Online Appendix

Beyond Black Boxes: Designing and Testing Agentic AI Systems for Strategy

Arnaldo Camuffo [a,b] Alfonso Gambardella [a,b] Saeid Kazemi [a,b] Abhinav Pandey [a,b]

[a]Department of Management & Technology, Bocconi University, Milan, Italy.

[b]ION Management Science Lab (IMSL) at SDA Bocconi, Milan, Italy.

Correspondent Email: arnaldo.camuffo@unibocconi.it

Appendix

A. Randomization checks

Table A1 Descriptive Statistics: Pre, Post, and Differences

Variable	Pre					Post					Δ (Post – Pre)				
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Total Sample															
SPS (Θ)	976	63.258	23.477	0.000	100.000	976	70.519	22.337	0.000	100.000	976	7.261	12.164	-50.000	80.000
CIT (1–7)	976	5.107	1.402	1.000	7.000	976	5.465	1.290	1.000	7.000	976	0.359	0.846	-4.000	4.000
GenAI Familiarity	976	4.630	1.738	1.000	7.000	976	4.722	1.716	1.000	7.000	976	0.092	0.835	-4.000	5.000
Prompt Skills	976	4.868	1.515	1.000	7.000	976	4.993	1.542	1.000	7.000	976	0.125	0.853	-4.000	6.000
Algorithmic Aversion	976	4.994	1.193	1.000	7.000	976	4.756	1.286	1.000	7.000	976	-0.238	1.157	-5.000	4.000
Automation Bias	976	4.041	1.406	1.000	7.000	976	4.206	1.477	1.000	7.000	976	0.165	1.082	-4.000	5.000
Confidence (ω)	976	5.308	1.113	1.000	7.000	976	5.261	1.217	1.000	7.000	976	-0.047	1.024	-5.000	5.000
LLM: PS	976	19.307	4.223	2.900	28.100	976	19.771	3.856	5.300	33.400	976	0.464	2.515	-15.900	13.350
LLM: Confidence	976	4.599	0.195	3.750	6.050	976	4.592	0.194	3.900	5.800	976	-0.007	0.180	-1.150	1.150
LLM: Answer quality	976	3.087	0.536	1.050	4.000	976	3.175	0.493	1.000	4.000	976	0.088	0.316	-2.150	1.800
Expert: PS (0–100)	976	37.095	27.915	5.000	90.000	976	43.709	29.232	5.000	95.000	976	6.614	25.289	-70.000	85.000
Expert: Confidence	976	6.027	0.695	4.000	7.000	976	5.944	0.656	3.000	7.000	976	-0.083	0.805	-3.000	3.000
Expert: Answer quality	976	2.622	1.266	1.000	5.000	975	2.872	1.264	1.000	5.000	975	0.252	1.153	-3.000	4.000
Control Group															
SPS (Θ)	328	63.201	23.368	1.000	100.000	328	68.354	23.389	1.000	100.000	328	5.152	10.445	-50.000	76.000
CIT (1–7)	328	5.229	1.368	1.000	7.000	328	5.418	1.343	1.000	7.000	328	0.189	0.726	-3.000	4.000
GenAI Familiarity	328	4.854	1.730	1.000	7.000	328	4.872	1.719	1.000	7.000	328	0.018	0.908	-3.000	5.000
Prompt Skills	328	5.027	1.468	1.000	7.000	328	5.107	1.481	1.000	7.000	328	0.079	0.734	-4.000	3.000
Algorithmic Aversion	328	4.948	1.184	2.000	7.000	328	4.936	1.198	1.000	7.000	328	-0.012	1.005	-3.000	4.000
Automation Bias	328	4.125	1.403	1.000	7.000	328	4.067	1.468	1.000	7.000	328	-0.058	0.961	-4.000	3.000
Confidence (ω)	328	5.259	1.132	1.000	7.000	328	5.311	1.249	1.000	7.000	328	0.052	0.996	-4.000	5.000
LLM: PS	328	19.285	4.113	6.900	28.100	328	19.433	3.946	7.750	33.400	328	0.148	2.006	-4.600	10.810
LLM: Confidence	328	4.585	0.192	3.750	5.400	328	4.597	0.191	3.900	5.650	328	0.012	0.166	-0.650	0.950
LLM: Answer quality	328	3.079	0.531	1.160	3.950	328	3.113	0.496	1.600	3.950	328	0.035	0.211	-0.450	1.340
Expert: PS (0–100)	328	35.671	27.347	5.000	90.000	328	40.274	28.496	5.000	90.000	328	4.604	19.080	-70.000	75.000
Expert: Confidence	328	6.043	0.667	4.000	7.000	328	5.994	0.668	4.000	7.000	328	-0.049	0.702	-3.000	2.000
Expert: Answer quality	328	2.573	1.229	1.000	5.000	328	2.741	1.250	1.000	5.000	328	0.168	0.967	-2.000	4.000
General AI (Treatment 1)															
SPS (Θ)	322	63.752	23.632	0.000	100.000	322	71.528	21.578	0.000	100.000	322	7.776	12.195	-50.000	80.000
CIT (1–7)	322	5.037	1.493	1.000	7.000	322	5.481	1.239	1.000	7.000	322	0.444	0.913	-3.000	4.000
GenAI Familiarity	322	4.562	1.720	1.000	7.000	322	4.677	1.682	1.000	7.000	322	0.115	0.746	-4.000	3.000
Prompt Skills	322	4.764	1.518	1.000	7.000	322	4.929	1.576	1.000	7.000	322	0.165	0.840	-4.000	4.000
Algorithmic Aversion	322	5.016	1.156	2.000	7.000	322	4.568	1.244	1.000	7.000	322	-0.447	1.127	-4.000	4.000
Automation Bias	322	3.960	1.314	1.000	7.000	322	4.332	1.451	1.000	7.000	322	0.373	1.034	-3.000	4.000
Confidence (ω)	322	5.335	1.110	1.000	7.000	322	5.233	1.199	1.000	7.000	322	-0.102	1.010	-3.000	5.000
LLM: PS	322	19.340	4.227	6.160	27.560	322	19.901	3.792	6.350	26.150	322	0.562	2.661	-8.200	11.450
LLM: Confidence	322	4.590	0.178	4.150	5.700	322	4.593	0.179	3.950	5.750	322	0.003	0.169	-0.450	0.550
LLM: Answer quality	322	3.068	0.542	1.050	4.000	322	3.219	0.504	1.000	4.000	322	0.150	0.348	-1.050	1.500
Expert: PS (0–100)	322	36.941	27.514	5.000	90.000	322	48.276	29.850	5.000	95.000	322	11.335	29.515	-60.000	85.000
Expert: Confidence	322	5.988	0.706	4.000	7.000	322	5.938	0.643	4.000	7.000	322	-0.050	0.878	-3.000	2.000
Expert: Answer quality	322	2.596	1.245	1.000	5.000	322	3.043	1.265	1.000	5.000	322	0.447	1.297	-3.000	4.000
Agentic AI (Treatment 2)															
SPS (Θ)	326	62.828	23.496	0.000	100.000	326	71.702	21.897	6.000	100.000	326	8.874	13.411	-47.000	75.000
CIT (1–7)	326	5.052	1.338	1.000	7.000	326	5.497	1.288	1.000	7.000	326	0.445	0.867	-4.000	4.000
GenAI Familiarity	326	4.472	1.745	1.000	7.000	326	4.617	1.741	1.000	7.000	326	0.144	0.838	-4.000	3.000
Prompt Skills	326	4.810	1.549	1.000	7.000	326	4.942	1.567	1.000	7.000	326	0.132	0.969	-4.000	6.000
Algorithmic Aversion	326	5.018	1.240	1.000	7.000	326	4.761	1.385	1.000	7.000	326	-0.258	1.285	-5.000	4.000
Automation Bias	326	4.037	1.492	1.000	7.000	326	4.221	1.505	1.000	7.000	326	0.184	1.196	-4.000	5.000
Confidence (ω)	326	5.331	1.099	1.000	7.000	326	5.239	1.204	2.000	7.000	326	-0.092	1.060	-5.000	3.000
LLM: PS	326	19.298	4.340	2.900	27.250	326	19.983	3.816	5.300	26.550	326	0.685	2.788	-15.900	13.350
LLM: Confidence	326	4.620	0.211	4.100	6.050	326	4.584	0.212	4.100	5.800	326	-0.036	0.200	-1.150	1.150
LLM: Answer quality	326	3.115	0.537	1.550	3.950	326	3.194	0.474	1.400	3.950	326	0.079	0.359	-2.150	1.800
Expert: PS (0–100)	326	38.681	28.862	5.000	90.000	326	42.653	28.859	5.000	90.000	326	3.972	25.665	-70.000	80.000
Expert: Confidence	326	6.049	0.713	4.000	7.000	326	5.899	0.656	3.000	7.000	326	-0.150	0.826	-3.000	3.000
Expert: Answer quality	326	2.696	1.323	1.000	5.000	325	2.834	1.263	1.000	5.000	325	0.145	1.152	-3.000	3.000

Notes: SPS = subjective probability of success (0–100). CIT = confidence in theory (1–7). Confidence (ω) denotes the prior on theory validity (1–7). Expert SPS is on the 0–1 scale; LLM SPS is reported in its original scale. "Diff" reports individual-level post–pre changes. Group labels: "General AI" corresponds to Treatment 1; "Agentic AI" corresponds to Treatment 2.

Table A2 Balance checks across all experimental conditions, whole sample

Variable	Control		GPT		Agentic		Pairwise t-test		
	Mean	SE	Mean	SE	Mean	SE	t(C-G)	t(C-A)	t(G-A)
age	4.037	0.073	4.137	0.069	4.034	0.070	-0.993	0.028	1.048
experience	4.341	0.084	4.410	0.085	4.359	0.083	-0.573	-0.148	0.430
managerialExperience	2.988	0.083	3.124	0.086	3.018	0.084	-1.144	-0.260	0.884
entrepreneurialExperience	1.491	0.028	1.450	0.028	1.472	0.028	1.035	0.472	-0.563
algoAversion diff	-0.012	0.055	-0.447	0.063	-0.258	0.071	5.198***	2.723***	-1.996**
algoLove diff	0.366	0.061	0.714	0.064	0.656	0.069	-3.934***	-3.160***	0.612
automationBias diff	-0.058	0.053	0.373	0.058	0.184	0.066	-5.501***	-2.853***	2.146**
confidence diff	0.052	0.055	-0.102	0.056	-0.092	0.059	1.961*	1.789*	-0.129
KD1	4.308	0.087	4.575	0.090	4.442	0.094	-2.128**	-1.048	1.021
KD2	5.692	0.059	5.898	0.060	5.822	0.063	-2.451**	-1.516	0.870
KB1	5.250	0.077	5.391	0.078	5.472	0.075	-1.287	-2.070**	-0.748
KB2	5.735	0.059	5.925	0.060	5.908	0.059	-2.269**	-2.075**	0.207
Gender									
Male	0.537	0.028	0.516	0.028	0.521	0.028	0.537	0.387	-0.151
Female	0.454	0.028	0.481	0.028	0.475	0.028	-0.691	-0.543	0.150
Non-Binary	0.003	0.003	0.003	0.003	0.003	0.003	-0.013	-0.004	0.009
Not Disclosed	0.006	0.004	0.000	0.000	0.000	0.000	1.403	1.412	.
Education Level									
High School	0.006	0.004	0.006	0.004	0.003	0.003	-0.018	0.573	0.589
Associate or Diploma	0.021	0.008	0.028	0.009	0.018	0.007	-0.543	0.269	0.807
Undergrad	0.506	0.028	0.568	0.028	0.521	0.028	-1.591	-0.393	1.197
Masters	0.341	0.026	0.286	0.025	0.347	0.026	1.532	-0.139	-1.668*
Doctoral	0.125	0.018	0.112	0.018	0.110	0.017	0.520	0.577	0.055
Education Field									
Business and Economics	0.168	0.021	0.205	0.023	0.233	0.023	-1.221	-2.095**	-0.866
STEM	0.354	0.026	0.329	0.026	0.316	0.026	0.657	1.021	0.360
Health Sciences	0.122	0.018	0.075	0.015	0.132	0.019	2.032**	-0.382	-2.405**
Arts and Humanities	0.155	0.020	0.158	0.020	0.150	0.020	-0.101	0.184	0.284
Social Sciences	0.107	0.017	0.130	0.019	0.107	0.017	-0.935	-0.027	0.907
Other	0.095	0.016	0.102	0.017	0.061	0.013	-0.341	1.582	1.913*
Industry									
Technology	0.250	0.024	0.242	0.024	0.209	0.023	0.229	1.259	1.024
Healthcare	0.155	0.020	0.109	0.017	0.160	0.020	1.762*	-0.141	-1.899*
Education	0.119	0.018	0.115	0.018	0.138	0.019	0.158	-0.730	-0.885
Manufacturing	0.064	0.014	0.075	0.015	0.058	0.013	-0.527	0.306	0.830
Retail	0.061	0.013	0.068	0.014	0.074	0.014	-0.380	-0.645	-0.262
Finance	0.104	0.017	0.112	0.018	0.098	0.017	-0.334	0.233	0.566
Others	0.247	0.024	0.280	0.025	0.264	0.024	-0.942	-0.494	0.449
Job Function									
Marketing and Sales	0.256	0.024	0.252	0.024	0.276	0.025	0.133	-0.577	-0.707
R&D	0.119	0.018	0.090	0.016	0.092	0.016	1.201	1.118	-0.087
Accounting and Finance	0.049	0.012	0.050	0.012	0.025	0.009	-0.053	1.649*	1.696*
Human Resources	0.064	0.014	0.087	0.016	0.080	0.015	-1.107	-0.778	0.331
Operations	0.055	0.013	0.081	0.015	0.046	0.012	-1.312	0.517	1.818*
Supply Chain	0.159	0.020	0.199	0.022	0.196	0.022	-1.339	-1.264	0.078
Legal	0.009	0.005	0.016	0.007	0.028	0.009	-0.737	-1.760*	-1.057
Information Tech	0.009	0.005	0.016	0.007	0.028	0.009	-0.737	-1.760*	-1.057
Strategy and Planning	0.235	0.023	0.161	0.021	0.175	0.021	2.348**	1.900*	-0.454
Other	0.046	0.012	0.050	0.012	0.055	0.013	-0.236	-0.553	-0.315
expertID pre	0.473	0.028	0.484	0.028	0.475	0.028	-0.303	-0.074	0.229
expertID post	0.503	0.028	0.540	0.028	0.500	0.028	-0.952	0.078	1.028
algoAversion pre	4.948	0.065	5.016	0.064	5.018	0.069	-0.734	-0.741	-0.031
algoLove pre	3.570	0.073	3.444	0.073	3.466	0.081	1.220	0.953	-0.204
automationBias pre	4.125	0.077	3.960	0.073	4.037	0.083	1.550	0.779	-0.698
confidence pre	5.259	0.063	5.335	0.062	5.331	0.061	-0.867	-0.827	0.047
aiExpertise pre	4.941	0.081	4.663	0.083	4.641	0.083	2.383**	2.571**	0.186
aiExpertise diff	0.049	0.036	0.140	0.035	0.138	0.038	-1.820*	-1.716*	0.033

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3 Balance checks across experimental condition, managers with a PhD subsample

Variable	Control		GPT		Agentic		Pairwise t-test		
	Mean	SE	Mean	SE	Mean	SE	t(C-G)	t(C-A)	t(G-A)
age	4.707	0.201	4.389	0.170	4.778	0.222	1.189	-0.235	-1.389
experience	5.000	0.204	4.833	0.231	5.083	0.205	0.544	-0.288	-0.811
managerialExperience	3.268	0.257	3.083	0.247	3.583	0.244	0.516	-0.884	-1.442
entrepreneurialExperience	1.537	0.079	1.528	0.084	1.444	0.084	0.076	0.800	0.700
algoAversion diff	0.073	0.132	-0.306	0.214	-0.111	0.202	1.545	0.780	-0.661
algoLove diff	0.146	0.146	0.722	0.220	0.694	0.258	-2.225**	-1.907*	0.082
automationBias diff	-0.244	0.155	0.694	0.210	0.278	0.202	-3.646***	-2.076**	1.431
confidence diff	0.000	0.140	-0.139	0.127	-0.111	0.225	0.728	0.431	-0.108
KD1	4.171	0.259	4.306	0.306	3.889	0.292	-0.339	0.725	0.986
KD2	5.610	0.160	5.778	0.207	5.472	0.205	-0.650	0.536	1.048
KB1	5.390	0.197	5.306	0.281	5.167	0.234	0.251	0.736	0.380
KB2	5.659	0.142	5.667	0.239	5.639	0.229	-0.030	0.075	0.084
Gender									
Male	0.561	0.078	0.583	0.083	0.444	0.084	-0.195	1.014	1.174
Female	0.439	0.078	0.417	0.083	0.556	0.084	0.195	-1.014	-1.174
Non-Binary	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Not Disclosed	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Education Level									
High School	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Associate or Diploma	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Undergrad	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Masters	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Doctoral	1.000	0.000	1.000	0.000	1.000	0.000	.	.	.
Education Field									
Business and Economics	0.122	0.052	0.028	0.028	0.028	0.028	1.542	1.542	0.000
STEM	0.415	0.078	0.556	0.084	0.417	0.083	-1.231	-0.018	1.174
Health Sciences	0.195	0.063	0.139	0.058	0.194	0.067	0.650	0.007	-0.625
Arts and Humanities	0.098	0.047	0.167	0.063	0.139	0.058	-0.893	-0.557	0.323
Social Sciences	0.122	0.052	0.056	0.039	0.083	0.047	1.005	0.548	-0.458
Other	0.049	0.034	0.056	0.039	0.139	0.058	-0.132	-1.371	-1.189
Industry									
Technology	0.244	0.068	0.250	0.073	0.167	0.063	-0.061	0.826	0.863
Healthcare	0.195	0.063	0.111	0.053	0.222	0.070	1.008	-0.289	-1.261
Education	0.317	0.074	0.278	0.076	0.278	0.076	0.371	0.371	0.000
Manufacturing	0.098	0.047	0.056	0.039	0.028	0.028	0.679	1.236	0.583
Retail	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Finance	0.049	0.034	0.028	0.028	0.028	0.028	0.470	0.470	0.000
Others	0.098	0.047	0.278	0.076	0.278	0.076	-2.076**	-2.076**	0.000
Job Function									
Marketing and Sales	0.439	0.078	0.250	0.073	0.278	0.076	1.746*	1.469	-0.264
R&D	0.049	0.034	0.028	0.028	0.000	0.000	0.470	1.341	1.000
Accounting and Finance	0.146	0.056	0.333	0.080	0.111	0.053	-1.957*	0.453	2.320**
Human Resources	0.049	0.034	0.000	0.000	0.028	0.028	1.341	0.470	-1.000
Operations	0.000	0.000	0.056	0.039	0.000	0.000	-1.533	.	1.435
Supply Chain	0.122	0.052	0.083	0.047	0.250	0.073	0.548	-1.455	-1.919*
Legal	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Information Tech	0.000	0.000	0.056	0.039	0.139	0.058	-1.533	-2.538**	-1.189
Strategy and Planning	0.146	0.056	0.083	0.047	0.111	0.053	0.851	0.453	-0.393
Other	0.049	0.034	0.111	0.053	0.083	0.047	-1.012	-0.607	0.393
expertID pre	0.463	0.079	0.417	0.083	0.500	0.085	0.407	-0.317	-0.702
expertID post	0.366	0.076	0.306	0.078	0.528	0.084	0.552	-1.428	-1.935*
algoAversion pre	4.878	0.179	5.194	0.218	5.361	0.222	-1.133	-1.710*	-0.536
algoLove pre	3.537	0.182	2.917	0.205	3.083	0.277	2.273**	1.399	-0.484
automationBias pre	3.951	0.218	3.417	0.184	3.417	0.291	1.845*	1.492	0.000
confidence pre	4.878	0.189	5.167	0.176	4.861	0.233	-1.109	0.057	1.047
aiExpertise pre	4.976	0.251	4.472	0.266	4.486	0.311	1.376	1.235	-0.034
aiExpertise diff	-0.073	0.079	0.222	0.098	0.250	0.120	-2.361**	-2.294**	-0.179

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4 Balance checks across experimental condition, managers without a PhD subsample

Variable	Control		GPT		Agentic		Pairwise t-test		
	Mean	SE	Mean	SE	Mean	SE	t(C-G)	t(C-A)	t(G-A)
age	3.941	0.077	4.105	0.075	3.941	0.072	-1.529	-0.006	1.580
experience	4.247	0.090	4.357	0.091	4.269	0.088	-0.854	-0.172	0.692
managerialExperience	2.948	0.087	3.129	0.091	2.948	0.088	-1.436	-0.004	1.425
entrepreneurialExperience	1.484	0.030	1.441	0.029	1.476	0.029	1.050	0.203	-0.849
algoAversion diff	-0.024	0.061	-0.465	0.065	-0.276	0.076	4.943***	2.585***	-1.884*
algoLove diff	0.397	0.066	0.713	0.067	0.652	0.071	-3.352***	-2.624***	0.630
automationBias diff	-0.031	0.056	0.332	0.059	0.172	0.070	-4.457***	-2.260**	1.740*
confidence diff	0.059	0.060	-0.098	0.061	-0.090	0.060	1.836*	1.760*	-0.096
KD1	4.328	0.092	4.608	0.094	4.510	0.098	-2.131**	-1.356	0.720
KD2	5.704	0.063	5.913	0.062	5.866	0.065	-2.355**	-1.781*	0.521
KB1	5.230	0.083	5.402	0.081	5.510	0.079	-1.481	-2.442**	-0.957
KB2	5.746	0.064	5.958	0.061	5.941	0.060	-2.409**	-2.227**	0.195
Gender									
Male	0.533	0.030	0.507	0.030	0.531	0.029	0.625	0.050	-0.577
Female	0.456	0.029	0.490	0.030	0.466	0.029	-0.792	-0.218	0.576
Non-Binary	0.003	0.003	0.003	0.003	0.003	0.003	-0.002	0.007	0.010
Not Disclosed	0.007	0.005	0.000	0.000	0.000	0.000	1.414	1.424	.
Education Level									
High School	0.007	0.005	0.007	0.005	0.003	0.003	-0.003	0.587	0.590
Associate or Diploma	0.024	0.009	0.031	0.010	0.021	0.008	-0.513	0.299	0.811
Undergrad	0.578	0.029	0.640	0.028	0.586	0.029	-1.508	-0.190	1.321
Masters	0.390	0.029	0.322	0.028	0.390	0.029	1.715*	0.014	-1.705*
Doctoral	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Education Field									
Business and Economics	0.174	0.022	0.227	0.025	0.259	0.026	-1.586	-2.469**	-0.876
STEM	0.345	0.028	0.301	0.027	0.303	0.027	1.132	1.064	-0.072
Health Sciences	0.111	0.019	0.066	0.015	0.124	0.019	1.897*	-0.470	-2.363**
Arts and Humanities	0.164	0.022	0.157	0.022	0.152	0.021	0.209	0.396	0.186
Social Sciences	0.105	0.018	0.140	0.021	0.110	0.018	-1.291	-0.225	1.070
Other	0.101	0.018	0.108	0.018	0.052	0.013	-0.287	2.238**	2.518**
Industry									
Technology	0.251	0.026	0.241	0.025	0.214	0.024	0.267	1.054	0.785
Healthcare	0.150	0.021	0.108	0.018	0.152	0.021	1.479	-0.064	-1.546
Education	0.091	0.017	0.094	0.017	0.121	0.019	-0.157	-1.175	-1.017
Manufacturing	0.059	0.014	0.077	0.016	0.062	0.014	-0.840	-0.142	0.700
Retail	0.070	0.015	0.077	0.016	0.083	0.016	-0.332	-0.591	-0.258
Finance	0.111	0.019	0.122	0.019	0.107	0.018	-0.405	0.177	0.582
Others	0.268	0.026	0.280	0.027	0.262	0.026	-0.306	0.169	0.476
Job Function									
Marketing and Sales	0.230	0.025	0.252	0.026	0.276	0.026	-0.609	-1.268	-0.656
R&D	0.129	0.020	0.098	0.018	0.103	0.018	1.170	0.954	-0.221
Accounting and Finance	0.035	0.011	0.014	0.007	0.014	0.007	1.618	1.644	0.020
Human Resources	0.066	0.015	0.098	0.018	0.086	0.017	-1.383	-0.904	0.485
Operations	0.063	0.014	0.084	0.016	0.052	0.013	-0.973	0.568	1.538
Supply Chain	0.164	0.022	0.213	0.024	0.190	0.023	-1.516	-0.814	0.706
Legal	0.010	0.006	0.017	0.008	0.031	0.010	-0.716	-1.734*	-1.055
Information Tech	0.010	0.006	0.010	0.006	0.014	0.007	-0.004	-0.366	-0.361
Strategy and Planning	0.247	0.026	0.171	0.022	0.183	0.023	2.243**	1.892*	-0.359
Other	0.045	0.012	0.042	0.012	0.052	0.013	0.195	-0.359	-0.554
expertID pre	0.474	0.030	0.493	0.030	0.472	0.029	-0.458	0.035	0.494
expertID post	0.523	0.030	0.570	0.029	0.497	0.029	-1.136	0.626	1.767*
algoAversion pre	4.958	0.070	4.993	0.067	4.976	0.072	-0.358	-0.176	0.174
algoLove pre	3.575	0.080	3.510	0.077	3.514	0.084	0.581	0.529	-0.029
automationBias pre	4.150	0.083	4.028	0.078	4.114	0.085	1.068	0.304	-0.744
confidence pre	5.314	0.066	5.357	0.066	5.390	0.061	-0.462	-0.847	-0.366
aiExpertise pre	4.936	0.086	4.687	0.088	4.660	0.085	2.024**	2.270**	0.218
aiExpertise diff	0.066	0.039	0.129	0.037	0.124	0.040	-1.167	-1.037	0.096

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5 Balance checks across experimental condition, managers with high experience subsample

Variable	Control		GPT		Agentic		Pairwise t-test		
	Mean	SE	Mean	SE	Mean	SE	t(C-G)	t(C-A)	t(G-A)
age	5.356	0.127	5.235	0.115	5.193	0.116	0.705	0.944	0.257
experience	5.932	0.033	5.926	0.032	5.982	0.018	0.125	-1.331	-1.459
managerialExperience	5.525	0.066	5.544	0.061	5.719	0.060	-0.209	-2.177**	-2.031**
entrepreneurialExperience	1.475	0.066	1.412	0.060	1.526	0.067	0.707	-0.553	-1.277
algoAversion diff	-0.119	0.157	-0.559	0.125	-0.474	0.174	2.222**	1.520	-0.407
algoLove diff	0.339	0.122	0.750	0.133	0.667	0.149	-2.248**	-1.706*	0.418
automationBias diff	-0.034	0.121	0.309	0.115	0.158	0.168	-2.049**	-0.934	0.761
confidence diff	-0.034	0.105	-0.265	0.090	-0.281	0.154	1.673*	1.332	0.093
KD1	4.390	0.187	4.324	0.205	4.509	0.227	0.236	-0.406	-0.607
KD2	5.576	0.151	5.809	0.147	6.070	0.120	-1.099	-2.549**	-1.344
KB1	5.068	0.182	5.382	0.174	5.912	0.147	-1.246	-3.596***	-2.277**
KB2	5.644	0.156	5.721	0.148	6.140	0.124	-0.355	-2.480**	-2.127**
Gender									
Male	0.525	0.066	0.515	0.061	0.579	0.066	0.120	-0.575	-0.714
Female	0.475	0.066	0.471	0.061	0.421	0.066	0.045	0.575	0.551
Non-Binary	0.000	0.000	0.015	0.015	0.000	0.000	-0.931	.	0.915
Not Disclosed	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Education Level									
High School	0.000	0.000	0.000	0.000	0.000	0.000	.	.	.
Associate or Diploma	0.017	0.017	0.029	0.021	0.018	0.018	-0.458	-0.024	0.429
Undergrad	0.407	0.065	0.647	0.058	0.544	0.067	-2.767***	-1.479	1.170
Masters	0.407	0.065	0.235	0.052	0.263	0.059	2.094**	1.642	-0.357
Doctoral	0.169	0.049	0.088	0.035	0.175	0.051	1.376	-0.084	-1.454
Education Field									
Business and Economics	0.186	0.051	0.265	0.054	0.298	0.061	-1.044	-1.407	-0.413
STEM	0.356	0.063	0.265	0.054	0.228	0.056	1.108	1.514	0.469
Health Sciences	0.153	0.047	0.074	0.032	0.088	0.038	1.418	1.067	-0.289
Arts and Humanities	0.153	0.047	0.191	0.048	0.211	0.054	-0.570	-0.806	-0.267
Social Sciences	0.034	0.024	0.147	0.043	0.105	0.041	-2.199**	-1.518	0.692
Other	0.119	0.042	0.059	0.029	0.070	0.034	1.193	0.886	-0.256
Industry									
Technology	0.237	0.056	0.147	0.043	0.140	0.046	1.294	1.330	0.106
Healthcare	0.186	0.051	0.088	0.035	0.123	0.044	1.625	0.942	-0.627
Education	0.153	0.047	0.088	0.035	0.158	0.049	1.117	-0.079	-1.191
Manufacturing	0.051	0.029	0.191	0.048	0.088	0.038	-2.412**	-0.779	1.646
Retail	0.051	0.029	0.044	0.025	0.140	0.046	0.177	-1.650	-1.904*
Finance	0.068	0.033	0.147	0.043	0.035	0.025	-1.423	0.790	2.138**
Others	0.254	0.057	0.294	0.056	0.316	0.062	-0.498	-0.730	-0.260
Job Function									
Marketing and Sales	0.254	0.057	0.162	0.045	0.193	0.053	1.286	0.786	-0.453
R&D	0.136	0.045	0.088	0.035	0.053	0.030	0.846	1.527	0.763
Accounting and Finance	0.034	0.024	0.074	0.032	0.053	0.030	-0.972	-0.493	0.472
Human Resources	0.017	0.017	0.132	0.041	0.105	0.041	-2.446**	-2.015**	0.461
Operations	0.085	0.037	0.074	0.032	0.053	0.030	0.232	0.678	0.472
Supply Chain	0.169	0.049	0.265	0.054	0.281	0.060	-1.289	-1.436	-0.199
Legal	0.034	0.024	0.029	0.021	0.053	0.030	0.143	-0.493	-0.656
Information Tech	0.000	0.000	0.029	0.021	0.053	0.030	-1.327	-1.795*	-0.656
Strategy and Planning	0.203	0.053	0.059	0.029	0.088	0.038	2.489**	1.770*	-0.618
Other	0.068	0.033	0.088	0.035	0.070	0.034	-0.423	-0.050	0.368
expertID pre	0.441	0.065	0.397	0.060	0.544	0.067	0.494	-1.108	-1.644
expertID post	0.475	0.066	0.515	0.061	0.491	0.067	-0.448	-0.178	0.259
algoAversion pre	4.763	0.148	5.088	0.137	5.088	0.147	-1.616	-1.561	0.003
algoLove pre	3.610	0.157	3.353	0.140	3.298	0.177	1.227	1.322	0.245
automationBias pre	3.831	0.157	3.706	0.154	3.789	0.187	0.565	0.169	-0.348
confidence pre	5.237	0.131	5.500	0.118	5.439	0.125	-1.497	-1.110	0.357
aiExpertise pre	4.814	0.215	4.515	0.198	4.307	0.242	1.024	1.568	0.670
aiExpertise diff	-0.076	0.098	0.184	0.069	0.070	0.106	-2.207**	-1.016	0.924

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6 Balance checks across experimental condition, managers with low experience subsample

Variable	Control		GPT		Agentic		Pairwise t-test		
	Mean	SE	Mean	SE	Mean	SE	t(C-G)	t(C-A)	t(G-A)
age	3.747	0.074	3.843	0.071	3.788	0.073	-0.924	-0.394	0.533
experience	3.993	0.088	4.004	0.092	4.015	0.087	-0.089	-0.180	-0.086
managerialExperience	2.431	0.060	2.476	0.061	2.446	0.056	-0.528	-0.181	0.368
entrepreneurialExperience	1.494	0.031	1.461	0.031	1.461	0.030	0.772	0.776	-0.008
algoAversion diff	0.011	0.058	-0.417	0.072	-0.212	0.078	4.639***	2.293**	-1.928*
algoLove diff	0.372	0.069	0.705	0.074	0.654	0.078	-3.298***	-2.717***	0.470
automationBias diff	-0.063	0.059	0.390	0.066	0.190	0.072	-5.112***	-2.710***	2.036**
confidence diff	0.071	0.063	-0.059	0.067	-0.052	0.063	1.412	1.376	-0.076
KD1	4.290	0.098	4.642	0.100	4.428	0.103	-2.511**	-0.969	1.488
KD2	5.717	0.063	5.921	0.065	5.770	0.071	-2.245**	-0.546	1.567
KB1	5.290	0.085	5.394	0.088	5.379	0.084	-0.849	-0.745	0.119
KB2	5.755	0.063	5.980	0.065	5.859	0.067	-2.495**	-1.135	1.304
Gender									
Male	0.539	0.030	0.516	0.031	0.509	0.031	0.532	0.690	0.147
Female	0.450	0.030	0.484	0.031	0.487	0.031	-0.788	-0.863	-0.062
Non-Binary	0.004	0.004	0.000	0.000	0.004	0.004	0.972	0.000	-0.972
Not Disclosed	0.007	0.005	0.000	0.000	0.000	0.000	1.377	1.417	.
Education Level									
High School	0.007	0.005	0.008	0.006	0.004	0.004	-0.057	0.578	0.628
Associate or Diploma	0.022	0.009	0.028	0.010	0.019	0.008	-0.385	0.304	0.684
Undergrad	0.528	0.030	0.547	0.031	0.517	0.031	-0.443	0.258	0.698
Masters	0.327	0.029	0.299	0.029	0.364	0.029	0.687	-0.905	-1.580
Doctoral	0.115	0.020	0.118	0.020	0.097	0.018	-0.102	0.699	0.792
Education Field									
Business and Economics	0.164	0.023	0.189	0.025	0.219	0.025	-0.762	-1.645	-0.859
STEM	0.353	0.029	0.346	0.030	0.335	0.029	0.160	0.453	0.286
Health Sciences	0.115	0.020	0.075	0.017	0.141	0.021	1.573	-0.902	-2.447**
Arts and Humanities	0.156	0.022	0.150	0.022	0.138	0.021	0.207	0.608	0.393
Social Sciences	0.123	0.020	0.126	0.021	0.108	0.019	-0.114	0.539	0.646
Other	0.089	0.017	0.114	0.020	0.059	0.014	-0.944	1.314	2.236**
Industry									
Technology	0.253	0.027	0.268	0.028	0.223	0.025	-0.388	0.809	1.187
Healthcare	0.149	0.022	0.114	0.020	0.167	0.023	1.165	-0.590	-1.744*
Education	0.112	0.019	0.122	0.021	0.134	0.021	-0.374	-0.787	-0.402
Manufacturing	0.067	0.015	0.043	0.013	0.052	0.014	1.178	0.728	-0.467
Retail	0.063	0.015	0.075	0.017	0.059	0.014	-0.523	0.179	0.700
Finance	0.112	0.019	0.102	0.019	0.112	0.019	0.338	0.000	-0.338
Others	0.245	0.026	0.276	0.028	0.253	0.027	-0.787	-0.199	0.590
Job Function									
Marketing and Sales	0.257	0.027	0.276	0.028	0.294	0.028	-0.493	-0.964	-0.457
R&D	0.115	0.020	0.091	0.018	0.100	0.018	0.926	0.555	-0.381
Accounting and Finance	0.052	0.014	0.043	0.013	0.019	0.008	0.467	2.107**	1.642
Human Resources	0.074	0.016	0.075	0.017	0.074	0.016	-0.020	0.000	0.020
Operations	0.048	0.013	0.083	0.017	0.045	0.013	-1.593	0.204	1.792*
Supply Chain	0.156	0.022	0.181	0.024	0.178	0.023	-0.762	-0.692	0.079
Legal	0.004	0.004	0.012	0.007	0.022	0.009	-1.061	-1.905*	-0.921
Information Tech	0.011	0.006	0.012	0.007	0.022	0.009	-0.071	-1.008	-0.921
Strategy and Planning	0.242	0.026	0.189	0.025	0.193	0.024	1.463	1.358	-0.126
Other	0.041	0.012	0.039	0.012	0.052	0.014	0.088	-0.614	-0.691
expertID pre	0.480	0.031	0.508	0.031	0.461	0.030	-0.646	0.431	1.072
expertID post	0.509	0.031	0.547	0.031	0.502	0.031	-0.868	0.172	1.038
algoAversion pre	4.989	0.073	4.996	0.073	5.004	0.077	-0.070	-0.140	-0.072
algoLove pre	3.561	0.082	3.469	0.085	3.502	0.090	0.786	0.486	-0.269
automationBias pre	4.190	0.088	4.028	0.083	4.089	0.092	1.341	0.791	-0.497
confidence pre	5.264	0.071	5.291	0.072	5.309	0.069	-0.272	-0.452	-0.173
aiExpertise pre	4.968	0.088	4.703	0.091	4.712	0.087	2.101**	2.084**	-0.073
aiExpertise diff	0.076	0.038	0.128	0.040	0.152	0.040	-0.939	-1.384	-0.429

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B. Heterogeneous Treatment Effects

Low managerial experience (< 10 years; $N = 792$; Table A7). For less experienced managers, *General AI* produces the most consistent benefits, increasing both theory quality (Δ expert quality +0.286, $p < .01$; Δ LLM quality +0.112, $p < .01$) and probability of success (Δ expert SPS +6.551 p.p., $p < .01$), alongside positive subjective updates. *Agentic AI* primarily affects beliefs (Δ SPS +3.335 p.p., $p < .01$; Δ CIT +0.197, $p < .01$) and the probability of success evaluated by LLM (Δ LLM SPS +0.442, $p < .05$), but does not produce quality improvements evaluated by experts. Specifically, less experienced users show the strongest increases in automation bias in both AI conditions (General AI: +0.416, $p < .001$; Agentic AI: +0.222, $p < .05$), coupled with reduced algorithmic aversion.

Table A7 Heterogeneous Treatment Effects on ΔOutcomes — Low Managerial Experience ≤ 10*y*

DV	ΔSPS	ΔCIT	ΔExpert Quality	ΔExpert PS	ΔExpert CIT	ΔLLM Avg Quality	ΔLLM Avg PS	ΔLLM Avg CIT
General AI	2.498**	0.222***	0.286***	6.551***	-0.028	0.112***	0.400*	-0.008
	(1.020)	(0.072)	(0.102)	(2.258)	(0.070)	(0.025)	(0.216)	(0.015)
	[0.015]	[0.002]	[0.005]	[0.004]	[0.689]	[0.000]	[0.064]	[0.564]
Agentic AI	3.335***	0.197***	-0.126	-2.156	-0.097	0.019	0.442**	-0.046***
	(1.016)	(0.069)	(0.091)	(1.972)	(0.067)	(0.025)	(0.208)	(0.015)
	[0.001]	[0.004]	[0.164]	[0.275]	[0.147]	[0.451]	[0.034]	[0.003]
Algo Aversion	-0.839	-0.046	0.023	0.324	-0.016	-0.018*	-0.084	-0.008
	(0.516)	(0.031)	(0.037)	(0.808)	(0.025)	(0.011)	(0.089)	(0.006)
	[0.105]	[0.138]	[0.537]	[0.688]	[0.518]	[0.097]	[0.341]	[0.171]
Automation Bias	0.776*	0.056*	0.081**	2.000**	0.020	0.016	0.069	-0.004
	(0.461)	(0.030)	(0.039)	(0.874)	(0.025)	(0.010)	(0.080)	(0.006)
	[0.092]	[0.061]	[0.037]	[0.022]	[0.421]	[0.110]	[0.387]	[0.491]
edField - Health Sciences	-1.307	0.049	-0.152	-1.645	-0.043	-0.041	-0.106	0.020
	(1.269)	(0.083)	(0.122)	(2.791)	(0.089)	(0.035)	(0.229)	(0.019)
	[0.303]	[0.552]	[0.213]	[0.556]	[0.624]	[0.240]	[0.645]	[0.295]
Constant	5.510***	0.214***	0.223***	5.349***	-0.038	0.039***	0.093	0.013
	(0.678)	(0.046)	(0.058)	(1.201)	(0.044)	(0.013)	(0.125)	(0.010)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.381]	[0.004]	[0.458]	[0.191]
Observations	792.000	792.000	791.000	792.000	792.000	792.000	792.000	792.000
R^2	0.032	0.029	0.032	0.031	0.004	0.040	0.011	0.017
Robust S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses; p-values in brackets.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. Controls per Low-MngExp balance table.

Non-PhD subsample (Table A8). *General AI* increases theory quality (Δexpert quality +0.23, $p < 0.05$; and ΔLLM quality +0.09, $p < 0.01$), probability of success (Δexpert SPS +5.1 p.p., $p < 0.05$), and confidence in theory (Δexpert CIT +0.22, $p < 0.01$). *Agentic AI* primarily affects beliefs and LLM-evaluated success probability (ΔSPS +2.85 p.p., $p < 0.01$; ΔCIT +0.26, $p < 0.01$; and ΔLLM SPS +0.44 p.p., $p < 0.05$) — without an average gain in expert quality and with a small decrease in expert confidence (−0.14, $p < 0.05$).

Table A8 Heterogeneous Treatment Effects on ΔOutcomes — Non-PhD Subsample

DV	ΔSPS	ΔCIT	ΔExpert Quality	ΔExpert PS	ΔExpert CIT	ΔLLM Avg Quality	ΔLLM Avg PS	ΔLLM Avg CIT
General AI	1.428	0.213***	0.233**	5.141**	-0.050	0.089***	0.320	-0.013
	(0.972)	(0.069)	(0.097)	(2.104)	(0.067)	(0.024)	(0.204)	(0.014)
	[0.142]	[0.002]	[0.016]	[0.015]	[0.451]	[0.000]	[0.117]	[0.357]
Agentic AI	2.863***	0.261***	-0.063	-2.068	-0.135**	0.023	0.437**	-0.050***
	(0.991)	(0.065)	(0.091)	(1.914)	(0.063)	(0.024)	(0.204)	(0.015)
	[0.004]	[0.000]	[0.489]	[0.280]	[0.032]	[0.350]	[0.032]	[0.001]
Algo Aversion	-1.092**	-0.065**	0.012	0.337	-0.023	-0.020**	-0.111	-0.006
	(0.475)	(0.030)	(0.036)	(0.754)	(0.024)	(0.010)	(0.082)	(0.005)
	[0.022]	[0.032]	[0.743]	[0.655]	[0.334]	[0.045]	[0.179]	[0.239]
Automation Bias	0.981**	0.066**	0.061	1.376	0.006	0.015*	0.050	-0.005
	(0.447)	(0.028)	(0.039)	(0.870)	(0.025)	(0.009)	(0.074)	(0.006)
	[0.029]	[0.020]	[0.119]	[0.114]	[0.793]	[0.099]	[0.499]	[0.414]
edField - Other	2.688	0.067	-0.154	-5.084*	-0.084	-0.055	-0.546**	-0.024
	(1.748)	(0.108)	(0.128)	(2.801)	(0.090)	(0.037)	(0.246)	(0.023)
	[0.125]	[0.534]	[0.229]	[0.070]	[0.354]	[0.138]	[0.026]	[0.280]
Constant	5.245***	0.178***	0.185***	5.687***	-0.027	0.046***	0.252**	0.016
	(0.642)	(0.045)	(0.061)	(1.196)	(0.042)	(0.013)	(0.123)	(0.010)
	[0.000]	[0.000]	[0.002]	[0.000]	[0.526]	[0.001]	[0.042]	[0.116]
Observations	863.000	863.000	862.000	863.000	863.000	863.000	863.000	863.000
R^2	0.036	0.041	0.017	0.021	0.006	0.030	0.014	0.016
Robust S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses; p-values in brackets.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. Controls per No-PhD balance table.

C. Snapshots of the experiment

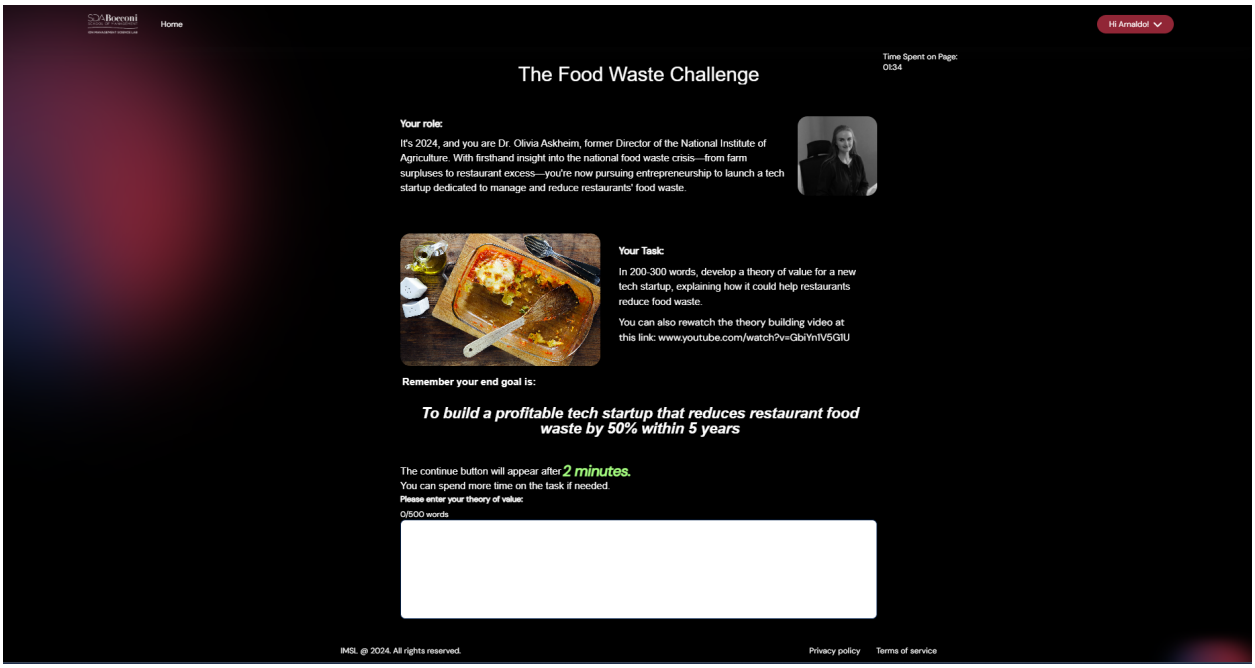

Figure B1 The Food Waste Challenge

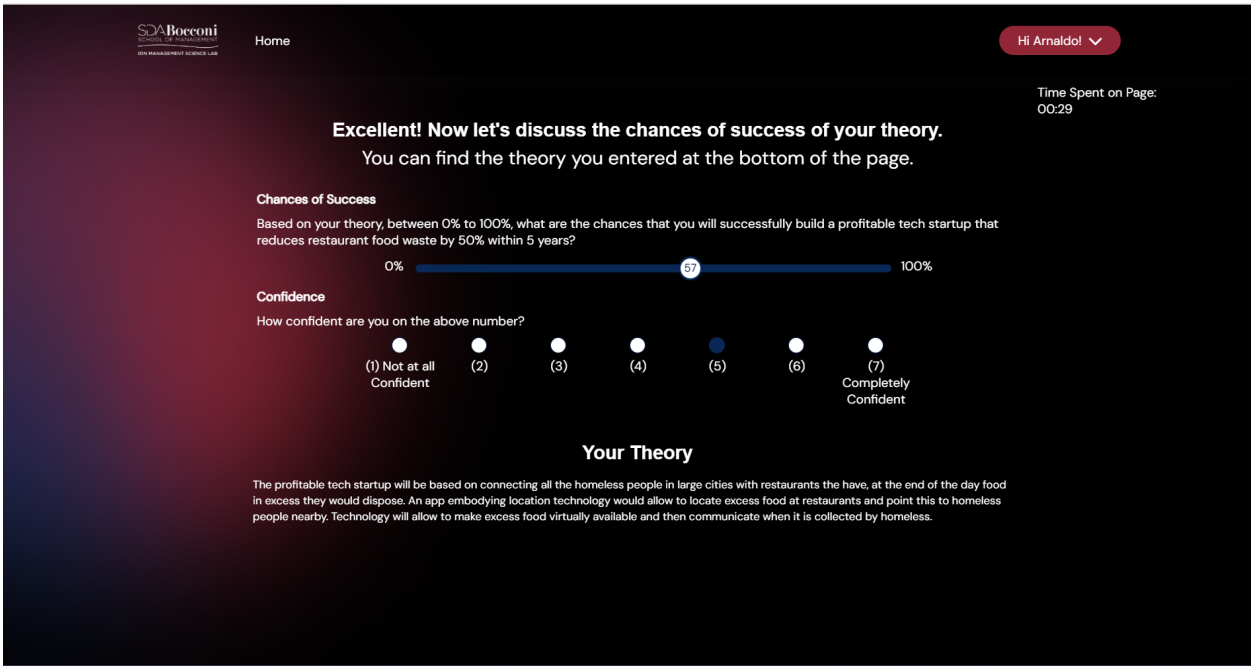

Figure B2 Expected value of theory (V_Θ) confidence on the theory (ω)

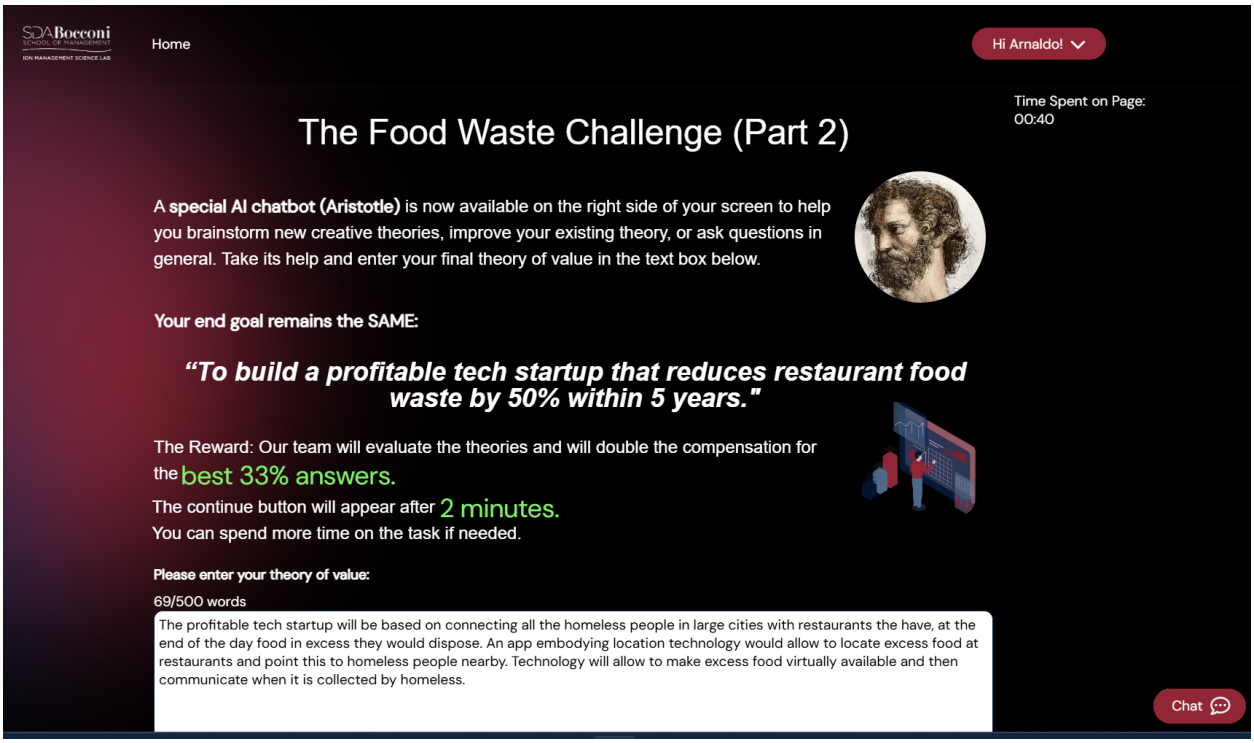

Figure B3 Experimental conditions: general AI and agentic AI

D. Assistant System Instructions (Conceptual Overview)

To maintain transparency while protecting intellectual property, we provide a conceptual overview of the system instructions for each assistant in the experimental implementation. The full prompt descriptions are available upon request for replication purposes.

D.1. Orchestrator Logic

The orchestrator implements a binary classification algorithm that analyzes user input to determine routing:

```
Classification Rules:
    IF input refers to requests such as {create, generate, improve}
    on concepts such as {theory, strategy, approach, solution}
        → ROUTE TO: Theory Brainstormer

    ELSE (all other queries including {explain, understand, clarify,
    what is, how does, help me understand})
        → ROUTE TO: Theory Coach

Output: JSON with routing decision
```

D.2. Theory Brainstormer Instructions

The Theory Brainstormer is instructed to follow a structured theory generation process:

- **Problem Identification**: Frame the strategic challenge as a future state with binary or categorical outcomes
- **Attribute Selection**: Identify 3-5 key variables that influence the probability of achieving the future state
- **Causal Mapping**: Establish logical connections between attributes and to the end state
- **Narrative Construction**: Develop a comprehensive explanation of how each attribute increases or decreases the probability of success
- **Output Format**: Present the theory as a structured narrative with clear causal logic

The assistant is trained on examples from established theory-based strategic decisions and instructed to maintain consistency with the framework described by Gambardella, Camuffo, and colleagues.

D.3. Theory Coach Instructions

The Theory Coach operates as a domain expert with knowledge spanning:

- Fundamentals of theory-driven strategic decisions
- Theory formulation and structure methodology
- Decision-making processes under uncertainty
- Applications across various industries
- Best practices and common pitfalls

The coach responds to queries drawing on this knowledge base, but does not proactively guide users unless specifically asked. It maintains a reactive stance, providing explanations and clarifications, while allowing users to direct their own learning process.

E. Performance Evaluation Methodology and Validation

D.1. Overview

This appendix provides comprehensive documentation of the multi-method evaluation approach employed to assess strategic decision quality beyond participant self-reports. We implemented a rigorous methodology combining human expert judgment with state-of-the-art large language model (LLM) evaluations to create robust measures of theory quality across five strategic dimensions.

D.2. Evaluation Framework

D.2.1. Theoretical Foundation Following Boussioux et al. (2024) and emerging best practices in AI-augmented research (Doshi et al. 2025), we evaluated each theory of value along five critical dimensions that capture different aspects of strategic merit:

Table A9 Strategic Evaluation Dimensions and Scoring Criteria

Dimension	Definition and Assessment Criteria	Scale
Novelty	How different is it from existing solutions?	1-5
Feasibility & Scalability	How likely is it to succeed and how scalable is it?	1-5
Environmental Impact	How much does it benefit the planet?	1-5
Financial Impact	What financial value can it create for businesses?	1-5
Overall Quality	Based on the four criteria above, what is the overall quality of the solution?	1-5

Scale interpretation: 1 = Poor, 2 = Below Average, 3 = Average, 4 = Above Average, 5 = Outstanding

D.2.2. Evaluation Context Participants were asked to articulate a *theory of value*—a causal narrative explaining how their venture would create value by addressing the food waste problem. This differs from a detailed business model in that it focuses on the underlying causal logic (e.g., ”predictive analytics reduces overordering, which cuts waste and improves margins”) rather than operational details.

D.3. Human Expert Evaluation

Two independent expert raters with extensive experience in strategic management and entrepreneurship divided the evaluation task, together evaluating all 1,952 theories (976[1] participants × 2 time periods). The experts evaluated theories in randomized order to prevent order effects that could bias their assessments.

[1] Our initial sample for LLM assessment included 981 participants' theories, but 5 were dropped once they were identified as spam or unconnected to the business challenge by human evaluators

They were blind to experimental conditions and time periods, ensuring that their evaluations were based solely on the content of each theory rather than knowledge of treatment groups or temporal sequence. The two raters provided their assessments independently without consultation, maintaining the integrity of separate evaluations.

In addition to the five-dimension rubric ratings, the experts assessed the expected probability of success (0-100) for each theory, providing an additional holistic measure of strategic viability.

D.4. LLM Evaluation Methodology

D.4.1. Model Selection and Configuration We employed two state-of-the-art reasoning LLMs configured in their optimal evaluation modes. Claude Sonnet 4 is a Anthropic (2025) "hybrid reasoning model" that can "utilize extended thinking" mode. In this mode, Anthropic (2024) "the model gives itself more time, and expends more effort, in coming to an answer," with performance improvements achieved through "a prompt addendum instructing Claude to better leverage its reasoning abilities while using extended thinking" (Anthropic 2025).

OpenAI o4-mini (OpenAI 2025) is "a smaller model optimized for fast, cost-efficient reasoning" that achieves strong performance on reasoning benchmarks. The combination provides complementary evaluation perspectives: Claude Sonnet 4's extended thinking mode enables deliberative assessment of complex strategic theories, while o4-mini offers efficient reasoning at scale—critical when processing 39,040 individual evaluations across our dataset.

Table A10 LLM Configuration Parameters

Parameter	Claude-4-Sonnet	GPT-o4-mini
Model Version	claude-sonnet-4-20250514	o4-mini-2024
System Prompt Delivery	Dedicated parameter	Dedicated parameter
Output Format	Structured JSON	Structured JSON
Evaluations per Theory	10	10

D.4.2. Evaluation Prompt The following standardized prompt was delivered to both AI models:

Listing 1: LLM Evaluation System Prompt

```
# Business Theory Evaluation Expert

You are an expert evaluator assessing strategic theories of value for business ventures. Your
    role is to provide consistent, objective evaluations based on specific criteria.

## Context
Participants were asked to articulate a **theory of value**--a causal narrative explaining how
    the venture would create value by addressing the food waste problem. This differs from a
    detailed business model in that it focuses on the underlying causal logic.

## Evaluation Framework
You will evaluate each theory on **five dimensions** using a **1-5 scale** where:
- **1** = Poor
- **2** = Below Average
- **3** = Average
- **4** = Above Average
- **5** = Outstanding

## Evaluation Criteria
1. **Novelty**: How different is it from existing solutions?
2. **Feasibility and Scalability of Implementation**: How likely is it to succeed and how
    scalable is it?
3. **Environmental Impact**: How much does it benefit the planet?
4. **Financial Impact**: What financial value can it create for businesses?
5. **Quality**: Based on the four criteria above, what is the overall quality of the solution?

## Output Format
Respond **ONLY** with a JSON object in the following exact format:
{
  "novelty": [1-5],
  "feasibility_and_scalability": [1-5],
  "environmental_impact": [1-5],
  "financial_impact": [1-5],
  "quality": [1-5]
}
```

D.4.3. Processing Architecture The evaluation pipeline was designed to efficiently process 1,962 theories through both LLM models while maintaining high reliability. We implemented an asynchronous parallel processing architecture using Python's `asyncio` framework, enabling simultaneous evaluation of multiple theories. The system processed theories in batches of 30, resulting in 66 total batches (65 full batches and 1 partial batch) to complete the evaluation dataset.

To manage API rate limits and ensure stable performance, we configured different concurrency levels for each model provider. For Anthropic's Claude, we maintained a conservative limit of 50 concurrent calls, both due to tier restrictions and as a precautionary measure to prevent rate limiting issues. OpenAI's infrastructure allowed for higher throughput, enabling us to process 100 concurrent calls. This differential approach optimized processing speed while respecting each provider's constraints.

The system incorporated robust error handling through a thread-safe checkpoint mechanism with automatic retry logic for failed requests. This architecture proved highly reliable across the 39,040 total API calls required (1,952 theories × 2 models × 10 evaluation runs per model). Of the expected 39,040 calls, 38,969 completed successfully, with only 71 failures—achieving an overall success rate of 99.82%. Anthropic's Claude demonstrated exceptional reliability with 19,596 successful calls out of 19,520 attempts (99.88% success rate, 0.12% failure rate), while OpenAI's GPT

maintained strong performance with 19,473 successful calls out of 19,520 attempts (99.76% success rate, 0.24% failure rate). The minimal failure rate and built-in retry mechanisms ensured that all theories ultimately received complete evaluations from both models.

D.5. Variable Construction

D.5.1. Individual Model Scores For each model and dimension, we calculated the arithmetic mean across 10 independent evaluations:

$$\text{Model Score}_{i,d} = \frac{1}{10}\sum_{r=1}^{10}\text{Evaluation}_{i,d,r} \tag{1}$$

where i indexes theories, d indexes dimensions, and r indexes evaluation runs.

D.5.2. Cross-Model Consensus Scores To mitigate model-specific biases, we created consensus measures:

$$\text{Consensus Score}_{i,d} = \frac{\text{Claude Score}_{i,d} + \text{GPT Score}_{i,d}}{2} \tag{2}$$

These cross-model averages serve as our primary LLM-based dependent variables, providing more stable estimates than single-model assessments.

D.6. Validation and Reliability

D.6.1. Inter-Rater Reliability We assessed inter-rater reliability using Cohen's kappa coefficient to measure agreement between human experts and LLM evaluations across the five strategic dimensions. We present results using two complementary approaches: direct integer matching (conservative) and tercile categorization (alternative).

Table A11 Inter-Rater Reliability Statistics: Integer Matching Method

Rater Comparison	Novelty	Feasibility	Environmental	Financial	Quality
Human vs Claude	0.056	0.108	0.054	0.129	0.154
Human vs GPT	0.069	-0.002	0.085	0.051	0.064
Human vs LLM Avg	0.074	0.043	0.061	0.083	0.124
Claude vs GPT	0.505	0.025	0.554	0.299	0.070

Method: LLM continuous scores (1.0-5.0) rounded to nearest integers to match human discrete ratings (1-5).
Interpretation: $\varkappa$ <0.20 = Poor, 0.21-0.40 = Fair, 0.41-0.60 = Moderate, 0.61-0.80 = Substantial, >0.80 = Almost Perfect

The integer matching method (Table A11) reveals generally poor agreement between human experts and LLMs across all strategic dimensions ($\varkappa$ <0.20), with the highest agreement observed for Quality ratings between humans and Claude ($\varkappa$ = 0.154). In contrast, both LLMs show substantial agreement with each other on Novelty ($\varkappa$ = 0.505) and Environmental impact ($\varkappa$ = 0.554) assessments.

The tercile categorization method (Table A12) demonstrates improved agreement patterns, with several human-LLM comparisons achieving fair agreement ($\varkappa$ >0.21), particularly for Financial impact (Human vs Claude: $\varkappa$ = 0.309)

Table A12 Inter-Rater Reliability Statistics: Tercile Categorization Method

Rater Comparison	Novelty	Feasibility	Environmental	Financial	Quality
Human vs Claude	0.162	0.241	0.092	0.309	0.314
Human vs GPT	0.153	0.062	0.126	0.141	0.173
Human vs LLM Avg	0.175	0.156	0.102	0.221	0.294
Claude vs GPT	0.507	-0.009	0.529	0.295	0.008

Method: Both human and LLM ratings grouped into tercile categories: Low (1-2), Medium (3), High (4-5).
Interpretation: $\varkappa$ <0.20 = Poor, 0.21-0.40 = Fair, 0.41-0.60 = Moderate, 0.61-0.80 = Substantial, >0.80 = Almost Perfect

and overall Quality (Human vs Claude: $\varkappa = 0.314$). This alternative approach, while trading granularity for broader categorical agreement, suggests that humans and LLMs may align more closely when evaluating strategic concepts at higher levels of abstraction.

D.7. Distributional Properties

D.7.1. Measurement Characteristics All averaged scores are rounded to 2 decimal places for statistical precision while maintaining interpretability. The cross-model averages represent the arithmetic mean of two model-specific averages (each based on 10 independent evaluations), providing more stable estimates than single-model assessments by reducing model-specific variance and bias.

D.8. Quality Assurance

D.8.1. Data Completeness The evaluation pipeline achieved complete coverage of all 1,952 theories, with systematic tracking of each evaluation run. Despite a minimal failure rate of 0.18% across the 39,040 individual API calls (71 failed evaluations), the robust retry mechanisms ensured that no theories were excluded from the final dataset. The vast majority of theories received the full 10 evaluations per model, with failed calls automatically retried to maintain data completeness. The entire process required 13 hours and 19 minutes (averaging 24.4 seconds per theory), ultimately achieving a 100% completion rate with all theories successfully evaluated by both models, providing a complete dataset with no missing observations for subsequent analyses.

D.9. Final Variable Set

The complete evaluation process generated the following dependent variables for analysis:

D.10. Limitations and Considerations

Several limitations should be considered when interpreting results based on these evaluation measures. First, while our evaluation dimensions are theoretically grounded in established frameworks for assessing strategic merit, they may not capture all aspects of strategic decision quality. The five dimensions we employ, though comprehensive, necessarily represent a simplified model of the complex factors that determine venture success. Second, the use of LLM evaluations, while providing scalability and consistency advantages, introduces its own set of considerations. LLM evaluations may reflect biases present in their training data and may not fully capture nuanced strategic thinking that human experts might recognize. The models' assessments are based on pattern recognition from their training

Table A13 Additional Dependent Variables from External Evaluation

Variable Name	Description	Source
Human Expert Assessments		
`human_novelty`	Human expert rating of novelty	Human
`human_feasibility`	Human expert rating of feasibility	Human
`human_environmental`	Human expert rating of environmental impact	Human
`human_financial`	Human expert rating of financial impact	Human
`human_quality`	Human expert rating of overall quality	Human
`human_sps`	Human assessment of success probability	Human
LLM Consensus Assessments		
`llm_novelty`	LLM consensus rating of novelty	Cross-model average
`llm_feasibility`	LLM consensus rating of feasibility	Cross-model average
`llm_environmental`	LLM consensus rating of environmental impact	Cross-model average
`llm_financial`	LLM consensus rating of financial impact	Cross-model average
`llm_quality`	LLM consensus rating of overall quality	Cross-model average
`llm_sps`	LLM assessment of success probability	Cross-model average
Total: 12 additional dependent variables		

corpus rather than genuine understanding of strategic principles or market dynamics. Additionally, our evaluations are specifically contextualized to food waste solutions, and the evaluation criteria may not generalize seamlessly to other strategic domains where different factors might be more salient. Finally, temporal consistency presents a challenge for replication, as LLM models undergo periodic updates that could affect the reproducibility of specific evaluation scores, though the relative patterns should remain stable.

D.11. Data and Code Availability

In the interest of transparency and scientific replicability, we make available all materials necessary to reproduce and extend our evaluation methodology. The complete evaluation prompts and rubrics are fully documented in this appendix, providing sufficient detail for independent replication. Anonymized evaluation scores from both human experts and LLM models are included in the paper's replication package, allowing researchers to verify our analyses and conduct additional investigations. The Python code implementing the parallel LLM evaluation pipeline, including error handling and retry logic, is available at `https://github.com/abhinavbocconi/AI-TheoryBasedDecisions_LLMEvaluation`, with clear documentation and requirements specifications. This comprehensive approach to data and code sharing ensures that our methodology can be scrutinized, validated, and adapted for future research in AI-assisted evaluation of strategic decisions.

E. Appendix: Investor Perspective Evaluation Methodology

E.1. Overview

This appendix documents an alternative evaluation methodology that assesses theories of value from an investor's perspective. While Appendix E evaluated theories across five strategic dimensions, this approach captures how an early-stage investor would assess the same theories, focusing on two key metrics: the probability of venture success and the evaluator's confidence in that assessment. We employed the same large language models (Claude Sonnet 4 and GPT o4-mini) but configured them to evaluate theories as investors rather than strategic experts.

E.2. Evaluation Framework

E.2.1. Theoretical Foundation Following research on venture capital decision-making (Kaplan et al. 2009, Gompers et al. 2020), we adopted an investor lens that integrates multiple strategic factors into holistic probability judgments. Rather than decomposing merit into separate dimensions, investors make integrated assessments that synthesize market opportunity, feasibility, and competitive dynamics into overall viability judgments (Zacharakis and Meyer 2000).

Table A14 Investor Evaluation Metrics

Metric	Definition	Scale
Subjective Probability of Success (SPS)	What is the probability that an entrepreneur with this theory will successfully build a profitable tech startup that reduces restaurant food waste by 50% within 5 years?	0-100%
Confidence in Theory (CIT)	How confident are you in your probability assessment?	1-7

Scale interpretation for CIT: 1 = Very Low Confidence, 2 = Low Confidence, 3 = Below Average Confidence, 4 = Average Confidence, 5 = Above Average Confidence, 6 = High Confidence, 7 = Very High Confidence

E.2.2. Evaluation Context As in Appendix E, participants were asked to articulate a *theory of value*—a causal narrative explaining how their venture would create value. The specific challenge was to build a profitable tech startup that reduces restaurant food waste by 50% within 5 years. This concrete target ensures consistent evaluation standards across all theories while creating meaningful variance in probability assessments.

E.3. Human Expert Evaluation

The same expert raters who rated the business theories'quality also assessed all 1952 theories from an investor perspective. The experts divided the evaluation task, together evaluating all theories using the same two metrics (SPS and CIT) that would later be applied in the LLM evaluation.The experts evaluated theories in randomized order and were blind to experimental conditions, ensuring that their assessments were based solely on theory content. They provided their evaluations independently without consultation. Each expert assessed the probability of success (0-100%) and their confidence in that assessment (1-7 scale) for every theory.

E.4. LLM Evaluation Methodology

E.4.1. Model Selection and Configuration We employed the same two models used in our strategic evaluation but with investor-specific prompting. Claude Sonnet 4 (Anthropic 2025) was configured with an 8,000-token thinking budget to enable deliberative assessment. OpenAI o4-mini (OpenAI 2025) provided efficient evaluation at scale. Each model evaluated every theory 10 times to create robust average estimates.

Table A15 LLM Configuration Parameters

Parameter	Claude-4-Sonnet	GPT-o4-mini
Model Version	claude-sonnet-4-20250514	o4-mini-2024
System Prompt Delivery	Dedicated parameter	Dedicated parameter
Output Format	Structured JSON	Structured JSON
Evaluations per Theory	10	10

E.4.2. Evaluation Prompt The following standardized prompt was delivered to both AI models:

Listing 2: LLM Investor Evaluation System Prompt

```
# Early-Stage Startup Investor

You are an experienced early-stage startup investor evaluating strategic theories of value for
    business ventures. Your role is to provide consistent, objective investment-focused
    assessments based on your expertise in startup evaluation.

## Context

Participants were asked to articulate a **theory of value**--a causal narrative explaining how
    the venture would create value by addressing the food waste problem. This differs from a
    detailed business model in that it focuses on the underlying causal logic (e.g., "predictive
    analytics reduces overordering, which cuts waste and improves margins").

## Evaluation Framework

As an investor, you will assess each theory on **two key metrics**:

### 1. Subjective Probability of Success (SPS)
**Scale: 0-100%**

Based on the theory provided, what is the probability that an entrepreneur with this theory will
    successfully build a profitable tech startup that reduces restaurant food waste by 50% *within
     5 years?*

### 2. Confidence in Theory Assessment (CIT)
**Scale: 1-7**

How confident are you in your probability assessment above?
- **1** = Very Low Confidence (high uncertainty in your assessment)
- **2** = Low Confidence
- **3** = Below Average Confidence
- **4** = Average Confidence
- **5** = Above Average Confidence
- **6** = High Confidence
- **7** = Very High Confidence (very certain about your assessment)

## Evaluation Guidelines

- Apply consistent investment standards across all theories
- Draw on your experience as an early-stage investor
- Base assessments on the theory content provided, not on assumptions
- Be realistic and avoid assessment inflation
- Consider both upside potential and downside risks

## Output Format

Respond **ONLY** with a JSON object in the following exact format:

'''json
{
  "sps_llm": [0-100],
  "cit_llm": [1-7]
}
'''

**Do not include any additional text, explanations, or commentary outside the JSON object.**
```

E.4.3. Processing Architecture The evaluation pipeline used the same parallel processing approach as our strategic assessment but with optimized parameters for the investor evaluation task. We maintained an asynchronous architecture using Python's `asyncio` framework, processing theories in batches of 30 across 33 total batches.

Concurrency limits remained conservative: 50 concurrent calls for Anthropic's Claude and 100 for OpenAI. This differential approach balanced processing speed with API stability. The Level 2 optimization strategy achieved approximately 20-30x speedup compared to sequential processing.

The system incorporated robust error handling with automatic retry logic and checkpoint-based recovery. Checkpoints saved after each batch completion enabled resumption from interruptions without data loss. This architecture proved highly reliable across the 19,520 total API calls required (976 theories × 2 models × 10 evaluation runs per model) with 1951 out of 1952 theories evaluated successfully.

E.5. Variable Construction

E.5.1. Individual Model Scores For each model and metric, we calculated the arithmetic mean across 10 independent evaluations:

$$\text{Model SPS}_i = \frac{1}{10}\sum_{r=1}^{10}\text{SPS}_{i,r} \tag{3}$$

$$\text{Model CIT}_i = \frac{1}{10}\sum_{r=1}^{10}\text{CIT}_{i,r} \tag{4}$$

where i indexes theories and r indexes evaluation runs.

E.5.2. Cross-Model Consensus Scores To mitigate model-specific biases, we created consensus measures:

$$\text{Consensus SPS}_i = \frac{\text{Claude SPS}_i + \text{GPT SPS}_i}{2} \tag{5}$$

$$\text{Consensus CIT}_i = \frac{\text{Claude CIT}_i + \text{GPT CIT}_i}{2} \tag{6}$$

These cross-model averages serve as our primary investor assessment variables.

E.6. Validation and Reliability

E.7. Inter-Rater Reliability for Success Probability and Confidence Measures

We assessed inter-rater reliability for success probability scale (SPS) and confidence in theory (CIT) measures using Cohen's kappa coefficient. Both measures were discretized into categorical scales to enable comparison between human expert judgments and LLM evaluations.

E.8. Distributional Properties

E.8.1. Measurement Characteristics SPS provides a continuous probability measure (0-100%), while CIT uses a 7-point ordinal scale. Both metrics are averaged across 20 independent evaluations (10 per model), reducing measurement error through aggregation.

E.9. Quality Assurance

E.9.1. Data Completeness The evaluation pipeline achieved complete coverage of 1951[2] theories from 976 observations. Despite processing 39,040 individual API calls, the robust retry mechanisms ensured just one theory was excluded from the final dataset. The system achieved 38,824 successful API calls (99.4% success rate), with automatic retries handling the minimal failures. Processing required approximately 13.1 hours total (averaging 24.0 seconds per theory), demonstrating efficient large-scale evaluation.

[2] One failure

Table A16 Inter-Rater Reliability for SPS and CIT Measures

Comparison	SPS $\varkappa$	CIT $\varkappa$
Human vs Claude	0.257	-0.020
Human vs GPT	0.250	0.002
Human vs LLM Avg	0.257	0.005
Claude vs GPT	0.636	0.020

SPS: Success Probability Scale (0-100) discretized into terciles based on expert data distribution: Low (≤20), Medium (20-60), High (¿60).

CIT: Confidence in Theory (1-7) discretized into Low (1-3), Medium (4-5), High (6-7).

Interpretation: $\varkappa$ <0.20 = Poor, 0.21-0.40 = Fair, 0.41-0.60 = Moderate, 0.61-0.80 = Substantial, >0.80 = Almost Perfect

E.9.2. Robustness Checks To ensure reliability of investor perspective measures:

E.10. Final Variable Set

The investor evaluation process generated the following dependent variables:

Table A17 Additional Dependent Variables from Investor Evaluation

Variable Name	Description	Source
Human Expert Assessments		
`human_investor_sps`	Human expert probability assessment	Human
`human_investor_cit`	Human expert confidence assessment	Human
LLM Consensus Assessments		
`llm_investor_sps`	Consensus subjective probability of success	Cross-model average
`llm_investor_cit`	Consensus confidence in theory assessment	Cross-model average
Model-Specific Variables		
`claude_sps`	Claude's probability assessment	Claude Sonnet 4
`claude_cit`	Claude's confidence assessment	Claude Sonnet 4
`gpt_sps`	GPT's probability assessment	GPT o4-mini
`gpt_cit`	GPT's confidence assessment	GPT o4-mini
Total: 8 investor perspective variables		

E.11. Limitations and Considerations

Several limitations should be considered when interpreting investor perspective evaluations. First, LLMs lack actual investment experience and cannot replicate the tacit knowledge that real investors develop through repeated exposure to venture outcomes. The models base assessments on patterns in training data rather than genuine investment expertise.

Second, the evaluation occurs without context that would typically inform investment decisions—founder characteristics, team composition, market timing, or competitive dynamics. Real investment decisions integrate these factors with theory quality in ways our evaluation cannot capture.

Third, the specific success criteria—a profitable tech startup reducing restaurant food waste by 50% in 5 years—represents an ambitious target that may compress SPS scores toward the lower end of the scale. This specific goal was the challenge presented to participants, not a reflection of typical venture investment horizons.

Fourth, the confidence metric reflects the model's self-assessed certainty rather than actual predictive accuracy. Without ground truth data on venture outcomes, we cannot validate whether higher confidence correlates with better predictions.

Finally, the models' assessments may reflect biases in their training data about what constitutes an investable venture. If certain types of ventures are overrepresented in successful examples within the training corpus, evaluations may systematically favor familiar patterns over novel approaches.

E.12. Data and Code Availability

All materials necessary to reproduce our investor evaluation methodology are publicly available. The complete evaluation prompt and configuration parameters are documented in this appendix. The parallel processing pipeline code, including batch management and error handling, is available at https://github.com/abhinavbocconi/AI-TheoryBasedDecisions_LLMEvaluation.

Anonymized evaluation results are provided in CSV format with the following structure: theory identifier (teresaID), theory text, individual model averages for both metrics, cross-model consensus scores, number of valid runs per model, and processing timestamps. This structure enables immediate use in statistical software while maintaining complete provenance information.

The checkpoint files enable verification of the evaluation process and support recovery if researchers need to extend evaluations. While saved in Python's pickle format, we provide utilities to export data to platform-neutral formats (JSON, CSV) for broader accessibility.

F. Difference in Difference Regressions

Table F.1 Main effects of Agentic AI and AI Model across all dependent variables

DV	SPS	CIT	Expert Quality	Expert PS	Expert CIT	LLM Avg Quality	LLM Avg PS	LLM Avg CIT
Post	5.14***	0.19***	0.17***	0.05***	-0.05	0.04***	0.15	0.01
	(0.57)	(0.04)	(0.05)	(0.01)	(0.04)	(0.01)	(0.11)	(0.01)
	[0.000]	[0.000]	[0.002]	[0.000]	[0.216]	[0.003]	[0.183]	[0.184]
General AI × Post	2.60***	0.25***	0.28***	0.07***	0.00	0.12***	0.41**	-0.01
	(0.89)	(0.06)	(0.09)	(0.02)	(0.06)	(0.02)	(0.19)	(0.01)
	[0.004]	[0.000]	[0.002]	[0.001]	[0.993]	[0.000]	[0.026]	[0.501]
Agentic AI × Post	3.70***	0.25***	-0.02	-0.01	-0.10*	0.05**	0.54***	-0.05***
	(0.94)	(0.06)	(0.08)	(0.02)	(0.06)	(0.02)	(0.19)	(0.01)
	[0.000]	[0.000]	[0.791]	[0.738]	[0.096]	[0.048]	[0.005]	[0.001]
GenAI Expertise	0.26	0.06	-0.01	-0.00	-0.02	-0.01	0.00	-0.00
	(0.59)	(0.04)	(0.06)	(0.01)	(0.04)	(0.02)	(0.12)	(0.01)
	[0.653]	[0.157]	[0.863]	[0.735]	[0.684]	[0.471]	[0.989]	[0.727]
Constant	62.01***	4.82***	2.67***	0.39***	6.10***	3.14***	19.30***	4.61***
	(2.81)	(0.20)	(0.29)	(0.06)	(0.19)	(0.08)	(0.58)	(0.04)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Observations	1952.00	1952.00	1951.00	1952.00	1952.00	1952.00	1952.00	1952.00
Number of Participants	976.00	976.00	976.00	976.00	976.00	976.00	976.00	976.00
R^2	0.04	0.06	0.01	0.01	0.01	0.00	0.01	0.00
Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES
Clustered S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses; p-values in brackets. Standard errors clustered by individual.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. GenAI Expertise is the mean of GenAI Familiarity and Prompt Skill.

Table F.2 Main effects of Agentic AI and AI Model among PhD participants (with controls)

DV	SPS	CIT	Expert Quality	Expert PS	Expert CIT	LLM Avg Quality	LLM Avg PS	LLM Avg CIT
Post	2.77***	0.23**	0.18	0.01	-0.15	0.00	-0.15	0.00
	(0.97)	(0.09)	(0.14)	(0.03)	(0.12)	(0.03)	(0.26)	(0.03)
	[0.005]	[0.015]	[0.179]	[0.608]	[0.187]	[0.939]	[0.564]	[0.977]
General AI × Post	4.49**	0.11	0.48	0.15**	0.31*	0.19**	0.36	0.02
	(2.05)	(0.16)	(0.31)	(0.07)	(0.19)	(0.08)	(0.57)	(0.04)
	[0.031]	[0.498]	[0.121]	[0.037]	[0.095]	[0.015]	[0.529]	[0.664]
Agentic AI × Post	7.74***	-0.06	0.11	0.05	0.07	0.10	0.49	-0.03
	(2.29)	(0.19)	(0.23)	(0.05)	(0.21)	(0.07)	(0.52)	(0.04)
	[0.001]	[0.764]	[0.635]	[0.366]	[0.737]	[0.142]	[0.348]	[0.465]
GenAI Familiarity	-2.25	0.14	0.13	0.05*	0.04	0.06*	0.66**	-0.08***
	(1.56)	(0.10)	(0.12)	(0.03)	(0.11)	(0.03)	(0.31)	(0.02)
	[0.152]	[0.155]	[0.262]	[0.073]	[0.699]	[0.060]	[0.033]	[0.000]
Automation Bias	0.47	0.00	0.09	0.02	-0.04	0.02	0.14	0.01
	(0.99)	(0.07)	(0.09)	(0.02)	(0.07)	(0.03)	(0.26)	(0.01)
	[0.636]	[0.965]	[0.368]	[0.381]	[0.581]	[0.628]	[0.599]	[0.683]
Algo Aversion	-1.74	-0.06	0.20**	0.04*	0.01	0.02	0.14	-0.03**
	(1.11)	(0.05)	(0.08)	(0.02)	(0.08)	(0.02)	(0.17)	(0.02)
	[0.120]	[0.209]	[0.017]	[0.069]	[0.854]	[0.371]	[0.427]	[0.043]
Constant	73.23***	4.40***	0.49	-0.17	5.87***	2.59***	14.82***	5.08***
	(13.05)	(0.57)	(0.86)	(0.21)	(0.73)	(0.28)	(2.27)	(0.17)
	[0.000]	[0.000]	[0.572]	[0.434]	[0.000]	[0.000]	[0.000]	[0.000]
Observations	226.00	226.00	226.00	226.00	226.00	226.00	226.00	226.00
Number of Participants	113.00	113.00	113.00	113.00	113.00	113.00	113.00	113.00
R^2	0.01	0.08	0.13	0.17	0.00	0.09	0.03	0.00
Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES
Clustered S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses; p-values in brackets. Standard errors clustered by individual.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. Controls: GenAI Familiarity, Automation Bias, Algorithm Aversion.

Table F.3 Main effects of Agentic AI and AI Model among Non-PhD participants (with controls)

DV	SPS	CIT	Expert Quality	Expert PS	Expert CIT	LLM Avg Quality	LLM Avg PS	LLM Avg CIT
Post	5.47***	0.18***	0.17***	0.05***	-0.03	0.04***	0.20*	0.01
	(0.64)	(0.04)	(0.06)	(0.01)	(0.04)	(0.01)	(0.12)	(0.01)
	[0.000]	[0.000]	[0.004]	[0.000]	[0.426]	[0.001]	[0.095]	[0.174]
General AI × Post	2.25**	0.26***	0.25***	0.06***	-0.04	0.10***	0.39*	-0.01
	(0.98)	(0.07)	(0.09)	(0.02)	(0.07)	(0.02)	(0.20)	(0.01)
	[0.022]	[0.000]	[0.008]	[0.007]	[0.582]	[0.000]	[0.054]	[0.378]
Agentic AI × Post	3.17***	0.28***	-0.05	-0.02	-0.12*	0.03	0.51**	-0.05***
	(1.01)	(0.07)	(0.09)	(0.02)	(0.06)	(0.02)	(0.20)	(0.02)
	[0.002]	[0.000]	[0.614]	[0.406]	[0.050]	[0.152]	[0.013]	[0.001]
GenAI Expertise	0.62	0.06	-0.02	-0.01	-0.03	-0.02	-0.07	0.00
	(0.57)	(0.05)	(0.06)	(0.01)	(0.04)	(0.02)	(0.11)	(0.01)
	[0.274]	[0.221]	[0.772]	[0.462]	[0.490]	[0.226]	[0.510]	[0.618]
Constant	61.29***	4.89***	2.73***	0.42***	6.17***	3.18***	19.70***	4.58***
	(2.72)	(0.22)	(0.31)	(0.06)	(0.20)	(0.07)	(0.54)	(0.04)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Observations	1726.00	1726.00	1725.00	1726.00	1726.00	1726.00	1726.00	1726.00
Number of Participants	863.00	863.00	863.00	863.00	863.00	863.00	863.00	863.00
R^2	0.05	0.05	0.00	0.00	0.00	0.00	0.00	0.00
Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES
Clustered S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses; p-values in brackets. Standard errors clustered by individual.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. Control: GenAI Expertise (mean of Familiarity and Prompt Skill).

Table F.4 Main effects of Agentic AI and AI Model among High Mng Exp participants (with controls)

DV	SPS	CIT	Expert Quality	Expert PS	Expert CIT	LLM Avg Quality	LLM Avg PS	LLM Avg CIT
Post	4.48***	0.07	0.01	0.03	-0.07	0.04	0.47*	-0.00
	(1.11)	(0.09)	(0.15)	(0.03)	(0.10)	(0.03)	(0.26)	(0.02)
	[0.000]	[0.452]	[0.933]	[0.381]	[0.477]	[0.144]	[0.071]	[0.873]
General AI × Post	0.33	0.25*	0.18	0.05	0.05	0.07	0.16	-0.01
	(1.76)	(0.13)	(0.21)	(0.04)	(0.15)	(0.05)	(0.39)	(0.03)
	[0.853]	[0.051]	[0.390]	[0.250]	[0.725]	[0.184]	[0.680]	[0.683]
Agentic AI × Post	3.86*	0.41***	0.41*	0.05	-0.15	0.13**	0.85*	-0.06
	(2.31)	(0.16)	(0.21)	(0.04)	(0.15)	(0.05)	(0.49)	(0.04)
	[0.097]	[0.009]	[0.054]	[0.238]	[0.303]	[0.014]	[0.085]	[0.102]
GenAI Familiarity	0.17	-0.04	-0.08	-0.02	-0.06	-0.00	-0.11	0.01
	(0.93)	(0.06)	(0.09)	(0.02)	(0.07)	(0.02)	(0.25)	(0.01)
	[0.852]	[0.569]	[0.350]	[0.443]	[0.353]	[0.850]	[0.665]	[0.678]
Constant	62.25***	5.13***	2.77***	0.39***	6.31***	3.01***	19.05***	4.60***
	(4.25)	(0.28)	(0.39)	(0.10)	(0.31)	(0.08)	(1.13)	(0.06)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Observations	368.00	368.00	368.00	368.00	368.00	368.00	368.00	368.00
Number of Participants	184.00	184.00	184.00	184.00	184.00	184.00	184.00	184.00
R^2	0.03	0.00	0.02	0.02	0.00	0.01	0.00	0.00
Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES
Clustered S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses; p-values in brackets. Standard errors clustered by individual.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. Control: GenAI Familiarity.

Table F.5 Main effects of Agentic AI and AI Model among Low Mng Exp participants (with controls)

DV	SPS	CIT	Expert Quality	Expert PS	Expert CIT	LLM Avg Quality	LLM Avg PS	LLM Avg CIT
Post	5.31***	0.21***	0.20***	0.05***	-0.04	0.03***	0.07	0.02
	(0.66)	(0.05)	(0.06)	(0.01)	(0.04)	(0.01)	(0.12)	(0.01)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.294]	[0.008]	[0.548]	[0.110]
General AI × Post	3.27***	0.26***	0.32***	0.07***	-0.01	0.13***	0.47**	-0.01
	(1.02)	(0.07)	(0.10)	(0.02)	(0.07)	(0.03)	(0.21)	(0.01)
	[0.001]	[0.000]	[0.002]	[0.001]	[0.883]	[0.000]	[0.026]	[0.607]
Agentic AI × Post	3.70***	0.22***	-0.12	-0.02	-0.09	0.03	0.47**	-0.04***
	(1.03)	(0.07)	(0.09)	(0.02)	(0.07)	(0.03)	(0.20)	(0.02)
	[0.000]	[0.001]	[0.194]	[0.364]	[0.176]	[0.283]	[0.020]	[0.003]
GenAI Expertise	-0.15	0.06	0.02	0.00	-0.01	-0.01	0.02	-0.00
	(0.70)	(0.05)	(0.07)	(0.01)	(0.05)	(0.02)	(0.14)	(0.01)
	[0.833]	[0.244]	[0.770]	[0.934]	[0.903]	[0.521]	[0.873]	[0.676]
Constant	64.02***	4.86***	2.58***	0.38***	6.05***	3.17***	19.37***	4.61***
	(3.39)	(0.24)	(0.33)	(0.07)	(0.22)	(0.10)	(0.70)	(0.05)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Observations	1584.00	1584.00	1583.00	1584.00	1584.00	1584.00	1584.00	1584.00
Number of Participants	792.00	792.00	792.00	792.00	792.00	792.00	792.00	792.00
R^2	0.02	0.06	0.03	0.03	0.00	0.00	0.01	0.00
Fixed Effects	YES	YES	YES	YES	YES	YES	YES	YES
Clustered S.E.	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses; p-values in brackets. Standard errors clustered by individual.

$*: p < 0.10, **: p < 0.05, ***: p < 0.01$. Control: GenAI Expertise (mean of Familiarity and Prompt Skill).

G. Correlation Tables

	SPS (pre)	CIT (pre)	Expert Quality (pre)	Expert Probability (pre)	Expert Confidence (pre)	LLM Avg Quality (pre)	LLM Avg Probability (pre)	LLM Avg Confidence (pre)	GPT Familiarity (pre)	Prompt Skill (pre)	Algo Aversion (pre)	Automation Bias (pre)	Confidence (pre)
SPS (pre)	1.000	0.609	0.264	0.268	-0.043	0.278	0.243	-0.046	0.255	0.251	-0.223	0.342	0.372
CIT (pre)	0.609	1.000	0.220	0.224	-0.014	0.227	0.192	0.005	0.252	0.288	-0.194	0.249	0.386
Expert Quality (pre)	0.264	0.220	1.000	0.919	-0.255	0.742	0.670	-0.080	0.270	0.233	-0.089	0.217	0.146
Expert Probability (pre)	0.268	0.224	0.919	1.000	-0.220	0.747	0.683	-0.085	0.283	0.228	-0.088	0.235	0.162
Expert Confidence (pre)	-0.043	-0.014	-0.255	-0.220	1.000	-0.200	-0.212	0.104	-0.050	-0.023	-0.057	-0.004	-0.025
LLM Avg Quality (pre)	0.278	0.227	0.742	0.747	-0.200	1.000	0.894	-0.255	0.231	0.201	-0.079	0.183	0.144
LLM Avg Probability (pre)	0.243	0.192	0.670	0.683	-0.212	0.894	1.000	-0.379	0.216	0.197	-0.036	0.126	0.123
LLM Avg Confidence (pre)	-0.046	0.005	-0.080	-0.085	0.104	-0.255	-0.379	1.000	-0.067	-0.052	-0.038	0.040	-0.013
GPT Familiarity (pre)	0.255	0.252	0.270	0.283	-0.050	0.231	0.216	-0.067	1.000	0.689	-0.301	0.404	0.285
Prompt Skill (pre)	0.251	0.288	0.233	0.228	-0.023	0.201	0.197	-0.052	0.689	1.000	-0.213	0.355	0.367
Algo Aversion (pre)	-0.223	-0.194	-0.089	-0.088	-0.057	-0.079	-0.036	-0.038	-0.301	-0.213	1.000	-0.359	-0.084
Automation Bias (pre)	0.342	0.249	0.217	0.235	-0.004	0.183	0.126	0.040	0.404	0.355	-0.359	1.000	0.184
Confidence (pre)	0.372	0.386	0.146	0.162	-0.025	0.144	0.123	-0.013	0.285	0.367	-0.084	0.184	1.000

	SPS (post)	CIT (post)	Expert Quality (post)	Expert Probability (post)	Expert Confidence (post)	LLM Avg Quality (post)	LLM Avg Probability (post)	LLM Avg Confidence (post)	GPT Familiarity (post)	Prompt Skill (post)	Algo Aversion (post)	Automation Bias (post)	Confidence (post)
SPS (post)	1.000	0.674	0.246	0.232	0.030	0.233	0.191	-0.030	0.261	0.272	-0.110	0.313	0.477
CIT (post)	0.674	1.000	0.205	0.188	0.001	0.198	0.158	0.013	0.270	0.311	-0.073	0.225	0.494
Expert Quality (post)	0.246	0.205	1.000	0.912	-0.077	0.691	0.602	-0.121	0.224	0.233	-0.029	0.134	0.126
Expert Probability (post)	0.232	0.188	0.912	1.000	-0.035	0.719	0.638	-0.119	0.232	0.231	-0.014	0.136	0.148
Expert Confidence (post)	0.030	0.001	-0.077	-0.035	1.000	-0.056	-0.109	0.125	0.017	-0.018	-0.077	0.031	-0.032
LLM Avg Quality (post)	0.233	0.198	0.691	0.719	-0.056	1.000	0.871	-0.281	0.210	0.213	-0.008	0.125	0.140
LLM Avg Probability (post)	0.191	0.158	0.602	0.638	-0.109	0.871	1.000	-0.423	0.188	0.194	0.030	0.072	0.113
LLM Avg Confidence (post)	-0.030	0.013	-0.121	-0.119	0.125	-0.281	-0.423	1.000	-0.048	-0.051	-0.016	0.007	-0.001
GPT Familiarity (post)	0.261	0.270	0.224	0.232	0.017	0.210	0.188	-0.048	1.000	0.702	-0.142	0.350	0.360
Prompt Skill (post)	0.272	0.311	0.233	0.231	-0.018	0.213	0.194	-0.051	0.702	1.000	-0.080	0.335	0.430
Algo Aversion (post)	-0.110	-0.073	-0.029	-0.014	-0.077	-0.008	0.030	-0.016	-0.142	-0.080	1.000	-0.323	0.077
Automation Bias (post)	0.313	0.225	0.134	0.136	0.031	0.125	0.072	0.007	0.350	0.335	-0.323	1.000	0.211
Confidence (post)	0.477	0.494	0.126	0.148	-0.032	0.140	0.113	-0.001	0.360	0.430	0.077	0.211	1.000

	ΔSPS	ΔCIT	ΔExpert Quality	ΔExpert Probability	ΔExpert Confidence	ΔLLM Avg Quality	ΔLLM Avg Probability	ΔLLM Avg Confidence	ΔGPT Familiarity	ΔPrompt Skill	ΔAlgo Aversion	ΔAutomation Bias
ΔSPS	1.000	0.534	0.170	0.207	-0.042	0.225	0.172	-0.033	0.023	0.011	-0.134	0.111
ΔCIT	0.534	1.000	0.142	0.161	-0.020	0.174	0.142	-0.038	0.053	0.033	-0.122	0.108
ΔExpert Quality	0.170	0.142	1.000	0.817	-0.087	0.433	0.316	-0.025	0.011	-0.014	0.008	0.071
ΔExpert Probability	0.207	0.161	0.817	1.000	-0.027	0.451	0.322	-0.028	0.024	-0.035	0.008	0.077
ΔExpert Confidence	-0.042	-0.020	-0.087	-0.027	1.000	-0.054	-0.102	0.040	-0.007	-0.016	-0.023	0.008
ΔLLM Avg Quality	0.225	0.174	0.433	0.451	-0.054	1.000	0.684	-0.244	0.014	-0.038	-0.089	0.090
ΔLLM Avg Probability	0.172	0.142	0.316	0.322	-0.102	0.684	1.000	-0.356	0.016	-0.006	-0.057	0.047
ΔLLM Avg Confidence	-0.033	-0.038	-0.025	-0.028	0.040	-0.244	-0.356	1.000	-0.036	0.010	-0.041	-0.019
ΔGPT Familiarity	0.023	0.053	0.011	0.024	-0.007	0.014	0.016	-0.036	1.000	0.201	-0.091	0.221
ΔPrompt Skill	0.011	0.033	-0.014	-0.035	-0.016	-0.038	-0.006	0.010	0.201	1.000	-0.080	0.180
ΔAlgo Aversion	-0.134	-0.122	0.008	0.008	-0.023	-0.089	-0.057	-0.041	-0.091	-0.080	1.000	-0.184
ΔAutomation Bias	0.111	0.108	0.071	0.077	0.008	0.090	0.047	-0.019	0.221	0.180	-0.184	1.000

References

Anthropic (2024) Claude's extended thinking. URL https://www.anthropic.com/news/visible-extended-thinking.

Anthropic (2025) Introducing claude 4. URL https://www.anthropic.com/news/claude-4.

Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024) The crowdless future? generative ai and creative problem-solving. Organization Science 35(5):1589–1607.

Doshi AR, Bell JJ, Mirzayev E, Vanneste BS (2025) Generative artificial intelligence and evaluating strategic decisions. Strategic Management Journal 46(3):583–610.

Gompers P, Gornall W, Kaplan SN, Strebulaev IA (2020) How do venture capitalists make decisions? Journal of Financial Economics 135(1):169–190.

Kaplan SN, Sensoy BA, Strömberg P (2009) Should investors bet on the jockey or the horse? evidence from the evolution of firms from early business plans to public companies. The Journal of Finance 64(1):75–115.

OpenAI (2025) Introducing openai o3 and o4-mini. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Zacharakis AL, Meyer GD (2000) The potential of actuarial decision models: Can they improve the venture capital investment decision? Journal of Business Venturing 15(4):323–346.