professor	0.02	(-0.16, 0.20)	0.20	36	0.03	1.000
	GPT-4 vs. Gemini 2.5	-0.05	(-0.17, 0.08)	-0.74	36	-0.12	1.000
	GPT-4 vs. Professor	0.08	(-0.09, 0.25)	0.99	36	0.16	1.000
	Gemini 2.5 vs. Professor	0.13	(-0.05, 0.30)	1.47	36	0.24	1.000
SA	Claude Sonnet 4 vs. DeepSeek v3.2	-0.20	(-0.35, -0.04)	-2.61	36	-0.43	0.130
	Claude Sonnet 4 vs. GPT-4	-0.12	(-0.28, 0.05)	-1.43	36	-0.24	1.000
	Claude Sonnet 4 vs. Gemini 2.5	-0.40	(-0.56, -0.23)	-4.80	36	-0.79	**<0.001**
	Claude Sonnet 4 vs. Professor	-0.32	(-0.49, -0.16)	-3.90	36	-0.64	**0.004**
	DeepSeek v3.2 vs. GPT-4	0.08	(-0.07, 0.23)	1.09	36	0.18	1.000
	DeepSeek v3.2 vs. Gemini 2.5	-0.20	(-0.41, 0.02)	-1.87	36	-0.31	0.699
	DeepSeek v3.2 vs. Professor	-0.13	(-0.33, 0.08)	-1.23	36	-0.20	1.000
	GPT-4 vs. Gemini 2.5	-0.28	(-0.45, -0.10)	-3.24	36	-0.53	**0.026**
	GPT-4 vs. Professor	-0.21	(-0.39, -0.02)	-2.28	36	-0.38	0.284
	Gemini 2.5 vs. Professor	0.07	(-0.08, 0.22)	0.97	36	0.16	1.000

Evaluation criteria are expressed as follows: A: Accuracy, C: Clarity, SA: Scientific adequacy, CI: Confidence interval, df: Degrees of freedom

Supplementary Table 6. Detailed Bonferroni-adjusted pairwise comparisons-laypersons only (criteria PR, C, PU)

Criterion	Comparison	Mean diff.	95% CI	t	df	Cohen's d	Bonf. p
PR	Claude Sonnet 4 vs. DeepSeek v3.2	0.14	(-0.02, 0.31)	1.77	36	0.29	0.844
	Claude Sonnet 4 vs. GPT-4	-0.17	(-0.33, -0.01)	-2.19	36	-0.36	0.351
	Claude Sonnet 4 vs. Gemini 2.5	-0.33	(-0.48, -0.18)	-4.49	36	-0.74	**<0.001**
	Claude Sonnet 4 vs. Professor	0.55	(0.34, 0.76)	5.27	36	0.87	**<0.001**
	DeepSeek v3.2 vs. GPT-4	-0.32	(-0.51, -0.12)	-3.32	36	-0.55	**0.021**
	DeepSeek v3.2 vs. Gemini 2.5	-0.48	(-0.67, -0.29)	-5.15	36	-0.85	**<0.001**
	DeepSeek v3.2 vs. Professor	0.41	(0.18, 0.63)	3.72	36	0.61	**0.007**
	GPT-4 vs. Gemini 2.5	-0.16	(-0.27, -0.06)	-3.08	36	-0.51	**0.040**
	GPT-4 vs. Professor	0.72	(0.55, 0.89)	8.77	36	1.44	**<0.001**
	Gemini 2.5 vs. Professor	0.88	(0.68, 1.08)	9.04	36	1.49	**<0.001**
C	Claude Sonnet 4 vs. DeepSeek v3.2	0.16	(-0.01, 0.33)	1.95	36	0.32	0.595
	Claude Sonnet 4 vs. GPT-4	-0.13	(-0.30, 0.05)	-1.48	36	-0.24	1.000
	Claude Sonnet 4 vs. Gemini 2.5	-0.17	(-0.35, 0.01)	-1.92	36	-0.32	0.628
	Claude Sonnet 4 vs. Professor	0.59	(0.36, 0.81)	5.26	36	0.86	**<0.001**
	DeepSeek v3.2 vs. GPT-4	-0.29	(-0.46, -0.11)	-3.36	36	-0.55	**0.019**
	DeepSeek v3.2 vs. Gemini 2.5	-0.33	(-0.53, -0.13)	-3.36	36	-0.55	**0.019**
	DeepSeek v3.2 vs. Professor	0.42	(0.17, 0.68)	3.41	36	0.56	**0.016**
	GPT-4 vs. Gemini 2.5	-0.05	(-0.22, 0.13)	-0.53	36	-0.09	1.000
	GPT-4 vs. Professor	0.71	(0.49, 0.94)	6.40	36	1.05	**<0.001**
	Gemini 2.5 vs. Professor	0.76	(0.52, 1.00)	6.41	36	1.05	**<0.001**
PU	Claude Sonnet 4 vs. DeepSeek v3.2	0.05	(-0.14, 0.23)	0.49	36	0.08	1.000
	Claude Sonnet 4 vs. GPT-4	-0.13	(-0.31, 0.06)	-1.40	36	-0.23	1.000
	Claude Sonnet 4 vs. Gemini 2.5	-0.26	(-0.43, -0.09)	-3.12	36	-0.51	**0.036**
	Claude Sonnet 4 vs. Professor	0.66	(0.42, 0.89)	5.68	36	0.93	**<0.001**
	DeepSeek v3.2 vs. GPT-4	-0.17	(-0.36, 0.02)	-1.79	36	-0.29	0.814
	DeepSeek v3.2 vs. Gemini 2.5	-0.31	(-0.48, -0.13)	-3.48	36	-0.57	**0.013**
	DeepSeek v3.2 vs. Professor	0.61	(0.37, 0.85)	5.19	36	0.85	**<0.001**
	GPT-4 vs. Gemini 2.5	-0.14	(-0.31, 0.04)	-1.61	36	-0.26	1.000
	GPT-4 vs. Professor	0.78	(0.55, 1.02)	6.82	36	1.12	**<0.001**
	Gemini 2.5 vs. Professor	0.92	(0.69, 1.15)	8.19	36	1.35	**<0.001**

Evaluation criteria are expressed as follows: PR: Perceived reliability, C: Clarity, PU: Practical usefulness, CI: Confidence interval, df: Degrees of freedom

Note for Supplementary Tables 4-6: Pairwise comparisons were performed using paired-samples t-tests on per-question source means (n=37; df=36). Mean difference was calculated as the first response source minus the second response source in each comparison. The 95% confidence interval (CI) represents the paired-samples CI for the mean difference. Cohen's d refers to dz, calculated as the mean paired difference divided by the standard deviation of paired differences. Bonferroni-adjusted p-values were calculated for 10 pairwise comparisons within each criterion and rater group. Bold values indicate p<0.05