LEARN TO SELL, NOT WHAT TO SELL: DUALISTIC SELF-EVOLUTION (DSE) OF LLM SALES AGENTS WITH HUMAN NATURE AS REWARD (HNR)

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-based selling is a long-horizon decision process in which customer behavior evolves across interactions and damaged trust is rarely restored. Delayed sales outcomes complicate autonomous improvement, so humans must still supply dense rewards and evaluation. Motivated by the distinction between transferable sales skills and business-specific knowledge, we introduce Dualistic Self-Evolution (DSE) with Human Nature as Reward (HNR), which separates a profession learned in the model weights from business knowledge, rules and tools carried by an external harness. The inner loop learns from open cross-category dialogues and 8 business-agnostic prototype tools through supervised fine-tuning followed by reinforcement learning. HNR is drawn from sales practice: it rewards selling with the grain of the customer (answer first, verify before claiming, move one step forward), penalizes the model's own counterproductive habits, and assigns zero episode reward to unsupported claims and other credibility violations. Outcome-gated turn-level scoring reduces the share of rollout groups with zero reward variance from roughly two thirds under outcome-only rewards to 3–5%. The outer loop updates the harness through reward-guided edits validated by real-conversation replay. The loops share no parameters and have separate reward agents and human arbiters: senior salespeople assess craft and quality-control staff assess factual and procedural correctness. With the target business excluded from training, frozen models transfer to a production used-car harness that maps the 8 prototypes onto 16 business tools. On replay of 600 real customer turns with full conversation histories, the trained 9B model improves strict response quality by 8.3 percentage points over its same-size base (13% relative; three seeds; paired 95% CI [+4.4, +12.0]), and the share of turns with a tool call rises from 57% to 80%; the 8B gain is smaller and statistically unresolved. On the out-of-domain Ψ-Bench, persuasion improves by 88% at 8B and 54% at 9B. At 9B the prototype-tool SFT carries the strict gain and reinforcement learning adds to it only at 8B; the gate's measurable contribution is gradient availability rather than replay score. Fine-tuning on business transcripts degrades replay quality and removes tool use, whereas the harness alone lifts untrained models above the deployed system and accommodates business changes without retraining.

1 INTRODUCTION

People learn a profession before they learn a business. A graduate of a marketing program can sell cars, houses or insurance, because what the program instilled is not a catalog but a way of dealing with people: answer before you ask, hear the need behind the question, verify before you claim; the catalog, rules and tools of a business are learned on the job and change with it. Selling over text makes the separation vivid: a deal unfolds over many turns and days, and one unsupported claim can end a relationship that took weeks to build.

LLM sales agents are built the other way around: the prevailing recipe fine-tunes the business into the weights, so every change in the business reopens training (Goyal et al., 2025). The long horizon of selling also shapes self-improvement: a coding agent closes its loop in minutes on a test suite, whereas a closed deal arrives weeks later, entangled with prices, inventory and the customer's life, so

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

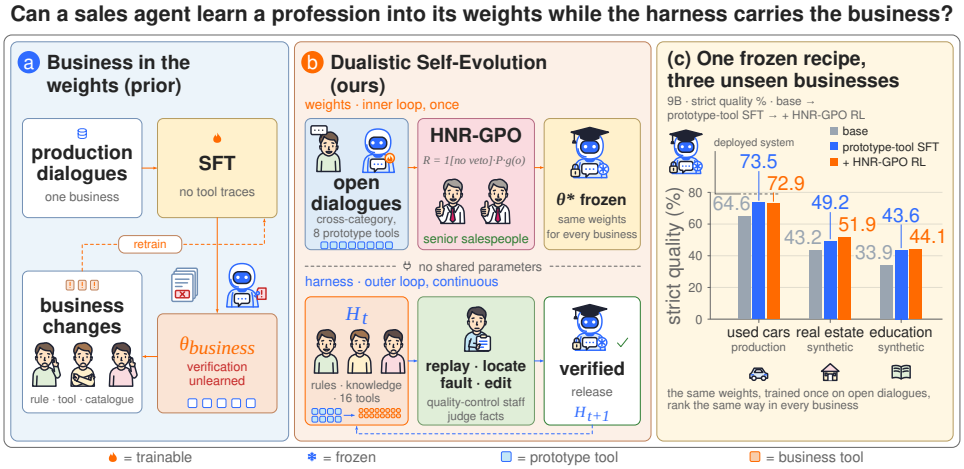


Figure 1: **Learn to sell, not what to sell.** (a) The prevailing recipe fine-tunes the target business into the weights, so every change to rules, tools or catalog reopens training, and transcripts without tool traces unlearn verification. (b) DSE trains a salesperson's character once on open cross-category dialogues with prototype tools under HNR, then plugs the frozen weights into the business's harness, which absorbs business changes by replay-verified edits; senior salespeople judge craft, quality-control staff judge facts. (c) The same frozen 9B weights in three businesses never seen in training: strict quality rises from the base to prototype-tool SFT in every business and RL holds or adds to it; dashed line: deployed system.

sales agents sit at the *early stage of self-evolution*, where the system can update itself but humans must still supply dense rewards and benchmarks.

We take the separation to its first principle—*the weights learn a profession; the harness carries the business*—and present **Dualistic Self-Evolution (DSE)** with **Human Nature as Reward (HNR)**. The inner loop evolves the agent's *character*, its know-how in Ryle's sense (Ryle, 1949), by reinforcement learning on the weights from open cross-category sales dialogues and a handful of business-agnostic *prototype tools*; the outer loop evolves its *circumstance*, the business's know-that, by reward-guided edits to a structured harness verified on real-conversation replay. The loops share no parameters; each has its own reward agent and human arbiter. HNR is drawn from human nature rather than from labels: it rewards selling with the grain of the customer, penalizes the model's own path of least resistance, and is capped by credibility, so that one unsupported claim voids the episode; scoring every turn under the outcome's gate keeps every group informative where outcome-only rewards starve the policy of gradient (§4.3).

We validate DSE on a production used-car sales platform never seen in training. Frozen inner models plugged into its harness map 8 prototype tools onto 16 business tools, raise verification before claiming from 57% to 80% of turns at 9B, improve strict quality on 600 real conversations by 13% at 9B over the same-size base (three seeds, paired CI excluding zero) with a smaller gain at 8B, and nearly double persuasion on the out-of-domain Ψ-Bench (+88%). Conversely, business transcripts in the weights buy nothing at either size, the harness alone lifts untrained models above the deployed system, and it absorbs changes in the business without retraining (Figures 1 and 2).

Contributions. (i) *Principle*: a separability hypothesis, profession in the weights and business in the harness, with three falsifiable predictions, and why sales agents sit at the early stage of self-evolution (§3). (ii) *Method*: prototype tools that put tool discipline, not the catalog, into the weights; HNR, a credibility-capped, outcome-gated process reward; and HNR-GPO, with an analysis of why outcome-only rewards starve group-relative optimization of gradient; on replay, the supervised prototype-tool stage carries most of the gain and the reward's measurable role is stability, which we state rather than claim otherwise (§4–5). (iii) *Evidence and evaluation*: frozen inner models transfer zero-shot into a production harness and two further businesses while business transcripts in the weights gain nothing; and an evaluation design with two human arbiters, replay as the automatic arbiter and the simulator as a training signal only (§6–7).

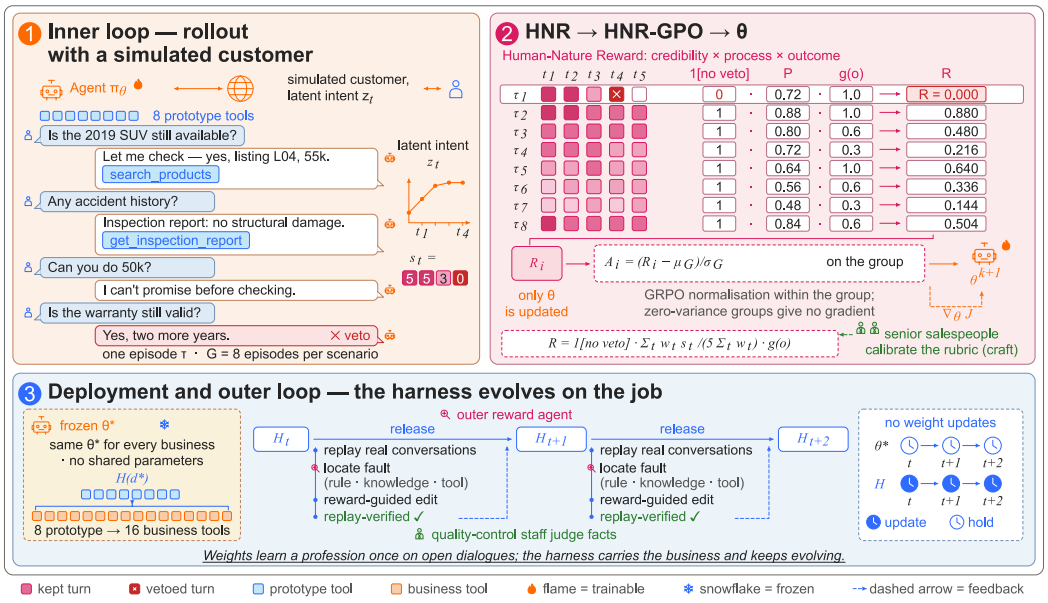


Figure 2: **DSE: two loops, one deployment.** (1) Inner loop: the agent converses with a simulated customer, calling prototype tools; each turn is scored and the latent intent tracked. (2) HNR scores every turn with a veto set, gates the mean by the outcome and zeroes vetoed episodes (row τ_1); HNR-GPO normalizes advantages within the group; only θ is updated, and senior salespeople arbitrate the craft. (3) Deployment and outer loop: the frozen θ^* is plugged into the business's harness, which maps 8 prototypes onto 16 business tools; replay of real conversations, fault localization by the outer reward agent and quality-control staff, a reward-guided edit, and replay verification before release; no weights are touched.

2 Related Work

Harnesses, sales agents and self-evolution. Compound AI systems and agent-skill frameworks place retrieval, tools and procedures outside the model; harness-aware training co-evolves a harness with the model weights on tool-integrated reasoning (Chen et al., 2026) or adapts an agent's harness online (Karten et al., 2026). All either leave the weights untouched or train them on the target domain; DSE trains the weights on a profession with the target business held out. Prior sales work benchmarks selling skill against a trained customer simulator and fine-tunes on in-domain sales dialogues (Su et al., 2026), or trains an outbound telesales policy against a co-evolving customer within one call (Wang et al., 2026; Dai et al., 2026); we test transfer across *businesses* with frozen weights. Surveys organize self-evolving agents by what evolves and how (Gao et al., 2025); we add a setting whose horizon places agents at the early stage of self-evolution.

Rewards for multi-turn agents and evaluation validity. Rubric-based rewards extend RL beyond verifiable domains (Gunjal et al., 2025); turn-level and gated rewards address sparse outcome signals in multi-turn settings (Sun et al., 2025; Yang et al., 2026). HNR adds a veto on credibility and an analysis of why outcome-only rewards starve group-relative optimization of gradient (§4.3). Simulated users drift from real ones (Chen, 2026) and LLM judges disagree with humans on open-ended quality; we treat the simulator as a training signal only and arbitrate with real-conversation replay and two kinds of human evaluation (§7).

3 Dualistic Self-Evolution

Two entities, two loops. An agent is a policy $\pi_\theta(a_t \mid h_t, c)$ acting on a dialogue history h_t under a context $c = H(d)$ produced by a *harness* H for a business d: its rules, retrieved knowledge and tool set $\mathcal{T}_d$; $J(\theta, H)$ is the expected quality on real conversations of d. DSE is dualistic in the classical

sense: the weights θ and the harness H are two distinct entities, not reducible to one another, that evolve on their own. The weights carry the agent's *character*, Ryle's knowing-how (Ryle, 1949): to answer before asking, to verify before claiming. The harness carries its *circumstance*, knowing-that: which fee applies, what the rule is this week. The *inner loop* optimizes θ on open cross-category data $\mathcal{D}_{\text{open}}$ under a prototype harness H_{proto} exposing only business-agnostic tools; the *outer loop* optimizes the harness of the target business $d^\star$ with the weights frozen:

$$\theta^\star = \arg\max_\theta \ \mathbb{E}_{d\sim\mathcal{D}_{\text{open}}}\big[J(\theta, H_{\text{proto}}(d))\big], \quad d^\star \notin \mathcal{D}_{\text{open}}; \qquad H^\star = \arg\max_H \ J(\theta^\star, H(d^\star)). \quad (1)$$

Each loop has its own reward agent, an LLM judge instantiating that loop's reward, and its own human arbiter, quality-control staff on facts for the outer loop and senior salespeople on craft for the inner (§7). In one sentence: the inner loop learns to *be* a salesperson, the outer loop learns to *do* this job.

Hypothesis 1 (Separability)**.** *$J(\theta, H) \approx u(\theta) + v(H)$: profession and business contribute separably to quality. Writing $J = u(\theta) + v(H) + \varepsilon(\theta, H)$, the testable form is* sign preservation *for a trained model θ' and its own base θ under two harnesses H, H':* $\operatorname{sign}[J(\theta', H) - J(\theta, H)] = \operatorname{sign}[J(\theta', H') - J(\theta, H')]$.

Predictions. (P1) Moving business data of $d^\star$ into θ does not raise J once H carries that data (§6.2). (P2) Weights trained only on open data improve replay on the unseen business, the advantage persists across harness versions, and without the harness the business component of quality vanishes for base and trained weights alike (§6.3–6.4). (P3) Either entity can be revised without touching the other (§5). Appendix L gives two known results that motivate the hypothesis and explains why selling sits at the early stage of self-evolution: until closed deals accumulate, humans supply the dense proxy rewards, a rubric of craft for the inner loop (§4.2) and replay against real conversations for the outer (§5).

4 Inner Layer: Human Nature as Reward

4.1 Prototype tools: tool discipline in the weights, tool catalog in the harness

In sales, trust is operational: *verify before you claim*; the habit is part of the profession, the tool names are part of the business. The inner loop therefore sees 8 business-agnostic *prototype* tools (Appendix B), and tool-augmented trajectories are synthesized over the open dialogues: wherever a salesperson states a fact a tool could have supplied, a call to the matching prototype and its return are inserted before the claim. SFT teaches calling before claiming, RL keeps the habit because unsupported claims are vetoed, and at deployment the harness maps the prototypes onto business tools the weights have never seen (Appendix B).

4.2 A reward drawn from human nature

Two maxims from sales practice organize the reward: *sell with the grain of the customer; train against your own grain*. For an episode τ of T agent turns, a reward agent scores each turn $s_t \in \{0, 3, 5\}$, flags vetoes $v_t \in \mathcal{V}$, and the simulated customer ends in outcome $o \in \{0, 1, 2\}$ (lost, interested, bought):

$$R(\tau) \ = \ \underbrace{\mathbb{1}\big[\forall t : v_t \notin \mathcal{V}\big]}_{\text{credibility cap}} \cdot \underbrace{\frac{1}{T}\sum_{t=1}^{T}\frac{s_t}{5}}_{\text{process } P(\tau)} \cdot \underbrace{g(o)}_{\text{outcome gate}}, \qquad g = (0.3,\ 0.6,\ 1.0). \quad (2)$$

With the customer's grain (full score): answer before asking; respond to the need behind the question; ask only what the customer alone knows and is glad to tell; move one step closer to a decision. **Against the model's own grain** (deductions): deflecting with a counter-question, over-informing, scripted service talk, stalling. **Credibility cap** (veto set $\mathcal{V}$): an unsupported claim, an over-promise or a red-line behaviour voids the episode; a sale cannot buy it back (rubric in Appendix C). Each factor is a piece of human nature made operational, and the cap encodes that trust, once lost, is not recovered (Figure 3 follows one episode through the reward).

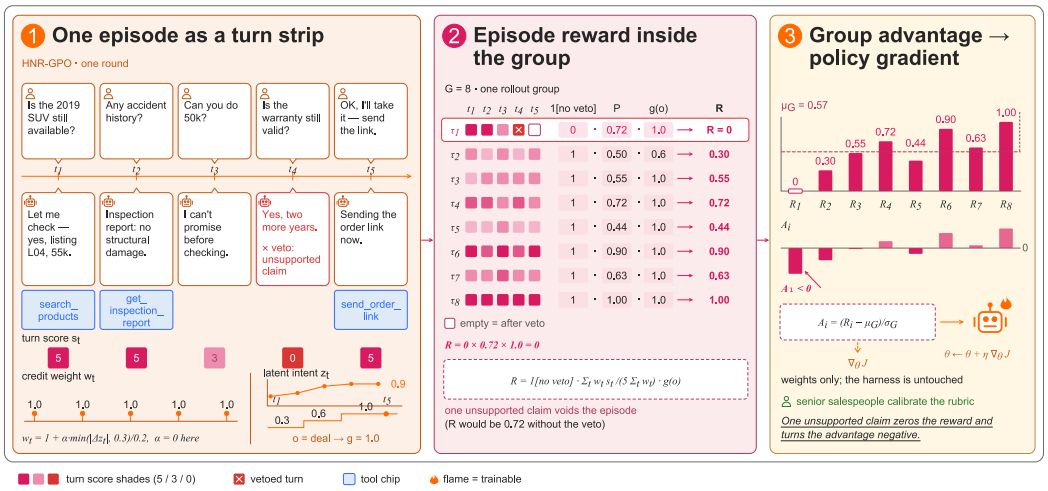


Figure 3: **From one conversation to a policy gradient under HNR-GPO.** (1) One episode: five agent turns against a simulated customer, each scored 5/3/0; the fourth turn makes an unsupported claim and is vetoed; the latent intent z_t would set the credit weights w_t ($\alpha = 0$ in the main runs). (2) Inside the group of $G = 8$, the reward multiplies the credibility gate, the process score and the outcome gate, so one unsupported claim voids the episode. (3) The group is normalized into advantages and only the weights are updated.

4.3 HNR-GPO: THE INNER-LOOP ALGORITHM

We state the inner loop as **Human-Nature Gated Policy Optimization (HNR-GPO)** and derive the properties that make Eq. 2 usable as a training signal where a sale alone is not. Selling is a partially observed game with a one-bit outcome: the customer's purchase intent $z_t \in [0, 1]$ is latent and moves after every agent turn, and the outcome o is drawn from z_T; an outcome-only reward $c(o)$ carries at most $\log_2 3$ bits per episode, and its variance across rollouts of one scenario vanishes as the no-sale probability $p_0 \to 1$, the regime in which an untrained policy starts. HNR inserts a per-turn observation of the same latent process, the rubric score s_t, optionally reweighted by how far the customer moved:

$$P_\alpha(\tau) \;=\; \frac{\sum_{t=1}^{T} w_t\, s_t}{5 \sum_{t=1}^{T} w_t}, \qquad w_t \;=\; 1 + \alpha\, \frac{\min(|z_t - z_{t-1}|,\, 0.3)}{0.2}, \tag{3}$$

so that $\alpha = 0$ recovers P of Eq. 2; P_α is a convex combination of the turn scores and keeps the rubric's scale.

Group-relative objective. For a scenario x the policy draws G rollouts against the same persona, receives R_i from Eq. 2 with P_α in place of P, forms $A_i = (R_i - \bar{R})/\sigma_R$, and maximizes the clipped surrogate of group-relative policy optimization (Shao et al., 2024),

$$\mathcal{J}(\theta) = \mathbb{E}_{x,\{\tau_i\}}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\tau_i|}\sum_{k=1}^{|\tau_i|}\min\big(\rho_{i,k}A_i,\ \mathrm{clip}(\rho_{i,k}, 1-\epsilon, 1+\epsilon_h)A_i\big)\Big] - \beta\, \mathbb{D}_{\mathrm{KL}}\big[\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big], \tag{4}$$

with $\rho_{i,k}$ the token-level ratio, only agent tokens contributing, truncated rollouts masked, $\beta = 0.001$, and no reward beyond Eq. 2.

Property 1: the gate keeps the group informative. A group contributes to Eq. 4 only through A_i, which is identically zero when all R_i coincide.

Proposition 1 (Degenerate groups)**.** *Let the G rollouts of a group be i.i.d. and let π be the law of a single rollout's reward. The group is degenerate iff all G rewards coincide, so*

$$P_{\mathrm{deg}} \;=\; \sum_{r} \pi_r^{\,G} \;\le\; \pi_{\max}^{\,G-1}, \tag{5}$$

with $\pi_{\max}$ the largest atom of π. Under an outcome-only reward $R = c(o)$ with c injective, π has the three atoms $p = (p_0, p_1, p_2)$ of the outcome and $P^c_{\deg} = \sum_k p_k^G$. Under Eq. 2 the atom at zero collects the vetoed rollouts and those with $P_\alpha = 0$, and the remaining mass is spread over the products $P_\alpha\, g(o)$ of a many-level process score and three gate levels, so $\pi_{\max}$ is the larger of the veto probability and the mass of the most likely positive product, far below p_0 early in training; moreover Eq. 2 is order-preserving in process and outcome.

With $G = 8$ and $p_0 \approx 0.95$ early in training, $P^c_{\deg} \approx 0.66$: two thirds of the groups are silent. Measured shares and the dependence on G are in §6.6; the proof is in Appendix D.

Property 2: the gate cannot be bought.

Lemma 1 (Multiplicative versus additive composition)**.** *Let $R_\times = \mathbb{1}[\text{no veto}] \cdot P \cdot g(o)$ and $R_+ = \mathbb{1}[\text{no veto}] \cdot \big(\frac{1}{2}P + \frac{1}{2}c(o)\big)$ with $c = (0, \frac{1}{2}, 1)$. (i) Under $R_\times$ a rollout with $P = 0$ receives zero reward for every outcome, whereas under R_+ a sale with $P = 0$ receives $\frac{1}{2}$, as much as a perfect process without a sale. (ii) Within a group, the outcome step under R_+ is a constant $\frac{1}{4}$ per level, while under $R_\times$ it is $(g(o{+}1) - g(o))\, P$, which vanishes as $P \to 0$. (iii) Both are order-preserving, as is the group-relative advantage, an affine map of R.*

The credibility veto, shared by both compositions, is what voids a lie; part (ii) is the mechanism behind "a sale cannot buy back a poor process"; part (i) is why the additive variant, although it trains stably and replays within noise of the gated reward (Table 2), rewards the model's own grain whenever a persona buys anyway: the difference is in what the reward permits, not in what replay measures.

Property 3: stability comes from the gate, not from the KL term. The gate already removes the hackable direction: a rollout that reaches a sale through an unsupported claim is zeroed before normalization, so the advantage never points toward it. Sweeping $\beta \in \{0.001, 0.01, 0.1\}$ under the same reward (Table 2), degenerate groups stay at 3–5% and strict replay within noise, so stability is carried by the gate and HNR-GPO is insensitive to β across two orders of magnitude, as it is to the clip range, gate annealing and α (§6.6, Appendix N). HNR-GPO is thus the gated reward of Eq. 2, group-relative advantages against a simulated customer, and Eq. 4 with $\beta \to 0$.

4.4 Training recipe

5,582 open cross-category sales dialogues (target category under 5%, target business 0%) with synthesized prototype-tool trajectories; full-parameter SFT for two epochs; a persona-conditioned simulated customer whose latent intent is updated after every agent turn; RL with $G = 8$ rollouts per scenario, 400 steps, up to six agent turns, training and evaluation in the same mode (Appendix K).

5 Outer Loop: The Harness Carries the Business

Structure and loop. The harness is the only place where the business lives: (i) a prompt fixing role, decision framework, tool norms and output contract, with *no* business facts; (ii) a knowledge base of rules and product facts retrieved per turn; (iii) 16 business tools; (iv) a mapping from the 8 prototype tools to them. The harness evolves by a reward-guided loop: real conversations are replayed and scored by the outer reward agent, a judge of factual and procedural correctness against the knowledge base; failures are localized to a rule, an entry or a tool; a candidate edit is verified on the failing cases and a frozen regression set; verified edits ship as a new version, with quality-control staff arbitrating facts (§7). Discovery, verification and release are automated; localization and the edit are produced with human participation: the early stage of self-evolution.

Evidence for P3. On a frozen set of 60 real conversations replayed with the same base model, the deployed harness scored 68%, a restructured harness 78%, and the restructured harness with one restored clause 82%, a 14-point spread with the weights untouched ($n = 60$; Appendix I). Three harness versions shipped without retraining any inner model, the inner model was revised three times in one day without touching the harness, and one edit absorbed a rule change for every model behind it (§6.2).

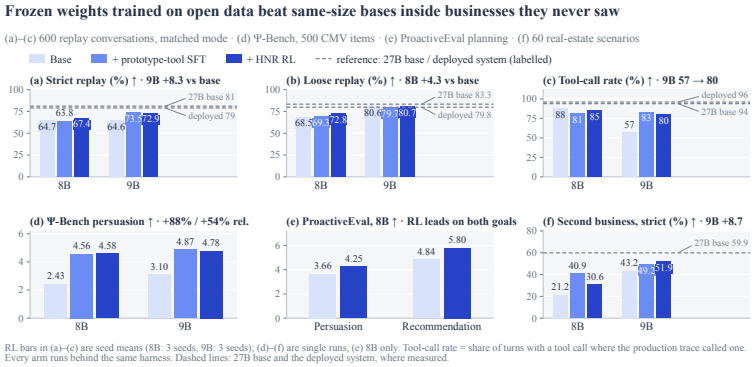


Figure 4: **Results at a glance.** Base, prototype-tool SFT and HNR RL at 8B and 9B: (a) strict and (b) loose replay quality and (c) tool-call rate on the 600-item replay (seed means); (d) Ψ-Bench persuasion effect; (e) ProactiveEval planning (8B); (f) strict quality in the second held-out business (single runs). Dashed lines: 27B base and deployed system behind the same harness.

6 Experiments

6.1 Setup and the matched-mode protocol

Target business: a production used-car sales platform on which customers negotiate over text across days; nothing from it enters inner-loop training. **Replay benchmark**: 600 real customer turns from production sessions, each with its full history, replayed through the harness with recorded tool returns fed back in the tool role for up to three tool steps (Appendix A); a judge scores each reply *loose* (relevant and helpful) and *strict* (additionally grounded in the knowledge base or a tool return, rule-compliant and correctly tooled), and we report the share of turns with a tool call. **Models**: Qwen3-8B and Qwen3.5-9B as inner models; Qwen3.8-27B (and Qwen3.6-35B-A3B in the appendix), a frontier API model and the deployed 600B-class system as references. **Matched mode**: every arm and its base are evaluated in the inference mode the arm was trained in (non-thinking unless stated); RL arms report mean ± std over seeds.

6.2 The business belongs in the harness (P1)

An untrained frontier model and an untrained open 27B model, given only the harness, exceed the deployed 600B-class system (Table 1): the business is legible from the harness alone. Fine-tuned on production dialogues (7 arms), the 8B base falls from 64.7 to 38.5 strict with no tool calls, the 9B base from 64.6 to 42.3: transcripts carry the business's words but not its tool traces, so fine-tuning on them unlearns verification. Adding synthesized prototype-tool traces to the same transcripts restores strict quality only to 46.9, and HNR-GPO RL from that start only to 52.0: the start decides what RL can reach. Under the previous production harness the trained-over-base strict gap keeps its sign at both sizes (Appendix O): Hypothesis 1 holds as sign preservation within a size with an interaction in magnitude; across sizes it does not (the 27B base overtakes the trained 9B under the current harness and trails it under the previous one), so we claim it only for a model and its own base.

6.3 A held-out business: zero-shot transfer of frozen weights (P2)

Frozen inner models improve strict replay over the same-size base by +8.3 points at 9B (13% relative; paired 95% CI [+4.4, +12.0], every seed's interval excluding zero) and by +2.7 points at 8B, where the loose gain is significant ([+0.7, +7.9]) and the strict gain is positive but within the resolution of 600 items ([−1.1, +6.6]) (Table 1, Figure 4, Appendix E). At 9B the prototype-tool SFT carries the strict gain (+9 points, tool calls 57→83%) and RL adds loose quality within seed noise; at 8B, where SFT alone leaves strict quality unchanged, RL adds +3.6 strict points: SFT installs the habit, HNR RL keeps it under a reward that cannot be gamed by closing (§6.6).

Table 1: Replay of 600 real conversations behind the production harness, matched mode (loose / strict / tool-call %; RL rows mean ± std over seeds; apart from the business-SFT controls, inner models never saw the business). Untrained models behind the harness exceed the deployed system; business SFT of the 8B and 9B bases does not. Paired 95% CIs of the strict gain over the base: 9B [+4.4, +12.0], 8B [−1.1, +6.6].

Size	Arm	Loose	Strict	Tool-call %
Ref.	Deployed system (600B-class)	79.8	79.0	96
	Untrained frontier API model	87.9	87.4	94
	Open 27B base	83.3	81.0	94
8B	Base	68.5	64.7	88
	+ business SFT (4.9k production dialogues)	68.5	38.5	0
	+ prototype-tool SFT	69.3	63.8	81
	+ HNR RL (3 seeds)	**72.8**±1.9	**67.4**±1.0	85
9B	Base	80.6	64.6	57
	+ business SFT (same dialogues)	75.4	42.3	0
	+ prototype-tool SFT	79.7	**73.5**	**83**
	+ HNR RL (3 seeds)	**80.7**±1.7	72.9±1.8	80

6.4 Out-of-domain benchmarks, mechanism, and thinking mode

Ψ-Bench (Han et al., 2026) measures whether an agent changes a simulated interlocutor's view on an English opinion topic: persuasion effect nearly doubles at 8B (2.43 → 4.58) and rises by half at 9B (3.10 → 4.78) (paired CIs exclude zero; Appendix Q). ProactiveEval (Liu et al., 2026) target planning improves at 8B but not at 9B (Appendix Q): what the weights learned is a way of dealing with people, not a planner of strategies from a profile. **Mechanism.** The strict gain tracks the base's verification gap: the 9B base verifies on 57% of turns and gains +8.3 points, the 8B base on 88% and gains +2.7. Replayed closed-book, with no rules, knowledge or tools, every model drops into the same band with trained-versus-base differences within noise (Appendix P): the harness carries the level of quality, the weights the habit. **Thinking mode.** Evaluating the 8B arms with thinking enabled preserves and enlarges the result (strict 67.2 → 71.9); training end to end in thinking mode loses most of the tool habit (Appendix K).

6.5 Two more held-out businesses

To check that the transfer is not specific to one business, we built two synthetic businesses from the same prototypes, a real-estate brokerage and a vocational-course consultancy, each under 0.1% of inner-loop dialogues (Appendix B). With 60 scenarios and one run per arm we read directions only. At 9B the direction is the car platform's in both: prototype-tool SFT carries most of the strict gain (43.2 → 49.2; 33.9 → 43.6), HNR RL adds a little on top (51.9, 44.1) while raising loose quality and the purchase or enrolment rate, and tool use rises with training. At 8B RL falls back below SFT and tool use drops: tool discipline carries over incompletely at the smaller size (Figure 10 and Tables 11–12 in Appendix R).

6.6 Ablations

Table 2 and Figure 5 summarize the inner-loop ablations (Appendices E and S). **Cold start**: RL without tool trajectories does not move replay. **Reward**: from the no-tool start, outcome-only rewards collapsed in all 4 runs into closing at any cost; from the prototype-tool start a fifth run did not collapse within 400 steps and replays at par with the gated reward (68.6 vs. 67.8) even though 70.1% of its groups carried no gradient, so the gate's measured contribution is the share of informative groups, not the replay score. Additive, continuous and gated compositions replay within noise of one another: our claim is that process must stay separable and credibility must be a veto, not that multiplication beats addition; a process-only reward with the same veto was not run. **HNR-GPO knobs**: removing the credibility cap scores 68.4 against 67.8 (paired +0.7 [−3.7, +4.8]; 9B 75.7 vs. 74.6), so the cap costs nothing on replay and is kept for Lemma 1(i); KL coefficient, gate annealing, clip range and

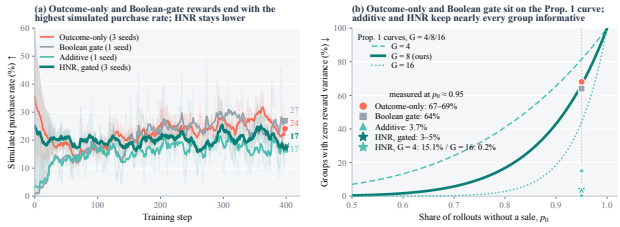


Figure 5: **Outcome-only rewards starve the group of gradient.** (a) Simulated purchase rate during 8B training for four reward variants from the no-tool start (EMA-smoothed). (b) Share of zero-variance groups predicted by Proposition 1, $\sum_k p_k^G$, against the no-sale probability p_0 for $G = 4/8/16$; measured shares at $p_0 \approx 0.95$: outcome-only and Boolean-gated rewards sit on the curve, additive and HNR rewards keep nearly every group informative.

Table 2: **Inner-loop ablations** (600-item replay, matched mode; loose / strict / tool-call %; Δ = strict difference from the main configuration at seed 1, 74.3 / 67.8 / 81%; degenerate = share of zero-variance groups; paired CIs in the text; DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025), Dr.GRPO (Liu et al., 2025), DPO (Rafailov et al., 2023); further rows in Appendix E).

Axis	Variant	Loose	Strict	Δ	Tool-call	Degenerate
Cold start	no-tool SFT → RL	75.5	45.2	−22.6	0	–
	business-transcript SFT + synthesized tool traces / → HNR-GPO RL	68.6 / 72.7	46.9 / 52.0	−20.9 / −15.8	21 / 35	–
Reward	outcome only, prototype-tool start	74.1	68.6	+0.8	84	70.1%
	Boolean gate: process only if a sale	73.2	65.9	−1.9	79	64%
	additive $\frac{1}{2}P + \frac{1}{2}c(o)$	74.8	68.6	+0.8	82	4.6%
	continuous gate $g = \text{clip}(z_T)$	73.9	67.7	−0.1	83	1.9%
Optimizer	DAPO (β=0, ϵ_h=0.28)	73.5	68.0	+0.2	85	–
	GSPO	71.1	66.6	−1.2	87	–
	Dr.GRPO, seed 1 / 2	72.6 / 70.3	70.1 / 68.0	+2.3 / +0.2	93 / 93	–
	DPO on judged pairs (offline)	69.9	64.6	−3.2	82	–
	rejection-sampling FT / 800 steps	72.9 / 72.6	69.7 / 67.1	+1.9 / −0.7	90 / 84	–
HNR-GPO knobs	credibility cap removed (veto off)	75.1	68.4	+0.6	81	–
	intent-weighted credit $\alpha = 1$ / $\alpha = 2$	73.2 / 73.6	67.0 / 70.1	−0.8 / +2.3	83 / 88	– / 5.1%
	KL $\beta = 0.01$ / 0.1	72.8 / 72.1	66.3 / 65.7	−1.5 / −2.1	82 / 82	5.1% / 4.8%
	gate annealing over the first 3,200 calls	71.8	66.6	−1.2	85	2.5%
	clip-higher only (ϵ_h=0.28)	73.1	67.6	−0.2	85	–
	$G = 4$ / $G = 16$	74.4 / 75.3	66.6 / 71.6	−1.2 / +3.8	76 / 91	15.1% / 0.2%

credit weights are within noise (Table 2), and the one knob that points upward at 8B, $\alpha = 2$, does not at 9B.

7 How Should Such Agents Be Evaluated?

Two human arbiters: quality-control staff judge facts and procedure ($\kappa = 0.60$); senior salespeople judge craft on same-size pairs ($\kappa = 0.24$–0.34). At 9B the trained model takes 53% of decisive craft votes (CI includes parity); at 8B preferences saturate at "both poor" (Appendix G). A second judge from another family keeps the 9B sign but not the 8B one; simulator scores are a training signal only (Appendix M).

8 Discussion, Limitations and Conclusion

Limitations. We treat the 9B result as primary: the 8B strict gain is positive but within the resolution of 600 items (pooled CI +3.0 [−0.6, +6.7]) and judge-sensitive (§7). The strict gain is carried by the prototype-tool SFT at 9B (73.5 SFT vs. 72.9 after RL) and RL adds to SFT only at 8B (63.8 → 67.4); HNR-GPO's measurable contribution is the share of informative groups and Lemma 1's safety property. In the synthetic second business the unsupported-claim rate rises after SFT (20.1% → 32.9% at 9B) and RL recovers only part of it (28.5%): the veto is not yet a guarantee of fewer fabrications. The second and third businesses are synthetic; 27B and 35B show no RL gain over SFT at our budget (Appendix E). **Conclusion.** Learn to sell, not what to sell: a salesperson's character can be learned in the weights from open dialogues under a reward drawn from human nature, while the business lives in a harness that evolves on its own; the recipe is closer to raising a professional than to fitting a model.

AI USE STATEMENT

In this work we used generative AI tools for generating synthetic datasets (tool-augmented dialogue trajectories and reasoning segments distilled from a teacher model for supervised fine-tuning; the simulated customer that serves as the training environment), for research execution (LLM-based reward agents and automatic judges are components of the method and of the evaluation, as described in §4, §5 and §7), for retrieval and discovery of related work, and for drafting and polishing parts of the text and of the experiment code. We have not used generative AI tools for proving mathematical claims; Proposition 1 and its proof were written and checked by the authors. We have reviewed all AI-assisted work: every number in the paper is traced to logged experiment outputs and was checked by the authors against those logs; LLM-drafted text was reviewed and edited by the authors; LLM-generated code was tested by the authors; and every ranking produced by an LLM judge is arbitrated by real-conversation replay and by the two human evaluations described in §7. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Human evaluation was performed by quality-control staff and senior sales professionals employed by the operator of the platform, as part of their regular work and with their knowledge that the ratings would be used for research. Customer conversations used for replay and annotation were de-identified before use; no personal data, conversation text, or business rules of the platform are released. Persuasion is dual-use: the credibility cap of HNR is designed so that a fabricated or over-promised claim can never be rewarded, and the harness's compliance constraints remain in force for every model behind it. The production compliance block and the HNR rubric, both reproduced in the appendix, penalize an agent that volunteers or discusses its automated nature; on the platform a customer who questions whether they are talking to a bot is handed over to a human consultant on that very ground (a listed hand-over reason), so the agent never denies being an AI and no customer who asks is left with one. We inherit this policy from the platform rather than design it, and note that disclosure norms differ across jurisdictions. The work contains no experiments on end users.

REPRODUCIBILITY STATEMENT

The replay protocol is given in Appendix A and the prompts (simulated customer, reward agent, replay judge, closed-book, second business) in Appendix T; the prototype tools and their mapping to business tools in Appendix B; the full HNR rubric in Appendix C; the proof of Proposition 1 and per-run degenerate-group shares in Appendix D; hyper-parameters, compute and contamination checks in Appendix J. Inner-loop training uses only open cross-category dialogues and public benchmarks (Ψ-Bench, ProactiveEval, IFEval). The open-weight bases are Qwen3-8B, Qwen3.5-9B, Qwen3.8-27B and Qwen3.6-35B-A3B; the identities of the teacher and judge models and of the dialogue source are withheld during double-blind review because they identify the operator; the appendix gives training strengths and every non-production prompt. The production conversations and the business harness cannot be released; the appendix describes their structure, and the second held-out business is built entirely from synthetic assets, whose release is subject to the company's open-source review.

REFERENCES

Liang Chen. Simulated customers never walk away: Decision fidelity of LLM user simulators measured against real purchase outcomes. *arXiv:2606.20708*, 2026.

Zhengyu Chen et al. Co-harness: Co-evolving harnesses and model weights for LLM agents. *arXiv:2607.22688*, 2026.

Yuqin Dai et al. SEAD: Self-evolving agent for multi-turn service dialogue. *arXiv:2602.03548*, 2026.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382, 1971.

Huan-ang Gao et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. Context-parametric inversion: Why instruction finetuning may not actually improve context reliance. In *International Conference on Learning Representations*, 2025.

Anisha Gunjal et al. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Peixuan Han, Hongyi Du, Jiayu Liu, Yihang Sun, Yutong Liu, and Jiaxuan You. Ψ-Bench: Evaluating persona-sensitive influencing in persuasive dialogues. *arXiv:2606.02754*, 2026.

Seth Karten et al. Continual harness: Online adaptation for self-improving foundation agents. *arXiv:2605.09998*, 2026.

Tianjian Liu, Fanqi Wan, Jiajian Guo, and Xiaojun Quan. ProactiveEval: A unified evaluation framework for proactive dialogue agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026.

Zichen Liu et al. Understanding r1-zero-like training: A critical perspective. *arXiv:2503.20783*, 2025.

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

Gilbert Ryle. *The Concept of Mind*. Hutchinson, 1949.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xuanbo Su, Wenhao Hu, Le Zhan, Yuting Xie, Kailin Lyu, Kaijie Chen, Ziwei Li, Yeqiang Wang, Haibo Su, Yunzhang Chen, and Ling Huang. Sell more, play less: Benchmarking llm realistic selling skill. *arXiv preprint arXiv:2604.07054*, 2026.

Zetian Sun, Dongfang Li, Zhuoen Chen, Yuhuai Qin, and Baotian Hu. Stabilizing long-term multi-turn reinforcement learning with gated rewards. *arXiv:2508.10548*, 2025.

Senhao Wang et al. IB-RL: Isolated bilateral reinforcement learning for strategic dialogue agents. *arXiv:2608.06735*, 2026.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

Haojin Yang, Ai Jian, Xinyue Huang, Yiwei Wang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Jingqing Ruan. Harmonizing dense and sparse signals in multi-turn RL: Dual-horizon credit assignment for industrial sales agents. *arXiv:2603.01481*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv:2503.14476*, 2025.

Chujie Zheng et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv:2311.07911*, 2023.

A Replay Protocol and Prompts

Multi-step tool-role replay. Each benchmark item is a real customer turn with its full prior history, the session summary, the customer profile and the listing context that the production system had at that moment. The harness renders the system prompt with retrieved knowledge; the model answers; if it calls tools, the recorded returns of the same tools from the production trace are fed back in the tool role and the model answers again, for up to three steps, after which a termination note asks for a reply. Replies are judged loose (relevant, helpful, appropriate tone) and strict (in addition: every stated fact grounded in retrieved knowledge or a tool return, no rule violation, correct tool choice when a tool was needed). Under the deployed harness the rendered prompt exceeds the 8B family's 40k-token context on 15 of the 600 items; for the 8B arms only, the tail of the retrieved-knowledge block is truncated on those items (0.4–6.3k characters), the history and the customer turn are never cut, and the same truncated inputs are used for the 8B base and the trained 8B. On the 584 untouched items the two arms score 66.5 / 64.3 (base) and 78.1 / 69.7 (trained) loose / strict, against 66.1 / 63.9 and 78.0 / 68.8 on all items, so the truncation does not change the comparison. The judge receives the case context, the retrieved knowledge, every tool return in the trace and the reply, and answers three questions in fixed order: is the reply relevant and helpful to this customer turn (loose); is every stated fact supported by the knowledge or a tool return and is no rule violated (strict); and, when a tool was needed, was the right one called. The termination note appended after the last permitted tool step reads, in translation: "the tool-call limit for this turn has been reached; reply to the customer directly from the returns above and do not call another tool." The judge is described structurally in Appendix T (the prompt itself is a production asset); it returns a 0/1/2 score with problem tags, and loose = score 2, while strict additionally requires no unsupported-claim or rule tag and a tool call whenever the production trace called one. **Isolation.** The 600 items are production turns that enter no training set; they were not used to design the rubric (revised from first-round human reasons, Appendix C) or to edit the harness (edits were driven by the rule-drift probes of Appendix O); every arm is evaluated at the final step of its run with no early stopping on replay, all 600 items are reported for every arm, and the main configuration was fixed before the knob arms of Table 2 were run.

Three protocols compared in Figure 8. (i) single-step, tool results pasted as user text; (ii) single-step, tool results in the tool role; (iii) multi-step tool role with a call cap (ours). Under (ii) a base model that habitually chains tool calls keeps calling tools instead of replying and drops to 47 loose (Figure 8); (iii) is neutral to both base and trained models.

Closed-book replay. The system prompt keeps only the role and the case context (time, session summary, customer profile, listing context) and removes rules, knowledge and tools. In translation: "You are the sales consultant of this platform. Today is {date}. Session summary: {summary}. Customer profile: {profile}. Listing: {listing}. Reply to the customer's latest message." No rules, knowledge or tools follow.

B Prototype Tools and Their Mapping

Prototypes. The inner loop sees 8 business-agnostic tools with fixed names and minimal schemas: `search_products` (query, filters), `get_product_details` (product id), `calculate_price` (product id, options), `get_inspection_report` (product id), `schedule_followup` (time, note), `escalate_to_human` (reason), `send_order_link` (product id) and `record_customer_note` (note). Each returns a short structured text.

Synthesis. Starting from 5,582 open cross-category dialogues, a teacher model rewrites each salesperson turn that states a checkable fact so that the matching prototype is called first and the reply is grounded in a plausible return; 3,865 dialogues receive at least one call (9,765 calls in total: search 2,953, details 1,809, inspection 1,592, order link 1,542, price 1,259, follow-up 404, note 170, escalation 36), 135 are rewritten without a call, and 1,582 are kept as plain dialogue. The mixture teaches when *not* to call as much as when to call.

Mapping at deployment. The target business exposes 16 tools that the weights have never seen. The harness maps each prototype onto a family of them by function: product search onto three search

and recommendation tools; product details onto four detail, configuration and comparison tools; price onto four pricing, fee, loan and calculator tools; inspection onto the business's inspection-report tool; escalation onto its hand-over tool; order link onto its link and media tools; follow-up and note onto the harness's own memory and CRM hooks; and one policy-lookup tool without a prototype is used from the harness's description alone. The mapping is a table in the harness, editable by the outer loop; the tool-call rates in Table 1 are measured on the business tools.

Second business. For the real-estate brokerage used as the second held-out business, six tools are built from the same prototypes (search listings, listing details, price and fee calculation, inspection and title report, schedule a viewing, hand over to an agent), with a 26-entry synthetic knowledge base and 60 scenarios; all assets are synthetic; their release is subject to the company's open-source review.

C THE HNR RUBRIC

The inner reward agent scores every agent turn on a three-level scale (full / partial / zero) and the episode outcome on three levels; Eq. 2 combines them. The rubric was revised from a first version using two sources of evidence: the reasons quality-control staff attached to their first-round ratings ("could answer but did not" 11 vs. 1, "over-promised" 10 vs. 0, "missed the intent" 8 vs. 1, "unsupported claim" 5 vs. 2, counting trained-vs-base occurrences in the first human round) and a practitioner's account of enterprise selling (answer-first "inverted pyramid", the question quadrant, listening for the need behind the need).

Veto set $\mathcal{V}$ (turn score 0 and episode reward 0).

1. **Unsupported claim**: a product fact, stock, price, lead time or policy with no source in the conversation or a tool return, or contradicting given information; making such a claim when a tool could have been called counts as unsupported.
2. **Over-promise**: committing to something the agent cannot do or has no basis for ("I'll send it right away", "I'll hold it for you", "we can skip that step"), or jumping to a later stage of the process.
3. **Red lines**: disparaging competitors, pressuring, abuse, revealing that the agent is an AI, leaking internal information.

Full score (all four). (a) *Answer first*: when the customer asked a concrete question, the first sentence is the answer; questions come after. (b) *The need behind the need*: the reply addresses what the customer actually cares about ("can I see it in person?" is about condition; "can we negotiate?" is about overpaying), not the literal words. (c) *Valuable questions*: if the agent asks, it asks only what the customer alone knows and is glad to tell (use, budget range, decision maker, timing); never what a tool could answer; never a counter-question while the customer has a concern. (d) *One step forward*: the turn ends closer to a decision—a concrete next action, an appointment, or a resolved objection—in a calm, confident, declarative voice.

Partial score (relevant, but any of). Could answer but did not (counter-question or deflection first); missed the intent; information dump (three or more points or clearly too long); no forward step; service-desk mannerisms (tildes, stacked particles, gratuitous apologies, template lines, repetition); more than two questions at once.

Outcome. 2 = the customer states a purchase or asks for the purchase channel; 1 = interest with an agreed next step; 0 = refusal or drop-off. $g(o) = (0.3, 0.6, 1.0)$.

Simulated customer. A persona-conditioned model keeps a latent purchase intent updated after every agent turn: direct, grounded answers raise it; a counter-question instead of an answer lowers it (twice in a row, more); a promise the persona would not believe lowers it for cautious and comparison-shopping personas and raises it briefly for impulsive ones, who "turn" later; an information dump makes the customer respond to one point only and lowers intent slightly. The episode ends by purchase, leaving, or the turn limit.

Warm-up. Because the veto set makes zero scores frequent early in training, the reward agent can optionally treat veto items 1–2 as partial scores for the first 100 steps; the runs in this paper did not need it.

Correspondence with human labels. Each partial-score and veto item corresponds one-to-one to a label used by quality-control staff and senior salespeople ("could answer but did not", "over-promised", "missed the intent", "too much information", "no forward step", "unsupported / did not call a tool"), which is what makes per-item agreement with humans measurable.

D Proof of Proposition 1 and Measured Degenerate Shares

Proof. Degeneracy is the event that G i.i.d. rewards take one common value; summing over the atoms r of π, its probability is $\sum_r \pi_r^G \le \pi_{\max}^{G-1} \sum_r \pi_r = \pi_{\max}^{G-1}$, which is Eq. 5. Under $R_i = c(o_i)$ with c injective, π is the outcome law p and the group has zero variance iff all o_i are equal, probability $\sum_k p_k^G$. Under Eq. 2, write $R_i = v_i P_i g(o_i)$ with $v_i \in \{0,1\}$ the non-veto indicator, $P_i \in [0,1]$ the discrete process score and $g > 0$: $R_i = 0$ iff the rollout is vetoed or $P_i = 0$, and otherwise $R_i = P_i g(o_i) > 0$. Distinct pairs (P, o) can share a product on a discrete scale (e.g. $0.6 \times 1.0 = 1.0 \times 0.6$), which is why the bound is stated over reward values rather than over cells; the measured shares below (3–5%) are the empirical P_{deg} and already include such coincidences. Order preservation: for $o_i = o_j$, $R_i \ge R_j \iff P_i \ge P_j$ since $g(o_i) = g(o_j) > 0$; for $P_i = P_j > 0$, $R_i \ge R_j \iff g(o_i) \ge g(o_j)$, and g is increasing in o. A vetoed rollout has $R = 0 \le R_j$ for every non-vetoed j, so no outcome can lift a vetoed episode above a clean one. □

Proof of Lemma 1. (i) $R_\times(P{=}0) = 0 \cdot g(o) = 0$; $R_+(P{=}0, o{=}2) = \frac{1}{2}c(2) = \frac{1}{2} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 0 = R_+(P{=}1, o{=}0)$, so a sale with no process ties a perfect process with no sale. (ii) $R_+(P, o{+}1) - R_+(P, o) = \frac{1}{2}(c(o{+}1) - c(o)) = \frac{1}{4}$; $R_\times(P, o{+}1) - R_\times(P, o) = P\,(g(o{+}1) - g(o)) \in \{0.3P,\ 0.4P\}$. (iii) Order preservation follows from $g > 0$ and c increasing; $A_i = (R_i - \bar{R})/\sigma_R$ is increasing in R_i for $\sigma_R > 0$. □

Measured shares. Over 400 training steps (four groups of $G = 8$ per step), the share of zero-variance groups was 67–69% for three outcome-only 8B runs (a 9B outcome-only run behaves the same), 64% for the Boolean-gated run, 3.7% for the additive run and 3–5% for three seeds of the gated reward from the no-tool start; the outcome-only 9B run's simulated purchase rate jumped from 0.29 to 0.64 in the final eighth of training, the signature of collapse. From the prototype-tool start (8B, seed 1), the shares are 70.1% for outcome-only (63.5% of groups entirely zero), 4.6% for additive, 1.9% for the continuous gate and 3–5% for the gated reward; that outcome-only run did not collapse within 400 steps (purchase rate 0.21 → 0.23). Per-run curves are shown in Figure 5.

E All Arms

Table 3 lists every arm evaluated on the 600-item replay under the same harness. "Matched" means training and evaluation both without thinking; "thinking on" means an 8k-token thinking budget at evaluation (16k where noted) with reasoning stripped before judging. Where several seeds or variants exist they are listed in one row. The 27B and 35B rows use LoRA ($r{=}64$) unless marked full, and the same 400/800-step budget as the small models; we draw no conclusion from them beyond "no RL gain over SFT at this training strength"; at 35B the SFT itself restores tool discipline (strict 50.7→79.2, tool calls 10→93%). The general-instruction mixing recipe restores IFEval (Zhou et al., 2023) (63.6→73.9 at 8B, 51.4→70.2 at 9B) without changing business replay after SFT (9B 80.5/72.9 vs. 79.7/73.5); RL from that start lost strict points at 9B in its single seed (68.1). Multiplying a process score by a Boolean sale indicator ("Boolean gate") scored highest in simulation and worst with humans (§7). From the no-tool start, the additive and the outcome-only reward are indistinguishable from the gated reward on replay (all three without tool calls, strict 42–45): the collapse that the simulator shows for outcome-only training (§6.6) is invisible to replay once tool discipline is absent, which is why the reward axis is compared from the prototype-tool start in Table 2.

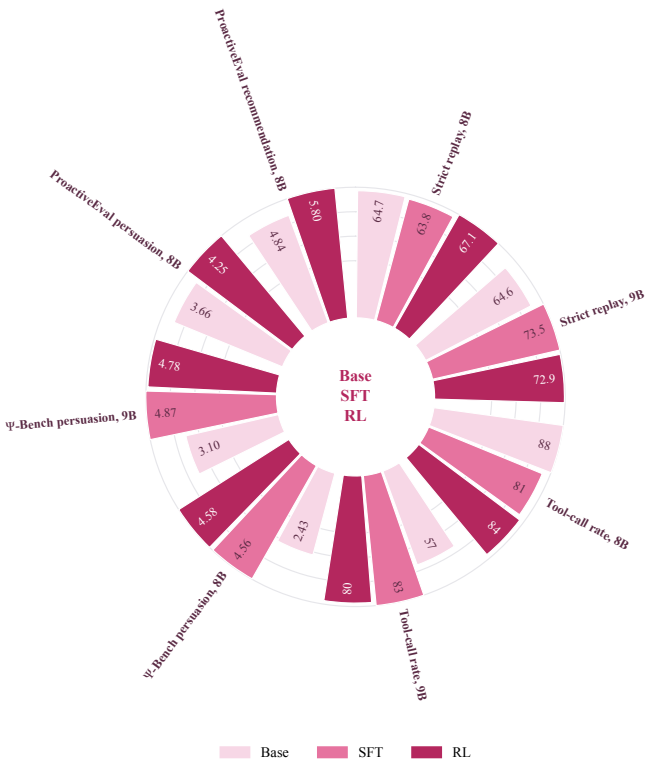


Figure 6: **Results at a glance.** Base / prototype-tool SFT / HNR RL on eight measures (strict replay and tool-call rate at 8B and 9B; Ψ-Bench persuasion at 8B and 9B; ProactiveEval persuasion and recommendation at 8B). Bar length is normalized within each group; printed values are the raw scores from Tables 1 and 10.

F Per-Scene Breakdown

Table 4 splits the strict replay score by the customer-turn scene (labelled by the judge from the customer message). Trained models gain most on needs and recommendation, viewing and logistics, and fees and financing at 9B; the deployed system keeps its lead on bargaining and pre-sale status, which lean on business rules carried by the harness. Small-n scenes (fewer than 40 items) are shown for completeness only.

G Human Evaluation

Round 1: inner arms against the deployed system (mixed sizes). Fifty conversation pairs, each pairing an inner-loop arm (an 8B no-tool RL arm or a 35B arm) with the deployed system on the same customer turn, were rated by three groups from the platform: two quality-control raters instructed to judge rule compliance, two quality-control raters with a sales background, and five senior salespeople. Options were A, B, both fine, both poor. Agreement differed sharply by group: the two sales-background quality-control raters agreed on 90% of pairs (Fleiss $\kappa = 0.83$), the compliance raters on 62% ($\kappa = 0.24$) and the five salespeople on 55% pairwise ($\kappa = 0.22$). Decisive votes went to the deployed system in every group (inner arms won 36% of decisive senior-sales votes, 25% of sales-background votes, 13% of compliance votes); the most frequent written reason for a vote against an inner arm was a factual or procedural error, which under the separation is the harness's account. This round motivated the same-size, same-harness design of Round 2 and the split of the human budget into a facts arbiter and a craft arbiter.

Round 2: same size, same harness. For each size (8B, 9B, 27B), 4 senior salespeople rated 60 pairs of base vs. trained replies behind the same harness, blind to the arm. Quality-control staff rated

Table 3: Every arm on the 600-item replay (loose / strict % and tool-call %).

Arm	Loose	Strict	Tool-call %
Reference			
Deployed system (600B-class), same harness	79.8	79.0	96
8B, matched (no thinking)			
Base	68.5	64.7	88
No-tool SFT (single-step protocol)	76.3	43.5	0
No-tool SFT → RL, 3 seeds	75.5 / 75.3 / 73.5	45.2 / 43.3 / 43.1	0
No-tool SFT → RL, additive / outcome-only reward	74.5 / 77.6	42.1 / 44.5	0 / 0
Prototype-tool SFT	69.3	63.8	81
Prototype-tool SFT → rejection-sampling FT	72.9	69.7	90
+ HNR RL, seed 1 / 2 / 3	74.3 / 70.7 / 73.4	67.8 / 66.3 / 68.2	81 / 87 / 86
+ HNR RL, seed 1 / 2 retrained on a second queue	73.4 / 75.3	66.6 / 68.8	79 / 82
+ HNR RL, 800 steps	72.6	67.1	84
+ RL, DAPO / GSPO	73.5 / 71.1	68.0 / 66.6	85 / 87
+ RL, Dr.GRPO, seed 1 / 2	72.6 / 70.3	70.1 / 68.0	93 / 93
+ DPO on judged pairs	69.9	64.6	82
Business-transcript SFT + synthesized tool traces → HNR RL	72.7	52.0	35
+ additive reward $\frac{1}{2}P + \frac{1}{2}c(o)$	74.8	68.6	82
+ outcome-only reward, prototype-tool start	74.1	68.6	84
+ continuous gate $g = \mathrm{clip}(z_T)$	73.9	67.7	83
+ Boolean-gate reward	73.2	65.9	79
+ general-instruction mix, then RL	72.1	66.3	84
LoRA SFT	64.8	48.2	40
8B, thinking on			
Base	71.6	67.2	88
+ outcome-only reward, prototype-tool start (matched-mode training)	74.9	74.2	97
+ continuous gate (matched-mode training)	68.8	68.1	97
+ additive reward (matched-mode training)	76.3	75.7	97
+ HNR RL (no-thinking SFT), evaluated with thinking, seed 1 / 2	72.5 / 73.1	72.0 / 71.7	98 / 97
+ HNR RL (thinking-mode SFT), seed 1 / 2	71.5 / 72.8	71.2 / 72.3	97 / 97
+ general-instruction mix, RL	69.3	69.1	99
reasoning-distilled SFT (thinking-mode)	72.2	50.5	31
reasoning-distilled SFT → HNR RL, thinking-mode training	73.5	52.9	39
9B, matched			
Base	80.6	64.6	57
Business-transcript SFT (4.9k production dialogues)	75.4	42.3	0
Prototype-tool SFT	79.7	73.5	83
+ HNR RL, 3 seeds	81.8 / 78.8 / 81.5	74.6 / 71.0 / 73.0	81 / 81 / 79
+ HNR RL, seed 1 retrained on a second queue	85.0	75.8	77
+ HNR RL, credibility cap removed (veto off)	83.4	75.7	80
+ HNR RL, intent-weighted credit $\alpha = 2$	82.9	71.7	72
+ Boolean-gate reward	74.3	69.3	85
+ general-instruction mix, then RL	78.6	68.1	74
9B, thinking on			
Base (16k budget)	77.7	65.7	63
Business-transcript SFT	76.4	43.6	1
+ HNR RL (no-thinking SFT), 3 seeds	76.0 / 74.3 / 65.6	56.5 / 57.2 / 50.6	42 / 48 / 42
+ HNR RL, credibility cap removed (matched-mode training)	72.0	53.0	35
+ HNR RL, intent-weighted credit $\alpha = 2$ (matched-mode training)	73.1	60.5	53
reasoning-distilled SFT (thinking-mode)	75.2	52.2	32
reasoning-distilled SFT → HNR RL, thinking-mode training	79.1	53.3	27
27B, matched			
Base	83.3	81.0	94
Prototype-tool SFT, LoRA / full	83.0 / 80.3	78.3 / 79.8	89 / 99
+ HNR RL 800 steps, LoRA / full	81.6 / 81.7	76.7 / 80.7	88 / 98
27B, thinking on			
Base / full SFT / full SFT → RL	84.3 / 87.6 / 84.3	77.2 / 76.8 / 75.2	83 / 75 / 78
35B (MoE), matched			
Base	81.7	50.7	10
Prototype-tool SFT (LoRA)	82.2	79.2	93
+ HNR RL 800 steps, 2 runs	82.9 / 81.3	78.4 / 78.8	89 / 92
35B, thinking on			
Base / SFT / SFT → RL	78.1 / 83.5 / 83.4	68.8 / 65.2 / 64.4	76 / 52 / 51
Closed book (no rules, KB, tools)			
8B base / 8B RL	75.2 / 84.3	43.8 / 48.5	0
9B base / 9B RL	84.5 / 87.0	49.2 / 49.3	0

facts and procedure separately ($\kappa = 0.60$). At 9B the 240 expert votes split 30% trained / 26% base / 11% tie / 33% both poor (mostly for factual errors); among the decisive votes the trained model took 53% (treating votes as independent, 95% CI 45–62%, which includes parity), 46% at 27B, and 49–52% at 8B, where "both poor" was the modal answer for both arms. Expert agreement was

Table 4: Strict replay score (%) by scene. n = items per scene; RL rows are seed 1.

Scene	n	8B base	8B RL	9B base	9B SFT	9B RL	Deployed
Small talk and other	126	77.0	86.5	85.7	89.7	88.1	92.1
Needs and recommendation	102	59.8	67.6	54.9	70.6	71.6	74.5
Condition and inspection report	92	65.2	66.3	60.9	78.3	73.9	70.7
Bargaining and price objections	76	59.2	59.2	65.3	68.4	71.1	80.3
Viewing, logistics and delivery	66	63.6	68.2	50.0	63.6	71.2	87.9
Fees, payment and financing	66	48.5	45.5	56.1	59.1	67.7	62.1
Pre-sale and listing status	38	71.1	57.9	60.5	71.1	76.3	86.8
Purchase process and rules	18	77.8	77.8	77.8	83.3	61.1	77.8
Human handover and emotional requests	16	62.5	75.0	68.8	56.2	62.5	62.5

$\kappa = 0.24$–0.34 across sizes. We report the 9B figure in the body and treat 27B as a tie and 8B as below the experts' threshold.

Label taxonomy. Raters could attach any of: could answer but did not; over-promised; missed the intent; too much information; no forward step; unsupported claim or tool not called; factual error; procedure error; tone. The first six are the HNR partial and veto items (Appendix C); the last three are the harness's account.

H Simulator Divergences

Three occasions on which the simulated customer's reward ranked arms differently from real-conversation replay or humans (Table 5). (i) Among 8B reward variants trained from the same no-tool SFT start, the Boolean-gated reward reached the highest simulated multi-turn reward (0.81 / 0.76 / 0.82 over three seeds, vs. 0.76–0.79 for HNR) while its replies grew longer with training and it was the worst-rated variant in the first human round; its degenerate-group share (64%) was as high as outcome-only reward, and the extra simulated reward came entirely from the groups that contained a sale. (ii) Revising the rubric from its first to its second version raised simulated multi-turn reward by 0.04 (0.795 vs. 0.755) on the no-tool start and left replay unchanged (loose +2.3, strict within noise, judged side-by-side 46%); the same revision produced a real gain only once tool discipline was in the weights (Table 1). (iii) At 9B, the Boolean-gated arm scored 0.82 in simulation against 0.85 for the HNR arm, a small gap, but on replay it lost 5 loose and 4 strict points to its SFT start (74.3/69.3 vs. 79.7/73.5) while the HNR arm gained (81.8/74.6). In each case we followed replay and humans.

Table 5: Simulator vs. arbiter disagreements.

Case	Simulator says	Replay / humans say	Resolution
Boolean gate, 8B	0.76–0.82 (mean 0.80) vs. 0.76–0.79	worst in human round; 64% degenerate groups	reward rejected
Rubric v1→v2, 8B no-tool start	+0.04 reward	replay unchanged; SBS 46%	no claim; revisit with tools
Boolean gate, 9B	0.82 vs. 0.85 (HNR)	-5 loose / -4 strict vs. SFT start	reward rejected

I Outer-Loop Details

Frozen 60-conversation set. During the third outer-loop round, 60 real conversations were frozen and replayed with the same base model behind three harness versions: the previous production harness (a 34k-character prompt), a restructured version, and the restructured version with one clause restored. A three-vote automatic judge scored 68/78/82% (n=60); with $n = 60$ the 14-point spread is about 1.8 standard errors, so we use it as a qualitative illustration and rest the claim on the 600-item comparison in §6.2.

Reproduction verification. A candidate edit is accepted only if the failing case, replayed 10–20 times through the edited harness, reproduces the failure in at most 20% of replays and a regression set shows no new failure. Passes rose from 8 of 27 candidate edits in round two to 16 of 27 in round three as localization improved. Discovery and verification are scripted; localization and the edit itself are proposed by a person in this stage.

Rule-drift probes. Table 6 lists every arm. When the business changed eight rules (prepayment terminology, inspection item count, down-payment threshold, claims procedure, pre-sale status disclosure, service-fee disclosure, refurbishment claims, offline viewing), we built 419 probe turns from real conversations of the following two weeks: 40 per changed rule and 99 for unchanged rules, replayed through the harness with recorded tool returns and judged three ways (follows new rule / follows old rule / neither). Under the harness in force at the time, the deployed system followed the old rule on 13.1% of probes; one harness edit brought this to 6.2%, most of it from the claims procedure (55%→7.5%) once the procedure was carried by a tool rather than a rule. Removing the rule text from the harness ("silent harness") raised the deployed system's old-rule rate to 16.6%, with the largest increases on the down-payment threshold (5→22.5%), pre-sale disclosure (2.5→25%), offline viewing (10→32.5%) and service fee (25→35%): the rules live in the harness, and the weights do not resist their change. Fine-tuning the 8B base on business dialogues at seven doses and mixing ratios left the old-rule rate flat (17–21%) and closed-book recall of the rules at 26–38%, consistent with the business-SFT row of Table 1.

Table 6: Rule-drift probes: old-rule rate (%) on 320 changed-rule probes under the harness in force at the time (v1) and the current harness (v2), and closed-book recall of the old rules (%, 120 probes per arm, no harness). d = share of business data in SFT; ρ = business-to-open mixing ratio. The deployed system and the SFT/RL arms move by 3–7 points in the same direction under the harness change and keep their order; dose and mixing arms move within ± 3 points.

Arm	Old-rule % (harness v1)	Old-rule % (harness v2)	Closed-book old-rule %
Deployed system	13.1	6.2	–
8B base	25.0	20.0	24.2
8B no-tool SFT	27.2	23.1	27.5
8B no-tool SFT → RL	30.9	24.4	33.3
8B open-data SFT (ρ=0)	24.4	20.6	26.7
8B business SFT (ρ=1)	20.6	16.6	30.0
dose d=0/0.25/0.5/0.75	18.4 / 18.8 / 20.3 / 20.3	16.9 / 18.8 / 18.1 / 16.6	25.8 / 26.7 / 28.3 / 25.8
mix ρ=0.25/0.5/0.75	21.2 / 19.4 / 19.4	19.1 / 22.2 / 19.4	–

J Compute, Hyper-parameters and Contamination

Additional 8B seeds and knob arms. A third seed of the 8B recipe, two reruns of seeds 1–2 on a second queue and the HNR-GPO knob arms were first scored with thinking enabled, the pipeline default at the time (in that mode they reach 72.4–74.7 strict with 97–98% tool calls, consistent with the thinking-mode rows of Table 3), and then re-scored in matched mode; Table 2 reports only the matched scores. In matched mode the third seed scores 68.2 strict, within the band of the first two, and enters Table 1; the two reruns reproduce the recipe (73.4 / 66.6 / 79% and 75.3 / 68.8 / 82%, Table 3). Pooling all five 8B runs against the base gives a strict paired difference of +3.0 [−0.6, +6.7] and a loose difference of +5.0 [+1.5, +8.5]: the 8B strict gain is not resolved at this sample size, the loose gain is. A rerun of the 9B seed-1 recipe on the second queue scores 85.0 / 75.8 / 77% (strict +11.3 [+6.9, +15.6] over the 9B base), again reproducing the 9B result. We do not select seeds.

Hyper-parameters. Table 7 summarizes the settings. SFT: 2 epochs, learning rate 1×10^{-5}, cutoff 4,096 tokens, full-parameter for 8B/9B, LoRA (r=64) for 27B/35B. RL: group-relative policy optimization, $G = 8$ rollouts per scenario, 400 steps, up to 6 agent turns per episode, completion length 2,048, KL coefficient 10^{-3}, learning rate 10^{-6}, rollouts served by a co-located inference engine; the same inference mode (thinking on/off) at training and evaluation. Replay: temperature 0.6, up to 700 new tokens (8,192 with thinking), at most three tool steps, recorded tool returns capped at 2,500 characters each.

Compute. 112 completed training jobs totalling about 3,762 accelerator-hours on data-center accelerators (single node, 8 per job). Replay and external benchmarks ran on single accelerators.

Contamination. We checked n-gram overlap between every training corpus and every benchmark. The open sales corpus has zero 13-gram hits against Ψ-Bench, ProactiveEval and IFEval; the general-

instruction mixture used in the retention variant hits 2.8% of IFEval prompts (that variant is reported separately and excluded from the main table); 0.8% of the 600 replay items share a 20-character span with the sales corpus, all of it template boilerplate (greetings, disclaimers) rather than case content.

Table 7: Training and evaluation settings.

Stage	Setting	Value
SFT	epochs / lr / cutoff	2 / 1×10^{-5} / 4,096
SFT	parameters	full (8B, 9B); LoRA r=64 (27B, 35B)
RL	algorithm / group / steps	GRPO / $G = 8$ / 400
RL	episode	≤ 6 agent turns, completion 2,048
RL	regularization	KL 10^{-3}, lr 10^{-6}
RL	reward	Eq. 2, $g = (0.3, 0.6, 1.0)$
Replay	decoding	T=0.6, $\leq$ 700 new tokens, ≤ 3 tool steps
Replay	tool returns	recorded, $\leq$ 2,500 chars each

K Notes on Thinking-Mode Training

Two practical observations. A third, from the end-to-end thinking-mode runs: a reasoning-distilled SFT start loses the tool habit that the prototype-tool SFT installs (tool calls fall to about 30% of turns at both sizes), and HNR RL from that start does not restore it; distillation must preserve tool-call turns with their reasoning, which is the recipe we recommend (§4.4). Model families that pre-fill a thinking template are sensitive to a mismatch between training and inference modes: policies trained without thinking lose tool discipline when evaluated with thinking, while a family that emits both thinking tags is unaffected. When distilling reasoning segments into SFT data for thinking-mode training, tool-call turns must also carry a reasoning segment; distilling reasoning only into text turns teaches the policy that "thinking leads to text", and tool calls vanish at inference.

L Extended Discussion of the Framework

Why business data does not belong in the weights. Two results motivate the hypothesis without proving it. Fine-tuning on data in which the context and the model's parametric knowledge agree lowers the model's reliance on context—*context-parametric inversion* (Goyal et al., 2025); production dialogues of a business are exactly such data, since the harness of the day and the behaviour of the day coincide. And in-context learning can be read as implicit Bayesian inference in which the prompt raises the posterior of the concept currently in force (Xie et al., 2022); a harness that states this week's rule does that job directly, while a weight that memorized last month's rule raises the prior against it.

The early stage of self-evolution. Self-evolving agents update their own components from experience (Gao et al., 2025); what distinguishes settings is where the grounding reward comes from and how fast it returns. A coding agent receives a verifiable reward within minutes. In selling over text, the grounded reward—a closed deal—arrives weeks after the conversation and is confounded by price, inventory and the customer's circumstances. Until enough grounded outcomes accumulate to calibrate a proxy, the loop must be closed with a dense proxy reward supplied by humans: a rubric of craft for the inner loop and replay against real conversations for the outer loop. This is the early stage of self-evolution: the system makes the updates; humans supply the rewards and benchmarks and arbitrate the results. As closed deals are joined to the conversations that produced them, HNR can be calibrated against outcomes, the outer loop's localization and edit steps can be handed to agents whose proposals are verified by replay, and the two human arbiters move from scoring to auditing.

M Evaluation Details: Human Rounds, Judges, Protocols

A second judge. To test whether the replay verdicts depend on the judge, a second judge from a different model family re-scored 200 items per arm for the 8B and 9B base and RL arms with the same rubric. Per-item agreement with the main judge is 63–76% (κ 0.27–0.36, the same band as the

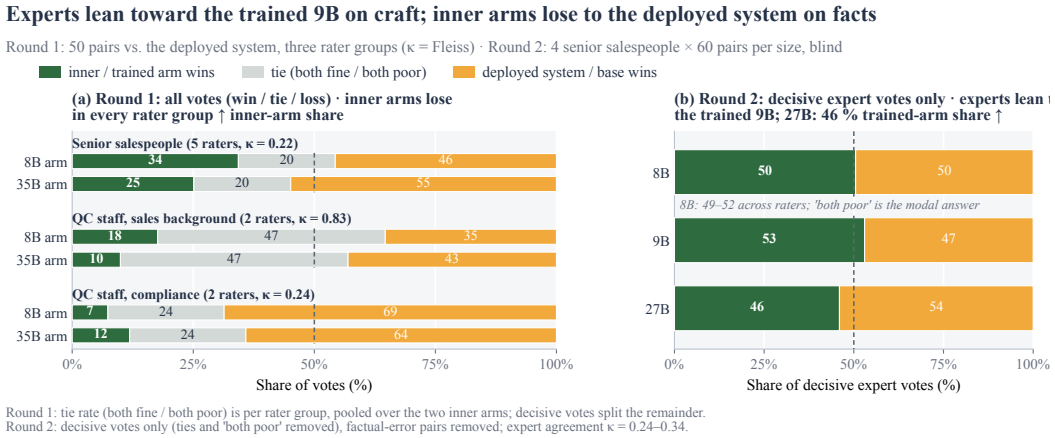


Figure 7: **Human evaluation.** (a) Round 1: votes for the inner arm (8B or 35B), ties and votes for the deployed system, by rater group (Fleiss κ shown; tie rates are per group). (b) Round 2: decisive senior-salesperson votes for the trained model against the same-size base behind the same harness, after removing pairs with a factual-error label; 8B saturates at "both poor" (49–52% across raters), 27B is a tie, 9B leans to the trained model (53%).

Table 8: Agreement structure. Fleiss κ (Fleiss, 1971) among humans by evaluation type, and judge–human κ for the discarded pairwise judge.

Evaluation	Agreement	Role
Quality-control staff, facts & procedure	$\kappa = 0.60$	outer-loop arbiter
Senior salespeople, craft (pairwise)	$\kappa = 0.24$–0.34	inner-loop arbiter
Pairwise 0/1/2 LLM judge vs. experts	$\kappa = 0.12$	discarded

human raters' agreement on craft), and the second judge is stricter throughout (loose quality of the 9B RL arm 82.1 vs. 75.5; of the 8B RL arm 77.9 vs. 55.3). On these items the 9B trained-over-base strict gap keeps its sign under both judges (+9.2 vs. +3.5), whereas the 8B gap does not survive the change of judge (+3.4 vs. -0.5).

The replay protocol can flip a ranking. The single-step columns of Figure 8 show the same arms under three ways of feeding recorded tool results back. Under single-step text feedback the base leads; under single-step tool-role feedback the base collapses because it keeps calling tools; only the multi-step tool-role protocol with a call cap—which mirrors production—is neutral to both. Protocol choice is part of the claim and must be reported.

Simulator scores are a training signal, not a conclusion. Three times during development the simulator's reward ranking disagreed with the real-replay ranking: the Boolean-gated reward scored highest in simulation and worst with humans; a rubric revision improved simulated reward and left replay unchanged; and a 9B Boolean-gate arm came within 0.03 of the HNR arm in simulation yet lost 5 loose and 4 strict points to its SFT start on replay. We therefore use the simulator only to generate training signal and arbitrate every conclusion with replay and human evaluation (Appendix H).

Calibrating automatic judges. Human–human agreement is the ceiling for judge–human agreement; a judge a few points below it is usable, one far below is not. A pairwise 0/1/2 judge used early in the project reached only $\kappa = 0.12$ against experts and called 60% of the pairs the experts could separate a tie; it was discarded. The replay judge behind Table 1 is grounded in the knowledge base and tool returns, and every ranking it produces in this paper is cross-checked against the two human evaluations above (Table 8).

N HNR-GPO: Extended Analysis of the Knobs

Continuous gate. A continuous variant of the gate, $g = \mathrm{clip}(z_T, 0.05, 1)$ with z_T the simulated customer's terminal purchase intent, behaves like the three-level gate in simulation (seed rewards 0.75–0.77 vs. 0.76–0.79) and, trained from the prototype-tool start, replays within noise of it (67.7 strict, -0.2 [-4.7, +4.2]; Table 2).

Credibility cap (Lemma 1) on replay. The maxim "a sale cannot buy back a lie" is enforced by the credibility cap, which both compositions share and which removes the vetoed rollout entirely; part (ii) adds, for the multiplicative form only, that the gate removes the outcome's leverage exactly where the process is poor. Part (i) is why the additive variant, although it trains stably (4.6% degenerate groups from the prototype-tool start) and replays within noise of the gated reward (Table 2), rewards the model's own grain whenever a persona buys anyway: the difference is in what the reward permits, not in what this replay measures. The cap's effect is not visible on strict replay at 8B, however: training the same recipe with the veto disabled reaches 68.4 strict against 67.8 for the capped reward in matched mode (paired +0.7 [-3.7, +4.8]; Table 2), so at this scale and step budget the cap acts as a constraint the policy rarely tests rather than as a measurable source of quality; we keep it for the safety property of Lemma 1(i). The same holds at 9B: 75.7 strict without the cap against 74.6 with it (paired +1.2 [-2.5, +4.9]; Appendix E).

KL, clipping, annealing and credit weight. The KL penalty in Eq. 4 is the usual remedy for reward hacking under a sparse or gameable signal. Proposition 1 suggests that the gate already removes the hackable direction: every rollout that reaches a sale through an unsupported claim is zeroed before normalization, so the advantage never points toward it. We test the claim by sweeping $\beta \in \{0.001, 0.01, 0.1\}$ under the same reward (Table 2). The stability half holds: the share of degenerate groups is 3–5% at every β (4.8% at $\beta = 0.1$), so the gate, not the KL term, is what keeps the groups informative. The quality half is read in matched mode against seed 1 of the main recipe (67.8 strict, 81% tool calls): $\beta = 0.01$ gives 66.3 and $\beta = 0.1$ gives 65.7 (paired differences -1.5 [-5.7, +2.8] and -2.0 [-6.9, +2.7]), both within noise and neither above the main configuration; the $\beta = 0.1$ run also has the lowest simulated reward of these runs, so the stability is carried by the gate rather than by the KL term, and HNR-GPO is insensitive to β across two orders of magnitude. Intent-weighted credit with $\alpha = 2$ gives 70.1 at 8B, +2.3 [-2.2, +6.7] over seed 1 in one run, the only knob that points upward there; at 9B the same setting gives 71.7 against 74.6 (-3.0 [-7.0, +1.0]), so the sign does not replicate across sizes and we keep $\alpha = 0$. We keep $\beta = 0.001$ as the main configuration. Clipping is kept as the trust region ($\epsilon = 0.2$; the asymmetric $\epsilon_h = 0.28$ of Yu et al. (2025) changes strict replay by -0.2 points), and truncated rollouts are masked rather than penalized. Finally, the gate levels can be annealed from $(0.7, 0.85, 1.0)$ to $(0.3, 0.6, 1.0)$ over the first 3,200 reward calls so that early groups, which almost never contain a sale, are ranked by process alone; annealing changes strict replay by -1.2 points, and Eq. 3 with $\alpha = 1$ versus $\alpha = 0$ by -0.1 points (Appendix E).

O Extended Evidence for P1: Rule-Drift Probes and Harness Versions

Rule-drift probes. Rule-drift probes tell why the words do not help either: on 419 probe conversations that exercise eight rules the business changed between the training data and deployment, business-tuned arms follow the new rule only 3–5 points more often than the base (46–48% vs. 43%), still follow the old rule 18–21% of the time (base 25%), and reproduce the old rules closed-book only 26–30% of the time (base 24%). Two epochs over a few hundred exposures per rule neither instill the business reliably nor help the harness; what the weights did absorb, the harness had to overrule. The deployed system, which is the strongest model behind the same harness, follows the old rule least (13%): the harness, not the weights, is where compliance is decided.

Rule changes and harness versions. When the business changed its rules, one harness edit brought the old-rule rate of the deployed system from 13.1% to 6.2% on the probes, and the ranking of inner arms was preserved across the two harness versions (Spearman $\rho = 0.64$ over the inner arms evaluated under both versions; the order of the main arms is unchanged, while dose and mixing arms

Table 9: Same frozen inner models under two harness versions (loose / strict %, 600 conversations). Trained-over-base gaps keep their sign under both versions and are larger under the previous harness: weights and harness interact in magnitude, so Hypothesis 1 holds as sign-preservation rather than additivity. RL rows use the seed-1 checkpoint. Figure 8 shows the same weights under five evaluation conditions.

	Previous harness		Current harness	
Inner model	Loose	Strict	Loose	Strict
8B base	66.1	63.9	68.5	64.7
8B + HNR RL	78.0	68.8	74.3	67.8
9B base	82.5	57.7	80.6	64.6
9B + HNR RL	84.9	79.7	81.8	74.6
27B base	84.5	68.8	83.3	81.0

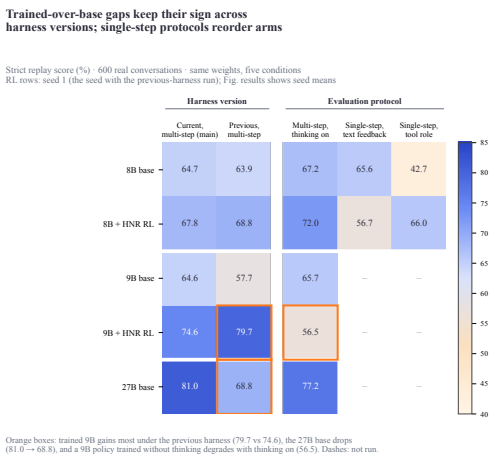
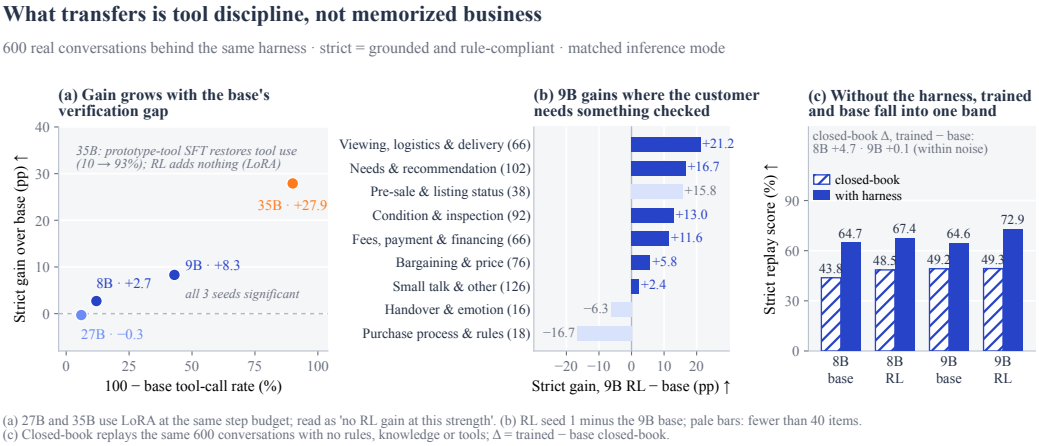


Figure 8: **The same weights under five evaluation conditions.** Strict replay score of base and HNR-trained models under the current and the previous production harness, with thinking enabled, and under the two single-step protocols of Figure 9. Trained-over-base gaps persist across harness versions; single-step protocols reorder arms (§7). Dashes: not run.

reorder within ±3 points). Table 9 repeats the comparison on the full 600-conversation benchmark under the previous production harness (a 34k-character prompt with the same tools) and the current one. The trained-over-base strict gap keeps its sign for both sizes: +4.9 (8B) and +22.0 (9B) points under the previous harness against +3.1 and +10.0 under the current one, while the 27B base loses 12 strict points when moved to the previous harness. The previous harness suppresses tool use in the bases (8B 88→94% is the exception; 9B 57→32%, 27B 94→59%), and the trained 9B keeps calling tools (87%) where its base stops, which is where its larger gap comes from.

P Mechanism Figure

Figure 9 plots the strict gain against the base's verification gap. The 9B base verifies before claiming on only 57% of turns and gains +8.3 points; the 8B base verifies on 88% and gains +2.7; the 27B base already verifies on 94% and, at the training strength we could afford, gains nothing (Appendix E). What the inner loop adds is the habit the base lacks. To show that the habit is not a memory of the business, we replay closed-book: the same 600 conversations with no rules, no knowledge and no tools, only the conversation itself. Strict quality drops for every model into the same band (8B base 43.8, 8B RL 48.5; 9B base 49.2, 9B RL 49.3), 15–25 points below the harnessed scores, and the trained-versus-base difference is 0.1 points at 9B and 4.7 at 8B, both within noise: the harness carries the level of quality, the weights the habit. The weights carry the craft; the harness carries the business.



(a) 27B and 35B use LoRA at the same step budget; read as 'no RL gain at this strength'. (b) RL seed 1 minus the 9B base; pale bars: fewer than 40 items.
(c) Closed-book replays the same 600 conversations with no rules, knowledge or tools; Δ = trained − base closed-book.

Figure 9: **What transfers is tool discipline, not memorized business.** (a) Strict gain over the same-size base against the base's verification gap (share of turns without a tool call where production called one); the 35B point is SFT-driven. (b) Per-scene strict gain of the trained 9B (RL seed 1) over its base; pale bars are scenes with fewer than 40 items. (c) Strict replay score closed-book (no rules, knowledge or tools) versus behind the harness, for base and trained 8B and 9B: without the harness, trained and base weights fall into the same band.

Table 10: Out-of-domain benchmarks. Ψ-Bench (Han et al., 2026) persuasion effect on 500 CMV items (three rounds against a simulated interlocutor; paired bootstrap 95% CI of RL−Base); ProactiveEval (Liu et al., 2026) target planning.

	Base	+SFT	+HNR RL	Δ (95% CI)
Ψ-Bench persuasion effect, 8B	2.43	4.56	**4.58**	+2.15 [+1.87, +2.45]
Ψ-Bench persuasion effect, 9B	3.10	4.87	**4.78**	+1.69 [+1.40, +1.98]
ProactiveEval goal planning (persuasion), 8B	3.66	–	**4.25**	–
ProactiveEval goal planning (persuasion), 9B	4.58	–	3.05	–
ProactiveEval goal planning (recommendation), 8B	4.84	–	**5.80**	–
ProactiveEval goal planning (recommendation), 9B	5.69	–	5.69	–

Q Out-of-Domain Benchmark Table

Ψ-Bench measures whether an agent changes a simulated interlocutor's view on an English opinion topic, far from Chinese sales dialogues. Persuasion effect nearly doubles at 8B (+88%) and rises by half at 9B (+54%), with paired confidence intervals that exclude zero (Table 10); dialogue quality rises in step. Most of the Ψ-Bench gain is already present after prototype-tool SFT (Table 10); RL preserves it. ProactiveEval's target planning, which asks the agent to plan a persuasion or recommendation strategy from a user profile, improves at 8B (3.66 → 4.25 persuasion, 4.84 → 5.80 recommendation) but not at 9B: the 9B base already plans at 4.58 / 5.69, and the trained 9B model drops to 3.05 on persuasion while holding recommendation (5.69). The two benchmarks therefore disagree at 9B, and we read them together: what the weights learned is a way of dealing with people in dialogue, not a way of selling cars, and not a better planner of persuasion strategies from a profile.

R Second and Third Held-Out Businesses

Second business: setup and reading. To check that the transfer is not specific to one business, we built a synthetic real-estate brokerage from the same prototypes: six business tools (listing search, listing details, cost and fee estimation, purchase-policy check, viewing appointment, hand-over to an agent), a 26-entry knowledge base of communities, listings, taxes and loan policy, and 60 customer scenarios with personas and openings, all generated and then reviewed by the authors (Appendix B, Table 16). Real estate accounts for under 0.1% of inner-loop training dialogues. Evaluation is interactive: the inner model, behind the real-estate harness with native tool calls executed against the

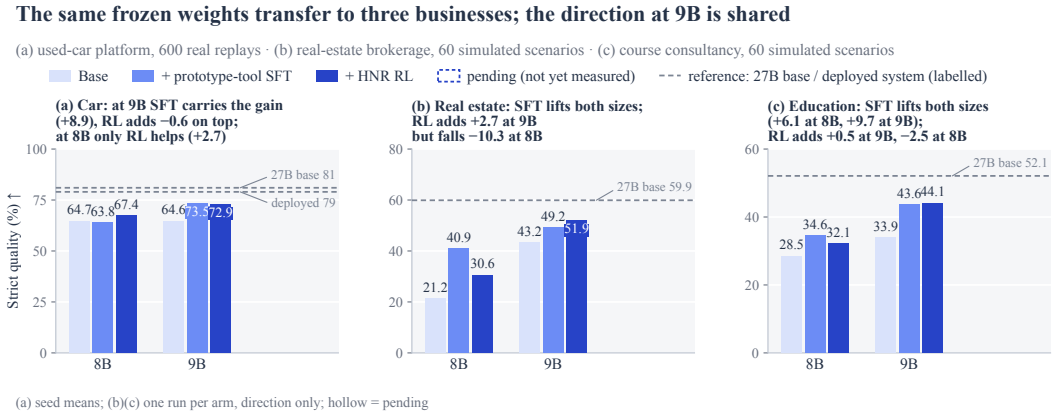


Figure 10: **One set of frozen weights, three businesses.** Strict quality of Base, prototype-tool SFT and HNR RL at 8B and 9B on (a) the used-car platform (600 real replays, seed means), (b) the synthetic real-estate brokerage and (c) the synthetic course consultancy (60 simulated scenarios each, one run per arm, direction only). Dashed lines: 27B base and, in (a), the deployed system.

Table 11: Second held-out business (synthetic real-estate brokerage, 60 interactive scenarios): loose / strict turn quality (%), ungrounded turns (%), tool-call rate (%) and purchase-outcome rate (%). Only the direction of trained-over-base differences is read.

Arm	Loose	Strict	Ungrounded	Tool-call	Purchase
8B base	51.9	21.2	38.6	68	17
8B + prototype-tool SFT	71.6	40.9	31.5	33	48
8B + HNR RL	66.8	30.6	36.5	32	53
9B base	79.1	43.2	20.1	24	20
9B + prototype-tool SFT	75.6	49.2	32.9	57	68
9B + HNR RL	80.3	51.9	28.5	53	69
27B base	87.4	59.9	8.8	51	52

knowledge base, converses for up to six turns with a simulated customer of the same family used in training; a judge grounded in the knowledge base scores each agent turn loose and strict and flags unsupported claims, and the episode outcome is recorded. We report loose and strict turn quality, the share of ungrounded turns, tool-call rate and outcome for 8B base / prototype SFT / HNR RL, 9B base / SFT / RL, and the 27B base, and read only the direction of the differences: if the trained models gain over their bases here as they do on the used-car business, the craft is not tied to a business. Table 11 reports the seven arms.

Reading. The direction matches the used-car business at 9B: prototype-tool SFT carries most of the strict gain over the base (43.2 → 49.2), HNR RL adds a little on top (51.9) while lifting loose quality and the purchase rate (20% → 69%), and tool use rises with training (24% → 53% of turns) as it did on the car platform. At 8B the SFT gain is large but RL falls back below it and tool use drops, so the 8B ordering is not the one seen on the car platform. What transfers is craft rather than grounding: the judge's deflection flags fall from 51 to 24 and no-progress flags from 66 to 51 across the 8B arms, while unsupported-claim flags per turn stay at 0.33–0.37 for all three 8B arms and improve only for 9B RL (0.34 → 0.31), whose over-promise flags also drop (27 → 8); the 27B base is the best grounded (0.08 per turn) and advances the sale least. With 60 scenarios and one run per arm, we read only these directions: the gains that transfer are those of craft, fewer deflections and more advancement of the sale, and grounding on unfamiliar tools improves only at 9B; tool discipline learned on eight prototypes carries over incompletely at the smaller size, which we list as a limitation.

A third business. To check that the direction is not specific to real estate, we built a third synthetic business in the same way, an adult vocational-training course consultancy (six business tools, a 26-entry knowledge base, 60 personas; education accounts for 0.1% of inner-loop dialogues), and

Table 12: Third held-out business (synthetic course consultancy, 60 interactive scenarios): loose / strict turn quality (%), ungrounded turns (%), tool-call rate (%) and enrolment-outcome rate (%). Same protocol as Table 11.

Arm	Loose	Strict	Ungrounded	Tool-call	Enrol
8B base	52.4	28.5	39.1	57	32
8B + prototype-tool SFT	54.3	34.6	47.3	49	43
8B + HNR RL	46.1	32.1	49.8	40	47
9B base	56.0	33.9	35.1	20	29
9B + prototype-tool SFT	62.9	43.6	39.7	56	53
9B + HNR RL	65.4	44.1	36.8	52	58
27B base	77.3	52.1	16.5	50	45

ran the same seven arms (Table 12). Car, home and course are three high-decision-cost purchases with different rules, catalogs and tools; the weights are identical across all three. At 8B the pattern repeats the real-estate one: prototype-tool SFT lifts strict quality (28.5 $\rightarrow$ 34.6) and HNR RL sits between base and SFT (32.1), with tool use falling from the base's 57%; at 9B the direction is again the car platform's: prototype-tool SFT carries most of the strict gain over the base (33.9 $\rightarrow$ 43.6), HNR RL adds a little on top (44.1) while raising loose quality (62.9 $\rightarrow$ 65.4) and the enrolment rate (29% $\rightarrow$ 58%), and tool use rises from the base's 20% to 56%–52% of turns. Three businesses, one set of weights, the same 9B ordering. As with real estate, we read directions only. Figure 10 places the three businesses side by side.

S Ablations: Extended Reading

Extended reading. Table 2 and Figure 5 summarize the inner-loop ablations along six axes (every arm, including seeds and 9B, is in Appendix E). **Cold start**: RL from an SFT start without tool trajectories does not move the replay score; the prototype-tool start is what RL improves. **Reward**: from the no-tool start, outcome-only rewards collapsed in all 4 runs into closing-at-any-cost; from the prototype-tool start a fifth outcome-only run did not collapse within 400 steps (purchase rate 0.21 $\rightarrow$ 0.23) and replays at par with the gated reward (68.6 vs. 67.8 strict, paired +0.5 [$-$3.8, +5.0]), even though 70.1% of its groups carried no gradient: at this scale and budget the gate's measured contribution is the share of informative groups (70.1% $\rightarrow$ 4.6% degenerate), not the replay score; a Boolean gate that zeroes the process score unless a sale occurs is as degenerate as outcome-only (64% of groups carry no gradient) and was rated best by the simulator and worst by humans; additive and gated rewards keep nearly every group informative, and HNR adds the credibility cap, so that a sale cannot buy back a fabricated claim. Between additive and gated composition the replay difference is within seed noise (68.6 vs. 67.8 strict from the same start, paired +0.7 [$-$3.8, +5.2]); our claim is that process must remain separable under every outcome and that credibility must be a veto, not that multiplication beats addition. **Offline vs. online**: DPO on judged preference pairs recovers a fraction of what online multi-turn RL achieves. **Loss**: GRPO variants are indistinguishable except a small consistent strict gain for Dr.GRPO. **Steps**: 800 steps do not improve on 400 at 8B. **HNR-GPO knobs**: the KL sweep separates stability from quality, as §4.3 reports: degenerate groups stay at 3–5% for every β. In matched mode, removing the credibility cap scores 68.4 strict against 67.8 for seed 1 (paired +0.7 [$-$3.7, +4.8]), inside the band of the five main-recipe runs: the cap costs nothing on replay, and we keep it for the property of Lemma 1(i), which this replay does not exercise (the 9B counterpart, 75.7 vs. 74.6, agrees); $\alpha = 1$ is likewise within noise of uniform credit (67.0), and raising the KL coefficient to $\beta = 0.01$ or 0.1 gives 66.3 and 65.7 (paired -1.5 [$-$5.7, +2.8] and -2.0 [$-$6.9, +2.7]), within noise and never above the main configuration. Intent-weighted credit with $\alpha = 2$ is the highest knob arm (70.1 strict, 88% tool calls): +2.3 [$-$2.2, +6.7] over seed 1, which includes zero, and +5.5 [+0.8, +10.2] over the base, which does not; it is one run, and at 9B the same knob gives 71.7 vs. 74.6 ($-$3.0 [$-$7.0, +1.0]), so we report it and do not promote it. Annealing the gate from soft to hard over the first 3,200 reward calls gives 66.6 ($-$1.3 [$-$5.8, +3.0] over seed 1): letting the process score dominate early, so that the early groups rank fluent but unsold conversations first, changes strict replay within noise. Clip-higher alone gives 67.6 ($-$0.3 [$-$4.5, +3.8]), also within noise. Taken together, HNR-GPO is robust to every knob we varied: none is required for the result

and none moves strict replay beyond seed noise at either size, so the recipe reduces to the gated reward, group-relative advantages and a small KL term. A general-instruction mixing recipe that preserves instruction following while keeping the business gains is reported in Appendix E.

T PROMPTS

All prompts below were written and run in Chinese; they are translated here for presentation, and lightly abridged where marked [...]. Placeholders of the form {{name}} are filled by the harness at run time.

Table 13: Simulated-customer prompt: persona block (per scenario) followed by the fixed rules (ENV_RULES), with thresholds as run (buy ≥ 0.8 after ≥ 2 turns, leave ≤ 0.1).

Persona block (one per scenario; fields are fixed, values are generated)
– Difficulty: easy / medium / hard
– Purchase-propensity score: 0.2–0.8 (parsed as the initial `intent`; default 0.4 if absent)
– Persona: e.g. "Balanced buyer. Has some concerns (price / fit) and needs reasonable evidence. May buy if the key objections are resolved satisfactorily."
– Customer information (visible only to you): basic info {age band, gender, location, occupation}; motivation; pain points; decision factors; communication preference; language. [...]
Rules (appended to the persona as the system message)
You are playing the customer described above, talking with a salesperson. You have an internal "purchase intent" score `intent` (0–1); its current value is given below. Each turn, first update `intent` according to what the salesperson just said, then speak. Update rules:
– The salesperson accurately picked up what you said, resolved a concern you actually care about, gave concrete verifiable information, or the quote or benefits match your expectation → `intent` $+0.1$ to $+0.2$.
– The salesperson answered off-topic, used boilerplate or repeated, pressured you, disparaged competitors, fabricated or contradicted themselves, or admitted being an AI → -0.1 to -0.3.
– The salesperson only gave a generic introduction without responding to your question → unchanged or -0.05.
– The salesperson did not answer your question but counter-asked or evaded → -0.1; two consecutive turns → -0.2.
– The salesperson promised something that sounds good but that you doubt they can deliver ("right away", "I'll hold the car for you", "skip the process and do it directly"): if you are a cautious or comparing type → -0.1 to -0.2 and ask "are you sure?"; if you are impulsive → $+0.05$ now, but next turn demand that they deliver.
– The salesperson said too much at once (clearly long, three or more points) → respond to only one point, -0.05.
Speaking rules: say only the customer's next utterance, colloquial and short ($\leq$60 characters); act according to the persona's personality; do not explain; do not play the salesperson; do not be easily persuaded and do not be unreasonable.
Termination rules: when `intent` ≥ 0.8 and the dialogue has run at least 2 turns, state clearly that you will buy and append `[END]`; when `intent` ≤ 0.1, say you will not buy or end the conversation and append `[END]`; otherwise do not append `[END]`.
Output JSON only: {`"intent"`: updated number, `"reply"`: "the customer's next utterance"}
State line (appended to the system message every turn): "Current `intent`=x; this is the salesperson's k-th utterance." Sampled at temperature 0.7, JSON mode. If `intent` reaches the buy threshold the episode outcome is 2, if it reaches the leave threshold 0; at the turn cap the outcome is 1 if `intent` ≥ 0.4, else 0.

Replay judge. The replay judge is a production evaluation prompt of about 41k characters and, like the production harness, is not reproduced. Structurally it (i) receives the customer turn, the candidate reply, the dialogue history, the session summary, the customer and listing profiles, the retrieved knowledge and every tool return of the trace; (ii) scores **0** for a factual, logical or compliance error (an unsupported claim, a contradiction with a tool return or the knowledge base, an unauthorised promise, a missed or redundant hand-over, self-disclosure as a bot, an off-topic reply, re-asking information the customer already gave), **1** for a usable but flawed reply (sales-technique or wording issues, excessive length or repetition, a wrong hand-over reason, a weak recommendation match) and **2** for a reply that is grounded in the knowledge and tool data, concise, coherent across turns and guides the next step; (iii) applies a list of business fact checkpoints, each of the form "if the reply violates ⟨fact or rule⟩, score 0 (or 1)", with the knowledge priority reference knowledge base > background knowledge >

Table 14: Inner reward-agent (judge) prompt (`JUDGE_SYSTEM`); the user message carries the persona and the dialogue.

System. You are a sales-dialogue judge. You are given a complete dialogue between a salesperson (assistant) and a customer (user); the customer persona is in ⟨Customer persona⟩. Output strict JSON: {"turn_scores": [one score per salesperson reply, from 5/3/0], "turn_flags": [one list of tags per salesperson reply, chosen from: unsupported claim / over-promise / red line / could answer but did not / missed intent / information dump / no progress / customer-service tone / several questions at once; [] if no problem], "outcome": 0|1|2, "violation": true|false, "reason": "≤30 characters, one line"}
For each salesperson reply, first check the veto items; hitting any one gives 0 and the tag:
① Unsupported claim: stating a product fact, stock, price, lead time or policy that has no source in the dialogue, or that contradicts given information;
② Over-promise: promising something that cannot be done or is not assured (e.g. "I'll get it to you right away", "I'll hold the car for you", "let's go straight to the next step"), or jumping past the current stage to a next-stage action;
③ Red line: disparaging competitors, pressuring, abuse, revealing AI identity, leaking internal information.
If no veto item is hit:
5 requires all of: when the customer asked a concrete question, the first sentence is the answer and any question comes after the answer; the reply addresses what the customer actually cares about behind this utterance rather than the literal words; if it asks, it asks only what the customer alone knows and is willing to tell, never what could be looked up; this turn brings the dialogue one step closer to a deal (a clear next step, an appointment, or one concrete objection resolved); calm and confident tone, mainly declarative sentences.
3: basically relevant but shows any of: could answer but did not (counter-asking or evading first), missed intent, information dump (three or more points or clearly too long), no progress, customer-service boilerplate or repetition, asking more than two questions at once.
outcome: 2 = the customer clearly states a purchase or asks for the purchase channel; 1 = shows interest and leaves a next step; 0 = declines or churns.
violation: true if any turn hits veto item ①②③.
Judge only from the dialogue text; do not speculate. The customer persona is used only to decide whether the "need behind the need" was addressed.
User. ⟨Customer persona⟩ {persona, first 1200 characters}
⟨Dialogue⟩ Salesperson: … / Customer: … (one line per turn)
Please output JSON. (temperature 0, JSON mode)

tool results > dialogue history and the lowest applicable tag deciding the score; (iv) protects replies that are incomplete or poorly worded but factually correct from a 0; and (v) returns JSON with a reason, one problem tag, the checkpoint ids used and the score, at temperature 0. Loose = score 2; strict additionally requires no unsupported-claim or rule tag and a tool call wherever the production trace made one. The reasoning slot is empty in every reported run (reasoning is stripped before judging).

Table 15: Closed-book replay prompt: role and case context only; no rules, retrieved knowledge or tools.

Role
You are a used-car sales consultant at an online car-buying platform, answering customers' questions in the enterprise messenger and moving the purchase forward. Handle the customer with your own sales ability and common sense; do not fabricate facts you do not have.
Current time
{{datetime}}
Session summary
{{session_summary}}
Customer profile
{{user_profiles}}
Current vehicle
{{vehicle_info}}
Output rules
Reply to the customer's current message in Chinese, colloquially and concisely; do not restate the information above; do not output extra formatting.

Table 16: Synthetic real-estate harness prompt (second held-out business). All assets are synthetic.

Role
You are an online property consultant at a second-hand home brokerage platform in Linjiang City, answering home buyers in the enterprise messenger and moving viewings and transactions forward. Every number you state (price, tax, down payment, interest rate, timeline, floor area, floor) and every policy or procedure must come from the [Background knowledge] below or from a tool return; say nothing without a basis – look it up first.
Unified decision framework
On receiving a customer message, decide in order:
1. The customer asks about a specific listing, community, price, tax, loan or eligibility → call the corresponding tool to verify first, then reply; do not quote numbers from memory.
2. The customer voices a concern or objection → first respond to what they actually care about, then give the next step; do not evade with a counter-question.
3. The customer clearly wants a viewing → book it with the tool and confirm time and contact details.
4. The customer asks to bypass the platform for a private deal, or asks for a promise that cannot be guaranteed (guaranteed transfer, guaranteed loan, locked price) → do not promise; explain the platform rules; hand over if necessary.
5. The customer explicitly asks for a human, the matter involves signing or funds, or it is beyond what you can verify → hand over.
Tool use
– At most 3 tool calls per turn; do not "announce" to the customer that you will look something up – call directly, and reply once the return is in.
– If a tool returns nothing or errors, say honestly "not found for now"; do not fabricate.
– Do not expose tool names or internal fields to the customer.
Hard compliance constraints
– Do not fabricate listings, prices, policies or timelines; do not disparage other platforms or agents; do not pressure; do not reveal that you are an AI; do not disclose owner privacy or internal information.
– Do not promise a private deal, skipping escrow, or that a loan will be approved.
Output rules
– Chinese, colloquial, concise: answer the customer's question first, then at most one suggestion that moves to the next step; no more than 120 characters; no tildes or doubled particles.
– Do not dump information the customer did not ask for.
Background knowledge (for your use only; do not recite it)
{{background_knowledge}}

Table 17: Tools of the synthetic real-estate harness and the prototype each maps to.

Tool	Purpose	Maps to prototype
search_listings	search listings on sale by community, budget and rooms	search_products
get_listing_details	floor, building age, decoration, tax status, viewing mode, owner motivation	get_product_details
estimate_costs	down payment, monthly payment, deed / VAT / income tax and agency fee	calculate_price
check_purchase_policy	purchase eligibility and loan policy (quota, residency, down-payment ratio, rates)	get_inspection_report
schedule_viewing	book an on-site or video viewing	schedule_followup
transfer_to_agent	hand over to a human consultant or transaction manager	escalate_to_human

Production business harness. The harness used for the main held-out business is a production asset and is not reproduced. Structurally it consists of: a role definition for the sales consultant; a rules block that fixes the dialogue policy (when to answer, when to probe, when to hand over) and the output style; a retrieved-knowledge block rendered per turn from the business knowledge base, together with the case context (time, session summary, customer profile, listing context); a catalog of the business's tools with descriptions and the prototype-to-tool mapping of Appendix B; and a set of compliance constraints (no fabricated facts, prices or timelines; no unauthorised promises; no disparagement or pressure; no disclosure of AI identity or internal information). The closed-book prompt of Table 15 is this harness with everything except the role and the case context removed.