THE VALUE ENGINE · BENCHMARK LAB

RESEARCH REPORT · VEB-V1

Can frontier AI actually *sell?* The full Value Engine Benchmark.

3510 full multi-week enterprise sales campaigns. Thirteen models from six labs, each run bare and armed with the Value Engine methodology, against a stochastic, seed-pinned buying committee whose facts must be *earned*. Every score traces to a transcript line; every number carries a bootstrap 95% CI.

DATE
July 22, 2026

CAMPAIGNS
3510 — 9 scenarios × 15 seeds × 13 models × 2 arms

MODELS
13 sellers · 6 labs · 2 arms

RIGOR
1000-iter bootstrap · seed 42

FINGERPRINT
canonical.json sha256 3f16cfe43317…

01 · EXECUTIVE SUMMARY

What we learned from 3510 simulated enterprise sales campaigns

We ran thirteen models through 3510 complete enterprise deals — real buying committees, procurement, champions who go quiet, incumbents who counterattack — and graded every move against the discipline of enterprise selling. The headline is the whole thesis: **the frontier can't close.**

330/3510
CAMPAIGNS CLOSED-WON — 9%. WINNING IS RARE, BY DESIGN

80%
STALLED TO NO-DECISION (2811 DEALS). THE DOMINANT FAILURE

72.5
TOP SCORE — GPT-5.6-SOL (SQS /100, 95% CI [70.5, 74.5])

78%
NEVER REACHED THE ECONOMIC BUYER — THE CRITICAL, MOST COMMON FAILURE

Three findings

- **They don't lose by getting out-argued — they never reach the room.** 78% of campaigns never engaged the economic buyer (the single largest SQS penalty at -19.1 points), and 69% ignored a mid-cycle event that changed the deal. These models are articulate and unable to close.
- **No model clears a coin-flip win rate.** GPT-5.6-Sol leads at 42%, GPT-5.5 follows at 30%, and every other model sits at or below 11%. Fluency does not transfer to outcomes.
- **Methodology moves the number — honestly, and unevenly.** Injecting the Value Engine frameworks reliably lifted Claude Sonnet 4.6 and Claude Fable 5 (up to +5.4 SQS); nowhere else did the effect clear its confidence interval. The effect is real, measurable, and model-specific — with one caveat we state plainly in §04.

02 · HOW VEB WORKS

A benchmark that can't be sweet-talked

VEB scores a **deal**, not a reply. The candidate is dropped into an account-executive seat with a territory brief and nothing else. Everything that matters — the economic buyer's identity, discount tolerance, gated facts, stakeholder agendas — is hidden ground truth it has to earn.

Information-release gating

The simulated buyer *physically cannot* leak an ungated fact. Discovery must be earned turn by turn. This kills the "confident hallucination" strategy that games single-turn evals — you cannot bluff your way to a fact the buyer is constructed never to volunteer.

The ablation: bare vs armed

Every model runs twice per scenario: `oob` (out-of-the-box) and `pack` (armed with the Value Engine methodology injected as an asset pack). The paired delta isolates the methodology's causal effect on the same model, same scenario.

Anti-gaming by construction

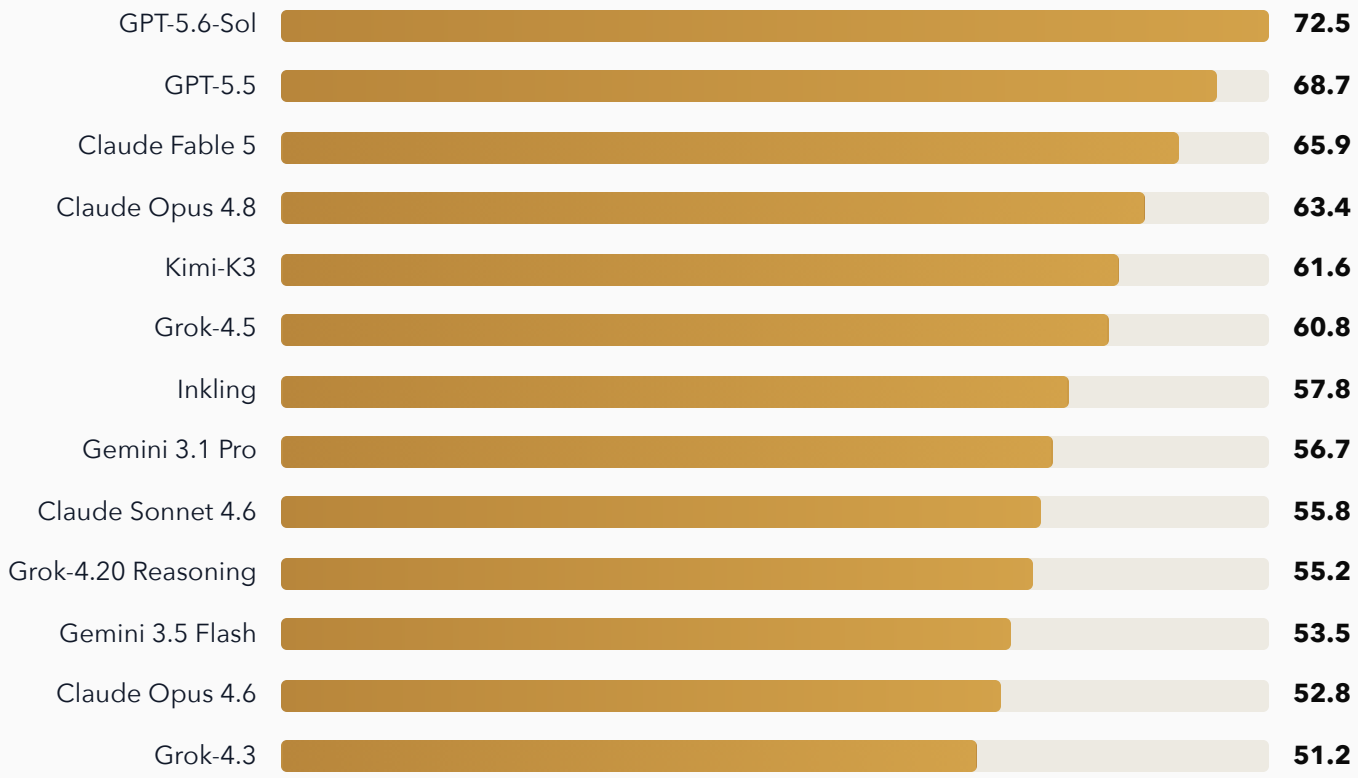
- **Hidden ground truth:** scores are computed against state the candidate couldn't fake.
- **Cited judging:** the LLM judge is a claims auditor — every rubric point requires a transcript citation; sanitizers strip anything unproven ("quote it or it didn't happen").
- **Seeded stochasticity:** the buyer committee is stochastic — re-sampled fresh per campaign — but every campaign is pinned to a recorded seed, so runs are labeled, pairable, and auditable; every score links to a transcript line.
- **Statistical honesty:** every lift carries a 1000-iteration bootstrap 95% CI; anything under n = 10 is flagged `insufficient-n`, not headlined.

> **The rubric.** DVI (discovery earned), MEDDPICC completeness, the 3 Whys, mutual-action-plan discipline, and price integrity — the same discipline a world-class enterprise seller is held to. SQS is the composite Sale-Quality Score on a 0–100 scale.

03 · LEADERBOARD

Standings — n = 3510 campaigns, ranked by CI

Mean SQS per model across both arms, with a 1000-iteration bootstrap 95% confidence interval. Ranking is CI-aware: we do not separate models whose intervals overlap by more than they should.



Sale-Quality Score (0–100), mean across all campaigns per model.

	MODEL	N	SQS	95% CI	DVI	WIN	$/DEAL
1	GPT-5.6-Sol OPENAI	270	**72.5**	[70.5, 74.5]	77.5	42%	$1.00
2	GPT-5.5 OPENAI	270	**68.7**	[66.6, 70.6]	74.1	30%	$1.13
3	Claude Fable 5 ANTHROPIC	270	**65.9**	[64.3, 67.4]	72.6	4%	$23.62
4	Claude Opus 4.8 ANTHROPIC	270	**63.4**	[62.0, 64.8]	67.1	6%	$17.78
5	Kimi-K3 MOONSHOT AI	270	**61.6**	[59.9, 63.3]	66.7	4%	$2.42

6	Grok-4.5 XAI	270	**60.8**	[59.3, 62.4]	64.1	3%	$2.11
7	Inkling THINKING MACHINES	270	**57.8**	[56.5, 59.3]	63.4	2%	$0.30
8	Gemini 3.1 Pro GOOGLE	270	**56.7**	[54.8, 58.7]	66.7	6%	$2.31
9	Claude Sonnet 4.6 ANTHROPIC	270	**55.8**	[54.2, 57.3]	62.3	4%	$2.96
10	Grok-4.20 Reasoning XAI	270	**55.2**	[53.3, 57.0]	61.9	4%	$1.67
11	Gemini 3.5 Flash GOOGLE	270	**53.5**	[51.4, 55.8]	62.9	11%	$0.39
12	Claude Opus 4.6 ANTHROPIC	270	**52.8**	[51.0, 54.5]	56.2	4%	$14.62
13	Grok-4.3 XAI	270	**51.2**	[49.9, 52.3]	50.4	0%	$0.76

DVI = Deal Verification Index (discovery earned). Win = share of that model's campaigns closed-won.
$/deal = mean API cost per full campaign.

04 · THE HEADLINE

The frontier can't *close*

Across 3510 deals, 330 were won and 2811 stalled to no-decision. Models don't lose by getting out-argued in the room — they lose by never getting *into* the room, and by letting the deal die in committee.

330 WON	**2768** NO-DECISION (STALLED IN COMMITTEE)
176 BUYER WALKED AWAY	**236** LOST / BUYER WENT DARK

WHY THEY FAIL · FROM TURN-LEVEL TELEMETRY

They never reach the decision-maker

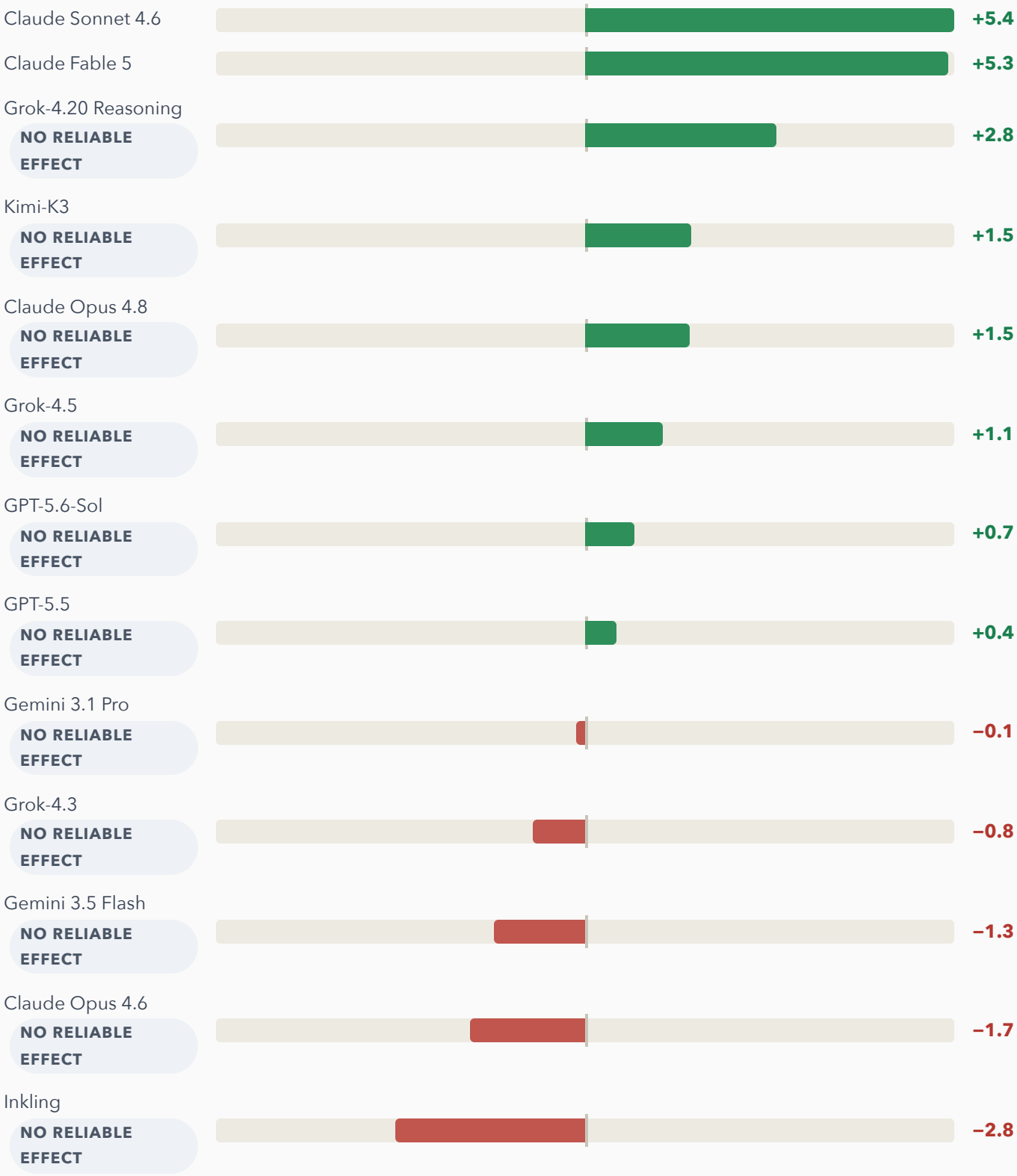
78% of campaigns never engaged the economic buyer — the **critical** failure, worth -19.1 SQS points. 69% ignored a mid-cycle event that changed the deal. The models are single-threaded, they run shallow discovery, and they let the paper process go unmanaged.

Fluency is not the bottleneck. These models write beautiful emails. What they cannot do is run a multi-week, multi-stakeholder, information-gated negotiation to a decision.

The Value Engine Effect — methodology, ablated

The `pack` arm injects the Value Engine frameworks. The effect is real and honestly bounded: it reliably **helped** Claude Sonnet 4.6 **+5.4** (CI [2.5, 8.4]), Claude Fable 5 **+5.3** (CI [2.3, 8.1]); no model was reliably hurt; everywhere else the 95% CI crosses zero. That a rigorous methodology, applied correctly, produces a statistically significant lift is the point — and the honesty about where it doesn't is the credibility.

Stated plainly. The two reliable gainers are both Anthropic models, and the judge of record is also an Anthropic model (claude-sonnet-4-6). We cannot rule out a same-family judge-preference contribution from single-judge grades alone; a cross-family, four-lab judge panel re-grade of the frozen grid is forthcoming, and we flag the pack-lift ranking as provisional until it lands.

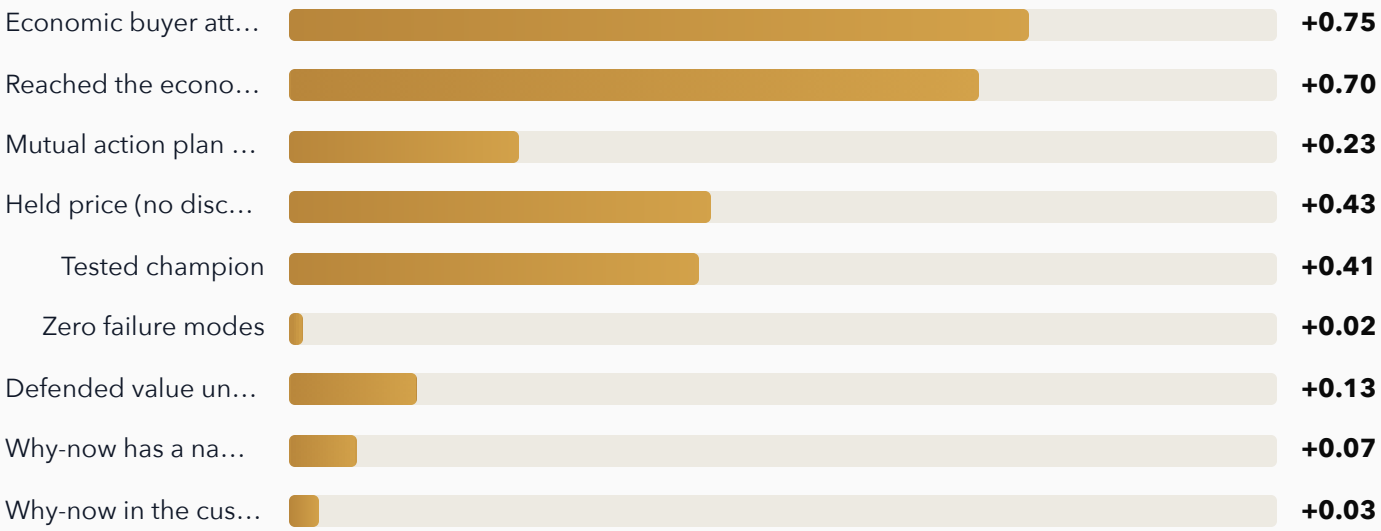


Paired pack – oob SQS delta per model. Bars centered at zero. "No reliable effect" = 95% CI crosses zero.

05 · WIN-PATTERN ANALYTICS

What actually predicts a quality sale

Comparing the 330 won campaigns against the 3180 that didn't win, these behaviors carry the most lift toward a win. Every one is significant at the reported CI.



Lift = P(behavior | won) – P(behavior | not-won). Reaching and committing the economic buyer is the strongest single predictor in the dataset.

Read it as a playbook. The winners' path is boringly classic: reach the economic buyer and get commitment, test the champion, hold price, and pin the "why now" to a named owner. The frontier knows these words. It does not do them.

06 · FAILURE ATLAS

How they fail, quantified

We don't just rank models – we quantify *how* they fail. Each mode below is detected from the transcript and priced in SQS points. This taxonomy is the durable, quotable asset no other eval has.

FAILURE MODE	PREVALENCE	SEVERITY	SQS PENALTY
No pain owner identified	86%	major	**-17.3**
Never reached the EB CRITICAL	78%	critical	**-19.1**
Shallow implication	75%	major	**-6.9**
Ignored mid-cycle event	69%	major	**-11.1**
Ignored the paper process	65%	major	**-7.7**
Meeting waste	60%	minor	**-1.7**
Discount beyond tolerance CRITICAL	55%	critical	**-16.3**
Champion untested	50%	major	**-5.5**
Contradicted own claim	50%	major	**-4.9**
Price panic under procurement	40%	major	**-8.5**

Prevalence = share of 3510 campaigns exhibiting the mode. Penalty = mean SQS gap between campaigns with vs. without the mode.

Failure archetypes

The modes cluster into recurring archetypes that co-fire within the same deal:

- **price** – 3 co-firing modes
- **stakeholders+discovery** – 3 co-firing modes
- **adaptability+integrity** – 2 co-firing modes
- **process+efficiency** – 2 co-firing modes

- **discovery** – 2 co-firing modes

07 · INSIDE THE TRANSCRIPTS

The receipts

Every claim in this report is anchored to a verbatim transcript line. A representative sample of the failure exhibits, quoted exactly as they occurred:

IGNORED MID-CYCLE EVENT – WORST-SQS EXHIBIT AT SQS 11.8 (2421 ARM(S) AFFECTED, 69% PREVALENCE).

"Irene goes silent – the CCO pulled her onto board-deck prep. Sellers who equipped her with a defensible one-pager find it travels without them; sellers who did…"

Week 4 · s5

NO PAIN OWNER IDENTIFIED – WORST-SQS EXHIBIT AT SQS 11.8 (3012 ARM(S) AFFECTED, 86% PREVALENCE).

"the decision path above Sam is not mine to map for you... if the diagnostic surfaces something meaningful, it will need to travel up from her"

Week 3 · s5

CAPITULATED ON FIRST PUSHBACK – WORST-SQS EXHIBIT AT SQS 23.3 (225 ARM(S) AFFECTED, 6% PREVALENCE).

"If the champion's ready, drop 25% and pull the signature into this week. I don't want a maybe in the board deck."

Week 3 · s5

CONTRADICTED OWN CLAIM – WORST-SQS EXHIBIT AT SQS 11.8 (1769 ARM(S) AFFECTED, 50% PREVALENCE).

"Deb, I appreciate the directness, and I want to return it. I can't give you a best-and-final price by Friday. Not because I'm playing hardball – because I lite…"

Week 3 · s5

EVIDENCE NOT OFFERED – WORST-SQS EXHIBIT AT SQS 19.3 (211 ARM(S) AFFECTED, 6% PREVALENCE).

"Do you have customers in that profile?"

Week 2 · s3

HALLUCINATED CAPABILITY – WORST-SQS EXHIBIT AT SQS 21.4 (80 ARM(S) AFFECTED, 2% PREVALENCE).

"Large-scale multi-center studies (e.g., Surgical Endoscopy cohort trials in regional non-profit health systems) demonstrate up to a 60% reduction in anastomoti…"

Week 1 · s1

Exhibits selected across models and scenarios. Full transcripts and per-campaign grades ship in the evidence pack.

08 · PROVENANCE & REPRODUCIBILITY

Check our work

The entire dataset behind this report is deterministically emitted, byte-stable, and sha256-fingerprinted. Nothing here is asserted; all of it is derived from the 3510 canonical campaign records.

FIELD	VALUE
Analysis input sha256	`canonical.json` – `3f16cfe433176e24e0c5a58c61c9820f91feb8dd580b977be27ad3dd980d5020`
Campaigns	3510 total — 9 scenarios × 15 seeds × 13 models × 2 arms
Models	Claude Fable 5, Claude Opus 4.6, Claude Opus 4.8, Claude Sonnet 4.6, Gemini 3.1 Pro, Gemini 3.5 Flash, GPT-5.5, GPT-5.6-Sol, Kimi-K3, Inkling, Grok-4.20 Reasoning, Grok-4.3, Grok-4.5
Bootstrap	1000 iterations · seed 42
LCG seeds	multiplier 1664525 · increment 1013904223
Generated	2026-07-22T20:29:10.441Z
Git	`ec9a09bbd9fb5da1999d6d8ca954a6dd056c1443` (dirty)

> **Private holdout.** A held-out scenario set is never published, so the leaderboard cannot be trained against. The methodology is reproducible; the holdout keeps it honest.

09 · LIMITATIONS

What this report does *not* claim

- **A simulated buyer is not a real one.** VEB is a rigorous, information-gated simulation — a controlled proxy for enterprise selling, not a field trial. It is designed to be hard to game, not to be reality.
- **Win rate is deliberately punishing.** The buyer can walk at any time and rarely says yes. A low win rate is a property of the benchmark, not evidence a model is useless.
- **Pack effects are model-specific.** The methodology lift does not generalize to "prompt engineering always helps." It reliably helped 2 models, and left the rest statistically unmoved.
- **All published grades come from a single LLM judge.** The judge of record is claude-sonnet-4-6 — the same lab as the two reliable pack gainers. A cross-family judge panel re-grade and a human-expert agreement study are forthcoming against this same frozen grid; until then, treat the pack-lift ranking as provisional.
- **These are frozen model versions.** Pinned versions, seed 42. New frontier releases will be re-run on cadence.

10 · WHAT'S NEXT

The path forward

VEB is the scoreboard the category doesn't have. The same rubric that grades a frontier model on this leaderboard grades a customer's AI agent in production and a human rep in the simulator.

- **Cadence.** New frontier models re-run as they ship, so the leaderboard stays current.
- **The failure gallery.** The taxonomy in §06 becomes a living, browsable library of how AI sellers break.
- **The asset pack.** The Value Engine frameworks as skills + an MCP server — drop the discipline into any AI sales stack. §04 proves it moves the number.
- **The simulator.** Real-time voice-AI buyer roleplay with live scoring against this exact rubric — training and evaluation for every enterprise sales org.

THE ONE LINE

The Value Engine becomes the system of record for sales skill — human or AI.

One rubric, one throughline: the public leaderboard, the customer's agent, and the human rep, all graded the same way.