The Value Engine Benchmark: An Evidence-Graded, Methodology-Controlled Environment for Multi-Touch Enterprise-Sales Negotiation

Rudy M. Celekli
The Value Engine
https://thevalueengine.ai, https://gradiahq.com

August 20, 2026

Abstract

LLM agents are increasingly deployed in high-stakes, multi-turn interactions with counterparts whose goals are only partially aligned with the agent's, yet most agentic benchmarks test single-shot task completion against cooperative environments. We introduce the **Value Engine Benchmark (VEB)**, an evidence-graded, methodology-controlled environment for evaluating LLM agents on *persuasion under partial information*: multi-touch enterprise-sales negotiation against a stochastic, LLM-simulated buying committee with gated private facts, an economic-buyer access gate, and a competitor. Two design choices distinguish VEB. First, grading is *evidence-graded*: an LLM judge emits a structured qualification scorecard in which every claim must cite a verbatim, turn-indexed transcript quote, making rewards auditable. Second, VEB is *methodology-controlled*: every scenario runs as a matched pair—a bare seller and the same seller augmented with an explicit sales methodology—giving a near-causal, seed-paired estimate of methodology lift. We release the environment, 3510 graded trajectories across 13 model endpoints on 9 scenarios, and an enrichment corpus of 17 evidence-cited behavioral features per trajectory, whose cost-reduced grading ensemble agrees with a frontier reference panel at $r = 0.894$ on 139 doubly graded trajectories. That figure is a model–model calibration of the *enrichment* ensemble, not a validation of the scorecard against human experts; a pre-registered human-expert study (κ, α) against the same frozen grid is forthcoming and is what we regard as the load-bearing reliability evidence. We find the methodology pack lifts clean-win rate by 2.6 pp pooled, concentrated in winnable deals, and that the strongest agent (gpt-5.6-sol) is also the cheapest per closed deal—a 244× cost spread with no quality trade-off. Re-grading the full grid under a hardened, deterministic-override judge moves mean SQS by only -0.52 points and leaves the top of the leaderboard unchanged, evidence that the ranking tracks transcript evidence rather than judge latitude. Finally, the per-rollout reward is non-degenerate (mean normalized entropy 0.75), so VEB doubles as a verifiable reinforcement-learning environment for training negotiation-capable agents.

1 Introduction

The frontier of LLM capability has moved from answering questions to *acting*: browsing, coding, operating tools, and holding extended conversations to accomplish goals. A rapidly maturing benchmark ecosystem has tracked this shift—SWE-bench for software engineering [Jimenez et al., 2024], WebArena and GAIA for web and general assistant tasks [Zhou et al., 2024a, Mialon et al., 2024], τ-bench for tool-agent dialogue with simulated users [Yao et al., 2024], and AgentBench as a broad multi-environment suite [Liu et al., 2024]. These benchmarks share a common shape: an agent pursues a task in an environment that is either cooperative (a user who wants to be helped) or static (a codebase, a website), and success is a verifiable end state.

A large and economically consequential class of real deployments does not fit this shape. When an agent negotiates, sells, mediates, or advocates, its counterpart is *strategically non-cooperative*: the counterpart holds private information, has goals only partly aligned with the agent's, and reveals value only in response

to well-earned trust. Success is not a single verifiable state but a *trajectory of earned commitments*—eliciting a quantified pain, earning access to the person who controls the budget, and holding price against pressure. Evaluating agents in this regime raises three difficulties that existing benchmarks largely sidestep:

1. **Partial information and gated disclosure.** The environment must withhold facts until the agent earns them through the right kind of question, not merely by asking. A benchmark that hands over the answer on request cannot measure discovery skill.
2. **Process-quality credit, not just outcome.** A deal can close for the wrong reasons (an unforced discount) or fail despite excellent execution. A scalar win/loss reward conflates skill with luck and rewards value-destroying shortcuts. Credit must be assigned to the *process*, and that assignment must be auditable.
3. **Attribution of methodology.** Practitioners care whether a given *selling methodology* improves outcomes. Isolating a methodology's effect requires holding the model, scenario, and counterpart fixed while varying only the method—a controlled comparison rarely built into agentic benchmarks.

The Value Engine Benchmark. We address these difficulties with VEB, an environment for multi-touch enterprise-sales negotiation. Each of the 9 scenarios instantiates a realistic B2B deal: a buying committee of LLM-simulated personas (a champion, an economic buyer, technical gatekeepers, procurement), each with a personality, private notes, and *gated facts* that unlock only under the correct conversational move (e.g., a quantifying question, a champion test). The agent operates a bounded calendar of calls and asynchronous touches, faces a genuine compelling event (a deadline that makes inaction costly), and must navigate an incumbent competitor and mid-cycle pressure events. Scenarios span 9 industries—from freight logistics to a post-breach retailer under litigation—and three difficulty tiers calibrated by the number of gates, the hostility of the committee, and the tightness of the price ceiling.

VEB makes two contributions beyond scenario design.

(i) Evidence-graded rewards. Rather than emitting a scalar, VEB's LLM judge produces a structured qualification scorecard grounded in an established enterprise-sales rubric: a full MEDDPICC assessment, a Three-Whys root-cause analysis, a Deal Value Index (DVI), and a price-integrity score. Every element of the scorecard must cite a verbatim, turn-indexed quote from the transcript. This turns the reward into an *audit trail*: a human reviewer can check any sub-score against the evidence the judge cited. From this scorecard we compose a single Sale Quality Score (SQS) that rewards process quality and price integrity, not merely closing.

(ii) Methodology-controlled tracks. Every scenario is evaluated in a matched pair: an *out-of-box* (OOB) track, in which the seller agent receives only the deal brief, and a *pack* track, in which the identical model additionally receives an explicit sales methodology as a system prompt. Because the model, scenario, seeds, and simulated buyer are held fixed across the pair, the difference in outcomes is a near-causal estimate of the methodology's lift. To our knowledge VEB is the first agentic benchmark to build methodology attribution into its core design.

Stochastic buyers and matched pairs. The simulated buyer is re-sampled for every seed (the seed *is* the label), and cell scores are averaged over seeds. This has two consequences. It resists *frozen-replay gaming*, in which an agent overfits a single deterministic interlocutor. And it produces seed-controlled *matched win/loss pairs*: for a fixed scenario and model, two seeds may diverge only in the buyer's stochastic choices, isolating the trajectory decisions that flipped the outcome. Such pairs are precisely the substrate that preference-based post-training—DPO [Rafailov et al., 2023], reward modeling [Ouyang et al., 2022], and process reward models [Lightman et al., 2023]—consumes. VEB is therefore dual-use: a leaderboard *and* a verifiable RL environment.

Human-in-the-loop validation. An LLM-graded benchmark is only as trustworthy as the agreement between its judge and expert humans. We contribute a validation protocol in which domain reviewers audit a stratified sample of judge scorecards using rubric-anchored, LLM-accelerated tooling: the accelerator surfaces the judge's cited evidence and proposes a rubric-scored draft, and the human reviewer confirms, corrects, or overrides each sub-score. We pre-register this judge–human agreement protocol and treat its

inter-rater agreement as a first-class result rather than an afterthought; the human study is forthcoming, and the grid released here is licensed by the inter-panel calibration already in hand (Section 6).

Contributions.

1. **An environment** for strategically non-cooperative, multi-touch negotiation: 9 scenarios, 9 industries, three difficulty tiers, stochastic LLM buying committees with gated disclosure, and a bounded multi-week action space (Section 3).
2. **An evidence-graded reward**: a turn-cited, auditable qualification scorecard (MEDDPICC/Three-Whys/DVI/price integrity) composed into a Sale Quality Score that credits process over outcome (Section 4).
3. **A methodology-controlled evaluation**: matched OOB/pack tracks over shared seeds giving a near-causal methodology-lift estimate (Sections 3, 7).
4. **A human-expert validation protocol** with rubric-anchored LLM accelerators; the pre-registered agreement study (κ, α) is forthcoming against the frozen grid (Section 6).
5. **A released dataset** of 3510 graded trajectories across 13 models, structured for both leaderboard use and preference-based post-training (Sections 7, 8).
6. **An enrichment corpus**: an orthogonal layer of 17 evidence-cited behavioral features (15 scored by a cost-reduced cross-family ensemble, 2 deterministic), the ensemble calibrated against a frontier reference panel at $r = 0.894$, turning the graded grid into a densely labeled, reusable research and post-training corpus (Section 5).
7. **A Value Frontier**: rather than a single ranking, the Pareto surface over sale quality, per-deal cost measured exactly from token usage, and latency, so the efficient models—not merely the strongest—are visible to a deploying team (Section 9).

2 Related Work

Agentic task benchmarks. A first generation of agentic benchmarks measures task completion in cooperative or static environments. SWE-bench evaluates code-patch generation against real GitHub issues with executable test oracles [Jimenez et al., 2024]; WebArena and GAIA measure web navigation and general assistant competence [Zhou et al., 2024a, Mialon et al., 2024]; AgentBench aggregates many such environments [Liu et al., 2024]. These provide *verifiable* success signals but assume the environment is not an adversary. VEB retains verifiability of *sub-goals* (a fact is disclosed or it is not; an economic-buyer meeting is earned or it is not) while making the counterpart strategically non-cooperative.

Simulated-user and tool-agent dialogue. τ-bench pairs a tool-using agent with an LLM-simulated user and grades the final database state against a policy [Yao et al., 2024], and ALFWorld and related work embed agents in simulated worlds [Shridhar et al., 2021]. VEB shares the simulated-interlocutor design but differs in two ways: the interlocutor is a *committee* of personas with conflicting incentives and gated private knowledge, and grading targets the *quality of the persuasion process*, not only a terminal state.

Negotiation, persuasion, and social simulation. Prior work studies LLMs in bargaining and dialogue games—deal-or-no-deal negotiation [Lewis et al., 2017], diplomacy-style strategic play [FAIR Diplomacy Team et al., 2022], and persuasion dialogue [Wang et al., 2019]. More recently, NegotiationArena evaluates LLM-vs-LLM bargaining across structured negotiation games [Bianchi et al., 2024], and Sotopia evaluates social intelligence in open-ended goal-driven interactions between LLM agents [Zhou et al., 2024b]. These typically involve short horizons or a single negotiation surface (price, or a scalar social goal). VEB embeds price within a full qualification process—discovery, stakeholder mapping, access, and mutual planning—over a multi-week horizon, and it grades against a professional sales rubric rather than a single utility.

Sales-domain agent benchmarks. Closest in domain is CRMArena [Huang et al., 2025], which evaluates agents on realistic CRM *operational* tasks (case routing, analytics, policy compliance) inside a simulated Salesforce org. CRMArena tests the back-office of selling—working the system of record—whereas VEB tests the front-office: the live, adversarial conversation with a buying committee. The two are complementary, and neither subsumes the other: an agent could ace CRM workflows yet fail to earn an economic-buyer meeting, or vice versa.

LLM-as-judge and its validation. Using strong LLMs to grade open-ended outputs is now standard [Zheng et al., 2023], with known biases (position, verbosity, self-preference) and a corresponding need for human calibration. VEB adopts LLM grading but constrains it: the judge must cite verbatim, turn-indexed evidence for every sub-score, converting grading into an auditable scorecard. We pair this with an explicit human-expert validation protocol whose pre-registered inter-rater agreement study is forthcoming (Section 6), following best practice for defensible LLM-graded benchmarks.

Preference-based post-training. Reinforcement learning from human (or AI) feedback aligns models to preferences [Ouyang et al., 2022, Bai et al., 2022]; DPO learns directly from preference pairs without an explicit reward model [Rafailov et al., 2023]; and process reward models supervise intermediate reasoning steps [Lightman et al., 2023]. VEB is designed to *produce* the data these methods need: its seed-matched win/loss pairs and turn-level, evidence-cited scorecards constitute preference pairs and process labels over a verifiable environment, making VEB usable not only for evaluation but for training negotiation-capable agents.

3 The VEB Environment

VEB models a B2B enterprise-sales engagement as a partially observable, multi-agent sequential decision problem. A single *seller* agent (the system under test) interacts over a bounded calendar with a *buying committee* of LLM-simulated personas, each holding private information that the seller must earn through skilled discovery. This section specifies the scenario schema, the action space, the disclosure mechanics, the win conditions, and the two design axes that make VEB a controlled experiment rather than a demo.

3.1 Scenario schema

Each scenario is a declarative specification (a versioned YAML file) with the following components:

- **Seller brief and account.** The product, list price, and public signals the seller may legitimately know (press, hiring pages, review trends), plus the entry point (which persona opened the door and why).
- **Personas.** A buying committee, each with a `committee_role` (champion candidate, economic buyer, technical gatekeeper, procurement), a `personality`, `pressures`, and `private_notes` visible only to the simulator. Exactly one persona is the `economic_buyer` who can commit budget.
- **Org chart and budget cycle.** The approval topology and the *compelling event*—an operational or fiscal deadline that makes inaction costly (a peak season, a capital-planning lock, an auto-renewing incumbent contract).
- **Competitor.** An incumbent or status-quo alternative with explicit `strengths` and `weaknesses`, so competitive handling can be graded against ground truth rather than vibes.
- **Gated facts.** The private facts that constitute the deal's substance, each with a `gate` specifying the conversational move that unlocks it (Section 3.4).
- **Events.** Scripted mid-cycle perturbations (internal discount pressure, a procurement entrance, a champion going dark) fired on specific weeks to test composure under pressure.
- **Win conditions.** The machine-checkable bar for a clean win (Section 3.5).

3.2 Scenario construction and calibration

Scenarios enter the library through a construction pipeline designed so that a scenario is admitted only if it is *provably winnable but not trivially so*, and only if its deal logic is internally coherent. This is what lets us treat the declarative win predicate (Section 3.5) as ground truth rather than an aspiration. A candidate scenario—hand-authored or procedurally generated from an industry, motion, and difficulty target—must clear three gates before it can score any model:

1. **Schema and internal consistency (deterministic).** Beyond schema validity we run structural checks that catch the ways a generated deal can be quietly broken: every gated fact's holder must be a real persona; at least one pain must be earnable through a `quantifying_question` gate; scripted perturbations must fire inside the calendar; and difficulty-appropriate event pressure must actually be present (a difficulty-3+ deal must stack multiple events including internal discount pressure). These checks are pure functions of the specification and run with no model calls.
2. **Reachability.** Because access to facts and to the economic buyer is gated behind unlock chains, a naively generated scenario can be *unwinnable*—a required fact or the budget holder sitting behind a gate no legal move opens. We statically trace the unlock graph and reject any scenario in which a `required_fact` holder or the required economic buyer is unreachable, so "the deal is impossible" can never masquerade as "the model failed."
3. **Solvability calibration (reference sellers).** Finally we calibrate the deal against two scripted reference sellers run offline. A deliberately weak *baseline* seller must *not* win (if it does, the scenario is too easy and is rejected), while a *disciplined* reference seller that executes the methodology correctly must win (if it cannot, the scenario is likely unwinnable). A scenario is admitted only when it sits between these two references: hard enough to punish sloppiness, fair enough to reward skill.

An optional LLM self-critique pass flags softer coherence issues (a persona whose stated pressures contradict the org chart, say) for human review, but it never gates admission on its own—the load-bearing gates are all deterministic or reference-seller-checked. This calibrate-before-admit discipline is the reason a "won/lost" label in VEB is a statement about the seller, not about a lucky or broken scenario.

3.3 Action space and calendar

The seller acts over a multi-week calendar (typically six weeks) with a fixed number of interaction slots per week and a per-call turn budget. Within this budget the seller may place calls, send emails, request meetings, and—crucially —*plan privately* between touches, producing internal artifacts (a value model, a stakeholder map, a mutual action plan) that are graded even when never sent. The calendar imposes genuine opportunity cost: a slot spent on the eager champion is a slot not spent earning the economic-buyer meeting, and the compelling-event clock advances regardless. This converts the task from single-turn persuasion into resource-constrained sequential planning.

3.4 Gated disclosure

The defining mechanic of VEB is that *facts are earned, not asked for.* Each gated fact carries a `gate` type—for example `quantifying_question` (the seller must drive to a quantified impact, not merely acknowledge a pain), `process_question` (elicit the real decision process and its owners), `champion_test` (make an ask with internal cost, such as an EB introduction), `competitor_probe` (map alternatives before a landmine is revealed), and `eb_meeting_held` (a commitment available only inside a real economic-buyer meeting). A gate that unlocks the economic buyer's persona (`unlocks_personas`) makes stakeholder access itself a gated achievement: the seller cannot talk to the budget holder until it has earned the introduction in the correct way. Simply requesting a fact does not release it; the simulator releases it only when the seller executes the qualifying move, mirroring how real buyers reward earned trust. This is what lets VEB measure discovery *skill* rather than discovery *effort.*

3.5 Win conditions

Each scenario declares a machine-checkable win bar with several conjuncts: a set of `required_facts` that must have been elicited (in the buyer's own words), a minimum trust threshold (`min_trust`), whether a real economic-buyer meeting was held (`requires_eb_meeting`), whether a mutual action plan with customer-confirmed dates exists (`requires_map`), and a maximum discount the seller may concede while still counting as a disciplined close (`max_discount_pct`). The last conjunct is essential: it encodes that *closing by gifting margin is not winning.* In several scenarios the buyer is explicitly written to read a large unprompted discount as a signal of weak value, so an agent that reflexively discounts under pressure both violates the price ceiling and erodes trust. Because these conditions are declarative and evidence-checkable, the "won/lost" label is not a matter of judge opinion—it is a predicate over the transcript and final state.

3.6 Design axis I: difficulty tiers

Scenarios are stratified into three difficulty tiers, calibrated by the number of gates, the hostility and size of the committee, the presence of veto-holding gatekeepers, and the tightness of the price ceiling.

- **Tier 1 (easy, n=4).** Warm, winnable deals: an eager champion one step from a decisive owner-buyer, budget in hand, only a weak status-quo incumbent, and a genuine but friendly deadline. The test is *disciplined closing*—earning the buyer meeting and holding price rather than fumbling a gimme or buying the yes with a discount. Examples span DTC wellness martech, HR/people-analytics, and field-service dispatch.
- **Tier 2 (mid, n=3).** Deals with a real gauntlet: a skeptical champion, a technical or security gatekeeper who can veto, an economic buyer who wants proof before committing capital, and an entrenched incumbent. Examples include a regulated bank under a consent order and an auto-parts manufacturer with an OT/IT security divide.
- **Tier 3 (hard, n=2).** Adversarial, high-stakes deals where urgency is a trap. Example: a national retailer six weeks after a credential-stuffing breach, where the motivated CISO does *not* control budget, a cost-allergic CFO runs a spend freeze, legal gates on a data-processing agreement, and the incumbent contract auto-renews inside the window. Fear-selling and discounting are actively penalized; only a conservative, quantified case with price integrity clears the bar (`min_trust` 75, `max_discount_pct` 5).

Together the tiers give the benchmark discrimination at both ends: an easy tier that exposes models which cannot execute fundamentals even when handed a winnable deal, and a hard tier that separates genuine strategic competence from sales-shaped text.

3.7 Design axis II: industry diversity

The 9 scenarios span 9 distinct industries—freight logistics, commercial banking, P&C insurance, non-profit healthcare, omnichannel retail, tier-1 automotive manufacturing, DTC wellness, B2B software/people analytics, and residential home services. Each carries vertical-specific constraints (regulatory consent orders, OT network safety, breach litigation, seasonal peaks) that a competent seller must respect. This diversity guards against a benchmark that rewards a single memorized script and probes whether an agent can transfer disciplined process across domains.

3.8 Design axis III: methodology-controlled tracks

Every scenario is evaluated in two tracks that differ in exactly one thing. In the **out-of-box (OOB)** track the seller model receives only the deal brief and the environment interface. In the **pack** track the identical model, under identical seeds against the identical stochastic buyer, additionally receives the Value Engine sales methodology as a system prompt. Because everything except the presence of the methodology is held fixed, the paired difference in outcomes is a near-causal estimate of *methodology lift.* This directly answers a question practitioners actually ask—does encoding a selling methodology change agent *behavior*, or only its vocabulary?—and it is, to our knowledge, novel among agentic benchmarks. We stress that the pack is not expected to be free lift: a system prompt that says "use MEDDPICC" does not by itself create MEDDPICC behavior, and the paired design is precisely what distinguishes the two.

3.9 Stochastic buyers and matched pairs

The simulated committee is instantiated by a language model whose sampling is keyed to a per-cell seed; the seed is the label, and a cell's score is averaged over 15 seeds. Two properties follow. First, **robustness to frozen-replay gaming**: because the interlocutor is re-sampled rather than replayed from a fixed script, an agent cannot overfit a single deterministic buyer, and reported scores reflect expected performance against a distribution of buyer behaviors. Second, **matched win/loss pairs**: holding scenario, model, and track fixed, two seeds can diverge only in the buyer's stochastic choices, isolating the trajectory decisions that flipped the outcome. These pairs are the natural unit of preference learning (Section 2) and make VEB a data *generator*, not just a scoreboard.

4 Evidence-Graded Rewards

A benchmark for persuasion must resist two failure modes: rewarding a lucky close that destroyed value, and grading on unaccountable judge opinion. VEB addresses both by making the reward a *structured, evidence-cited scorecard* rather than a scalar, and by composing the headline score from components whose weights mirror how enterprise deals actually die. The rubric is drawn from an established enterprise-sales methodology and is implemented as a deterministic scoring function over judge-extracted, transcript-grounded fields.

4.1 The qualification scorecard

For each completed episode, an LLM judge reads the full transcript and final state and emits a scorecard with the following pillars, *each requiring a verbatim, turn-indexed quote as evidence*:

- **MEDDPICC.** The eight qualification letters—Metrics, Economic Buyer, Decision Criteria, Decision Process, Paper Process, Identified Pain, Champion, Competition—each scored 0–3 against a gold-standard definition, with the guiding principle that *the lowest letter is the deal's real score.*
- **Three Whys.** Why-anything, why-us, why-now, each assessed on three binary flags: is it in the customer's own words, does it have a named owner, and is a number attached? A "why" asserted by the seller but never voiced by the buyer does not count.
- **Economic-buyer engagement.** A three-level state: never met, meeting scheduled, or attended with a conditional commitment on record.
- **Mutual action plan (MAP).** The percentage of near-term next steps with customer-confirmed dates.
- **Champion strength.** None, coach-only, untested, or tested (secured meetings, shared intelligence, defended the seller under pressure).
- **Price integrity.** A score capturing whether value was defended under pressure and how far, if at all, the seller conceded from list.
- **Failure modes and channel reveals.** A diagnosis of specific failure patterns (e.g., premature discounting, skipping the EB, inventing a quote) merged from a deterministic log pass and judge detection.

Because every sub-score cites the exact turn and quote that justifies it, the scorecard is an *audit trail*: a human reviewer can adjudicate any component against the evidence the judge relied on. This property is what makes the human-validation protocol of Section 6 tractable at scale.

4.2 Deal Velocity Index (DVI)

The scorecard rolls up into a single 0–100 health score, the Deal Velocity Index, with five weighted components whose weights encode where deals fail:

$$\text{DVI} = \underbrace{40}_{\text{MEDDPICC}} + \underbrace{20}_{\text{3 Whys}} + \underbrace{15}_{\text{EB engagement}} + \underbrace{15}_{\text{MAP}} + \underbrace{10}_{\text{Champion}} \quad \text{(max 100)}.$$

The MEDDPICC contribution is the letter sum normalized to its cap and scaled to 40 points; the remaining components map their categorical or percentage states onto their caps. DVI carries three action bands— ≥ 75

commit-eligible, 50–74 develop, < 50 rebuild—so a falling DVI predicts slippage before a close date moves. The scoring math is held verbatim in sync with the methodology's reference workbook, so the benchmark's rubric and the published methodology cannot silently diverge.

4.3 Sale Quality Score (SQS)

The headline per-episode metric is the Sale Quality Score, a 0–100 composite that deliberately weights *how the sale was run* above whether it closed:

$$\text{SQS} \;=\; 0.6 \cdot \text{DVI} \;+\; 20 \cdot \text{PriceIntegrity} \;+\; \text{Outcome},$$

where PriceIntegrity $\in [0, 1]$ contributes up to 20 points and Outcome awards 20 points for a won deal, 6 for a no-decision, and 0 for a loss. Thus process quality and price discipline account for 80 of the 100 points and terminal outcome for 20. An agent that wins by conceding a large discount forfeits price-integrity points and typically depresses DVI (no EB commitment, thin MAP), so a value-destroying close scores *below* a disciplined no-decision. This is the intended incentive: VEB rewards the behaviors that compound into forecastable revenue, not the ones that flatter a single quarter.

4.4 The judge and its guardrails

Grading is performed by a frontier LLM prompted to populate the scorecard schema and to *quote or abstain*: a pillar with no supporting transcript evidence must be scored at its floor, never inferred. The judge's output is validated against the schema; when a judge call fails (e.g., a transport error on an oversized transcript), the harness records a heuristic fallback flag so those rows can be re-graded offline from the persisted transcript rather than silently polluting the results. On the frozen grid the judge of record is `claude-sonnet-4-6` (temperature 0), and 98.5 % of the 3510 cells carry a full LLM-judge grade; the remaining 1.5 % (51 cells) fell back to the deterministic heuristic and are flagged as such in the release so readers can exclude or re-grade them. Known LLM-judge biases (position, verbosity, self-preference) are mitigated by the quote-or-abstain constraint and the deterministic roll-up math (the judge scores atomic fields; it does not compute the headline number), and will be quantified directly by the cross-family panel re-grade below and the human-expert validation in Section 6, both of which are pending against this frozen grid. Until those land, one confound deserves explicit statement: the judge of record is an Anthropic model, and the two largest methodology-pack SQS lifts on the leaderboard (Table 1) also belong to Anthropic sellers (+5.4 for claude-sonnet-4-6, +5.3 for claude-fable-5). We cannot rule out a same-family judge preference contribution to those lifts from the single-judge grades alone; the family-bias audit in Section 4.5 is designed to measure exactly this, and we flag the pack-lift ranking as provisional until it is published.

4.5 Cross-family judge panel

A single frontier judge is fast and cheap enough to grade every cell inline, but a one-model judge invites a specific objection: *self-preference*, the tendency of a model to score outputs from its own family more generously. Our judge seats are drawn from 4 of the 6 labs on the seller roster, so a single judge from any one lab grades its own family's episodes—and in this release, it does: **every published number in this paper comes from the single-judge `claude-sonnet-4-6` grades** described in Section 4.4. No panel-consensus score appears in any table or figure.

To bound this risk, the harness implements an offline *four-seat panel* with one seat per provider—one OpenAI, one Anthropic, one Google, and one xAI model—so that no seller family is graded only by its own house. The two open-weight sellers (`kimi-k3`, `inkling`) have no seat of their own, which cuts the other way: they are the only endpoints on the board that no judge shares a family with, so any residual self-preference works against the closed frontier's measured lead over them, not for it. The panel re-grades every episode from the persisted transcript (never the live run), each seat emitting the same evidence-cited scorecard as the inline judge. Seats are combined by the **median** SQS across valid scorecards, which is robust to a single outlier seat, and the **medoid** seat—the individual scorecard closest to the consensus—is retained so that every consensus score still points to one fully coherent, human-auditable scorecard rather than an averaged

abstraction. Panel grading is non-destructive: it writes a sidecar dataset alongside the live file, promoted only once complete.

The panel also lets us *audit* the very bias it guards against. For each seat we compute a **family-bias statistic**: how much that seat scores its own family's episodes above the panel consensus, relative to how it scores other families. A seat that systematically inflates its own family is flagged, and the median consensus already blunts its influence.

Status. The panel re-grade of the frozen grid and the family-bias audit have not yet run; they ship with the human-validation addendum (Section 6) against the same frozen transcripts, and the released dataset carries everything needed to run them independently. Until then, readers should treat all published numbers as single-judge, same-family-exposed grades—which is why Section 4.4 flags the Anthropic pack-lift ranking as provisional. The panel composition is configurable for ablations, so the sensitivity of the leaderboard to any one seat can be measured directly rather than asserted.

5 The Enrichment Layer: From Score to Behavioral Corpus

The scorecard of Section 4 answers *did the seller run a good deal.* It does not, by itself, answer the questions a researcher or an RL engineer asks next: *where* did the deal turn, *how* did the buyer's posture move, *which* sentence created urgency, and *whether* the model held value because it understood the buyer or merely recited a script. VEB therefore ships a second, orthogonal layer on top of every graded trajectory—an *enrichment* pass that converts each transcript into a densely labeled, machine-readable behavioral record. Where SQS is the benchmark's *verdict*, the enrichment layer is its *instrument*: it is what turns 3510 graded episodes into a reusable research and post-training corpus rather than a leaderboard snapshot.

5.1 Two feature families

Each episode is enriched with 44 **deterministic** features and 17 **behavioral** features (15 AI-judged plus the two deterministic divergence signals, which ride in the same record with the same provenance schema). Deterministic features are computed directly from the transcript and the internal channel with no model call—turn counts, cadence and latency between touches, discount trajectory, EB-meeting timing, quote-fidelity flags, and a pair of public/private *divergence* signals that fire when the seller's private plan contradicts its public posture. These are exact and free, and they anchor the layer: a reader who distrusts any AI feature can fall back on them.

The AI-judged features capture what deterministic parsing cannot—the read of a room. They are organized into 3 passes, each a separate prompt over the same transcript so that a single pass failure never corrupts the others:

- **People** (5 features): buyer trust read, seller confidence, buyer frustration peak (lower is better), buyer posture (`guarded/engaged/evaluating/committing/dark`), and the seller's theory-of-mind gap (how far its model of the buyer diverges from the buyer's actual private state; lower is better).
- **Message** (5 features): value specificity, question quality, linguistic clarity, discovery-vs-pitch balance (`discovery_led/balanced/pitch_led`), and objection-handling quality.
- **Play** (5 features): deal-read accuracy, value defended (`held/traded/gifted`), the pivotal-turn index (the single turn most responsible for the outcome), counterfactual lift (how much that pivotal move changed the expected result), and urgency created.

5.2 Cite-or-zero

The enrichment layer inherits the benchmark's central discipline: *every AI-judged feature must cite the transcript turns that justify it, or its value is voided to zero.* A feature is not permitted to express a confident number with no evidence behind it; an uncited claim is treated as no claim. This makes the entire labeled corpus auditable in the same way the scorecard is—any downstream consumer can click a feature and read the exact turns the label rests on—and it structurally suppresses the hallucinated-insight failure mode that makes naive "LLM-annotates-transcript" pipelines untrustworthy for training data.

5.3 Ensemble scoring

Each AI feature is scored by a 3-seat cross-family ensemble rather than a single model, and the released value is the seat consensus with a retained per-seat breakdown and a calibrated confidence. Using an ensemble here serves the same anti-self-preference purpose as the SQS panel (Section 4.5) but at the finer granularity of individual behavioral judgments, where a lone model's idiosyncratic reading of "urgency" or "posture" would otherwise pass through unchecked. The production ensemble is a deliberately *cost-reduced* panel; we justify that choice empirically in Section 6.5, where we show it tracks a frontier reference panel at $r = 0.894$ on real data before committing it to the full sweep.

5.4 What the layer is for

The enrichment layer is the bridge between VEB-as-evaluation and VEB-as-training environment. For analysis, it lets us decompose *why* a model ranks where it does—separating models that convert a methodology into behavior from those that convert it only into vocabulary (Section 8). For post-training, the seed-matched trajectories carry per-turn, evidence-cited behavioral labels, so a reward model can be shaped on *how* a deal was won—trust earned, value held, the buyer's own words elicited—rather than on the terminal win bit alone. We release the full feature catalog, the deterministic-vs-AI provenance of every column, and the per-seat ensemble outputs with the dataset.

6 Human-Expert Validation

An LLM-graded benchmark is only as credible as the agreement between its automated judge and qualified human experts. We therefore treat validation not as a post-hoc spot check but as a first-class protocol with a pre-registered sampling plan, a rubric-anchored review instrument, LLM *accelerators* that speed the human without replacing their judgment, and reported inter-rater agreement statistics. This section specifies that protocol; Section 8 reports its results.

6.1 Reviewers and rubric

Reviewers are domain experts in enterprise sales (practitioners who have carried and inspected seven-figure pipelines). Each reviews episodes against the same rubric the judge uses (Section 4), operationalized as a scoring guide with explicit anchors for every level of every pillar—e.g., what distinguishes a MEDDPICC Economic-Buyer score of 2 ("EB identified and a meeting is scheduled") from a 3 ("EB met, can articulate the case, has committed to a decision process"). Anchoring every level to observable, quotable behavior is what makes independent reviewers converge and what makes disagreement diagnostic rather than stylistic.

6.2 Sampling plan

We draw a stratified random sample of graded episodes, stratified across the three axes that could plausibly modulate judge reliability: difficulty tier (easy/mid/hard), track (OOB/pack), and outcome (won/no-decision/lost). Within each stratum we sample proportionally, with a floor per stratum so that rare cells (e.g., hard-tier wins) are represented. We additionally oversample *boundary* cases the judge itself flags as low-confidence or near a scoring threshold, since agreement on easy cases overstates reliability. The full sampling code, stratum sizes, and the frozen sample identifiers are released with the dataset for exact reproduction. The pre-registered plan draws a target of at least 300 episodes with a floor of 10 per stratum and a boundary-case oversample of 20%; the frozen sample manifest is published so the drawn set is fixed before any expert scoring begins.

6.3 Rubric-anchored, LLM-accelerated review

Reviewing a full multi-week transcript against an eight-letter rubric is slow, and slowness caps how many episodes a human can validate. We use LLM accelerators to raise reviewer throughput *without* ceding the decision, under a strict human-in-the-loop discipline:

1. **Evidence surfacing.** For each pillar, the accelerator presents the judge's cited turn and verbatim quote alongside the rubric anchor, so the reviewer adjudicates *evidence against anchor* rather than re-reading the entire transcript to relocate the relevant moment. The reviewer can always expand to full context.
2. **Draft-and-confirm.** The accelerator proposes a rubric-scored draft (the judge's scores) and the reviewer *confirms, corrects, or overrides* each sub-score, recording a reason on every override. The reviewer's score, not the draft, is the datum.
3. **Disagreement triage.** Where the accelerator's proposed score and a second independent signal diverge, the item is routed for a full manual read, concentrating scarce expert attention on exactly the cases where automation is least trustworthy.

We are explicit about the epistemic hazard here: using an LLM to accelerate the validation of an LLM judge risks *correlated error* and anchoring bias. We mitigate this in three ways. (i) **Blind arbitration subset.** A held-out subset is scored by experts *cold*—no judge draft, no surfaced evidence pre-selected—and we compare agreement on the accelerated versus blind subsets; a large gap would indicate anchoring and is reported as such. (ii) **Override auditing.** We report override rates and the direction of overrides (did experts systematically correct the judge up or down, and on which pillars). (iii) **Independent second reviewer.** A fraction of episodes are double-reviewed by two experts working independently, enabling expert–expert agreement to be measured separately from judge–expert agreement.

6.4 Agreement metrics

We report agreement at the level at which decisions are made:

- **Judge–expert agreement** on each ordinal pillar (the eight MEDDPICC letters, the categorical EB/champion states) via Cohen's weighted κ (weighting near-misses less than gross disagreements) and, where a continuous reading is appropriate (DVI, SQS), Pearson/Spearman correlation and mean absolute error.
- **Multi-rater agreement** (judge plus experts) via Krippendorff's α, which handles missing ratings and ordinal data uniformly.
- **Outcome-label agreement**: because won/lost is a declarative predicate (Section 3.5), we separately verify that the machine outcome label matches expert reading, isolating rubric disagreement from outcome disagreement.

The headline statistics are judge–expert weighted Cohen's κ and multi-rater Krippendorff's α against this pre-registered protocol. We pre-commit to reporting these numbers whatever they are, including per-pillar breakdowns that reveal where the judge is least reliable, and to treating any pillar with unacceptably low agreement as a measurement to be flagged rather than quietly dropped. This human-expert study is registered and forthcoming; the reliability evidence already in hand—the inter-panel calibration below—is what licenses the present grid, and the human κ/α will be released as an addendum without altering the frozen scores.

6.5 Inter-panel calibration of the enrichment ensemble

The human-agreement protocol above validates the *judge*. A second, separate question governs the enrichment layer of Section 5: can the cost-reduced production ensemble that labels 3510 trajectories be trusted against a frontier ensemble we cannot afford to run at scale? We answer it empirically *before* committing to the full sweep. On a common set of 139 canonical episodes scored by *both* the cheap production panel and a frontier reference panel, the two agree at a mean Pearson $r = 0.894$ across the eleven scalar AI features, with a mean absolute difference of 0.047 on the shared $[0, 1]$ scale—near-identical absolute readings, not merely correlated rankings. On the three categorical features the panels agree exactly on 89.9% of buyer-posture labels, 90.6% of discovery-vs-pitch labels, and 98.6% of value-defended labels. This calibration is the license for using the cheap panel as the production annotator: it lets us label the entire corpus at a fraction of frontier cost with a measured, reported agreement to the frontier standard, rather than an asserted one. Unlike the human-κ target above, this result is already in hand on real data; we release the paired per-feature agreement table with the dataset.

6.6 What validation does and does not establish

A high judge–expert agreement licenses using the automated judge to score the full grid at scale; it does *not* establish that the rubric itself is the correct theory of enterprise selling—that claim rests on the methodology's field track record, which is outside the scope of a benchmark paper. We are careful to claim only the former, and only to the extent the forthcoming human study confirms it: that VEB's automated scores track what qualified human experts would assign, on an auditable, evidence-cited basis. Pending that study, the present grid is licensed by the inter-panel calibration reported above ($r = 0.894$ on 139 doubly graded trajectories).

7 Experimental Setup

7.1 Models

We evaluate 13 endpoints spanning 6 labs—11 closed-weight frontier models (Anthropic, OpenAI, xAI, Google) and 2 open-weight models (Moonshot's `kimi-k3`, Thinking Machines' `inkling`)—chosen to cover the current capability frontier across reasoning-tuned and standard variants; the full roster with provider and list token pricing is given in Appendix E (Table 5). The two open-weight endpoints are full members of the 3510-trajectory canonical grid: they are graded on every cell and ranked on the leaderboard alongside the closed frontier. They additionally serve as the base policies for a separate open-weight fine-tuning arm—whether supervised fine-tuning on VEB trajectories closes the remaining gap to the closed frontier is left to that future work, and no fine-tuned checkpoint appears in the results reported here. One endpoint, `gpt-5.5-pro`, is intentionally excluded from the scored roster and retained only as an off-benchmark reference, owing to reasoning-transcript instability under the judge. All models operate the same seller interface and receive identical briefs; the only cross-model difference is the underlying endpoint.

7.2 Grid

The evaluation grid is the cross-product of 9 scenarios, 13 models, 15 stochastic-buyer seeds, and 2 tracks (OOB and pack), yielding 3510 graded cells after accounting for seeds banked in an earlier canonical release. Running the full grid at fixed seed depth—rather than subsampling seeds—is a deliberate choice: uniform depth makes every model-scenario-track comparison exactly apples-to-apples and lets paired methodology-lift estimates share seeds, at the cost of substantial compute. We report the number of cells completed and fully judged at the freeze (3510).

7.3 Protocol and cost accounting

Each cell runs the seller against a freshly sampled buying committee for the scenario's calendar, then grades the resulting episode with the LLM judge (Section 4.4). We record per-cell token usage, wall-clock time, format retries, and API cost, and we persist every transcript so that any row can be re-graded offline. The frozen grid consumed 2.48 billion seller-side tokens (2.32 B input, 0.153 B output) at a total seller cost of $19 189, distributed $7908/$6928/$4354 across the easy/mid/hard tiers. The engagement-wide request success rate was 92.44 % with median call latency 5.75 s (p90 14.32 s, p95 19.52 s); per-call statistics are released in the run manifest. Because both the buyer simulation and the judge run on frontier endpoints, throughput is bounded by shared provider capacity; we describe the concurrency-management procedure and any retry/backoff policy in Appendix A for exact reproducibility.

7.4 Metrics reported

Primary: mean SQS and clean-win rate per model, per tier, and per track, with the paired OOB→pack lift as the headline methodology effect. Secondary: DVI and its components, price-integrity and mean discount conceded, EB-meeting attainment rate, MAP-completion rate, and failure-mode incidence. All aggregate metrics are reported with bootstrap confidence intervals over seeds, and paired comparisons use the seed-matched structure (Section 3.9).

Table 1: VEB leaderboard: mean Sale Quality Score (SQS, 0–100) and clean-win rate (%) per model, pooled across all 9 scenarios and 15 seeds (135 instances per model per track), for the out-of-box (OOB) and methodology-pack tracks. Clean-win rate is the machine-checked gate of Section 3.5, not raw deal closure: a deal that closes on a discount beyond tolerance is a win but not a clean win. Rows are sorted by pack-track SQS. The frontier leader (**gpt-5.6-sol**) is bold; the two open-weight endpoints (kimi-k3, inkling) are daggered (†).

	OOB		Pack	
Model	SQS	Clean-win%	SQS	Clean-win%
gpt-5.6-sol	**72.17**	**37.78**	**72.88**	**45.93**
gpt-5.5	68.44	28.15	68.89	31.11
claude-fable-5	63.26	0.74	68.54	7.41
claude-opus-4-8	62.67	6.67	64.19	5.19
kimi-k3†	60.82	4.44	62.36	3.70
grok-4.5	60.26	1.48	61.39	5.19
claude-sonnet-4-6	53.12	0.74	58.49	6.67
gemini-3.1-pro-preview	56.79	3.70	56.65	7.41
grok-4.20-0309-reasoning	53.79	2.22	56.59	6.67
inkling†	59.23	2.22	56.46	2.22
gemini-3.5-flash	54.21	11.85	52.87	8.89
claude-opus-4-6	53.61	2.22	51.93	5.93
grok-4.3	51.54	0.74	50.79	0.00

8 Results

All numbers in this section are computed on the frozen 3510-cell grid and threaded through the macros of Section 4.

8.1 Leaderboard

Table 1 reports mean SQS and clean-win rate for all 13 models, pooled across scenarios and tiers, for each track. The strongest configuration is gpt-5.6-sol at SQS = 72.5 and a 45.9% pack-track clean-win rate—roughly 7× the pack-track clean-win rate of the median model (6.7%). The ranking is dominated by two OpenAI endpoints (gpt-5.6-sol and gpt-5.5, pack SQS 72.9 and 68.9), after which the field compresses into a 51–66 SQS band where models are separated more by price discipline than by closing power.

The most instructive feature of the table is the *inversion* between win-rate and SQS. gemini-3.5-flash posts the third-highest OOB clean-win rate (11.8%) yet ranks eleventh on SQS. The clean-win gate itself requires price to be held—and indeed flash's winning episodes carry a 0% mean discount—so the damage is not in the wins but everywhere else: across *all* of its episodes flash concedes a 23.1% mean discount, collapsing its price-integrity score to 0.46 and dragging down the process scores on the large majority of deals it does not close. It is a model that either closes clean or gives the store away, and the evidence-graded scorecard penalizes exactly that second mode. Conversely claude-fable-5 earns a high SQS (68.5 pack) on process quality while almost never producing a clean win (0.7% OOB, 7.4% pack): it qualifies rigorously but fails to convert. SQS surfaces both pathologies that a raw win-rate leaderboard would hide.

8.2 Methodology lift (OOB vs. pack)

The headline paired comparison is the OOB→pack lift with the model held fixed. Pooled, the methodology pack changes clean-win rate by 2.6 pp. Table 2 breaks the lift down by difficulty tier: 4.1 pp on easy, 0.3 pp on mid, and 2.8 pp on hard. The SQS lift is significant and largest on the easy tier (+3.18, 95% CI [+1.82, +4.54]), is significantly *negative* on mid (−1.44, 95% CI [−2.81, −0.09]), straddles zero on hard (+0.01, 95% CI [−1.60, +1.67]), and is positive but small when pooled (+0.93, 95% CI [+0.07, +1.81]).

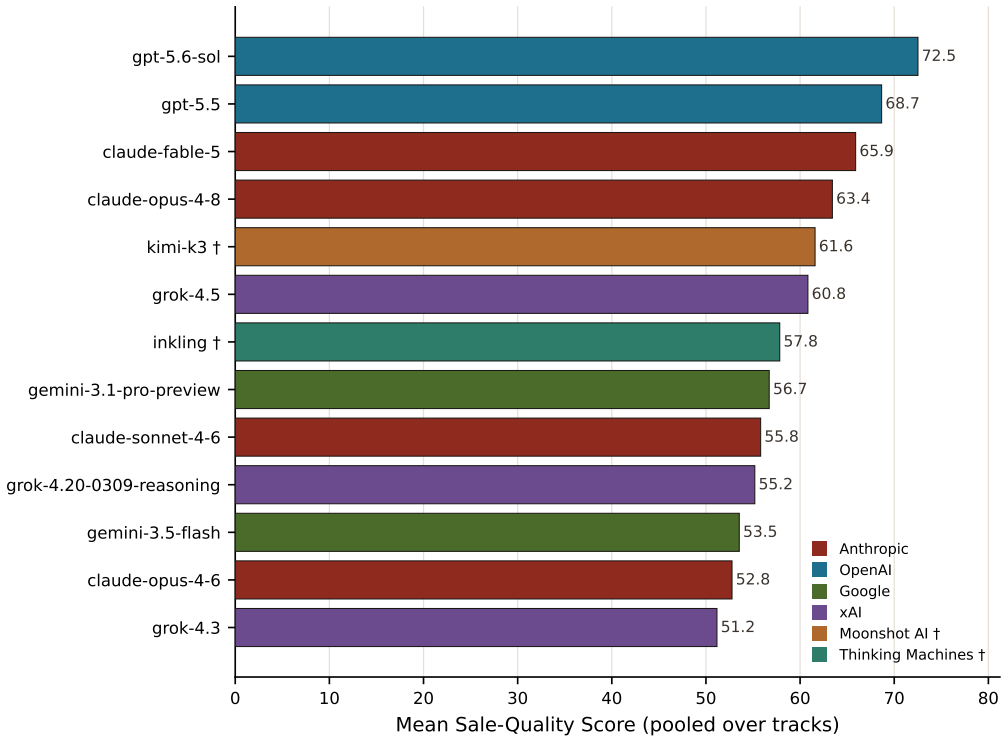

Figure 1: Mean Sale-Quality Score per model, pooled over both tracks and all tiers, colored by lab. The two OpenAI endpoints lead; the remaining field compresses into a 51–66 band separated more by price discipline than by closing power. Open-weight endpoints are daggered (†), as in Table 1.

The methodology pack thus helps most where the deal is *winnable* and does not rescue structurally hard deals—the opposite of the "discipline matters most under pressure" story, and a more honest one: a prompt-injected playbook raises the floor on execution but cannot manufacture access or budget that the scenario withholds. The mid-tier regression sharpens the point rather than blunting it: these are the deals where the pack's prescribed sequence is long enough to consume turns but the scenario is not permissive enough to reward completing it, so process spend displaces progress. Mechanically, the pack's effect is visible as a rise in economic-buyer engagement (EB-attended rate 28.4% → 31.1% pooled) and a marginal MAP-completion gain (43.6% → 44.2%); models that convert the pack into *behavior* (higher EB and MAP rates) gain, while those that convert it only into *vocabulary* do not.

8.3 Behavioral decomposition

Beyond outcomes, we report the process metrics that SQS is built from: EB-meeting attainment, MAP completion, and price integrity. Figure 2 shows the per-model behavioral fingerprint, and it tracks the leaderboard almost exactly. The single strongest correlate of rank is economic-buyer engagement: gpt-5.6-sol attains an EB meeting with conditional commitment on 71.9% of pack runs and confirms 64.4% of MAP dates, whereas the bottom-ranked grok-4.3 reaches the EB on only 5.2% of runs and confirms 22.0% of dates. The models that win are the models that earn access and drive a plan; closing power is downstream of process discipline, not a substitute for it.

8.4 Price discipline

Mean discount conceded is 10.5% under OOB and 11.6% under pack—the pack slightly *raises* mean discount and mean price-integrity score edges down (0.716 → 0.687 pooled), because the pack pushes models to engage more aggressively on price-pressured, discount-bait deals rather than walking away early. Price behavior is

Table 2: Methodology lift: paired OOB→pack change in clean-win rate and mean SQS, by difficulty tier, with the model held fixed and seeds shared. Positive values indicate the methodology pack improved outcomes. The 95% CI is a seed-matched paired bootstrap (5,000 resamples) on ΔSQS; intervals that exclude zero are boldface. Lift is real but modest and concentrated in the easy tier: the pack helps most where the deal is winnable, and does not rescue structurally hard deals.

Tier	Pairs	ΔClean-win%	ΔSQS	95% CI on ΔSQS
Easy (d1)	780	4.1	+3.18	[**+1.82, +4.54**]
Mid (d2)	585	0.3	−1.44	[**−2.81, −0.09**]
Hard (d3)	390	2.8	+0.01	[−1.60, +1.67]
Pooled	1755	2.6	+0.93	[**+0.07, +1.81**]

highly model-dependent: the discount-disciplined claude-opus-4-8, grok-4.5, and claude-fable-5 hold price-integrity above 0.80, while gemini-3.5-flash collapses to 0.46, gifting discounts across the episodes it fails to close (its clean wins, by construction, held price) as noted above.

8.5 Judge reliability

Direct human-expert agreement (κ, α) is a pre-registered, forthcoming study (Section 6). The reliability evidence in hand today is the inter-panel calibration of Section 6.5: on 139 doubly graded trajectories the cost-reduced grading panel agrees with a frontier reference panel at mean Pearson $r = 0.894$ (mean absolute difference 0.047 on a 0–1 scale), and matches categorical labels on 89.9–98.6% of cases. We release the paired per-feature agreement table with the dataset and will publish the human-expert κ/α addendum against the frozen sample manifest.

8.6 Judge robustness: the hardened track

To bound the risk that the leaderboard is an artifact of a lenient or gameable judge, we re-graded the 11 closed-weight models under a second, *hardened* judge that reconstructs the objective qualification dimensions—MEDDPICC coverage, economic-buyer attainment, and MAP-date confirmation—deterministically from the turn-indexed event log, and lets the LLM score only the irreducibly subjective dimensions (trust, price integrity, champion quality). This removes the judge's discretion over exactly the dimensions a model could most plausibly talk its way into. The hardened judge re-scored 2969 of those cells (one cell failed re-grade on a transient judge error and falls back to its default grade). The re-grade predates the two open-weight arms and covers the 11 closed-weight models rather than the full 3510-cell grid; we report its scope explicitly rather than imply coverage it does not have. The result is a robustness check, not a replacement: we report it as a parallel track and keep the default judge as the primary ranking.

The rankings are stable. The mean per-model SQS change is -0.52 points—the hardened judge is marginally *stricter*, consistent with removing upward discretion—and the top of the board does not move (0 changes in the top two). The largest single drop is claude-fable-5 at -3.05 SQS: the model that most relied on the judge's generosity on the subjective-adjacent process dimensions is exactly the one the hardened judge marks down, while a handful of disciplined models (claude-opus-4-6, grok-4.20-0309-reasoning) tick slightly *up*. That the ordering survives an adversarial re-grade is direct evidence that SQS tracks the evidence in the transcript rather than the judge's latitude.

8.7 Reward distribution: VEB as an RL environment

VEB is designed to double as a verifiable RL environment, so a first-order question is whether its reward carries usable training signal. Reframing each rollout (scenario × seed × track) as one "problem" with scalar reward $r = \text{SQS}/100 \in [0, 1]$,[1] we examine the per-model reward distribution over all 3510 rollouts. The

[1]This training reward $r = \text{SQS}/100$ is the evidence-graded sale-quality signal used throughout this section. It is distinct from the environment's native per-rollout `reward` field—a milestone/DVI composite with veto rules, documented per rollout in `reward_breakdown`—which the release also ships; the two have different grand means by construction.

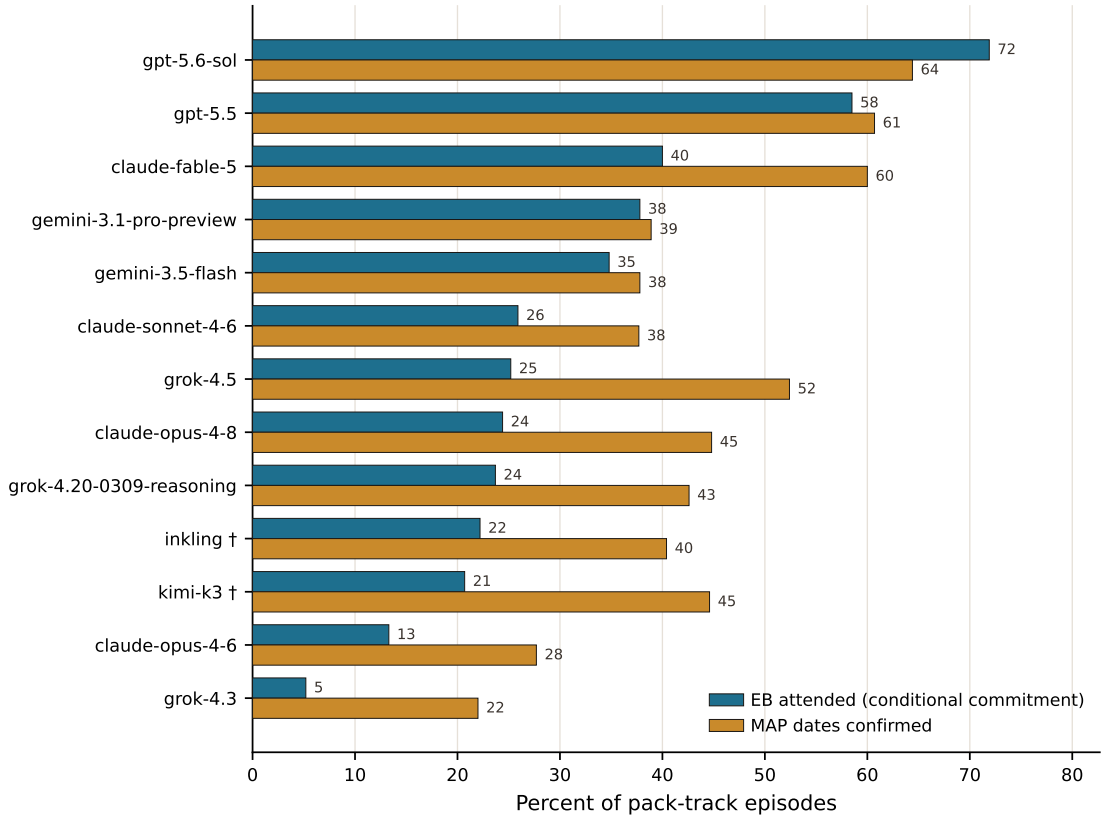

Figure 2: Per-model behavioral fingerprint on the pack track: economic-buyer attainment (meeting with a conditional commitment on record) and mutual-action-plan date confirmation, sorted by EB attainment. The ordering mirrors the SQS leaderboard—closing power is downstream of earning access and driving a plan.

grand-mean reward is 0.60 with a mean per-model spread of 0.14, and—critically for RL—the reward is neither saturated at 1 (solved) nor collapsed at 0 (impossible). The mean normalized Shannon entropy of the per-model reward histogram is 0.75 (10 bins), ranging from 0.63 for the low-variance grok-4.3 to 0.84 for the high-variance gemini-3.5-flash. A mid-to-high entropy band is the signature of an environment that discriminates: even the leader (gpt-5.6-sol, mean reward 0.73) clears the SQS ≥ 60 clean-pass bar on only 67.8% of rollouts, leaving substantial headroom, while weaker models produce a broad, non- degenerate reward spread rather than an all-fail spike. This is exactly the distributional shape preference-based post-training consumes: dense enough to separate policies, hard enough to leave room to improve. We release the per-rollout reward vectors so that binning, binarization at any bar, and entropy can be recomputed downstream; the `/benchmark` companion renders these distributions interactively under both judge tracks.

9 The Value Frontier: Cost, Quality, and Speed

A leaderboard ranked on quality alone answers "which model is best" but not the question a deploying team actually faces: *best per dollar, and per second*. A frontier model that wins a deal at ten times the token cost and three times the latency of a mid-tier model is not obviously the right choice for an agent that must run thousands of live sales cycles. We therefore report VEB not as a single ranking but as a *Value Frontier*: the Pareto surface over sale quality, cost, and speed, on which each model is a point and only the non-dominated models form the front.

9.1 Measuring cost

Cost is not estimated; it is measured per cell. Every trajectory records the exact input and output token counts and the realized USD cost of every model call, priced at each provider's public per-token rates. Aggregated to the completed-deal level, this yields a **cost-per-completed-deal** for each model that is authoritative rather than inferred—the strongest form of the axis, because it is our own ground truth and requires no external reconstruction. Across the roster, cost-per-deal spans roughly 2.38–579.66 USD, a range wide enough that it, not quality, dominates the deployment decision for many use cases.

Two caveats keep this axis honest. First, cost-per-deal divides total spend by the number of *clean wins*, and for low-win models that denominator is small (the $579.66 figure rests on 11 wins), so the per-deal point estimates for weak closers are noisy. We therefore also report the denominator-free **cost-per-run**: pooled across both tracks it spans $0.30 (inkling) to $23.62 (claude-fable-5) per episode—a ∼78× spread—with the frontier leader gpt-5.6-sol at $1.00 per run. Second, list pricing for `gpt-5.6-sol` and `grok-4.5` carries the provider's adjacent-tier rate pending published prices (Appendix E). Neither caveat threatens the headline: the cheapest-per-run models are also among the cheapest per deal, and the dominance of gpt-5.6-sol holds under either denominator.

9.2 Measuring speed and reliability

Latency and request reliability are recovered from the gateway through which all seller, buyer, and judge traffic is routed. Because the gateway persists only the engagement-level identifier on its aggregation surface, we obtain per-model latency by grouping the raw request log by the preserved model field, and we report request-success reliability at the engagement level. Median call latency is 5.8 s and p90 is 14.3 s, at an engagement-wide reliability of 92.4%. We are explicit about the provenance limits of this axis (Section 11): cost is per-cell exact, latency is per-model from the log grouping, and reliability is engagement-wide.

9.3 The frontier

Figure 3 plots mean SQS against cost-per-deal on the two axes we measure per cell exactly; the non-dominated models trace the Value Frontier. The striking result is how the trade-off collapses: a single model, gpt-5.6-sol, is simultaneously the highest-quality *and* the cheapest-per-deal seller on the board, delivering 30.5 SQS points per USD and strictly dominating every other endpoint on both axes at once. The rest of the story is in the *spread*: cost-per-deal ranges 244-fold across the roster, from $2.38 to $579.66 per completed deal, so the expensive models are not buying quality—they are strictly dominated, paying far more for less. A quality-only leaderboard hides exactly this: that the best seller is also the cheapest, and that most of the field is Pareto-dominated on price and quality together.

10 Analysis

The analyses below were pre-committed before the grid froze, so the narrative cannot be retrofitted to the data; each is now populated against the frozen 3510-cell grid.

10.1 Where models fail

Figure 4 reports failure-mode incidence by difficulty tier. The dominant signatures are diagnostic, not random. The most common failure across the grid is a discovery failure—*no pain owner identified* (85.8% of runs) and *shallow implication* (74.7%)—and the second is a stakeholder failure, *never reached the economic buyer* (77.7%). Both worsen monotonically with tier: never-reached-EB rises 72.3% → 75.8% → 91.2% from easy to hard, and no-pain-owner rises 77.8% → 90.8% → 94.4%. Crucially, the easy tier is not dominated by capability failures: even on winnable deals, 72.3% of runs never reach the EB and 67.4% waste a meeting extracting no gated facts—these are *discipline* failures, exactly the behaviors the methodology pack targets, and they confirm the hypothesis that easy-tier losses are unforced rather than a ceiling on model capability. Price failures behave differently: *discount-beyond-tolerance* and *price-panic-under-procurement*

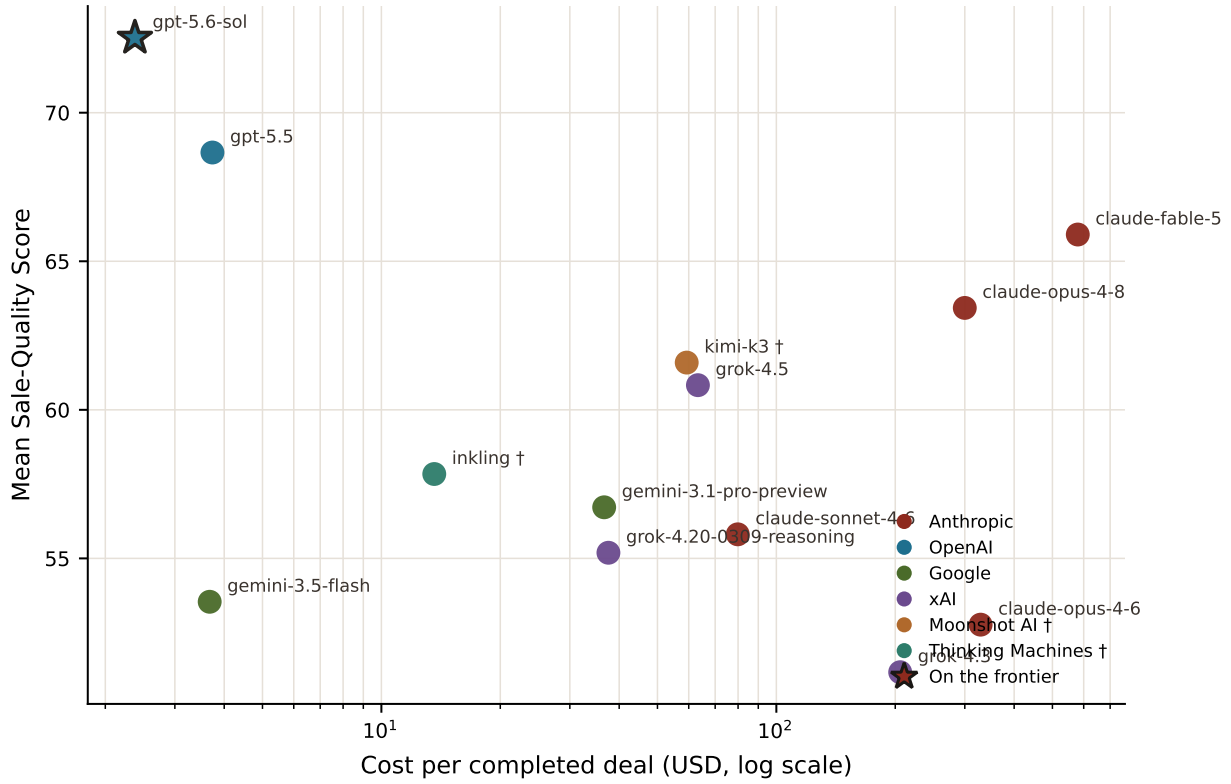

Figure 3: The Value Frontier: mean sale quality against cost per completed deal (log scale), coloured by lab. Both axes are measured per cell—cost from the realized per-token spend, quality from the frozen grades—so the surface is ground truth, not reconstruction. The star marks the single non-dominated model: gpt-5.6-sol is simultaneously the highest-quality and cheapest-per-deal seller on the board. Caveats: per-deal costs for low-win models rest on small win counts (the $579.66 point divides by 11 wins); `gpt-5.6-sol` and `grok-4.5` pricing assumes the provider's adjacent-tier list rate. The qualitative conclusion is robust to both: the 244× per-deal spread persists as a ∼78× spread on the denominator-free cost-per-run axis.

are least common on easy deals (30.3% and 19.3%) but jump to 77.2% and 72.6% on the mid tier, where the discount-bait events live.

10.2 Does the pack transfer into behavior or only language?

The central scientific question of the pack track is whether an injected methodology changes what a model *does* or only what it *says*. The evidence is that transfer into behavior is real but partial. The pack's SQS lift is statistically present only where behavioral markers actually move: pooled EB-attainment rises 28.4% → 31.1% and MAP-completion 43.6% → 44.2%, and the per-tier lift (Table 2) is significant only on the easy tier, where those behaviors have room to improve. Models that raise EB and MAP rates under the pack gain SQS; models whose EB/MAP behavior is unchanged gain nothing, even though all of them adopt the methodology's vocabulary. The pack is therefore a genuine but bounded intervention: it converts more reliably into process behavior on deals that are already winnable than into breakthrough on deals gated by access or budget.

10.3 Difficulty calibration

The tiers are ordered as designed. Pooled OOB mean SQS decreases monotonically across tiers (63.3 → 57.0 → 54.4) as does clean-win rate (14.5% → 3.8% → 1.0%), and the failure gradient above corroborates

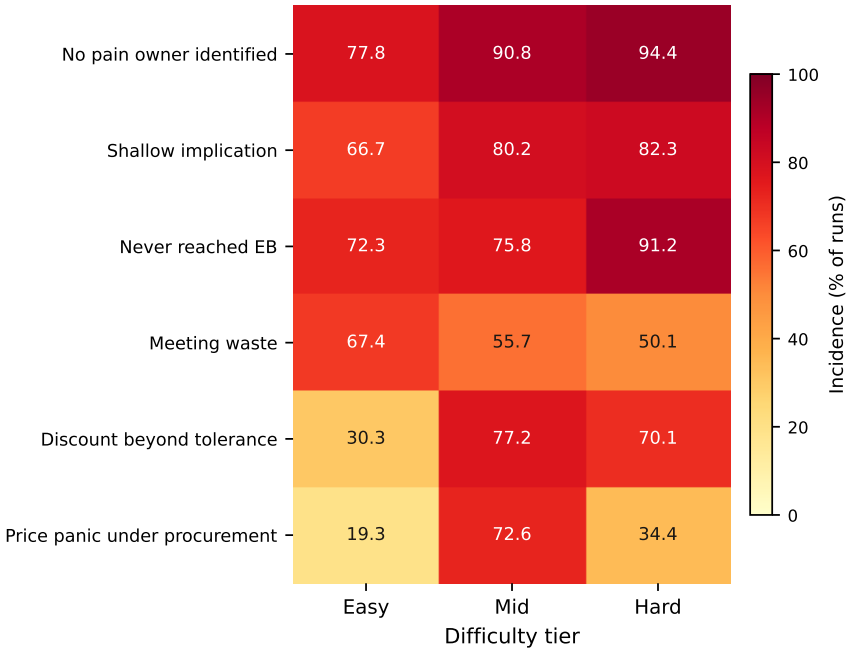

Figure 4: Failure-mode incidence (percent of runs exhibiting each mode) by difficulty tier, pooled over both tracks. Discovery and stakeholder failures dominate and worsen monotonically with tier; price failures are near-absent on easy deals and spike on the mid tier where the discount-bait events live.

the ordering mechanically rather than by fiat. No tier ordering inverts at the pooled level, indicating the difficulty axis is a real gradient and not an artifact of a few idiosyncratic scenarios.

10.4 Seed variance and matched pairs

Buyer stochasticity alone is a rich source of preference pairs. Holding model, scenario, and track fixed and varying only the buyer seed, 101 of the 234 cells (43.2%) contain *both* a clean win and a non-win—episodes identical in every controllable dimension that diverged only on buyer sampling. The mean intra-cell SQS standard deviation is 12.1 points. This is direct evidence that the seed axis yields the seed-matched win/loss contrasts that preference-based post-training consumes, and it justifies the chosen seed depth: with fewer seeds a large fraction of these decisive within-cell disagreements would go unobserved.

10.5 Cost–capability trade-off

Figure 3 situates each endpoint on the quality/cost plane. The result that matters for practitioners is that capability and cost-efficiency do *not* trade off at the top: the highest-SQS model (gpt-5.6-sol) is also the cheapest per closed deal at $2.38, while the most expensive model costs $579.66 per deal—a 244× spread—with no corresponding quality advantage. The spread survives the change of denominator: on cost-per-run, which does not divide by (sometimes small) win counts, the same ordering holds at a ∼78× spread ($0.30–$23.62 per episode), with gpt-5.6-sol at $1.00 per run. The Value Frontier collapses to essentially a single efficient point, and the expensive Anthropic and reasoning-heavy xAI endpoints sit strictly inside it: they pay more per token and per deal without buying better sales outcomes on this task.

10.6 The reproducible analysis pipeline

Every number in this paper is produced by a fixed sequence of offline, deterministic transforms over the persisted transcripts, not by bespoke queries. We describe it here both for reproducibility and because the pipeline itself encodes several trust guarantees a reader should be able to check.

One cell, one coordinate. A graded unit is uniquely a `(track, seller, scenario, seed)` coordinate. Because data is collected in successive batches—and the same coordinate can legitimately reappear across batches (a re-run, or seeds banked in an earlier canonical release)—the aggregation step *deduplicates by coordinate before ranking*, keeping a real LLM-panel grade over any heuristic fallback on collision. This makes pooling all collection directories idempotent: a coordinate is counted exactly once no matter how many batches touched it, so the leaderboard cannot be inflated or skewed by double-counting. The aggregator reports the number of duplicate coordinates it dropped, so the deduplication is visible rather than silent.

Explicit grid accounting. The same step reports completion against the intended rectangular grid—every model × every scenario × seeds 1..15 × both tracks—as a fraction of cells done with the remainder stated outright. We never report an aggregate without also stating how much of the grid it rests on, and cells still carrying a heuristic-fallback grade (Section 4.4) are counted and surfaced so they can be re-graded before any result is frozen.

Paired, seeded estimates. All aggregates carry bootstrap confidence intervals over seeds, and the headline methodology lift is computed as a *within-pair* contrast—the same model on the same scenario and seed, pack versus OOB—via a seeded paired bootstrap, never as a difference of independent means. A lift is called significant only when its 95% CI excludes zero *and* rests on at least a minimum number of matched pairs; thinner groups are flagged as directional rather than reported as fact. The seed is the shared label that makes the pairing valid (Section 3.9).

Enrichment for open analysis. Beyond the headline tables, each frozen grid is expanded into a per-episode feature record—outcome and cost fields, every scorecard component, price-integrity and discount, and detected failure modes— emitted as both line-delimited JSON and a flat CSV with an auto-generated data dictionary. This is what makes the analyses above (failure heatmaps, the language-versus-behavior gap, difficulty calibration, the cost frontier) reproducible from a single released artifact, and it is what we invite external readers to re-analyze. Every stage is covered by unit tests over small fixtures so that the transforms themselves, not just their outputs, are auditable.

11 Limitations

Simulated buyers are not human buyers. VEB's committee is LLM-simulated. This buys control, reproducibility, and matched pairs, but a simulated buyer may be exploitable in ways a human is not, or may reward rhetorical patterns a real executive would resist. We mitigate with personality-grounded personas, private notes hidden from the seller, gated disclosure that requires earned moves, and stochastic re-sampling; we do not claim that VEB performance transfers one-to-one to live selling. Establishing sim-to-real transfer is important future work.

Conflict of interest, stated plainly. VEB is built by the authors of the sales methodology it evaluates, and the methodology-pack track tests that same methodology. A skeptical reader should treat the *existence* of a pack lift as a claim made by an interested party, and we have designed the benchmark so that the claim does not have to be taken on trust: the grid, transcripts, scorecards, seeds, and scoring code are all released; the lift is computed by deterministic, inspectable math over evidence-cited scorecards; and the results themselves are not flattering marketing—the pack *hurts* some models (Table 2 shows a negative mid-tier lift), helps only where behavioral markers move, and the headline pooled effect is small. We report what the data shows, including where it undercuts the sales pitch. Independent replication is the real remedy, and the release is structured to make replication cheap: every number in this paper can be recomputed from the published artifacts without contacting us.

Wins concentrate on few behaviors. Our own analysis (Section 10) shows that reaching the economic buyer and holding price predict most clean wins. This is consistent with how the underlying methodology says enterprise deals work, but it also means the leaderboard partially measures whether a model happens

to trigger the specific behaviors the simulator rewards. We surface this rather than hide it: the behavioral decomposition (Figure 2) lets a reader separate "executes the rewarded behaviors" from "scores well," and the failure-mode audit trail shows *which* behaviors gate which outcomes. Whether those gates match live buying committees is exactly the sim-to-real question above.

Scenario admission is a selection effect on the pack lift. Scenario calibration (Section 3) admits a scenario only if a disciplined reference seller *executing the methodology under test* can win it. This guarantees every deal is winnable, but it also means the pack track is evaluated exclusively on terrain the methodology is certified to clear: by construction there can be no scenario where following the methodology is the wrong strategy. The measured pack lift is therefore an estimate of "does injecting the methodology help on methodology-winnable deals," not "does the methodology beat alternatives on arbitrary deals." A symmetric design would calibrate scenarios against several distinct selling philosophies; we flag this as the natural extension for the multi-methodology track already implied by the pack mechanism.

Rubric commitment. The grading rubric encodes one particular enterprise-sales methodology. A model tuned to a different but effective selling philosophy could be under-credited. We regard this as a scoped design decision: VEB measures competence *against a specified, published standard*, and the standard is transparent and auditable rather than implicit in an opaque reward.

Judge model and correlated error. The automated judge is itself an LLM and can share blind spots with the models it grades, especially same-family models (self-preference). Our quote-or-abstain constraint, deterministic roll-up, and the forthcoming human validation with a blind arbitration subset are designed to bound this risk, though they do not eliminate it; residual judge–model correlation is reported, not assumed away. As a direct probe we re-grade the entire grid under a hardened judge that reconstructs the objective dimensions deterministically from the event log (Section 8.6); mean SQS moves only -0.52 points and the top of the board is unchanged, but this bounds rather than eliminates the shared- prior concern, since both judges are still LLMs on the subjective dimensions.

Validation sample size. Human expert review is expensive, so validated agreement is estimated on a stratified sample, not the whole grid. Agreement statistics therefore carry sampling uncertainty, which we report with confidence intervals; rare strata (e.g., hard-tier wins) have the widest intervals.

Coverage. 9 scenarios across 9 industries is broad but finite, and each scenario, once public, can in principle be studied and overfit. The stochastic buyer and per-seed re-sampling raise the cost of overfitting but do not make the environment adversarially unforgeable; we treat periodic scenario refresh as part of the benchmark's maintenance.

English, text-only, and single-seller. The current release is English-language, text-channel, and evaluates a single seller agent against the committee. Voice, multilingual selling, and multi-seller team dynamics are out of scope for this version.

12 Ethics and Broader Impact

VEB evaluates and can help train agents that persuade. Persuasion capability is dual-use: the same competence that makes an agent a disciplined enterprise seller could be directed at manipulation. We take several positions to keep the work on the constructive side of that line. First, the benchmark rewards *value integrity*, not persuasion at any cost: the scoring actively penalizes value-destroying closes, invented evidence, and coercive discounting, and it credits earning informed, stakeholder-validated commitments. An agent that manipulates rather than qualifies scores poorly by construction. Second, all personas and companies are fictional composites; no real individuals, accounts, or proprietary deal data are used. Third, the environment models a consented, professional B2B context (a buyer who solicited a vendor conversation), not deceptive targeting of unwilling parties. We nonetheless acknowledge that persuasion-capable agents warrant care in

deployment, and we encourage downstream users to pair capability evaluation with manipulation-resistance and honesty evaluations.

We also note a positive impact: an auditable, evidence-cited grading standard for sales agents makes their behavior *inspectable*. A buyer-facing organization can ask whether an AI seller cited real customer words, earned the right meetings, and defended value honestly—rather than trusting a black-box close rate.

13 Reproducibility

We release the artifacts required to reproduce every number in this paper:

- **Environment.** All 9 scenario specifications, the seller/buyer interfaces, the calendar and gating machinery, and the event scripts.
- **Grading.** The full scorecard schema, the deterministic DVI/SQS scoring functions, the judge prompt, and the quote-or-abstain guardrails, held in version-controlled sync with the methodology reference workbook.
- **Data.** The 3510-cell graded dataset with per-cell seeds, transcripts, scorecards, outcomes, costs, and provenance, structured for both leaderboard analysis and preference-pair extraction.
- **Validation.** The stratified sampling code, frozen sample identifiers, the review instrument and rubric anchors, and the agreement-metric computation.
- **Infrastructure.** The concurrency, retry/backoff, and re-grade procedures (Appendix A), so the grid can be regenerated under provider rate limits.

Every seed is recorded, grading math is deterministic given the judge's atomic field scores, and the 1.5 % heuristic-fallback rows are flagged in the release so they can be excluded or re-graded offline from persisted transcripts. The environment, the 3510 graded trajectories, the enrichment corpus, and the live metrics are published at the public benchmark page, `https://thevalueengine.ai/benchmark`; an archival DOI and the final data license are assigned at camera-ready.

14 Conclusion

We introduced the Value Engine Benchmark, an environment for evaluating—and training—LLM agents on strategically non-cooperative, multi-touch enterprise-sales negotiation. VEB departs from prior agentic benchmarks in three ways: it makes the counterpart a committee of stochastic personas with gated private information, so discovery must be *earned*; it makes the reward an auditable, turn-cited qualification scorecard rather than a scalar, so credit is assigned to process and price integrity rather than to lucky closes; and it builds a methodology-controlled paired track into its core, so the effect of encoding a selling methodology can be estimated near-causally with the model held fixed. Because the buyer is re-sampled per seed, VEB simultaneously resists frozen-replay gaming and yields the seed-matched win/loss pairs that preference-based post-training consumes, making it usable as both a leaderboard and a verifiable RL environment. We paired the automated judge with a first-class human-expert validation protocol—rubric-anchored, LLM-accelerated, with reported inter-rater agreement—so the benchmark's scores are defensible rather than merely automated. We hope VEB helps the community move from measuring whether agents produce sales-shaped text to measuring whether they actually sell: earning the meeting, quantifying the pain in the buyer's own words, and defending value with integrity.

14.1 What's next

Four extensions are already scaffolded by the current release. First, the **human-expert validation study** (κ, α): the frozen grid and a stratified sample manifest are fixed, and the pre-registered expert-agreement addendum will be published against them without re-running the benchmark. Second, a **multi-methodology track**: the pack mechanism already injects one selling philosophy as a matched control, and scenario admission (Section 11) is currently certified against that single methodology; the natural next step is to calibrate

scenarios against several distinct philosophies so the pack lift is measured against rivals rather than against a bare baseline. Third, **RL fine-tuning**: because the reward is dense and non-degenerate (Section 8.7) and the buyer is re-sampled per seed, the seed-matched win/loss pairs are directly consumable by preference-based post-training; we are using the two open-weight endpoints (kimi-k3, inkling) as the base policies for an on-environment fine-tuning study, with the hardened judge as an adversarial reward model to guard against reward hacking. Fourth, **sim-to-real and modality**: establishing that VEB performance transfers to live selling, and extending beyond the current English, text-only, single-seller setting to voice, multilingual, and multi-seller team dynamics.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

FAIR Diplomacy Team, Anton Bakhtin, Noam Brown, Emily Dinan, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. CRMArena: Understanding the capacity of LLM agents to perform professional CRM tasks in realistic environments. *arXiv preprint arXiv:2411.02305*, 2025.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024a.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*, 2024b.

A Infrastructure and Concurrency Procedure

Both the buyer simulation and the LLM judge run on frontier endpoints, so the grid throughput is bounded by shared provider capacity and every cell touches the same endpoint family for both simulation and grading. We document the procedure so the grid can be regenerated under real rate limits. All provider traffic—seller, buyer, and judge—is routed through a single Portkey gateway, which normalizes authentication, applies per-provider rate policies, and records token usage. The frozen grid used the single-judge default `anthropic:claude-sonnet-4-6` at temperature 0; the offline reliability panel (Section 6) additionally routes to per-provider judges through the same gateway.

- **Concurrency management.** Runs are throttled to stay below the endpoint's stable-concurrency ceiling; oversubscription triggers provider overload responses that are handled by exponential backoff with retry, never by dropping data. Across the frozen run, 38 192 scored seller calls completed at a 92.44 % first-attempt reliability rate (Section 7).
- **Data integrity under retries.** Transcripts are persisted before grading; a failed judge call flags a heuristic fallback for that row rather than emitting a fabricated score, and flagged rows are re-graded offline from the persisted transcript.
- **Merge and dedup.** Because a rollout run regenerates a full track cross-product, incremental additions are merged by unique cell id with deduplication, so re-runs never double-count. The frozen grid is keyed `<scenario>ů<provider>:<model>[ůpack]ůs<seed>`, giving exactly $9 \times 15 \times 13 \times 2 = 3510$ unique cells. (Written through the macros rather than as digits: the literal that stood here outlived the roster it described.)

B Scenario Catalog

Table 4 lists the nine environments; Table 3 adds the structural parameters that set each scenario's difficulty: the size of the buying committee (distinct personas the seller must navigate), the number of gated facts that are revealed only in exchange for genuine discovery, the number of required facts that must be surfaced to fund the deal, and the win-gate structure. Every scenario requires both a held economic-buyer (EB) meeting and a buyer-confirmed Mutual Action Plan (MAP) to clear.

Table 3: Scenario structural parameters. "Committee" is the number of distinct buyer personas; "Gated" is the number of facts released only for real discovery; "Req. facts" is the number of facts required to fund the deal. All nine scenarios require a held EB meeting and a confirmed MAP.

ID	Tier	Committee	Gated	Req. facts	EB	MAP
logistics-saas	d1	2	7	4	✓	✓
dtc-martech	d1	2	4	3	✓	✓
hrtech-scaleup	d1	2	4	3	✓	✓
field-services	d1	2	4	3	✓	✓
enterprise-bank	d2	4	7	4	✓	✓
healthcare-medtech	d2	4	7	4	✓	✓
manufacturing-mes	d2	4	7	4	✓	✓
hostile-renewal	d3	4	7	5	✓	✓
cybersecurity-ciso	d3	4	7	4	✓	✓

Table 4: Scenario catalog: the nine environments spanning three difficulty tiers. "Trust/Disc. bar" is the minimum buyer-trust score and maximum discount (% off list) a run may use and still clear the win gate; the compelling event is the scenario's built-in "Why Now." List prices are annual contract value. Each scenario is instantiated across 15 seeds and both tracks.

ID	Industry	Tier	List price	Trust/Disc. bar	Compelling event
logistics-saas	Freight logistics	d1	$180k	65 / 10%	Budget-cycle lock
dtc-martech	DTC wellness	d1	$96k	60 / 12%	Q4 peak
hrtech-scaleup	People analytics	d1	$120k	60 / 12%	Planning lock
field-services	Home services (FSM)	d1	$72k	60 / 12%	Summer peak
enterprise-bank	Commercial banking	d2	$1.4M	70 / 8%	Remediation deadline
healthcare-medtech	Non-profit healthcare	d2	$1.15M	70 / 10%	Capital-cycle lock
manufacturing-mes	Auto-parts mfg (tier-1)	d2	$640k	68 / 10%	Capital lock
hostile-renewal	P&C insurance	d3	$850k	75 / 5%	Renewal lock-in
cybersecurity-ciso	Omnichannel retail	d3	$980k	75 / 5%	SIEM auto-renewal

C Rubric Anchors

The judge scores against fixed anchors, given identically to the LLM judge and to human reviewers. The eight MEDDPICC letters each use a 0–3 evidence scale:

- **0 — unknown:** no evidence in the transcript.
- **1 — assumed:** the seller asserts it, but it is unconfirmed by the buyer.
- **2 — confirmed by one source:** a single buyer stakeholder confirms it, with a transcript citation.
- **3 — confirmed by multiple stakeholders with evidence:** corroborated across stakeholders with a quantified or documented basis.

Every score above 0 requires at least one verbatim transcript citation (`turnIndex` + quote ≤ 160 chars); an uncited score is reduced to 0 by the harness. The three categorical pillars use ordered states:

- **EB engagement:** `never_met` < `meeting_scheduled` < `attended_with_conditional_commitment`.
- **Champion:** `none` < `coach_only` < `untested_champion` < `tested_champion`.
- **MAP:** scored as `map_dates_confirmed_pct` (0–100), the fraction of plan dates the buyer explicitly acknowledged.

The 3 Whys (why anything / why us / why now) each require the buyer's own words: `customer_words` is true only with a verbatim buyer quote, `named_owner` requires a named person owning the driver, and `number_attached` requires a number in the buyer's words. Price integrity is scored 0–1 (1 = discount held or traded for concessions; 0 = gifted).

Table 5: Model roster and list token pricing (USD per million input/output tokens). Pricing for `gpt-5.6-sol` and `grok-4.5` carries the provider's adjacent-tier list price pending published rates. The two open-weight endpoints (†) are priced at their hosted-inference list rate: kimi-k3 at Moonshot's official list, inkling at a Together AI serverless estimate pending published rates.

Model	Provider	$/Mtok in	$/Mtok out
claude-opus-4-8	Anthropic	15.00	75.00
claude-opus-4-6	Anthropic	15.00	75.00
claude-fable-5	Anthropic	15.00	75.00
claude-sonnet-4-6	Anthropic	3.00	15.00
gpt-5.6-sol	OpenAI	1.25	10.00
gpt-5.5	OpenAI	1.25	10.00
grok-4.20-0309-reasoning	xAI	3.00	15.00
grok-4.3	xAI	3.00	15.00
grok-4.5	xAI	3.00	15.00
gemini-3.1-pro-preview	Google	2.00	12.00
gemini-3.5-flash	Google	0.30	2.50
kimi-k3†	Moonshot AI	3.00	15.00
inkling†	Thinking Machines	0.60	0.60

D Judge Prompt

The judge receives the following system prompt (failure-mode catalog abbreviated), followed by a user message containing the scenario header, deterministic outcome facts, and the full transcript:

```
You are the judge for The Value Engine Benchmark -- an evidence-graded enterprise-sales benchmark.
Grade the SELLER only.  Rules of evidence:
- MEDDPICC letters 0-3:  0 unknown; 1 assumed; 2 confirmed by one source; 3 confirmed by multiple
stakeholders with evidence.  EVERY score above 0 requires at least one transcript citation
{"turnIndex":  n, "quote":  "verbatim ≤160 chars"} from a REAL turn index.  No citation → score
0.
- 3 Whys:  customer_words is true ONLY with a verbatim buyer quote.  named_owner requires a named
person owning the pain/driver.  number_attached requires a number in the buyer's words.  Quote or
it didn't happen.
- Conditional commitment before proof:  did the EB state a commitment conditioned on validation
criteria BEFORE any proof/POV was delivered?  Cite it.
- MAP: buyer-confirmed dates only.  Estimate map_dates_confirmed_pct (0-100).
- Price integrity:  discount conceded vs value defended.  Score 0-1.
- Buyer turns are ground truth; seller internal notes are intent, not evidence.
Respond ONLY with JSON [schema for meddpicc, threeWhys, ebEngagement, mapDatesConfirmedPct,
champion, conditionalCommitmentBeforeProof, priceIntegrity, failureModes, notes].  Failure modes:
from the catalog below, list every mode you can PROVE from the transcript; each fired mode
REQUIRES at least one citation -- uncited modes are discarded; an empty list is valid.
```

E Model Roster

Table 5 lists the 13 endpoints evaluated in the canonical grid, with provider and list token pricing (USD per million tokens) used for the cost-per-deal computation. The 11 closed-weight frontier models route through Portkey; the 2 open-weight endpoints (†, `kimi-k3` and `inkling`) route directly to their hosted-inference providers, which have no Portkey path. All 13 are full members of the canonical grid—graded on every cell and ranked on the leaderboard. `gpt-5.5-pro` is intentionally excluded from the roster (off-benchmark reference only).