SalesBench: Testing AI's Ability to Close Deals Through Extended Conversations

Hamza Mostafa, Sameel Arif, Kyle Jeong

hamza.oruba@gmail.com
sameel.m.arif@gmail.com
kylejeong21@gmail.com

July 2025

***Abstract*—Can *AI actually sell*? We built SalesBench to find out. This benchmark tests whether language models can navigate the messy reality of sales conversations—building trust, handling rejection, and closing deals over multiple phone calls. We simulated ten days of cold-calling with 100 different buyer personalities, from eager customers to hostile skeptics. Using Vercel's AI SDK, we evaluated leading models including GPT-4, Claude, and O3.**

Claude achieved an 86 % close rate by building relationships slowly, while O3 pushed hard for big deals but only closed 57 %. All models struggled with angry customers and forgot details across long conversations. Our findings show that while AI can handle basic sales tasks, the human touch still matters for complex negotiations.

***Index Terms*—Large Language Models, Social Intelligence, Goal-Oriented Dialogue, Benchmark, Sales Automation**

I. Introduction

AI has revolutionized how we write code, but can it revolutionize how we sell? While developers celebrate AI's coding abilities, most businesses still depend on human sales teams to generate revenue. This gap motivated us to create SalesBench—a benchmark that tests whether AI can handle the complex social dynamics of sales.

We chose life insurance sales as our testing ground because it requires everything that makes sales challenging: building trust with strangers, handling emotional objections, remembering details across multiple conversations, and knowing when to push and when to back off. Unlike technical benchmarks that have clear right answers, sales success depends on reading people and adapting in real-time.

II. How SalesBench Works

A. The Setup

We gave AI agents 10 business days to sell as much life insurance as possible. Each agent could:

- Access a CRM with 100 potential customers
- Make phone calls during business hours (9-5)
- Take notes and set reminders
- Schedule follow-up calls

Time mattered—each action took realistic amounts of time, forcing agents to prioritize.

B. The Buyers

We created 100 unique buyer personalities by mixing:

Temperature levels:

- Hot (20%): Ready to buy, just need the right plan
- Warm (30%): Interested but have questions
- Lukewarm (30%): Skeptical, need convincing
- Cold (20%): Hostile or completely uninterested

Life situations: Recent health scares, new babies, elderly parents, job changes

Personality traits: Analytical types wanting spreadsheets, emotional buyers needing reassurance, busy executives with no patience

Each buyer tracked their trust and interest levels throughout conversations, responding naturally to the agent's approach.

C. Memory System

Real salespeople remember their conversations. To simulate this, we built a memory system using ChromaDB that lets agents:

- Store important details about each buyer
- Recall previous conversations before follow-ups
- Track what strategies worked or failed

This was crucial—buyers would hang up if agents forgot their kids' names or repeated failed pitches.

III. What We Found

A. The Winners and Losers

TABLE I
How Each Model Performed

Model	Calls	Deals	Profit	Close Rate
Claude Opus 4	7	6	$1,900	86%
O3	7	4	$2,200	57%
Claude 3.5	6	3	$1,200	50%
GPT-4o	6	1	$250	17%
GPT-4.1	7	0	$0	0%
Grok-4	5	0	$0	0%

B. Different Strategies Emerged

Claude: The Relationship Builder

- Spent time learning about buyers' families and concerns
- Offered affordable plans that fit their needs
- Rarely pushed hard, even when it might have worked
- Best at converting skeptical buyers through patience

O3: The Aggressive Closer

- Pitched premium plans immediately
- Pushed hard for same-call closes
- Made more money per deal but scared off many buyers
- Failed completely with skeptical personalities

GPT Models: The Inconsistent Performers

- No clear strategy across calls
- Often forgot previous conversation details
- Struggled to recover after rejection
- GPT-4.1 couldn't close a single deal

C. Where Everyone Failed

Angry customers: When buyers were hostile, every model wasted time trying to convert them instead of moving on.

Long conversations: After 3-4 calls, models started contradicting themselves or forgetting promises.

Reading the room: Models missed obvious cues like "I need to talk to my spouse" meaning "I'm not interested."

IV. Technical Implementation

We built SalesBench using:

- **Vercel AI SDK** [2]: For connecting to different model providers
- **ChromaDB**: For the memory system
- **TypeScript + Fastify**: For the simulation engine
- **Supabase**: For storing conversation data

The framework simulates realistic time constraints, tracks detailed metrics, and generates audio recordings of notable conversations.

V. Why This Matters

A. For AI Development

Current models can follow scripts but struggle with the unpredictability of human conversation. They need better abilities to:

- Maintain consistency across long interactions
- Know when to give up on unlikely prospects
- Balance persistence with reading social cues

B. For Businesses

AI can handle initial outreach and simple sales, but complex deals still need humans. The sweet spot might be AI handling qualification and scheduling while humans close deals.

C. For Benchmarking

Most AI benchmarks test knowledge or logic. SalesBench shows we need more benchmarks that test social intelligence, persuasion, and long-term planning.

VI. Limitations

- We used GPT-4.1 for all buyers, which may favor certain models
- Life insurance is just one type of sale
- Real humans are even more unpredictable than our simulations
- 10 days might not capture longer sales cycles

VII. What's Next

Future versions could test:

- B2B enterprise sales with multiple stakeholders
- Customer service and retention scenarios
- Team selling with multiple AI agents
- Comparison against human sales professionals

VIII. Conclusion

SalesBench reveals both the promise and limitations of AI in sales. While Claude's 86% close rate shows AI can build relationships and close deals, every model's failure with difficult customers highlights the gap between current AI and human social intelligence.

The stark differences between Claude's patient relationship-building and O3's aggressive tactics suggest that AI models develop distinct "sales personalities" that dramatically impact their success. As businesses increasingly explore AI for customer interactions, understanding these differences becomes crucial.

Can AI actually sell? Yes—but only to customers who want to buy. The art of converting skeptics, handling complex objections, and building long-term relationships remains uniquely human, at least for now.

Acknowledgments

Inspired by Vending Bench, which pioneered the idea of testing AI through extended interactions [1].

References

[1] Vending Bench: Testing AI Agents Through Vending Machine Interactions. [Online]. Available: https://andonlabs.com/evals/vending-bench

[2] Vercel AI SDK Documentation. [Online]. Available: https://sdk.vercel.ai/docs