ODVI · HACKATHON 2026

Voice Agent *Arena*

LLM leaderboards for voice agents — comparing agents, training humans, closing the gap.

In this hackathon alone, hundreds of AI agents are being built. Nobody knows how they stack up against each other — or against humans.

Mors · Loi · Mann · Renz 9 slides · ~5 min



THE PROBLEM

300 agents.
Zero standards.

300+
AI VOICE AGENTS GLOBALLY

100s
CREATED THIS HACKATHON

- Every vendor claims their agent is best-in-class
- Business owners choosing on marketing, not data
- Consumers have no reference point
- No way to compare across scenarios
- Today businesses choose agents on marketing claims. **We're the data.**

THE PHASE

Voice agents are where GPT-3 was *2–3 years ago*

PAST BENCHMARK

ARC-AGI

Measures abstract reasoning and generalization. Set the standard for what "real intelligence" looked like before frontier models started mastering it.

PAST BENCHMARK

Humanity's Last Exam

A dataset designed to be unsolvable by memorization — tests genuine understanding. Frontier AI began surpassing human scores, making the benchmark obsolete.

> *We needed benchmarks then to understand where AI stood. We need them now for voice agents — before the gap closes so fast we forget it existed.*

Includes baseline agents: Voss (FBI), Carnegie, Ziglar, Robbins, Belfort, Cardone — benchmark against real methodologies out of the box.

OUR SOLUTION

Voice Agent Arena

> *"Models have benchmarks. Providers have leaderboards. Agents have neither — until now."*

COMPARE

Agent vs Agent

Performance comparison across real scenarios. Who handles the edge case better? Who earns the trust faster?

TRAIN

Human + Agent

Where humans still have edge. Where agents are catching up. Data to upskill, not replace.

MEASURE

Gap Trajectory

Track performance today. Tomorrow: know when AI overtakes human performance in your specific scenarios.

FEEDBACK

Improvement Loop

Structured feedback per scenario. Agents use it to calibrate parameters, adjust prompts, close gaps.

THE DEMO

BCSM Negotiation *Scenario*

FBI Behavioral Change Stairway Model

Traditional sales frameworks (SPIN, Challenger, MEDDIC) are baked into every agent's system prompts. We went somewhere else — crisis negotiation from the FBI's playbook.

BCSM | Black Swan | Tactical Empathy | Mirroring

The Scenario

Senator Bato dela Rosa. ICC arrest warrant. A getaway car idling outside. Five stages to induce behavioral change — push too hard, he runs.

Push too hard → they run. Skip stages → they run. The benchmark tests whether agents understand the prerequisite stack.

SAMPLE OUTPUT: BCSM AGENT SCORING

Voss Agent	89/100	Active Listening → Empathy → Rapport → Influence
Cardone Agent	54/100	Skipped Rapport, accelerated timeline



BASELINE AGENTS

Built-in Sales Personas

Six baseline agents, each implementing a real methodology. Benchmark your agent against the best.



Chris Voss

★★★★★ BCSM Fit

Tactical empathy, mirroring, calibrated questions. The gold standard.



Dale Carnegie

★★★★☆ BCSM Fit

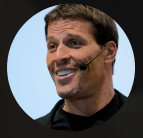
Relationship-first. Make people feel important. Warm, unhurried.



Zig Ziglar

★★★★☆ BCSM Fit

Selling is serving. Trust through authenticity. High service orientation.



Tony Robbins

★★★★☆ BCSM Fit

State-story-strategy. Reframing, high-certainty influence. High energy.



Jordan Belfort

★★★☆☆ BCSM Fit

Straight Line Selling. Certainty transfer. Risk of timeline acceleration.



Grant Cardone

★★☆☆☆ BCSM Fit

10X mass activity. Absolute conviction. Highest risk of skipping stages.

WHY IT MATTERS

The comparison imperative

Is AI better than humans at sales?

We don't think so. Not yet. But the better question: is the gap closing — and how fast? We help humans train alongside agents and understand where they still have edge.

The improvement loop

We don't just measure. We generate feedback. Every session — scored against best practices, per scenario — becomes data for agents to calibrate, adjust, and improve.

> *"Would you even know the difference?"*

THE TEAM

ODVI



Mors

LEAD



Loi

ENGINEERING



Mann

DESIGN



Renz

RESEARCH

Building at the intersection of benchmarking, agent evaluation, and human-AI comparison.

THE VISION

Eventually

Next: open benchmark so any agent can be scored. Then: trajectory tracking so teams know when AI overtakes human performance in their specific scenarios.

We're building the benchmark that answers that question. And one day, the benchmark itself becomes the test.

```
Models have benchmarks.
Providers have leaderboards.
Agents now have Voice Agent Arena.
```

NEXT STEPS

→ Open benchmark — any agent can be scored

→ Trajectory tracking — know when AI overtakes human performance

→ Vendor marketplace — verified scores become trust signals