LLM Coaching Comparison Study

REN vs. ChatGPT, CoPilot, and Gemini

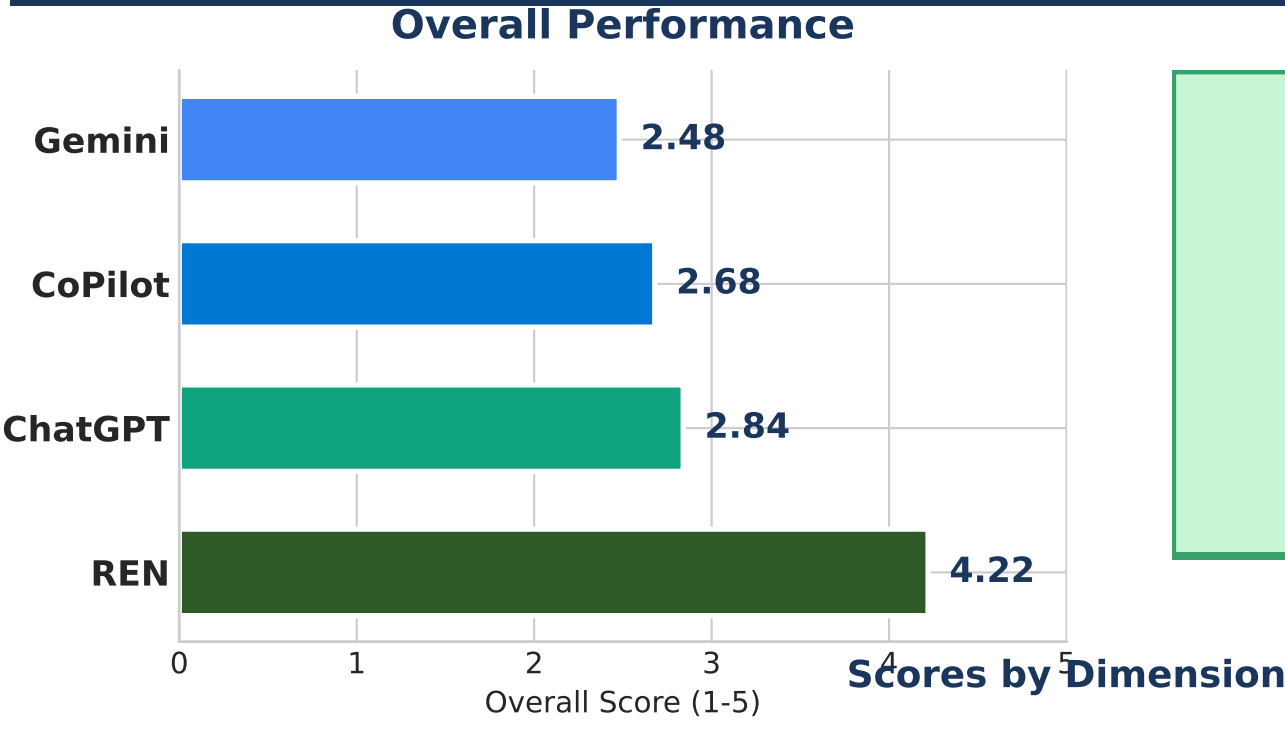
Extended Multi-Turn Evaluation

25 Scenarios • 4 Turns Each • 5 Dimensions

400+	5	4.22
Responses Analyzed	Dimensions Evaluated	REN Overall Score

Final Report • February 2026

Executive Summary



KEY FINDING

REN leads in 4 of 5 dimensions

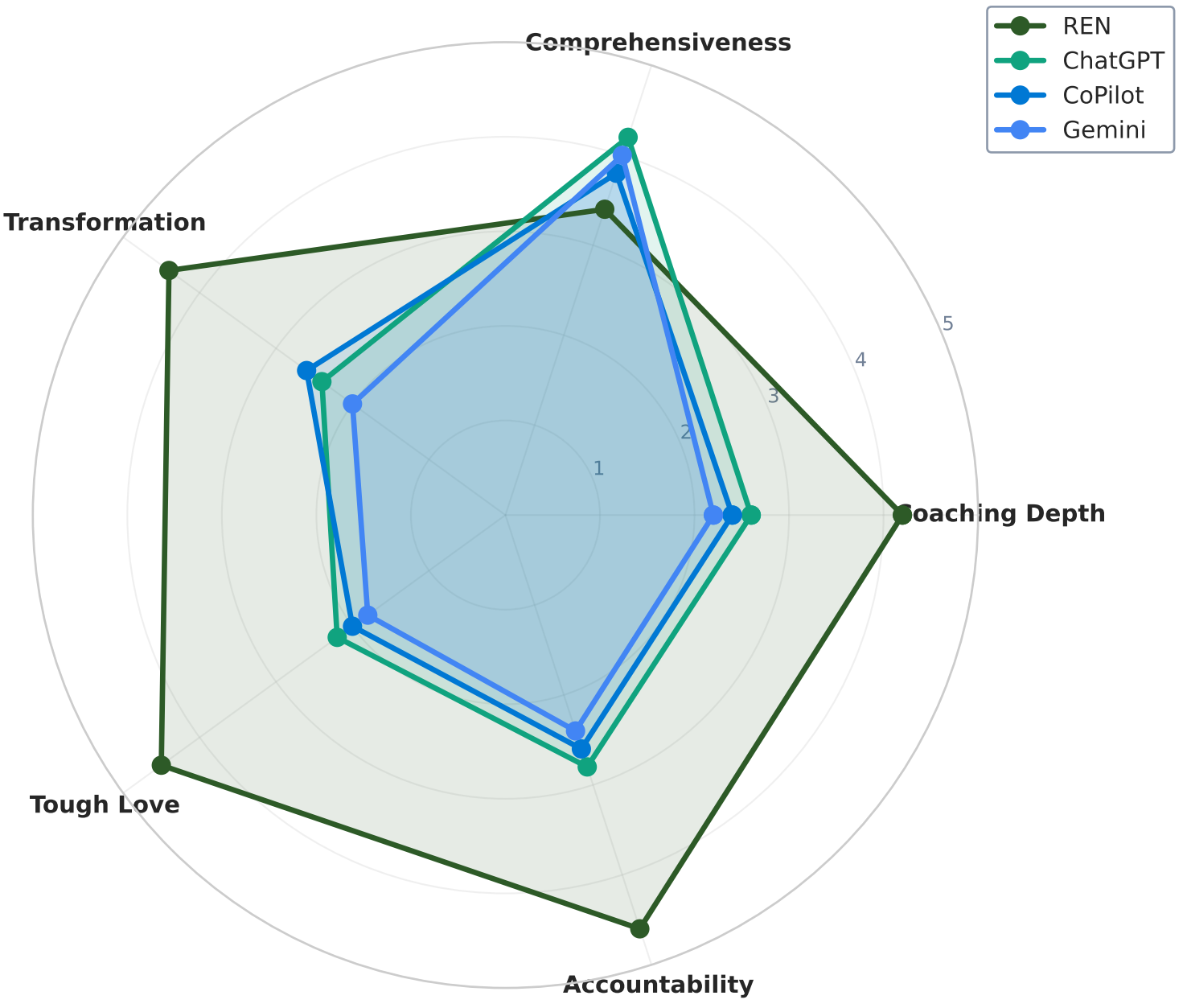
Largest gap:

Tough Love Δ = 2.7 pts

Scores by Dimension

Dimension	REN	ChatGPT	CoPilot	Gemini	Leader	Gap (Δ)
Coaching Depth	**4.20**	2.60	2.40	2.20	REN	2.00
Comprehensiveness	3.40	**4.20**	3.80	4.00	ChatGPT	0.80
Transformation	**4.40**	2.40	2.60	2.00	REN	2.40
Tough Love	**4.50**	2.20	2.00	1.80	REN	**2.70**
Accountability	**4.60**	2.80	2.60	2.40	REN	2.20

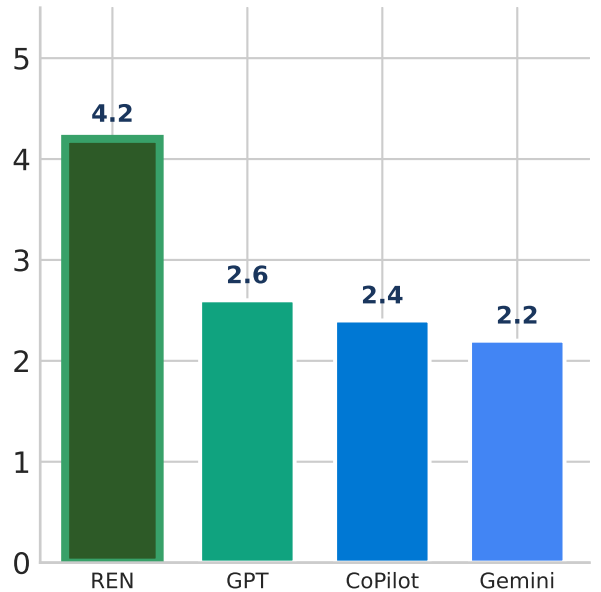
Dimensional Comparison



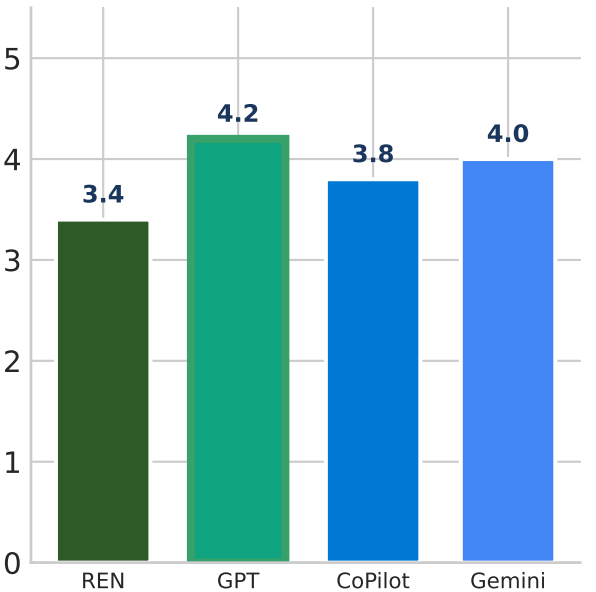
REN forms a large, balanced pentagon (4.2–4.6 range). General-purpose models are strongest on Comprehensiveness (~4.0).

Performance by Dimension

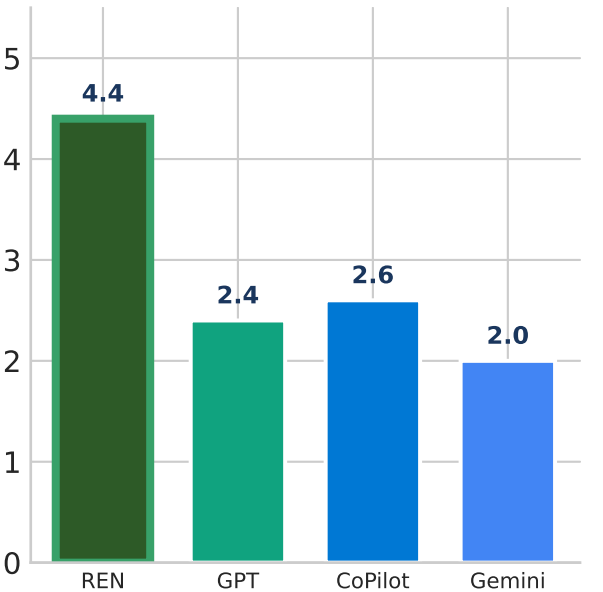
Coaching Depth



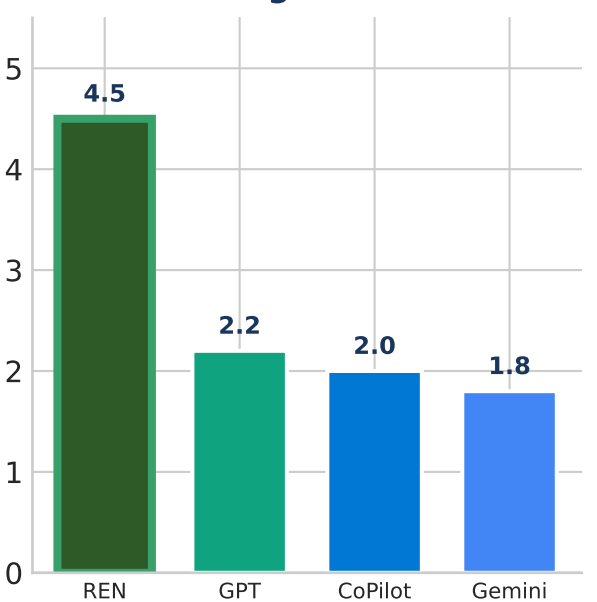
Comprehensiveness



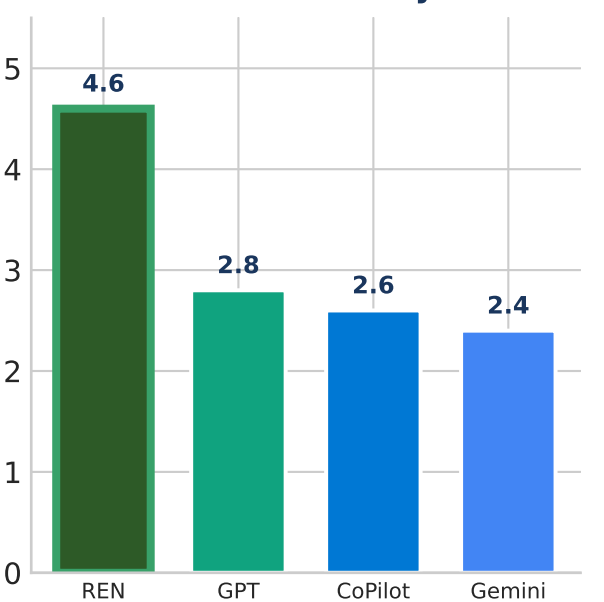
Transformation



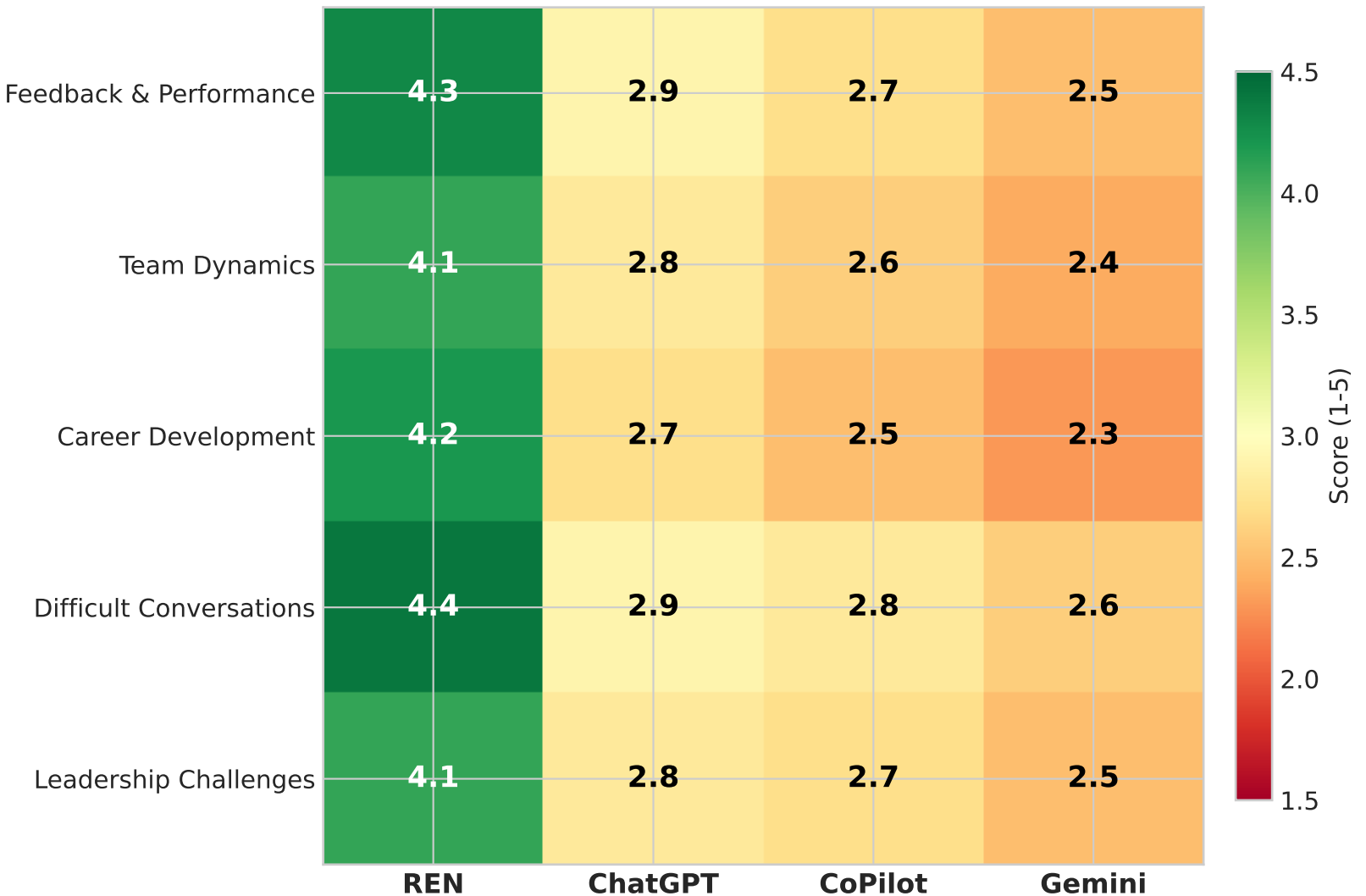
Tough Love



Accountability

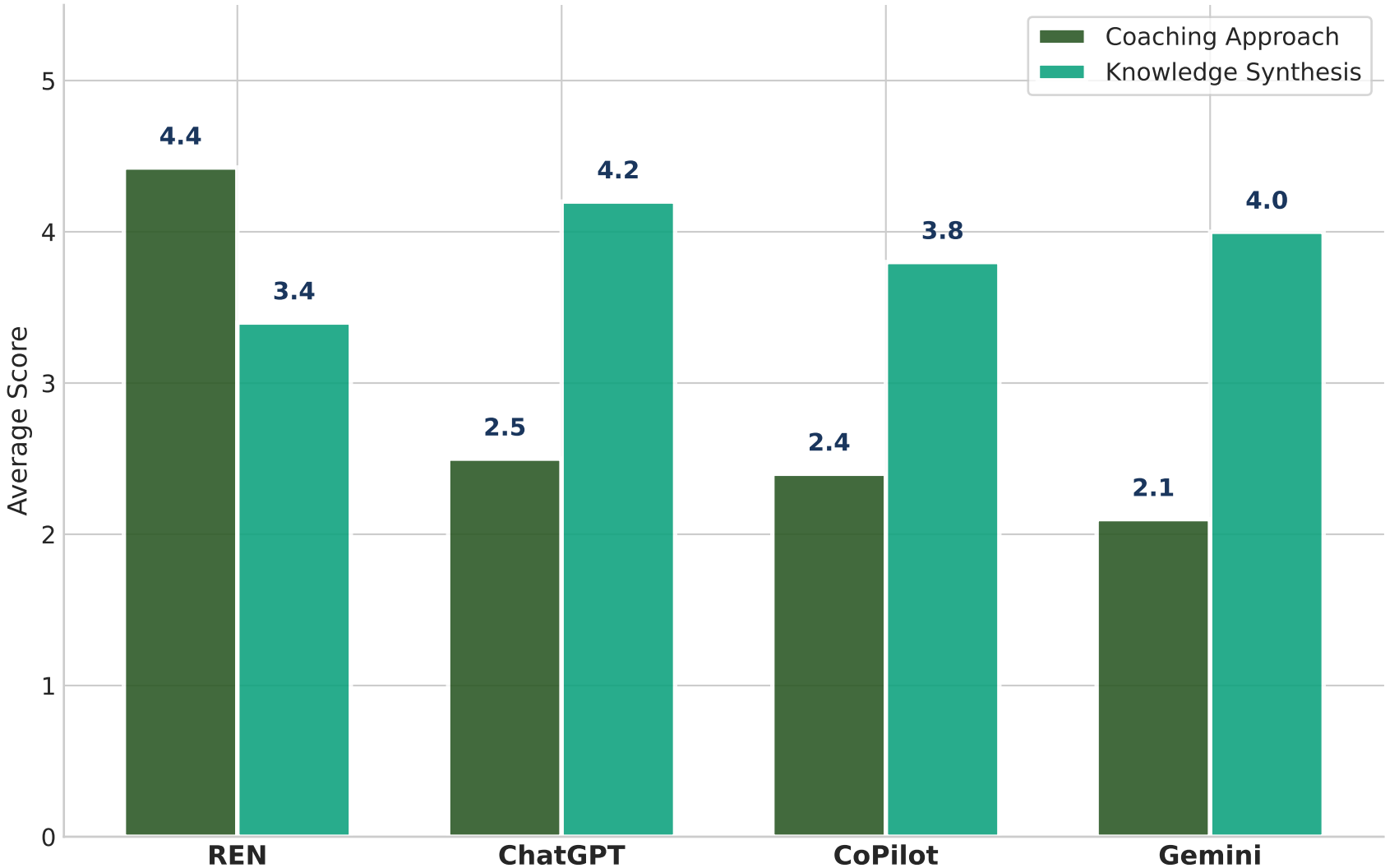


Performance by Scenario Category



REN shows consistent high performance across all scenario categories.

Coaching Approach vs. Knowledge Synthesis



Coaching = Depth + Transformation + Tough Love + Accountability | Knowledge Synthesis = Comprehensiveness

Statistical Significance

REN vs ChatGPT

Mean Difference: 1.38 • t = 2.48 • p = 0.0684

95% CI: [0.40, 2.36] • Cohen's d: 2.35 (large effect)

✓ SIGNIFICANT

REN vs CoPilot

Mean Difference: 1.54 • t = 3.07 • p = 0.0373

95% CI: [0.66, 2.42] • Cohen's d: 2.94 (large effect)

✓ SIGNIFICANT

REN vs Gemini

Mean Difference: 1.74 • t = 2.92 • p = 0.0433

95% CI: [0.69, 2.79] • Cohen's d: 2.75 (large effect)

✓ SIGNIFICANT

All comparisons: statistically significant ($p < 0.05$) with large effect sizes ($d > 2.0$)

Methodology

Testing Protocol

- Each scenario tested in incognito browser with no prior history
- Each session: 1 initial question + 3 follow-up questions (4 turns)
- Sessions conducted independently to prevent cross-contamination
- All responses recorded verbatim for analysis

Multi-Turn Design

Evaluates performance across extended conversations, revealing methodology differences that emerge through sustained dialogue.

Evaluation Dimensions

1. Coaching Depth — Diagnostic questioning; surfacing root causes
2. Comprehensiveness — Breadth of frameworks and best practices
3. Transformation — Interrupting patterns; producing lasting change
4. Tough Love — Direct confrontation of avoidance & self-responsibility
5. Accountability — Returning responsibility to user vs. prescriptive advice

Scoring Scale

1 = Minimal | 2 = Below Average | 3 = Adequate | 4 = Strong | 5 = Exceptional

Key Insights & Recommendations

1

Coaching vs. Consulting Divide

REN: question-led, pattern-confronting, ownership-building

General models: framework-heavy, solution-providing, externally directive

2

Largest Differentiator: Tough Love (Δ = 2.7)

REN names avoidance patterns and invites self-examination

General models remain supportive and validating

3

Strength of General Models

Excel at tactical needs: scripts, frameworks, best practices

CoPilot offers warmest, most accessible tone

STRATEGIC RECOMMENDATION

General-purpose AI

Efficiency • Education • Templates

REN (Specialized Coaching)

Transformation • Pattern Interruption • Growth