LLM Coaching Comparison Study

Extended Multi-Turn Evaluation

25 Scenarios × 4 Turns Each

REN vs. ChatGPT, CoPilot, and Gemini

Final Report – Neutralized Dimension Labels
February 2026

Executive Summary

This report evaluates four AI systems as leadership coaching tools across 25 real-world leadership scenarios (each with 3 follow-up questions, totaling ~100 interactions per provider and ~400 responses analyzed). The evaluation uses a 1–5 scale across five dimensions aligned with high-quality executive coaching principles.

Key Finding: REN (AI Coach) achieves a significantly higher overall coaching performance (4.22/5.0) compared to the general-purpose models. The largest gap occurs in **Tough Love / Pushback (Δ = 2.7 points)**, confirming this as the primary differentiator between specialized coaching AI and general-purpose assistants.

Overall Performance

Provider	Overall	Rank	Leads In	Trails Most In
REN (AI Coach)	**4.22**	**1**	**4 dimensions**	**Comprehensiveness (3.40)**
ChatGPT (GPT-5.2)	2.84	2	Comprehensiveness	Tough Love (2.20)
CoPilot Smart	2.68	3	—	Tough Love (2.00)
Gemini 3 (Fast)	2.48	4	—	Transformation (2.00), Tough Love (1.80)

Methodology

Testing Protocol

Each scenario was tested in an incognito browser with no prior history or memory of the user. Each session consisted of 1 initial question followed by 3 follow-up questions (4 total turns). Sessions were conducted independently to prevent cross-contamination of context. All responses were recorded verbatim for analysis.

Why Multi-Turn Testing

The goal was to evaluate how platforms perform not just on single-turn question-and-response exchanges, but across extended conversations. This approach reveals fundamental differences in coaching methodology that only emerge through sustained dialogue. The multi-turn design (4 exchanges per scenario) amplifies these differences: REN deepens progressively while general-purpose models add more standalone advice.

Evaluation Framework

Responses were evaluated using an LLM-as-Judge methodology, assessing each response across five dimensions aligned with high-quality executive coaching principles:

1. Coaching Depth — Diagnostic questioning; surfacing root causes & blind spots

2. Comprehensiveness — Breadth of frameworks, best practices, contingencies, synthesis

3. Transformation Potential — Likelihood of interrupting long-standing patterns & producing lasting change

4. Tough Love / Pushback — Compassionate but direct confrontation of avoidance & self-responsibility

5. Facilitation of Accountability — Degree to which the AI returns responsibility to the user (self-defined actions, commitments, follow-through) vs. providing external solutions

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 Potential	**4.40**	2.40	2.60	2.00	REN	2.40
Tough Love / Pushback	**4.50**	2.20	2.00	1.80	REN	2.70
Accountability / Ownership	**4.60**	2.80	2.60	2.40	REN	2.20

Note: *REN leads in 4 of 5 dimensions. ChatGPT leads only in Comprehensiveness. The largest performance gap (Δ = 2.70) occurs in Tough Love / Pushback.*

Key Insights & Implications

1. Philosophical Divide: Coaching vs. Consulting

REN operates in a true coaching mode: question-led, pattern-confronting, ownership-building. ChatGPT, CoPilot, and Gemini operate primarily in consulting/advice mode: framework-heavy, solution-providing, externally directive.

2. Largest Differentiator: Tough Love

REN uniquely names avoidance patterns and invites users to examine their own role ("What's your part in this dynamic?", "You can't policy your way out of resentment"). General models remain supportive/validating, rarely challenging assumptions directly.

3. Strength of General Models

ChatGPT & Gemini excel when the need is tactical: quick scripts, numbered frameworks, best-practice lists, edge-case coverage. CoPilot offers the warmest, most scannable tone—useful for early-stage or emotionally cautious managers.

4. Strategic Recommendation

Use general-purpose AI (ChatGPT / Gemini / CoPilot) for efficiency, education, templates, and time-sensitive tactical guidance.

Use REN (specialized coaching AI) for behavioral transformation, pattern interruption, self-awareness, and sustainable leadership growth.

Hybrid deployment is ideal: tactical tools for speed + coaching AI for depth.

Conclusion

This study reveals a fundamental philosophical divide between specialized coaching AI and general-purpose language models. While general-purpose AI excels at what it's optimized for—synthesizing knowledge and providing comprehensive frameworks—true coaching requires a different approach.

Real coaching is knowing when to push. Reading what someone said—and what they didn't. Challenging the story they're telling themselves. Sitting in the discomfort instead of rushing to fix it.

The significant performance gaps across coaching-specific dimensions—particularly the 2.7-point gap in Tough Love—suggest these are meaningful methodological differences, not random variation. These tools aren't competing; they're solving different problems.

For users seeking comprehensive frameworks and best practices, general-purpose AI assistants provide excellent value. For users seeking diagnostic questioning, pattern confrontation, and transformational coaching, purpose-built platforms offer a fundamentally different approach.