AI Coaching Simulation Report (Multi-Judge)

Generated on: 2025-04-29 17:06:16

AI Coach Configuration

Coach LLM: `gemini-2.5-flash-preview-04-17`

RAG Enabled: True

Coach System Prompt Enabled: True

AI Judge Configuration

Judge 1: `openai / gpt-4o`

Judge 2: `anthropic / claude-3-5-sonnet-20241022`

Judge 3: `google / gemini-2.5-pro-preview-03-25`

Overall AI Performance Metrics (Combined Average)

Metrics averaged across all judges and all successfully assessed scenarios (3 scenarios).

Overall Average Composite Score

3.81 / 5.00

Overall Average Competency Scores

Competency	Average Score
Accountability	**3.67**
Action Planning	**4.33**
Active Listening	**4.11**
Building Rapport and Trust	**4.00**
Continuous Learning	**3.56**
Ethical Foundation	**2.67**
Facilitating Insight	**4.33**
Goal Clarification	**3.33**
Overall Effectiveness	**4.22**
Powerful Questioning	**3.78**

Overall Run Summary

Metric	Value
Total Scenarios Processed	3
Total Runs Attempted (All Scenarios)	3
Total Runs Completed (Conversation OK)	3

Total Runs With >=1 Valid Assessment	3
Total Valid Individual Assessments	9
Total Failed Runs (Critical/Conv/Assess)	0
Total Runs Where Assessment Failed (All Judges)	0

Configuration Used (Summary)

Parameter	Value
`AI_JUDGE_COMPETENCY_WEIGHTS`	Dict (Keys: ['Ethical Foundation', 'Building Rapport and Trust', 'Active Listening', 'Powerful Questioning', 'Goal Clarification', 'Facilitating Insight', 'Action Planning', 'Accountability', 'Continuous Learning', 'Overall Effectiveness'])
`AI_JUDGE_MAX_TRANSCRIPT_LEN`	25000
`COACH_MAX_TOKENS`	4000
`COACH_MODEL`	gemini-2.5-flash-preview-04-17
`COACH_PROMPT_ENABLED`	True
`COACH_PROVIDER`	google
`COACH_TEMPERATURE`	0.2
`ENABLE_AI_JUDGE`	True
`ENABLE_RAG`	True
`ENABLE_RAG_SUMMARY`	True

GENERATE_REPORT	True
MAX_TURNS	12
MIN_TURNS_FOR_EARLY_EXIT	3
NUM_RUNS	1
OUTPUT_DIR	coaching_simulation_output
RAG_CONTEXT_WINDOW	1
RAG_ENDPOINT	https://magic.neuropower.ai/vectorsearch
RAG_K_RESULTS	1
RAG_TIMEOUT	15
SAVE_TRANSCRIPTS	True
SCENARIOS	['procrastination', 'conflict', 'career']
SCENARIO_GOALS	Dict (Keys: ['procrastination', 'conflict', 'career'])
SUMMARY_MAX_RAW_LENGTH	20000
SUMMARY_MAX_TOKENS	1500
SUMMARY_MODEL	gemini-2.0-flash
SUMMARY_PROVIDER	google
SUMMARY_TEMPERATURE	0.2
SUMMARY_THRESHOLD_LENGTH	20000
USER_MAX_TOKENS	1500
USER_MODEL	gemini-2.0-flash
USER_PROVIDER	google

USER_TEMPERATURE	0.2

Performance by Scenario

Scenario: career

Run Summary

Metric	Value
Runs Attempted	1
Runs Completed (Conversation OK)	1
Runs With >=1 Valid Assessment	1
Total Valid Individual Assessments	3
Failed Runs (Total)	0
Runs Where Assessment Failed (All Judges)	0

Performance Statistics (Across 1 Completed Run)

Metric	Value
Turn Count (Avg)	12.0
Turn Count (Stdev)	0.00

Turn Count (Range)	12 - 12
Duration (Avg Secs)	101.2
Duration (Stdev Secs)	0.00
Avg Errors Logged per Run	0.00

Combined Coaching Scores (Avg across 3 valid assessments)

Competency	Combined Avg Score
Accountability	**3.67**
Action Planning	**4.33**
Active Listening	**4.00**
Building Rapport and Trust	**4.00**
Continuous Learning	**3.33**
Ethical Foundation	**2.33**
Facilitating Insight	**4.00**
Goal Clarification	**3.33**
Overall Effectiveness	**4.00**
Powerful Questioning	**4.00**
Combined Composite Score (Avg)	**3.72**
Combined Composite Score (Stdev)	**0.08**

Per-Judge Score Breakdown (Average for this Scenario)

Competency	openai/gpt-4o (1 assess.)	anthropic/claude-3-5-sonnet-20241022 (1 assess.)	google/gemini-2.5-pro-preview-03-25 (1 assess.)
Accountability	3.00	4.00	4.00
Action Planning	4.00	4.00	5.00
Active Listening	4.00	4.00	4.00
Building Rapport and Trust	4.00	4.00	4.00
Continuous Learning	3.00	3.00	4.00
Ethical Foundation	3.00	3.00	1.00
Facilitating Insight	4.00	4.00	4.00
Goal Clarification	3.00	3.00	4.00
Overall Effectiveness	4.00	4.00	4.00
Powerful Questioning	4.00	4.00	4.00
Composite Score (Avg)	**3.64**	**3.72**	**3.80**

Termination Reasons (Across 1 Completed Run)

Reason	Count	Percentage
Maximum turns (12) reached	1	100.0%

Scenario: conflict

Run Summary

Metric	Value
Runs Attempted	1
Runs Completed (Conversation OK)	1
Runs With >=1 Valid Assessment	1
Total Valid Individual Assessments	3
Failed Runs (Total)	0
Runs Where Assessment Failed (All Judges)	0

Performance Statistics (Across 1 Completed Run)

Metric	Value
Turn Count (Avg)	12.0
Turn Count (Stdev)	0.00
Turn Count (Range)	12 - 12
Duration (Avg Secs)	115.9
Duration (Stdev Secs)	0.00
Avg Errors Logged per Run	0.00

Combined Coaching Scores (Avg across 3 valid assessments)

Competency	Combined Avg Score
Accountability	**3.33**
Action Planning	**4.00**
Active Listening	**4.00**
Building Rapport and Trust	**4.00**
Continuous Learning	**3.67**
Ethical Foundation	**2.67**
Facilitating Insight	**4.33**
Goal Clarification	**3.00**
Overall Effectiveness	**4.00**
Powerful Questioning	**3.67**
Combined Composite Score (Avg)	**3.69**
Combined Composite Score (Stdev)	**0.15**

Per-Judge Score Breakdown (Average for this Scenario)

Competency	openai/gpt-4o (1 assess.)	anthropic/claude-3-5-sonnet-20241022 (1 assess.)	google/gemini-2.5-pro-preview-03-25 (1 assess.)
Accountability	3.00	3.00	4.00

Action Planning	4.00	4.00	4.00
Active Listening	4.00	4.00	4.00
Building Rapport and Trust	4.00	4.00	4.00
Continuous Learning	3.00	4.00	4.00
Ethical Foundation	3.00	3.00	2.00
Facilitating Insight	4.00	4.00	5.00
Goal Clarification	3.00	3.00	3.00
Overall Effectiveness	4.00	4.00	4.00
Powerful Questioning	3.00	4.00	4.00
Composite Score (Avg)	**3.52**	**3.72**	**3.82**

Termination Reasons (Across 1 Completed Run)

Reason	Count	Percentage
Maximum turns (12) reached	1	100.0%

Scenario: procrastination

Run Summary

Metric	Value
Runs Attempted	1

Runs Completed (Conversation OK)	1
Runs With >=1 Valid Assessment	1
Total Valid Individual Assessments	3
Failed Runs (Total)	0
Runs Where Assessment Failed (All Judges)	0

Performance Statistics (Across 1 Completed Run)

Metric	Value
Turn Count (Avg)	12.0
Turn Count (Stdev)	0.00
Turn Count (Range)	12 - 12
Duration (Avg Secs)	93.3
Duration (Stdev Secs)	0.00
Avg Errors Logged per Run	0.00

Combined Coaching Scores (Avg across 3 valid assessments)

Competency	Combined Avg Score
Accountability	**4.00**
Action Planning	**4.67**

Active Listening	**4.33**
Building Rapport and Trust	**4.00**
Continuous Learning	**3.67**
Ethical Foundation	**3.00**
Facilitating Insight	**4.67**
Goal Clarification	**3.67**
Overall Effectiveness	**4.67**
Powerful Questioning	**3.67**
Combined Composite Score (Avg)	**4.03**
Combined Composite Score (Stdev)	**0.37**

Per-Judge Score Breakdown (Average for this Scenario)

Competency	openai/gpt-4o (1 assess.)	anthropic/claude-3-5-sonnet-20241022 (1 assess.)	google/gemini-2.5-pro-preview-03-25 (1 assess.)
Accountability	4.00	4.00	4.00
Action Planning	4.00	5.00	5.00
Active Listening	4.00	5.00	4.00
Building Rapport and Trust	4.00	4.00	4.00

Continuous Learning	3.00	4.00	4.00
Ethical Foundation	3.00	3.00	3.00
Facilitating Insight	4.00	5.00	5.00
Goal Clarification	3.00	4.00	4.00
Overall Effectiveness	4.00	5.00	5.00
Powerful Questioning	3.00	4.00	4.00
Composite Score (Avg)	**3.60**	**4.30**	**4.18**

Termination Reasons (Across 1 Completed Run)

Reason	Count	Percentage
Maximum turns (12) reached	1	100.0%