Results

Paired Samples Contingency Tables

Here, the effect size of the paired groups was determined with pairwise McNemar tests. A Chi-square value was calculated with the application of continuity correction to account for the low number of n=100 itmes.

For all pairs, a phi coefficient was calculated by applying the following formula:

$$phi = SQRT(X2/n)$$

Where X2 = Chi Square value

Contingency Tables

	Humans (MV)		
Gemini Pro	**0**	**1**	**Total**
0	2	34	36
1	1	63	64
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	31.1	1	< .001
χ^2 continuity correction	29.3	1	< .001
N		100	

$$phi = SQRT(29.3/100) = 0.541$$

Paired Samples Contingency Tables

Contingency Tables

	Humans (MV)		
Gemini Ultra	**0**	**1**	**Total**
0	1	31	32
1	2	66	68
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	25.5	1	< .001
χ^2 continuity correction	23.8	1	< .001
N		100	

$$phi = SQRT(23.8/100) = 0.488$$

Paired Samples Contingency Tables

Contingency Tables

	Humans (MV)		
LLMs (AVG)	0	1	Total
0	1	24	25
1	2	73	75
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	18.6	1	< .001
χ^2 continuity correction	17.0	1	< .001
N	100		

$phi = SQRT(17/100) = 0.412$

Paired Samples Contingency Tables

Contingency Tables

	Humans (AVG)		
Gemini Pro	0	1	Total
0	7	29	36
1	10	54	64
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	9.26	1	0.002
χ^2 continuity correction	8.31	1	0.004
N	100		

$phi = SQRT(8.31/100) = 0.288$

Paired Samples Contingency Tables

Contingency Tables

	Humans (MV)		
GPT-4	0	1	Total
0	2	20	22
1	1	77	78
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	17.2	1	< .001
χ^2 continuity correction	15.4	1	< .001
N	100		

phi = SQRT(15.4/100) = 0.392

Paired Samples Contingency Tables

Contingency Tables

	Mistral Large		
Gemini Pro	0	1	Total
0	15	21	36
1	2	62	64
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	15.7	1	< .001
χ^2 continuity correction	14.1	1	< .001
N	100		

phi = SQRT(14.1/100) = 0.375

Paired Samples Contingency Tables

Contingency Tables

	GPT-3.5		
Gemini Pro	0	1	Total
0	16	20	36
1	3	61	64
Total	19	81	100

McNemar Test

	Value	df	p
χ^2	12.6	1	< .001
χ^2 continuity correction	11.1	1	< .001
N	100		

phi = SQRT(11.1/100) = 0.332

Paired Samples Contingency Tables

Contingency Tables

	Humans (MV)		
LLMs (MV)	0	1	Total
0	2	18	20
1	1	79	80
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	15.2	1	< .001
χ^2 continuity correction	13.5	1	< .001
N	100		

$phi = SQRT(13.5/100) = 0.367$

Paired Samples Contingency Tables

Contingency Tables

	LLMs (MV)		
Gemini Pro	0	1	Total
0	17	19	36
1	3	61	64
Total	20	80	100

McNemar Test

	Value	df	p
χ^2	11.6	1	< .001
χ^2 continuity correction	10.2	1	0.001
N	100		

$phi = SQRT(10.2/100) = 0.319$

Paired Samples Contingency Tables

Contingency Tables

GPT-3.5	Humans (MV)		
	0	1	Total
0	2	17	19
1	1	80	81
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	14.2	1	< .001
χ^2 continuity correction	12.5	1	< .001
N	100		

phi = SQRT(12.5/100) = 0.354

Paired Samples Contingency Tables

Contingency Tables

Gemini Ultra	Humans (AVG)		
	0	1	Total
0	2	30	32
1	15	53	68
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	5.00	1	0.025
χ^2 continuity correction	4.36	1	0.037
N	100		

phi = SQRT(4.36/100) = 0.209

Paired Samples Contingency Tables

Contingency Tables

Gemini Ultra	Mistral Large		
	0	1	Total
0	15	17	32
1	2	66	68
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	11.8	1	< .001
χ^2 continuity correction	10.3	1	0.001
N	100		

phi = SQRT(10.3/100) = 0.321

Paired Samples Contingency Tables

Contingency Tables

	GPT-4		
Gemini Pro	0	1	Total
0	17	19	36
1	5	59	64
Total	22	78	100

McNemar Test

	Value	df	p
χ^2	8.17	1	0.004
χ^2 continuity correction	7.04	1	0.008
N	100		

phi = SQRT(7.04/100) = 0.265

Paired Samples Contingency Tables

Contingency Tables

	Humans (MV)		
Humans (AVG)	0	1	Total
0	1	16	17
1	2	81	83
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	10.9	1	< .001
χ^2 continuity correction	9.39	1	0.002
N	100		

phi = SQRT(9.39/100) = 0.306

Paired Samples Contingency Tables

Contingency Tables

	Humans (MV)		
Mistral Large	0	1	Total
0	2	15	17
1	1	82	83
Total	3	97	100

McNemar Test

	Value	df	p
χ^2	12.3	1	< .001
χ^2 continuity correction	10.6	1	0.001
N	100		

$phi = SQRT(10.6/100) = 0.326$

From here on, comparisons are conducted between participants with p > 0.05

Paired Samples Contingency Tables

Contingency Tables

	LLMs (MV)		
Gemini Ultra	0	1	Total
0	17	15	32
1	3	65	68
Total	20	80	100

McNemar Test

	Value	df	p
χ^2	8.00	1	0.005
χ^2 continuity correction	6.72	1	0.010
N	100		

$phi = SQRT(6.72/100) = 0.259$

Paired Samples Contingency Tables

Contingency Tables

	LLMs (AVG)		
Gemini Pro	0	1	Total
0	12	24	36
1	13	51	64
Total	25	75	100

McNemar Test

	Value	df	p
χ^2	3.27	1	0.071
χ^2 continuity correction	2.70	1	0.100
N		100	

phi = SQRT(2.7/100) = 0.164

Paired Samples Contingency Tables

Contingency Tables

	GPT-4		
Gemini Ultra	0	1	Total
0	17	15	32
1	5	63	68
Total	22	78	100

McNemar Test

	Value	df	p
χ^2	5.00	1	0.025
χ^2 continuity correction	4.05	1	0.044
N		100	

phi = SQRT(4.05/100) = 0.201

Paired Samples Contingency Tables

Contingency Tables

	Humans (AVG)		
LLMs (AVG)	0	1	Total
0	12	13	25
1	5	70	75
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	3.56	1	0.059
χ^2 continuity correction	2.72	1	0.099
N		100	

phi = SQRT(2.72/100) = 0.165

Paired Samples Contingency Tables

Contingency Tables

	Mistral Large		
LLMs (AVG)	**0**	**1**	**Total**
0	6	19	25
1	11	64	75
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	2.13	1	0.144
χ^2 continuity correction	1.63	1	0.201
N		100	

phi = SQRT(1.63/100) = 0.128

Paired Samples Contingency Tables

Contingency Tables

	LLMs (AVG)		
Gemini Ultra	**0**	**1**	**Total**
0	5	27	32
1	20	48	68
Total	25	75	100

McNemar Test

	Value	df	p
χ^2	1.04	1	0.307
χ^2 continuity correction	0.766	1	0.381
N		100	

phi = SQRT(0.766/100) = 0.088

Paired Samples Contingency Tables

Contingency Tables

LLMs (AVG)	GPT-3.5		
	0	1	Total
0	6	19	25
1	13	62	75
Total	19	81	100

McNemar Test

	Value	df	p
χ^2	1.13	1	0.289
χ^2 continuity correction	0.781	1	0.377
N		100	

$phi = SQRT(0.781/100) = 0.088$

Paired Samples Contingency Tables

Contingency Tables

GPT-4	Humans (AVG)		
	0	1	Total
0	3	19	22
1	14	64	78
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	0.758	1	0.384
χ^2 continuity correction	0.485	1	0.486
N		100	

$phi = SQRT(0.485/100) = 0.070$

Paired Samples Contingency Tables

Contingency Tables

| | Mistral Large | | |
GPT-4	0	1	Total
0	16	6	22
1	1	77	78
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	3.57	1	0.059
χ^2 continuity correction	2.29	1	0.131
N	100		

phi = SQRT(2.29/100) = 0.151

Paired Samples Contingency Tables

Contingency Tables

| | LLMs (MV) | | |
LLMs (AVG)	0	1	Total
0	7	18	25
1	13	62	75
Total	20	80	100

McNemar Test

	Value	df	p
χ^2	0.806	1	0.369
χ^2 continuity correction	0.516	1	0.472
N	100		

phi = SQRT(0.516/100) = 0.072

Paired Samples Contingency Tables

Contingency Tables

| | Gemini Ultra | | |
Gemini Pro	0	1	Total
0	17	19	36
1	15	49	64
Total	32	68	100

McNemar Test

	Value	df	p
χ^2	0.471	1	0.493
χ^2 continuity correction	0.265	1	0.607
N	100		

phi = SQRT(0.265/100) = 0.051

Paired Samples Contingency Tables

Contingency Tables

	Gemini Ultra		
GPT-3.5	0	1	Total
0	17	2	19
1	15	66	81
Total	32	68	100

McNemar Test

	Value	df	p
χ^2	9.94	1	0.002
χ^2 continuity correction	8.47	1	0.004
N	100		

phi = SQRT(8.47/100) = 0.291

Paired Samples Contingency Tables

Contingency Tables

	Humans (AVG)		
LLMs (MV)	0	1	Total
0	3	17	20
1	14	66	80
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	0.290	1	0.590
χ^2 continuity correction	0.129	1	0.719
N	100		

phi = SQRT(0.129/100) = 0.036

Paired Samples Contingency Tables

Contingency Tables

	GPT-3.5		
GPT-4	0	1	Total
0	18	4	22
1	1	77	78
Total	19	81	100

McNemar Test

	Value	df	p
χ^2	1.80	1	0.180
χ^2 continuity correction	0.800	1	0.371
N	100		

phi = SQRT(0.800/100) = 0.089

Paired Samples Contingency Tables

Contingency Tables

	Mistral Large		
LLMs (MV)	0	1	Total
0	17	3	20
1	0	80	80
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	3.00	1	0.083
χ^2 continuity correction	1.33	1	0.248
N	100		

phi = SQRT(1.33/100) = 0.115

Paired Samples Contingency Tables

Contingency Tables

	GPT-4		
LLMs (AVG)	**0**	**1**	**Total**
0	7	18	25
1	15	60	75
Total	22	78	100

McNemar Test

	Value	**df**	**p**
χ^2	0.273	1	0.602
χ^2 continuity correction	0.121	1	0.728
N		100	

phi = SQRT(0.121/100) = 0.035

Paired Samples Contingency Tables

Contingency Tables

	LLMs (MV)		
GPT-4	**0**	**1**	**Total**
0	19	3	22
1	1	77	78
Total	20	80	100

McNemar Test

	Value	**df**	**p**
χ^2	1.00	1	0.317
χ^2 continuity correction	0.250	1	0.617
N		100	

phi = SQRT(0.25/100) = 0.050

Paired Samples Contingency Tables

Contingency Tables

	Humans (AVG)		
GPT-3.5	**0**	**1**	**Total**
0	3	16	19
1	14	67	81
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	0.133	1	0.715
χ^2 continuity correction	0.0333	1	0.855
N	100		

phi = SQRT(0.03/100) = 0.02

Paired Samples Contingency Tables

Contingency Tables

	Mistral Large		
GPT-3.5	0	1	Total
0	16	3	19
1	1	80	81
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	1.00	1	0.317
χ^2 continuity correction	0.250	1	0.617
N	100		

phi = SQRT(0.25/100) = 0.050

Paired Samples Contingency Tables

Contingency Tables

	GPT-3.5		
LLMs (MV)	0	1	Total
0	19	1	20
1	0	80	80
Total	19	81	100

McNemar Test

	Value	df	p
χ^2	1.00	1	0.317
χ^2 continuity correction	0.00	1	1.000
N	100		

phi = SQRT(0100) = 0

Paired Samples Contingency Tables

Contingency Tables

Mistral Large	Humans (AVG)		Total
	0	1	
0	2	15	17
1	15	68	83
Total	17	83	100

McNemar Test

	Value	df	p
χ^2	0.00	1	1.000
χ^2 continuity correction	0.00	1	1.000
N	100		

phi = SQRT(0100) = 0

References

[1] The jamovi project (2022). *jamovi*. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

[2] R Core Team (2021). *R: A Language and environment for statistical computing*. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).