NOT EVERY QUESTION IS YOURS: LEARNING WHOM TO ANSWER IN MULTI-PARTY DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

A dialogue agent can give a relevant answer and still speak out of turn. We study *selective participation*: answering when addressed while leaving questions for other participants to answer. In a controlled multi-party text simulator, episode scores can improve while an agent still answers questions meant for others, a pattern we call *scoreboard masking*. We make participation an explicit speak-or-wait decision and evaluate it on shared contexts at matched answering rates. On identical training rows, four supervision schemes all cut other-addressed responses from 0.548 to 0.205–0.249 over five seeds at a fixed own-address answering rate, whereas appending the norm to the base model's instructions does not. The aggregate hides what was learned: on addressee-swap minimal pairs, decision-level credit widens the own-versus-other gap from 0.31 to about 0.70, whereas episode return leaves it at 0.22. A late decision state carries this selectivity: in Qwen3-VL-8B, inserting a trained adapter's state at one decision position raises the gap by up to 0.21, matching the adapter; the reverse removes 0.22, and the transferred gap follows the named addressee under swaps and renaming. Five further trainees, a full-duplex speech model among them, lower other-addressed speaking, and closed-set capability stays within 2 points. Dialogue agents should be evaluated on whom they answer, not only on how often they speak.

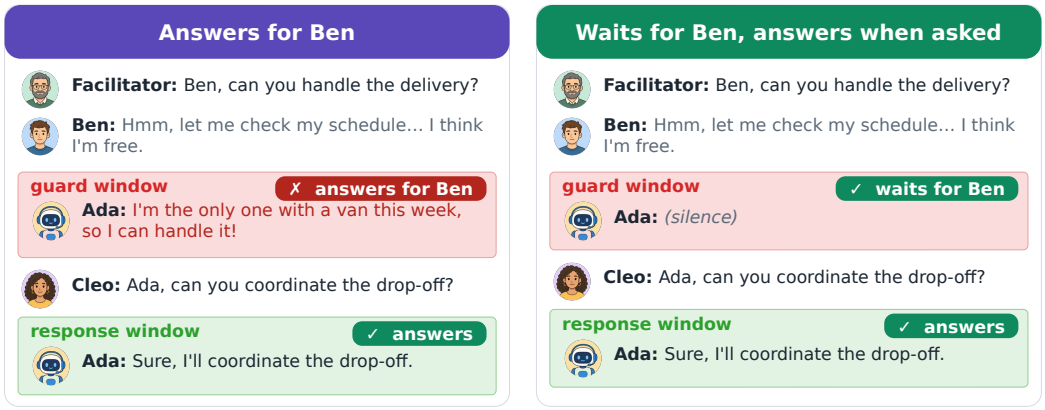

Figure 1: A relevant answer can still come from the wrong participant. One conversation, two participation policies: answering a question meant for Ben, or waiting for Ben and answering when Ada is addressed. Both can raise an episode scoreboard; the decision to answer for someone else separates them.

1 INTRODUCTION

Assistants now join meetings, group channels, and shared voice rooms where people talk to one another. Consider an assistant, Ada, in a conversation with two people, Ben and Cleo. A facilitator asks, "Ben, can you handle the delivery?" Ada knows the answer. Should Ada speak (Fig. 1)? An

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.

assistant that answers here is correct, yet unwelcome: in group conversations, knowing what to say does not mean it is your turn to speak.

Turn-taking research describes how speakers claim and yield the floor (Ekstedt & Skantze, 2020; 2022; Elmers et al., 2025; Sacks et al., 1974; Schegloff, 2000; Skantze, 2021; Umair et al., 2024). Multi-party systems identify addressees (Bohus & Horvitz, 2011; Gu et al., 2021; Johansson & Skantze, 2015; Ouchi & Tsuboi, 2016), while streaming models treat speaking as an action the model chooses (Défossez et al., 2024; Roy et al., 2026; Yao et al., 2026). However, these agents are trained on episode outcomes, per-decision labels, or fine-grained rewards (Li et al., 2026; Nama et al., 2026; Yao et al., 2026), and none of these signals asks whether the learned cue is the addressee: an episode can end well even if the agent repeatedly answers questions meant for others. We call this blind spot *scoreboard masking* (Section 3.1), the participation analogue of the outcome–process gap (Lightman et al., 2024; Uesato et al., 2022). When2Speak and Speak or Stay Silent label the speak-or-stay-silent decision and fine-tune on it, and ModeratorLM conditions it on an assigned role (Bhagtani et al., 2026; Mitra et al., 2026; Nama et al., 2026); all three score one operating point, whereas we hold the answering rate fixed and ask whom the agent answers.

We address this gap by treating participation as an independent decision. In a controlled three-party simulator, polled every two seconds, Ada chooses to speak or wait at annotated moments: questions for Ada, questions for others, requests for thinking time, and bad news that calls for a response. We train this choice and compare four ways to assign credit, in the spirit of process supervision that credits individual decisions (Arjona-Medina et al., 2019; Chan et al., 2024; Wu et al., 2023). We evaluate every policy on the same logged contexts at the same willingness to answer, which prevents models from appearing selective simply by staying quiet. We then move a single internal state between the trained network and its base to test whether that state carries the learned distinction, following activation patching and representation interventions (Chang et al., 2026; Wu et al., 2024; Zhang & Nanda, 2024) (survey: Appendix AK).

When trained on the participation decision, the model answers questions meant for others less than half as often while continuing to answer its own, under every credit construction we tried, and the decision-level constructions let the name decide the action when the addressee is swapped, which episode return does not. Four further text-reading trainees lower their other-address elevation under the same recipe, and a full-duplex speech model learns to answer by name. A single late decision state transfers the learned pattern in both directions, and the pattern tracks the addressee under swaps and renaming. On a public benchmark, retraining on 3,000 of its conversations lifts macro F1 from 0.299 to 0.680, close behind systems fine-tuned on the full split, while closed-set capability stays within 2 points (§§3.3–5).

We make three contributions. A protocol that scores whom an agent answers: shared contexts, a matched answering rate, and addressee-swap minimal pairs that change only who is asked. A controlled study on a fixed 8B model showing that every credit construction cuts the matched-rate violation where an appended norm does not, and that decision-level credit makes the name decisive for the action, which episode return does not. A causal diagnostic that locates the trained readout in a late decision-position state.

2 Participation as an explicit decision

2.1 A simulator with annotated moments

The environment contains the trainee, two LLM-driven peers, and a scripted facilitator; an event stream advances at 1 Hz, and we query the trainee every 2 s, with a 4 s refractory period after speaking. Each query shows the conversation so far, with explicit silence, and requests `speak` or `wait`; if the trainee speaks, a second call generates the utterance, separating the participation decision from answer quality. The transcript isolates that decision from perception: every text-model trainee reads it alone (vision encoders unused), a text-only LLM of the same family reproduces the result, and the full-duplex PersonaPlex hears TTS renderings of such scenarios; Figure 2 traces one decision from transcript to update.

Four annotated windows define the training signal and the evaluation: *own address* (a response is wanted within 16 s), *other address* (a question to another participant; defer for 6 s), *pause* (a

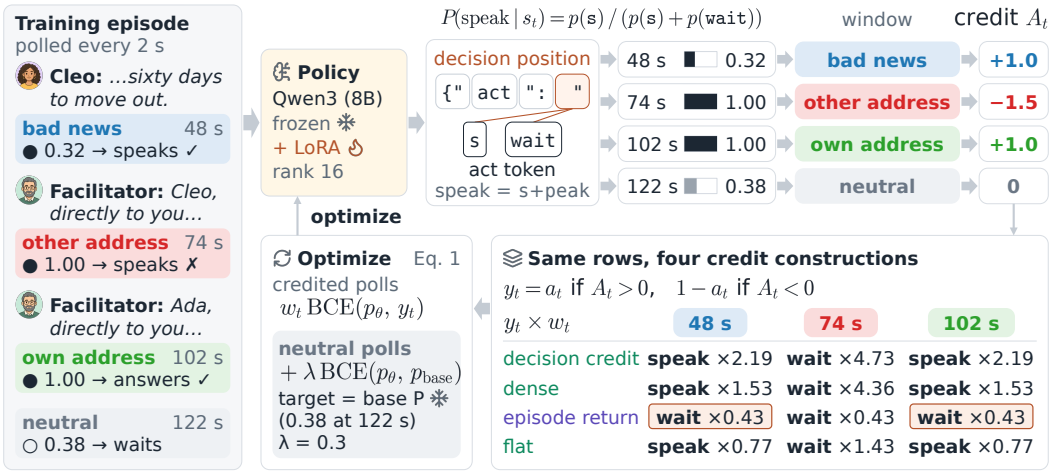

Figure 2: How one decision becomes an update. One training episode, polled every 2 s (filled marks: the base spoke); readouts from Qwen3-VL-8B. The LoRA reads $P(\text{speak} \mid s_t)$ at the act token; each poll's window sets the credit A_t, hence y_t and w_t, under four constructions. Outlined cells differ from decision credit's target: under episode return, whose return falls below the scenario mean, both rewarded responses become *wait*. Neutral polls anchor to the frozen base (Eq. 1).

request for thinking time; wait for 8 s), and *bad news* (a contribution is wanted within 20 s); all other opportunities are *neutral*, with no prescribed speaking rate. Windows are timed from the annotated utterance's end and blind to who holds the floor; speaking inside the two waiting windows, the *guards*, is the violation. Other address is a deliberate conflict: the trainee holds a private fact while a question to someone else goes unanswered for a beat, and the transcript, not the fact, drives the pull to speak (Appendix Q). Deferring there is a specified policy, not a universal rule, since a hesitating addressee can license another speaker (Lerner, 2003; Sacks et al., 1974; Stivers & Robinson, 2006); Appendix AA trains the complementary norm of speaking up, and Appendix C gives the simulator, prompts, and scenarios.

2.2 Training the speak-or-wait decision

We give waiting an explicit training target, just as speaking has one: let s_t be the context at opportunity t, $a_t \in \{0, 1\}$ its logged action, and $p_\theta(s_t)$ the speaking probability read at the decision position (token handling: Appendix B). A signed credit A_t selects the training target, the logged action or its opposite: $y_t = a_t$ if $A_t > 0$ and $y_t = 1 - a_t$ if $A_t < 0$. In the controlled comparison we optimize

$$\mathcal{L}(\theta) = \frac{\sum_{t \in \mathcal{C}} w_t \, \mathrm{BCE}(p_\theta(s_t), y_t)}{\sum_{t \in \mathcal{C}} w_t} + \frac{\lambda}{|\mathcal{N}|} \sum_{t \in \mathcal{N}} \mathrm{BCE}(p_\theta(s_t), p_{\text{base}}(s_t)), \tag{1}$$

where $\mathcal{C}$ holds credited decisions, w_t is the balanced absolute credit, and $\mathcal{N}$ holds neutral contexts anchored to the base model ($\lambda = 0.3$) to discourage indiscriminate shifts in speaking probability.

Four credit constructions supply A_t: *decision credit* uses the signed reward at that decision; *dense* centers and scales it per scenario; *episode return* assigns the scenario-normalized episode return to each selected decision; and *flat* keeps only its sign; the reward favors the first response in a response window and waiting in a guard window. Under decision credit and flat, y_t is the window's prescribed action, so the two differ only in weight; dense and episode return flip 77 and 607 targets. All four are offline, single-round label constructions on base-policy states, not on-policy policy gradient: a baseline here moves targets.

Two protocols. In the *controlled study* (Table 2, Fig. 4a) we train this action-only objective on identical rows over five seeds. The *pipeline study*, the historical sequence-level protocol (Table 1; Figs. 3, 4b–c, 8) also imitates the utterance of rewarded speaking, lets each construction set its own rows and anchors, reports one trained checkpoint per arm, adds a *critic* (return-to-go minus a cross-fitted value) and a GAE(0.5) arm (Schulman et al., 2016) (Table 4), and produced the seed-0

adapter of Section 4. Bracketed ranges are 95% cluster-bootstrap intervals, by scenario unless a caption says otherwise.

2.3 Measuring selectivity without rewarding silence

Episode outcomes record whether an agent answered its own question and whether it answered for someone else; their difference is episode selectivity. **Shared-context scoring**: we present exactly the same logged contexts to every model, so differences in conversational history drop out of the comparison. **Matched answering rates**: we sweep one speaking threshold per model and compare other-addressed response rates at the same own-address answering rate; because test-set matching traces a trade-off curve, we also apply development-selected thresholds unchanged to test, where the violation rate still falls to 0.144–0.165 (Appendix S). Three statistics recur: *other-address elevation*, a probability (other-addressed speaking probability minus its neutral-context value) that diagnoses one arm; the *violation rate*, the thresholded response rate on other-addressed contexts, compared at matched answering rates; and the *own − other gap* $S = \mathbb{E}[p_\theta \mid \text{own}] - \mathbb{E}[p_\theta \mid \text{other}]$, the probability difference used for the state interventions, where addressee swaps hold the rest of the input fixed. Arms are ranked by the violation rate at a matched rate; the swap pairs ask whether the cue is the name; elevation diagnoses one arm, since a rising neutral rate also lowers it. Section 3 first asks what current models do at these moments.

3 Diagnosing and training whom to answer

3.1 Scoreboard masking

We first probed eleven instruction-tuned checkpoints, the trainee base among them, on a fixed set of simulator contexts (Appendix J). All eleven are more willing to answer a question meant for someone else than to speak after a pause request (nine resolvably), and nine of eleven speak more there than in neutral contexts, four resolvably. The pull is one we built in: a peer-addressed question, the addressee's hesitation, and a prompt that asks Ada to contribute.

An episode-level improvement need not remove this error: on Qwen3-VL-8B, episode-return training removes most but not all of the base's other-address elevation on every sealed set (set 9, re-run: $+0.288$ to $+0.075$; the others in Table 1). On JoyAI-VL-Interaction, whose base sits near neutral ($+0.016$), episode return lifts the aggregate through own-address answering alone (0.40 to 0.65) and leaves the guard where it was (scenario-adjusted change $+0.004$; Appendix V). *Scoreboard masking* is an episode aggregate that improves while per-poll other-address elevation stays above neutral, as on every Qwen3-VL-8B set (Fig. 3). Because other-address elevation is taken against each arm's own neutral rate, it diagnoses an arm but does not rank arms; Section 3.4 ranks them at a matched answering rate. A general rating under-prices the error, by only 0.36 between-episode standard deviations for a 32B observer once scenario, source, and speaking rate are controlled, so we evaluate the action at the moment it is taken.

3.2 Six trainees, one direction

In the pipeline study we train Qwen3-VL-8B (Bai et al., 2025), JoyAI-VL (a proactive model already trained with per-second RL to take turns), InternVL3-8B, MiniCPM-o-2.6, and the full-duplex PersonaPlex-7B, all through LoRA adapters (Hu et al., 2022). Every text-model trainee shows lower other-address elevation under decision credit on every set (Table 1; JoyAI-VL builds on Qwen3-VL, InternVL3 and MiniCPM-o on Qwen2.5). We train PersonaPlex through the PAD-or-word choice of its text stream, and it learns to answer by name: on 24 minimal pairs own-name answering rises from 0.833 to 0.979 while other-addressed speaking falls from 0.771 to 0.542 (-0.229), widening the own − other gap from $+0.06$ to $+0.44$ (one training seed, evaluated open-loop; neutral, pause, and bad-news rates in Appendix Y).

Nor does the result need a vision-language trainee: on text-only Qwen3-8B the same rows and trainer cut the violation rate from 0.913 to 0.452 at its own answering rate of 0.968 (2 seeds), and the full sequence-level recipe lowers its other-address elevation against that base by 0.549 on the primary seed, a result four further seeds confirm (Table 1; Appendix X).

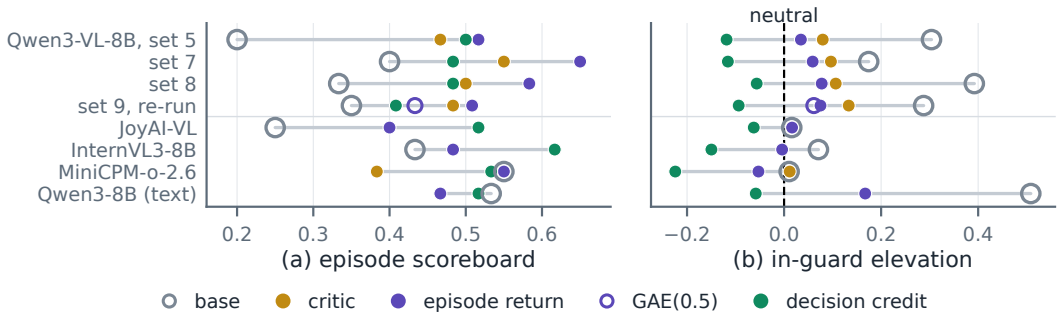

Figure 3: A better episode score need not fix the participation decision. One row per evaluation set; arms as in Section 2. (a) The episode scoreboard, own answered − other violation, end-anchored (App. AE): trained arms improve on their base except on MiniCPM-o, whose base (0.55) no arm beats, and on text-only Qwen3-8B, whose arms trail its base here but lead it under symmetric anchoring. (b) Other-address elevation per poll: decision credit ends below neutral on every set; episode return does so only on InternVL3 and MiniCPM-o (-0.004, -0.053; decision credit -0.150, -0.224).

Table 1: Pipeline study: decision credit lowers other-address elevation on five trainees and many sets. Elevation is other-address minus each arm's neutral speaking probability; contrasts are scenario-adjusted differences from each comparator (negative favors decision credit), scenario-clustered intervals. Symbols, dashes, shading, and the arms' configurations: App. AI. Contrasts diagnose each pipeline; arms are ranked at a matched rate (§3.4, set 9), where the vs. base and vs. critic columns hold and the vs. episode-return column does not.

	other-address elevation above neutral = answers for the addressee				**contrast: decision credit − comparator** more negative = lower elevation under decision credit		
trainee, sealed set	base	critic	episode return	**decision credit**	vs. base	vs. critic	vs. episode return
Qwen3-VL-8B							
set 5†	+0.304	+0.080	+0.035	−0.119	−0.418 [-0.55,-0.29]	−0.198 [-0.25,-0.14]	−0.159 [-0.22,-0.10]
set 7†	+0.174	+0.097	+0.059	−0.116	−0.285 [-0.45,-0.11]	−0.198 [-0.28,-0.11]	−0.157 [-0.24,-0.07]
set 8	+0.392	+0.106	+0.078	−0.056	−0.471 [-0.61,-0.33]	**−0.175*** [-0.25,-0.10]	**−0.138*** [-0.22,-0.05]
set 9	+0.251	+0.115	+0.050	−0.112	−0.367 [-0.51,-0.22]	**−0.228*** [-0.29,-0.17]	**−0.159*** [-0.22,-0.10]
set 9, tuned	+0.251	+0.061	+0.072	−0.112	= set 9	**−0.172*** [-0.24,-0.11]	**−0.179*** [-0.24,-0.12]
set 9, re-run	+0.288	+0.133	+0.075	−0.094	−0.371 [-0.50,-0.24]	**−0.226*** [-0.28,-0.17]	**−0.164*** [-0.22,-0.11]
Qwen3-8B (text)	+0.508	—	+0.167	−0.058	**−0.549*** [-0.68,-0.42]	—	**−0.222*** [-0.32,-0.13]
JoyAI-VL	+0.016	—	+0.016	−0.062	−0.076 [-0.13,-0.03]	—	**−0.080*** [-0.12,-0.04]
InternVL3-8B	+0.071	—	−0.004	−0.150	**−0.224*** [-0.28,-0.17]	—	**−0.135*** [-0.17,-0.10]
MiniCPM-o-2.6	+0.010	+0.012	−0.053	−0.224	**−0.237*** [-0.31,-0.17]	**−0.238*** [-0.29,-0.19]	**−0.173*** [-0.22,-0.13]

3.3 Identical rows: the same guard, different cues

In the controlled study the arms differ only in the credit construction (target and weight on the same annotated rows, so what varies is what a construction puts on a moment): we train each arm on 1,564 credited contexts and 782 neutral anchors, with the same initialization, row order, LoRA configuration, and 441 updates per seed, over five seeds. We evaluate on 2,884 shared contexts re-used from sealed set 9 (24 families; 115 other-addressed, 186 own-addressed); rows and recorded actions are identical

Table 2: Controlled study (five seeds). Speaking rates on shared set-9 test contexts (never trained on), every arm held at the base model's own-address answering rate (0.624); means over five seeds (seed SD in small type), base one frozen checkpoint. Last column: the planned contrast decision credit − arm with its scenario-and-seed bootstrap interval; it resolved for no arm. Shading: primary outcome; middle columns: secondary outcomes. The violation column measures the guard; the swap-gap column, the own − other gap on the 41 addressee-swap pairs of Section 4 (22 families; seed means; Appendix U), measures whether the name rather than the script position decides it. Last row: mixed credit, decision credit and episode return added at equal magnitude on the same rows (Appendix T).

construction	**other addressed** ↓	pause ↓	bad news ↑	neutral	swap gap ↑	contrast [CI]
base	0.548	0.600	0.531	0.287	+0.308	—
episode return	0.205 ±0.03	0.510 ±0.04	0.669 ±0.04	0.195 ±0.03	+0.216 ±0.02	+0.043 [-0.08, +0.16]
dense	0.249 ±0.04	0.633 ±0.09	0.868 ±0.05	0.522 ±0.08	+0.695 ±0.06	+0.000 [-0.05, +0.04]
flat	0.228 ±0.04	0.627 ±0.03	0.849 ±0.03	0.495 ±0.06	+0.733 ±0.04	+0.021 [-0.03, +0.07]
decision credit	0.249 ±0.04	0.627 ±0.05	0.850 ±0.04	0.511 ±0.06	+0.702 ±0.02	—
mixed credit	0.127 ±0.05	0.480 ±0.05	0.706 ±0.03	0.248 ±0.04	+0.737 ±0.09	5 seeds, no planned contrast

across arms. The violation rate falls from 0.548 to 0.205–0.249 with own-address answering held at the base rate, a gain of 0.30–0.34 over the base (Table 2, Fig. 4a); the pairwise differences among the four are unresolved (last column).

Every construction cuts the violation on these rows, but they do not learn the same cue. The 41 addressee-swap pairs of Section 4 (22 families) hold wording, position, and time fixed and change only who is asked. Decision credit, dense, and flat widen the base's own − other gap from 0.308 to 0.695–0.733 and rank the own-addressed member first in every pair and seed; episode return ranks it first in 197 of 205 but does not widen the gap (0.216; swap-gap column of Table 2). At each arm's matched threshold, the decision-level arms answer the own-addressed member and wait on the other in 0.610–0.693 of pairs, against the base's 0.366, while episode return stays at the base's level (0.351; decision credit − episode return +0.298 [+0.100, +0.484]). Its scenario-level return flips 82 of the 183 own-address targets away from the recorded action, so part of its set-9 edge comes from where a question falls in the generator's fixed moment order, not from the name.

Adding decision credit and episode return at equal magnitude (*mixed* credit, pre-registered after the controlled study) keeps the guard and removes decision credit's neutral cost, as pre-registered: violation rate 0.127 and neutral speaking 0.248 against 0.249 and 0.511 (Table 2, last row; 5 seeds). The cost falls on bad-news answering, 0.706 against 0.850. Two minimal untrained alternatives fall short (Appendix T): an appended norm quiets the base yet leaves its matched-rate violation rate at 0.557, and a last-utterance vocative rule that never violates the guard answers only 0.167 of own-addressed polls. Scale is a separate lever: the untrained Qwen3-VL-32B reads 0.139 at the same answering rate and a swap gap of 0.570, and the norm does nothing for it either; scale buys much of the name cue, and decision-level supervision gives it to a model a quarter the size.

In the other windows the policies diverge (Table 2, middle columns; Fig. 15): episode return speaks less in neutral contexts (0.195 against decision credit's 0.511). It also keeps the sharpest bad-news selectivity, while decision credit, dense, and flat answer bad news more and leave pause responses near the base's: what they learn is the addressing cue, not a general yielding of the floor (rates and ratios in Appendix S).

3.4 Selectivity at a matched answering rate

In Figure 4b we sweep one speaking threshold per sequence-trained arm of Table 1 over the same contexts: at the base model's own-address answering rate of 0.629 (bf16), decision credit cuts the violation rate by 0.252 against the base and by 0.165 against the critic. Against episode return the reading changes: Table 1's other-address elevation tells the arms apart and the scoreboard ranks episode return first, but at a matched rate their difference is +0.043 [−0.053, +0.111], as in the controlled study. Held at the base's neutral rate instead, decision credit reads 0.157 against episode return's 0.461 (Appendix AI): the leader depends on the rate held fixed, and every arm stays below the base.

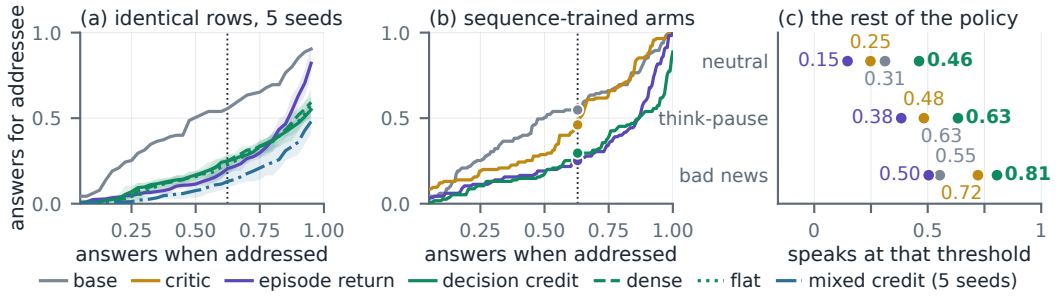

Figure 4: Selective at the same answering rate, different elsewhere. One speak threshold is swept per arm on shared held-out contexts; dotted: the base model's own-address answering rate. (a) Five-seed identical-support study (seed means, seed ranges; dash-dot: mixed credit, Table 2's last row). (b) Pipeline arms (sequence-trained Qwen3-VL-8B, Table 1; bf16, matched at 0.629, where Table 2 is fp32 at 0.624), dots at the matched rate. (c) At that threshold, how often each arm speaks in neutral, pause, and bad-news contexts: episode return withdraws; decision credit speaks more than the base in neutral and bad-news contexts and as often after a pause.

Figure 4c shows why a matched result is only part of the policy: among the pipeline arms, decision credit answers bad news at 0.805 against episode return's 0.504 and speaks more in neutral contexts (three contexts are read decision by decision in Fig. 8).

4 A LATE DECISION STATE CARRIES THE SELECTIVITY

Training changes many parameters; to ask whether one internal position carries the distinction, we follow cross-model patching (Jain et al., 2024; Prakash et al., 2024): the base Qwen3-VL-8B and the seed-0 decision-credit adapter of the pipeline study run on identical inputs, and we replace the residual state at the decision position, where the act token is predicted, leaving other positions and later layers to the recipient model.

A late decision state transfers the trained change. Among candidate decoder blocks $4, 8, \ldots, 32$ of 36, we select block 32 on a development split of 112 contexts by the gain in the gap, and we evaluate it on 227 re-used set-9 contexts (24 families) in fp32 (Appendix U).

Transferring the trained state into the base lowers mean speaking probability on other-addressed questions from 0.678 to 0.305, and on own-addressed questions by less than half as much (Table 18). The larger drop on other-addressed questions widens the gap by 0.210 ($[0.049, 0.366]$), and the patched base's gap closely matches the trained model's (Figure 5b). The reverse intervention, a base state inserted into the trained model, narrows the gap by 0.218 ($[-0.341, -0.096]$); the two directions show this state's causal contribution to the learned behavior. The five decision-credit seeds of the controlled study replicate the transfer ($+0.205$ at block 32), and an episode-return adapter's late state transfers a heterogeneous gap as well ($+0.278$): the transfer locates where training changes the decision, and the swap pairs test whether that change is the addressee.

In the layer sweep, patching at blocks 4–20 moves the gap by at most 0.007 in either direction, and transfer appears from block 24 on and grows to block 32 (Fig. 5a): the learned distinction reaches the decision position between blocks 20 and 24. Only that position is replaced; no neuron or head is identified.

The effect follows the addressee. To hold wording, topic, position, and time fixed, we build 41 addressee-swap pairs from 22 scenario families: a question to Ben or Cleo becomes one to Ada at the same point in the script. On these pairs the base already shows a gap (0.308); transferring the pipeline adapter's state raises it to 0.633, matching the adapter's own 0.571 (overlapping intervals) and well above the output bias's 0.334 (Figure 5c; Table 19). When we rename Ada/Ben/Cleo Mira/Noah/Iris throughout, the base and patched gaps are 0.309 and 0.574. The base ranks the own-addressed version first in 0.854 of the 41 pairs, the patched base in all of them: the transferred gap follows the named addressee, not the script position or the particular names.

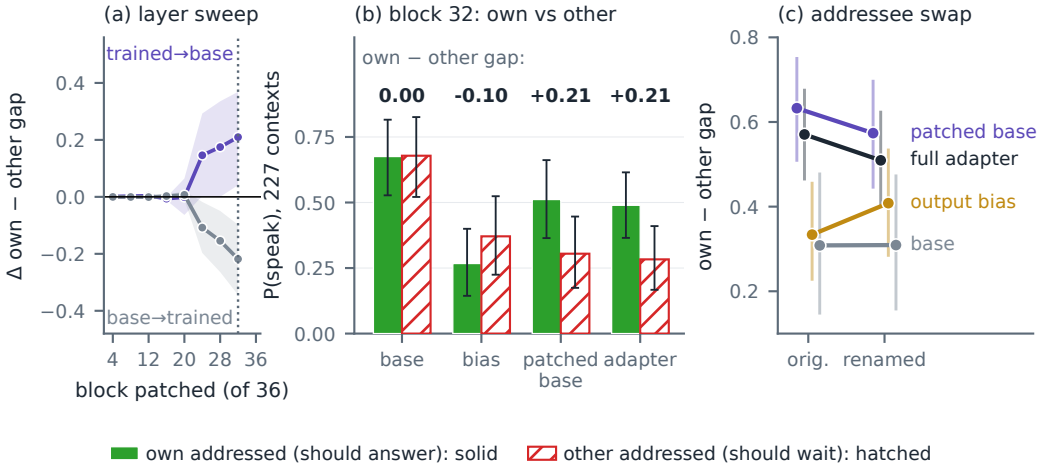

Figure 5: A late residual state carries the learned selectivity. The own − other gap is $P(\text{speak} \mid \text{own addressed}) - P(\text{speak} \mid \text{other addressed})$. (a) Trained state patched into the base at one decoder block ($+0.146$, $+0.175$, $+0.210$ at blocks 24, 28, 32), and the reverse; dotted: the development-chosen block. (b) At that block, own- and other-addressed speaking probability (solid, hatched) for the base, the base with a calibrated output bias or carrying the trained state, and the adapter, with the gap above each group; a mean shift and a random direction open none. (c) On addressee-swap pairs (Table 19), the same gap before and after renaming every participant. Bands and whiskers are scenario-bootstrap intervals.

A general reduction in speaking does not explain the transfer. Three simpler interventions fit on development data all fall short of the state transfer: a constant output-logit bias calibrated to the patched model's other-address speaking probability, the mean trained-minus-base state difference, and a random direction with the mean difference's norm. Because a constant logit shift preserves every context's ranking, the bias leaves the own-versus-other AUROC at the base value (0.472), whereas transfer raises it to 0.687 ($+0.214$); the mean difference, the random vector, and a self-patch move it by at most 0.003. In probability space the controls leave the gap near or below the base's, whereas transfer adds 0.210, and its advantage over each control excludes zero (Table 18). These controls rule out a general quieting, and the swap pairs show that the transferred state tracks each input's addressee.

The cue precedes training. On swap pairs the base already favors the own-addressed member (0.308) although its pooled own and other speaking probabilities nearly coincide (0.675 against 0.678), and linear probes on the final-layer decision state recover window type in three lineages (Qwen3-VL-8B other-addressed recall 0.967). Decision-level training enlarges this existing gap rather than creating it (Appendix R).

5 Public Benchmarks and the Scope of the Result

Two participation benchmarks built by others. When2Speak (Nama et al., 2026) supplies independent speak-or-silent labels for multi-party conversations. We retrain rather than transfer: on a 3,000-conversation subset of its training split we train decision credit as per-decision labels (plain label SFT, since correctness is binary) and episode return as one credit per conversation. On the full 21,687-context test set, macro F1 is 0.299 for the base and 0.680 and 0.584 for the two arms. Both arms beat an always-silent predictor (0.466) and every published zero-shot model, and, read at the default speak threshold, fall short of models fine-tuned on the full training split (Fig. 6a; Table 6). The readings we pre-registered on a frozen 3,810-context subsample (Appendix E) hold on the full test set, and they show per-decision supervision to be more selective rather than merely quieter: it cuts unsolicited-question speaking by 0.181 (0.068 net of statements) and separates speak from silent contexts with an AUROC of 0.879 against 0.688, 0.884 against 0.618 on unsolicited questions. Both arms turn conservative, trading recall for fewer false interruptions (Table 6).

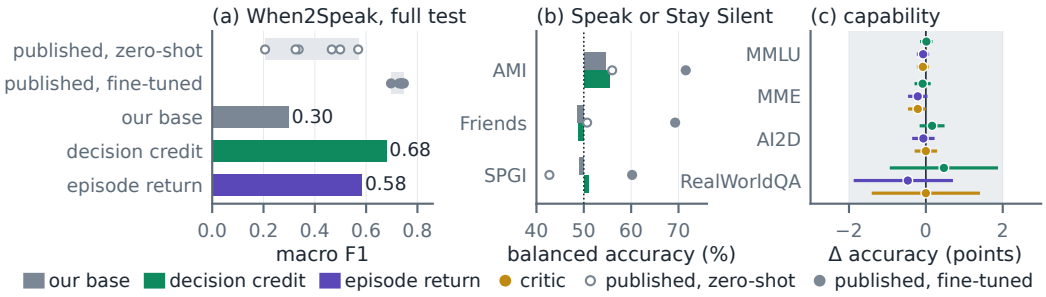

Figure 6: Public benchmarks built by others. (a) When2Speak's full test set: macro F1 of our base and its two arms trained on 3,000 conversations (green: decision credit; purple: episode return) against six published models, zero-shot (hollow) and fine-tuned on the full split (filled). (b) Speak or Stay Silent: balanced accuracy from chance (dotted) for our base and its simulator-trained decision-credit adapter, neither trained on these corpora, beside published Qwen3-8B results. (c) Accuracy change against the base for decision credit (green), episode return (purple) and the critic (gold), with intervals; shaded: the pre-registered 2-point margin.

Speak or Stay Silent (Bhagtani et al., 2026) labels whether a participant speaks next in AMI meetings, *Friends* episodes, and SPGI earnings calls. Zero-shot transfer is partial: the adapter's addressee gap (speaking when addressed minus when another is addressed) moves in the trained direction on *Friends*, where it closes an inverted gap by 0.103, and on SPGI, whose addressee labels are heuristic, but not on AMI. Its balanced accuracy changes by at most 1.9 points, well short of models fine-tuned on them (Fig. 6b).

On 1,909 contexts of 22 AMI meetings with manual addressee labels (Carletta, 2007), the adapter shows no zero-shot transfer (-0.011). The simulator's failure mode is absent from AMI, where the base's other-address elevation is already negative (-0.090; -0.069 on When2Speak), so there is nothing to correct (Appendix I).

Closed-set capability is retained. We score Qwen3-VL-8B's sequence-trained decision-credit, episode-return, and critic adapters on MMLU (Hendrycks et al., 2021), MME (Fu et al., 2023), AI2D (Kembhavi et al., 2016), and RealWorldQA's letter-option subset (xAI, 2024): all change closed-set accuracy against the base within the pre-registered 2-point margin, intervals included (Fig. 6c; Table 8).

Where the result applies. The simulator uses text transcripts with silence markers, 2 s polls, and leading vocative addresses, and held-out sets share the training generator's fixed moment order; the swap pairs break that order in one direction, by re-addressing a peer's question to Ada. The contrast keeps its sign at 0.25–1 Hz polling (about half as large at 1 Hz) and 2–6 s guards (Appendix AE); training needs annotated windows, although on our generator a keyword regex recovers 84% of the gain, and we leave group, implicit or non-verbal addressing, ASR noise, on-policy RL, and PersonaPlex's output audio to future work. Addressee selectivity is not general social competence: pause handling can worsen as addressed-question behavior improves, decision credit trails the critic on a norm needing remembered state, the LLM peers rate the trained Ada lower at similar speaking rates, and a blind judge rates her matched utterances 0.454 lower on a nine-point scale (Appendix H); the results hold in the controlled setting.

6 Conclusion

Knowing the answer does not make it an agent's turn to speak. Explicit speak-or-wait training improves this distinction, which episode scores can mask (scoreboard masking); at a matched answering rate every credit construction cuts the violation, but on addressee-swap pairs decision-level credit makes the name decide the action, which episode return does not, and in Qwen3-VL-8B a late decision-position state transfers the learned own − other gap in both directions. Whom an agent answers deserves the same scrutiny as how often it speaks.

REPRODUCIBILITY STATEMENT

The anonymized supplementary archive contains scenario generators, logged decision contexts, training configurations, analysis scripts, and per-experiment results. The controlled comparison includes five seeds per primary arm; intervention results include development selection, test estimates, and control conditions. Appendices distinguish earlier planned experiments, exploratory analyses, and follow-up experiments on existing evaluation sets. Quantitative tables and figures are rendered from stored artifacts; the document build checks numerical traceability, references, and rendered floats. The public-benchmark scorers, their frozen pre-registrations (Appendix F), and the rendered benchmark contexts are included. Experiments ran on shared nodes of eight NVIDIA A100-SXM4-80GB GPUs; one controlled-study training run takes about 50 minutes on an idle GPU, so the twenty runs behind Table 2's four primary arms cost about 17 GPU-hours of training before scoring; the sequence-level pipelines, live evaluations, and PersonaPlex runs are not accounted to the hour. The controlled study resamples scenario families and training seeds; state-transfer intervals resample scenario families. Development selection and exploratory test-layer analyses are recorded separately. An anonymized supplementary archive carries the pre-registration files with their commit times and SHA-256 digests.

ETHICS STATEMENT

Training conversations are synthetic, and every participant in them, every rater and every observer is a language model; no person took part in any experiment reported here. The AMI corpus (CC BY 4.0), When2Speak (CC BY 4.0), Speak or Stay Silent (Apache 2.0; its SPGI transcripts originate in SPGISpeech 2.0, which Kensho distributes under non-commercial research terms, and its *Friends* excerpts are copyrighted television scripts used here only as short evaluation text), and the four capability benchmarks are used under their stated licenses, as text transcripts without speaker identities beyond the corpora's own labels. The studied deference rule is an explicit task policy and may not match every user's preference or conversational culture; a policy tuned to defer can also suppress a needed contribution, which the bad-news window and the second norm (Appendix AA) exist to measure and deployments should monitor. The proposed human assessment has a prepared instrument and analysis plan, uses synthetic transcripts, and will proceed only after institutional review-board approval, with informed consent and with participants told that the transcripts are synthetic and model-generated. Applications should evaluate both unwanted interruptions and missed opportunities to respond. A deployed assistant that speaks unprompted should disclose that it is an AI to everyone on the call. The scripted facilitator allocates turns and the norm defers to that allocation; the pause guard gives a hesitating or non-native speaker room to finish, and both choices should stay visible in deployment. Every peer, observer and judge is a Qwen3-VL-family model, so the ratings carry one family's preferences.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with literature discovery and language refinement. In particular, these tools were used to help identify potentially relevant literature and to improve the clarity, readability, and presentation of the manuscript. The authors independently reviewed relevant sources and verified the accuracy and appropriateness of the AI-assisted revisions. All scientific content, interpretations, citations, and conclusions were reviewed and finalized by the authors, who take full responsibility for the final manuscript.

REFERENCES

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. RUDDER: Return decomposition for delayed rewards. In *NeurIPS*, 2019.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Robert M. Bell and Daniel F. McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–181, 2002.

Kratika Bhagtani, Mrinal Anand, Yu Chen Xu, and Amit Kumar Singh Yadav. Speak or Stay Silent: Context-aware turn-taking in multi-party dialogue. *arXiv preprint arXiv:2603.11409*, 2026.

Dan Bohus and Eric Horvitz. Decisions about turns in multiparty conversation: From perception to action. In *ICMI*, 2011.

Jean Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

Alex J. Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. Dense reward for free in reinforcement learning from human feedback. In *ICML*, 2024.

Cheng-Kuang Chang, Kai-Wei Chang, Alexander H. Liu, and James Glass. Overcoming state inertia in full-duplex spoken language models via activation steering. *arXiv preprint arXiv:2606.11386*, 2026.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

Erik Ekstedt and Gabriel Skantze. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of EMNLP*, 2020.

Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In *Interspeech*, pp. 5190–5194, 2022. doi: 10.21437/Interspeech.2022-10955.

Mikey Elmers, Koji Inoue, Divesh Lala, and Tatsuya Kawahara. Triadic multi-party voice activity projection for turn-taking in spoken dialogue systems. In *Proc. Interspeech*, pp. 3015–3019, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Agustín Gravano and Julia Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, 2011.

Raymond Grossman, Taejin Park, Kunal Dhawan, Andrew Titus, Sophia Zhi, Yulia Shchadilova, Weiqing Wang, Jagadeesh Balam, and Boris Ginsburg. SPGISpeech 2.0: Transcribed multi-speaker financial audio for speaker-tagged transcription. In *Proc. Interspeech*, pp. 4048–4052, 2025. Dataset released by Kensho Technologies: `https://huggingface.co/datasets/kensho/SPGISpeech2.0`.

Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *ACL*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *ICLR*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Martin Johansson and Gabriel Skantze. Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 305–314, 2015. doi: 10.18653/v1/W15-4642.

Natasa Jovanović, Rieks op den Akker, and Anton Nijholt. Addressee identification in face-to-face meetings. In *EACL*, 2006.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Gene H. Lerner. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society*, 32(2):177–201, 2003.

Yuxin Li, Donghang Wu, Guan-Ting Lin, Hung-yi Lee, Chengwei Qin, Zhehuai Chen, and Chen Chen. Decoupling conversational dynamics in full-duplex spoken models through reinforcement learning. *arXiv preprint arXiv:2607.07148*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2024.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, and Hongsheng Li. OmniInteract: Benchmarking real-world streaming interaction for real-time omnimodal assistants. *arXiv preprint arXiv:2605.26485*, 2026.

Soumyajit Mitra, Prabhat Pandey, Abhinav Jain, Shanmukha Sahith, and K V Vijay Girish. Adaptive turn-taking for real-time multi-party voice agents. *arXiv preprint arXiv:2606.13544*, 2026.

Vihaan Nama, Shreya Mendi, Zian Ye, and Brinnae Bent. When2Speak: A dataset for temporal participation and turn-taking in multi-party conversations for large language models. *arXiv preprint arXiv:2605.05626*, 2026.

Rieks op den Akker and David Traum. A comparison of addressee detection methods for multiparty conversations. In *Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, 2009.

Hiroki Ouchi and Yuta Tsuboi. Addressee and response selection for multi-party conversation. In *EMNLP*, 2016.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *ICLR*, 2024.

Antoine Raux and Maxine Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *NAACL-HLT*, 2009.

Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. PersonaPlex: Voice and role control for full duplex conversational speech models. *arXiv preprint arXiv:2602.06053*, 2026.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

Emanuel A. Schegloff. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63, 2000.

Donald J. Schuirmann. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *ICLR*, 2016.

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.

Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67, 2021.

Tanya Stivers and Jeffrey D. Robinson. A preference for progressivity in interaction. *Language in Society*, 35(3):367–392, 2006.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Muhammad Umair, Vasanth Sarathy, and J. P. de Ruiter. Large language models know what to say but not when to speak. In *Findings of EMNLP*, pp. 15503–15514, 2024. doi: 10.18653/v1/2024.findings-emnlp.909.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *ACL*, 2024.

Xintao Wang, Sirui Zheng, Hongqiu Wu, Weiyuan Li, Jen-tse Huang, Minghao Zhu, Can Zu, Qi Deng, Jiawei Wang, Qianyu He, Heng Wang, Xiaojian Wu, and Yunzhe Tao. Agentopia: Long-term life simulation and learning in agent societies. *arXiv preprint arXiv:2606.07513*, 2026.

Matthew D. Webb. Reworking wild bootstrap–based inference for clustered errors. *Canadian Journal of Economics*, 56(3):839–858, 2023. doi: 10.1111/caje.12661.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *NeurIPS*, 2023.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

xAI. RealWorldQA benchmark (Grok-1.5 Vision Preview). `https://x.ai/news/grok-1.5v`, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui

Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Xiangyu Zeng, Yifei Li, Haowen Hou, Zheming Liang, Congcong Wang, Kaiwen Tuo, Jun Zhang, Yuhan Zhu, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026.

Yuan Yao, Tianyu Yu, Ao Zhang, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ICLR*, 2024.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. Addressee and response selection in multi-party conversations with speaker interaction RNNs. In *AAAI*, 2018.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A EXPERIMENT INDEX

How to read the appendix. Apps. B and C give the pipeline and the environment that every study shares. Part I holds the studies the main text rests on: the identical-support comparison (App. S), the untrained baselines (App. T), state patching (App. U), the text-only trainee (App. X), PersonaPlex (App. Y), the external benchmarks and cases (Apps. E, F, G), and the rating audit and wild diagnosis (Apps. H, I). Part II is the historical campaign that preceded them (Apps. D and J–R, V, W, Z–AC, summarised in App. AI). Part III holds the registries and audits (Apps. AD–AH); App. AJ collects the additional figures and tables and App. AK the extended related work. Naming: lockbox-vN is sealed set N; a codename's PREREG file is the plan frozen before its data existed, and its RESULT file carries the outcome.

Names used for the same arms. Appendix tables and verdict files keep the codenames under which the arms were run: *decision credit* is also *local*, *moment credit*, *PD* (per-decision, on When2Speak) and *ours*; *episode return* is also *ER*, *GRPO* (which here names the offline scenario-normalised episode-return construction, not on-policy group RL) and *CS* (conversation-sum, on When2Speak); *critic* subtracts a cross-fitted value baseline from the return-to-go; *GAE(0.5)* is the generalised-advantage arm; *dense* centres and scales decision credit within a scenario, *flat* keeps its sign; *DLA* is divergence-localised credit and *FREE* its unlocalised counterpart (Appendix R). *e14_W*, where verdict files and App. I name it, is the seed-0 sequence-level decision-credit adapter. The *per-decision labels* row of Table 4 and Fig. 7 is the separate SYM arm (`grpo_dense`, App. N): episode return handed decision credit's labels, not decision credit itself.

The historical experiments carry repository codenames, with analysis plans and result files. Table 3 indexes that campaign; the controlled training and state-intervention follow-up studies are detailed in Appendices S and U. Table 3 maps each to its role, trainee, evaluation set, pre-registered primaries and disposition.

B THE PIPELINE

Decision credit and the sequence-level objective. The first speak in an own-address or bad-news response window receives $+1.0$; each preceding wait receives -0.15. Speaking in an other-address window receives -1.5, and speaking in a pause window receives -0.75. Waiting in these two windows receives $+0.375$ and $+0.1875$, respectively; neutral decisions receive zero credit. For the historical sequence-level trainer, positive-credit targets retain the logged action and, for speaking, its utterance. Negative-credit targets contain the opposite action without an utterance. The supervised term is the token-averaged negative log-likelihood weighted by absolute credit; neutral anchors use the same binary cross-entropy form as Eq. 1. The speaking probability $p_\theta(s_t)$ is read at the decision position by summing the supported `speak` token variants and renormalizing against `wait`. The historical implementation normalizes credit weights within a micro-batch, whereas the controlled action-only comparison uses fixed dataset denominators. These two training protocols are reported separately.

The historical arms use offline signed credit-weighted SFT (advantage-weighted regression (Peng et al., 2019) with signed weights, a negative weight supervising the opposite action) on shared source collections. Credit construction also changes retained rows, anchor counts, and update counts; Table 4 lists each arm's construction, rows, anchors, learning rate and steps; Figure 2 draws the pipeline, and Figure 7 shows, for one training episode, the weight each historical construction placed on every poll. Shading in Table 1: elevation cells are warm above neutral and teal below, deeper with magnitude, with the decision-credit column highlighted in green; contrast cells are teal, deeper with the reduction.

Table 3: Codename index. Verdict files carry the exact numbers; pre-registrations carry the frozen wording.

codename	what it is (trainee / set)	primaries → disposition	main text	where
V5, V7, V8, V9	sealed sets; decision credit vs base / critic / episode return (Qwen3-VL-8B)	decision-level contrasts → confirmed (v7 partial); v5 episode-level primary → REFUTED (Table 5)	§3, §3.2	App. N
KBASE	episode-return and cross-fitted-critic arms (Qwen3-VL-8B / V9)	ours−critic, ours−episode return → confirmed	§3.2 (Tab. 1)	App. N
SYM	episode return handed per-decision labels (V9)	two co-primaries → confirmed	—	App. N
TUNE	LR-tuned baselines selected on DEV elevation (V9)	both primaries → survive	§3.2 (Tab. 1)	App. N
ESC	escape-decode fix, full re-run (V9)	V9 co-primaries → replicate	§3.2 (Tab. 1)	App. AB
GAE	GAE(0.5) on the cross-fitted V (V9)	ours−GAE → confirmed	§3 (Fig. 3)	App. N
SEEDS, SEEDS A1	training-seed variance; seed-inclusive interval (V9)	descriptive → all intervals exclude 0	—	App. M
FLAT	graded vs flat ± 1 credit under a dose-preserving trainer	addressee window → confirmed	—	App. L (Tab. 10)
PFIX, DOSE	pause-guard rescale, twice	pause propensity → both variants missed their frozen rule (stopped)	—	App. AC
J2, J3	JoyAI-VL: replication on three collections (lockbox-v6); episode-return baseline on the same collections (lockbox-10)	J2 selectivity, guard ratio → confirmed; J3 ours−episode return, both levels → confirmed; episode return leaves the guard at base	§3.1, §3.2	App. V
IVL, MCPM	InternVL3-8B, MiniCPM-o-2.6, fresh sets	decision-level primaries → three of four; confirmed	§3.2	App. W
CRIT2	late critic arms on MiniCPM-o and InternVL3	critic contrasts → MCPM confirmed; IVL voided	§3.2	App. W
U, C	uncertainty/stakes and audience-conditioned disclosure norms	→ partial; boundary	§5 (C only)	App. AA, App. Z
DIAG, DIAGCI	eleven-checkpoint probe and its intervals	ordering → holds in all 11 (CI excludes 0 in 9)	§3.1	App. J
TEMPT	private fact swapped on logged contexts (V9)	elevation without the fact → holds	§2	App. Q
PROBE2, PROBE3	representation-probe replications, three lineages	accuracy, recall → PROBE2 missed recall; PROBE3 3/3	§4	App. R
ATTRIB	utterance occlusion (exploratory)	— → log-odds re-analysis	—	App. R
DLA S2, S3	divergence-localised credit: overlap; training	overlap gate; vs episode return, vs free floor → S2 confirm; S3 P1 fails, P2 holds	—	App. R
W2S	When2Speak external evaluation (frozen test subsample)	two co-primaries → confirmed	§5	App. E
WILD	AMI probe, six checkpoints + shipped adapter	elevation, transfer → base elevation negative; adapter does not transfer	§5	App. I
PPX, PPX-TRAIN (+A1)	PersonaPlex probe and training (audio instrument)	P1 other, P2 own → confounded pool missed P1; balanced pool holds	§3.2	App. Y
WINDUR	guard-window duration re-scoring, 2/3/4/6 s (V9)	descriptive → all contrasts exclude 0 at every length	§5	App. AE
ISO	matched-answering-rate sweep on the frozen contexts (V9)	ours $<$ base, $<$ critic → confirmed; ours $<$ episode return → FAIL	§3.4	App. AI
CAP (+A1)	general capability on four public benchmarks (Qwen3-VL-8B arms)	Δacc within 2 points → all three arms retained	§5	App. F
PUB	When2Speak full test; Speak or Stay Silent zero-shot	W2S P1, P2 → replicate; SoSS addressee gap → transfers on *Friends* and SPGI, not AMI	§5	App. F
CASES (`PREREG_FIX`)	rule-chosen single decisions from the frozen contexts (V9)	descriptive	§3.4	App. G (Fig. 8)
INFER	norm instruction and vocative rules, untrained (V9 controlled contexts)	prompted base at the matched rate → does not move the guard; rules descriptive	§3.3	App. T
MIX	$\lambda = 1$ anchor and mixed credit on the controlled rows (V9 contexts)	P1 guard kept, P2 neutral cost removed → the $\lambda = 1$ anchor: guard kept, neutral cost not removed; mixed credit: guard kept, neutral cost removed	§3.3	App. T
LLM	text-only Qwen3-8B on the controlled rows and trainer (V9 contexts)	matched-rate drop in other-address violation → replicates	§3.2	App. X
LLMSEQ	full sequence-level recipe on text-only Qwen3-8B (lockbox-v9; seed 0 primary, seeds 1–4 replications; no critic arm)	A1 ours vs base, A2 ours vs episode return → confirmed, confirmed; replication seeds 1: A1 confirmed, A2 confirmed, 2: A1 confirmed, A2 confirmed, 3: A1 confirmed, A2 confirmed, 4: A1 confirmed, A2 confirmed	§3.2	App. X
UTT	blind pairwise utterance ratings, base against three arms on the own-addressed and bad-news test contexts	mean rating difference, decision credit primary → decision credit -0.454, below the 0.50 criterion, interval excludes zero; episode return -0.701, cost; critic -0.461, below the 0.50 criterion, interval excludes zero	§5	App. H
human	Study A (180 items), Study B (60 pairs), ≥ 5 annotators	agreement, Q2 win rate → instrument and analysis frozen; collection pending	—	—
CONTROLLED	five-seed identical-support study: four credits on identical rows, anchors and update count (Qwen3-VL-8B / V9 contexts); protocol fixed before training, after the historical set-9 results	matched-rate contrasts among the constructions → unresolved on the guard; every arm below the base	§3.3 (Tab. 2)	App. S
PATCH, SWAP	bidirectional state patching at the decision position; 41 addressee-swap pairs (Qwen3-VL-8B / v8 development, v9 test)	separation transfers at late blocks → replicates on five seeds; swap gap → decision-level arms widen it, episode return does not	§4	App. U

Table 4: Every arm is offline signed credit-weighted SFT on fixed base collections with the loss of the sequence-level variant described in Section 2, learning rate 10^{-4}: LoRA rank 16, α=32, dropout 0.05 on all attention and MLP projections; batch 2×8 accumulation; warm-up 5%; bf16; KL anchor weight λ=0.3; seed 0 unless a seed study says otherwise. "Local scale, mass" is the shared rescaling of mean $|w|$ to decision credit's, mass-balancing of pro- and anti-speaking rows, and the $|w| < 0.05$ drop, applied identically to every episode-level construction; decision credit takes the mass balance and the drop and sets the scale. Anchors are drawn by one rule, $0.5\times$ the credit rows from base-rollout neutral contexts, capped by the pool at 2,225. The episode-style sets carry 3.3–3.5$\times$ the rows and are trained one epoch, so their optimizer steps land within 16% of decision credit's 441; the two divergence-localised arms, built from a smaller pool, received fewer. Decision credit's stored weights after the balance are -2.16 / -1.08 on guard speaks, $+0.54$ / $+0.27$ on guard waits, $+1.0$ on first responses and -0.15 on waits before them. Every value here is derived from the dataset files and the trainer's logs by `audit_config.py` (`CONFIG_RESULT`).

arm	per-decision credit A_t	scaling	rows (credit + anchors)	epochs, steps
decision credit (ours)	window sign and magnitude (the per-window values listed above)	mass, drop	1,564 + 782	3 ep, 441
episode return, global	episode return, global z-score	local scale, mass	5,967 + 2,225	1 ep, 512
episode return	$(R_i - \bar{R}_g)/\mathrm{sd}_g$, g = scenario (GRPO's normalisation (Shao et al., 2024))	local scale, mass	5,923 + 2,225	1 ep, 510
critic	$G_t - V(s_t)$, 5-fold cross-fitted ridge	local scale, mass	5,744 + 2,225	1 ep, 499
GAE(0.5)	$\sum_k 0.5^k \delta_{t+k}$, same V	local scale, mass	5,422 + 2,225	1 ep, 478
per-decision labels	structural credit, group-normalised	local scale, mass	1,487 + 743	3 ep, 420
divergence-localised (DLA)	$\pm\Delta R\, v_t$, $v_t \propto 1 - \cos(h_t^A, h_t^B)$	local scale, mass	2,968 + 1,484	1 ep, 279
free-floor control	$\pm\Delta R\, v_t$, v_t uniform on free-floor ticks	local scale, mass	3,988 + 1,994	1 ep, 374

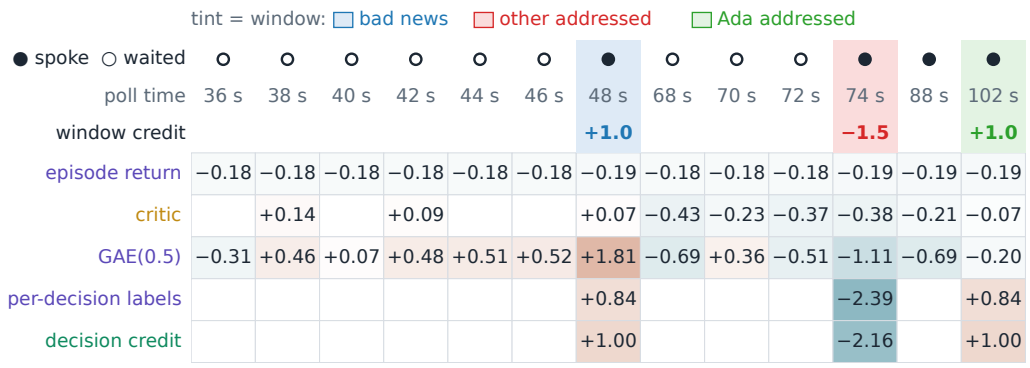

Figure 7: One training episode, poll by poll, under each historical credit construction. Columns are the trainee's polls in one training episode (vendor scenario), tinted by the annotated window they fall in, with the trainee's act and the window's signed credit above. Cells are the stored weight on the act token read from each arm's dataset file; a blank cell means that construction built no training row for the poll. Episode return weights every poll alike, the critic and GAE spread signed weight over neutral polls, and per-decision labels and decision credit touch only the credited polls. Cell tint: warm for a positive stored weight, teal for a negative one; stored weights include the mass balance of Table 4, which maps the window's -1.5 to -2.16.

The appendix carries the per-battery accounts, the robustness audit, and the boundary map. Every number in it is produced by the same committed scripts as the main text and is referenced from the section that relies on it.

C THE ENVIRONMENT AND ITS SCENARIOS

Scenario schema. A scenario specifies a meeting goal, one private fact per participant (Ada's is phrased "Something only you know (bring it up when relevant): ..."), a facilitator opener, a scripted sequence of utterances with start times, and four indices into that script marking the annotated moments: a question to another participant (other address), that participant's hesitation (pause), a question to Ada (own address), and a piece of bad news. Scenarios were written by a language model from a topic list and kept only if a structural validator accepted them (four moments present and ordered, speaker names consistent, utterance lengths within a verbosity envelope); each sealed set was generated and frozen before any tested policy was run on it and holds scenario families that no training collection used (App. D), and Table 5 lists their provenance.

Trainees. The participation decision studied here belongs to assistants that are always on: proactive, streaming, and full-duplex models that must decide whether to speak at all, not only what to say. Such assistants ship from vision-language and speech lineages, which is why the trainees are Qwen3-VL-8B (whose lineage includes the streaming Qwen3-Omni), JoyAI-VL-Interaction (built on Qwen3-VL; a proactive model trained with per-second RL to take turns), InternVL3-8B, MiniCPM-o-2.6 (a streaming omni model), and the full-duplex PersonaPlex-7B; with the text-only Qwen3-8B of App. X, six trainees on three backbones (Qwen3, Qwen2.5, Moshi); every text trainee is Qwen-derived (App. W). The simulator reaches every text-model trainee through the same text transcript, so their vision encoders are never used and the decision is isolated from perception; the text-only Qwen3-8B of the same family reproduces the decision-credit result under both protocols (App. X), so nothing rests on a vision tower. PersonaPlex hears TTS renderings of new scenarios from the same generator and decides frame by frame (App. Y). JoyAI-VL already decides when to speak, and decision credit still takes its other-address elevation below neutral (Table 1). All scenarios are English and use the names Ada/Ben/Cleo (Mira/Noah/Iris in the renamed control of App. U); the peers' own turns are unlabelled: only the four scripted moments open windows, and every other poll is neutral.

Engine. Time advances at 1 Hz for 160 s. Utterances last $\max(2, \lceil \text{words}/2.5 \rceil)$ s; a gap of at least 4 s is rendered as an explicit silence line, an utterance in progress carries the marker `[still speaking]` and shows only the words spoken so far (a share of its words proportional to elapsed time, followed by an ellipsis), so no poll sees text before it is uttered; starting to speak over one adds `[interrupting]`. Ada is the trainee. Ben and Cleo are peers driven by JoyAI-VL under the same prompt format and decision head, so they react to what Ada does; the facilitator and the four moment utterances are scripted, which is why the moments occur at the same times in every episode of a scenario. Ada is queried every 2 s and not for 4 s after speaking; at each query the model sees the system prompt below and the transcript so far followed by "It is now [mm:ss]. You are Ada and not currently speaking. Decision (JSON only):". The decision head reads the next-token distribution after the prefix `{"act":` : in this tokenizer `speak` is two sub-tokens (`s`, `peak`) and `wait` is one, so $P(\text{speak})$ is the probability mass of the `speak`-initial variants renormalised against `wait`; the action is sampled from that probability, and only a sampled `speak` triggers a second call that generates the utterance (prefix `{"act": "speak", "text": "`, temperature 0.7, at most 70 tokens). After the episode, the two peers rate each other and Ada, and a separate judge rates the transcript.

System prompt (verbatim, Ada's seat). *You are Ada, a colleague on a three-person voice call with Ben and Cleo, moderated by a facilitator. Goal of the call: [goal]. Something only you know (bring it up when relevant): [fact]. You will repeatedly be shown the conversation so far, with timestamps and silences. Each time, decide whether to start speaking RIGHT NOW or stay silent. Behave like a real person: contribute actively, react, answer when the facilitator addresses you, speak up when you know something relevant, interrupt only if you genuinely would, and stay quiet when staying quiet is the socially right thing. If someone is marked [still speaking], starting now means interrupting them. Never repeat a point you have already made. Answer ONLY with JSON: {"act": "wait"} or {"act": "speak", "text": "one or two natural sentences"}.* Figure 8 shows the tail of three real queries.

Window rules. Each annotated utterance opens a window at its audio end: 6 s after the other-addressed question and 8 s after the pause request (guards, in which speaking is the violation), 16 s after the question to Ada and 20 s after the bad news (response windows, in which the first speak is the response). A poll that falls in more than one window is assigned to the first of own address, other address, pause, bad news, in that order, both in credit assignment and in scoring; the pause window

opens only when the addressee's hesitation is a scripted utterance, so it never precedes the guard it follows. Polls outside every window are neutral and receive no credit; they supply the anchors of Eq. 1, drawn from base-policy episodes only. The temptation probe (Appendix Q) varies the private fact, and Appendix AE(v) re-scores the guard at 2–6 s. The windows are defined by time alone and do not condition on who holds the floor: on the 2,884 controlled-study test contexts the last transcript line carries `[still speaking]` in 0.629 of own-address, 0.548 of other-address, 0.683 of pause, 0.265 of bad-news and 0.247 of neutral contexts (`STILL_SPEAKING_RESULT`), so own-address polls are mid-turn more often than other-address polls, and a policy that merely avoided talking over someone would answer less, not more, when addressed.

D THE SEALED SETS

Every set below was generated and frozen with its analysis plan and permitted wording before any tested policy was run on it, holds scenario families that no training collection used, and was given one pre-registered evaluation before any result on it was seen; sealing fixes when a set is frozen and implies no different generator (lockbox-v5, for instance, was designed after the v4 results were read and frozen before any policy ran on it, `PREREG_V5`). Later follow-up studies (TUNE, GAE, SYM, SEEDS, ESC, TEMPT, WINDUR, ISO, CASES, INFER, MIX, LLM, LLMSEQ, UTT, the peer ratings, the controlled study and state patching) re-use v8/v9 contexts under their own protocols; those re-uses are listed in Table 3 and are not counted as sealed confirmations (App. AI); every choice is fixed on DEV or v8 before v9 is read, and mixed credit is the one design written after a v9 result. The list is generated from the artifacts by `sealed_sets.py`, which requires a scenario module, a dedicated one-shot evaluation runner, and a frozen pre-registration for each entry—the count in Table 5's caption (fifteen sets, 191 families) is derived from that script rather than maintained by hand.

Table 5: The fifteen sealed sets, 191 scenario families in total. No frozen decision-level other-address contrast against the base missed on any set, and four of the fifteen sets (lockbox-v2, v6, v8 and lockbox-10) confirmed every frozen primary. Elsewhere one primary or co-primary missed: lockbox-v3's first JoyAI-VL run (by 0.0161; App. V), lockbox-v4's author-shift primary (App. P), lockbox-v7 (PARTIAL, App. N: the vs-critic contrast resolved, the vs-episode-return one did not), the episode-level co-primary families of lockbox-v9 and lockbox-12, and lockbox-11's episode contrast against episode return (three of its four primaries confirmed; Apps. N, W), and the second-norm U set, whose three decision-level contrasts resolve while its episode co-primary misses ($+0.267$ $[-0.015,\ +0.549]$, p=0.054; App. AA). Lockbox-v5's episode-level primary against the learned baselines was REFUTED (App. N) while its author-shift primaries against the base passed (App. P), lockbox-13 (`PREREG_PFIX`) missed its co-primary (App. AC), and the third-norm set (`PREREG_C`) missed all three of its primaries (App. Z); the last two are counted among the nine stopped experiments of App. AG. The fifteenth (lockbox-14) carried two frozen batteries: `PREREG_DOSE` missed its co-primary (the ninth stopped experiment), and under `PREREG_FLAT` the graded credit beats the flat clone once the trainer preserves the dose (App. L, Table 10; under the identical-support study of §3.3, flat and decision credit are not resolved).

set	pre-registration	families	what it tested
`lockbox-v2`	`V2`	10	six independent training runs; run-level inference
`lockbox-v3`	`V3`	12	coverage; JoyAI first run
`lockbox-v4`	`V4`	12	vs episode return (GRPO-normalised) and cross-fitted critic
`lockbox-v5`	`V5`	13	scoreboard masking, first statement
`lockbox-v6`	`J2`	12	JoyAI replication, three collections
`lockbox-v7`	`V7`	12	decision-level primary, model-authored
`lockbox-v8`	`V8`	12	selection-free estimator frozen as primary
`lockbox-v9`	`V9`	24	largest set; symmetric episode family
`lockbox-10`	`J3`	12	episode-return baseline on JoyAI
`lockbox-11`	`IVL`	12	InternVL3-8B lineage
`lockbox-12`	`MCPM`	12	MiniCPM-o-2.6 lineage
`lockbox-13`	`PFIX`	12	pause-guard rescale (self-rejected)
`U-test`	`U`	12	second norm: uncertainty-conditioned
`C-test`	`C`	12	third norm: audience-conditioned disclosure (FAILED)
`lockbox-14`	`DOSE`	12	pause dose FAILED; magnitudes matter once delivered

E THE EXTERNAL BENCHMARK BATTERY, IN FULL

PREREG_W2S froze the design before any probe: the frozen evaluation subsample (3,810 of When2Speak's 21,687 test contexts: all 970 where the agent is solicited by name, all 840 unsolicited questions, 2,000 statements; seed 0, committed before any model saw them), the two training arms on one frozen 3,000-conversation subset, and both readings. Amendment A1 (pre-run, on the record) replaced a degenerate conversation-sum construction—label CE with a per-conversation learning rate—with the paper's ladder grammar: reward content is the initial policy's per-decision correctness (± 1), PD carries each decision's own sign, CS its conversation's centred mean, and a negative weight supervises the opposite of the taken label token. Training (train_w2s.py): one pass over the 56,473 decisions of the 3,000-conversation subset, one seed per arm, LoRA r=16, α=32, dropout 0.05 on the attention and MLP projections, AdamW at 10^{-4} with 32-decision accumulation (about 1,765 steps); no simulator data enters either arm. Full frozen-set numbers (W2S_SM_RESULT): base macro F1 0.331, MIR (missed-intervention rate, the share of SPEAK-labelled contexts read as SILENT) 0.406, recall 0.594; PD 0.784/0.474/0.526; CS 0.638/0.578/0.422. Stated plainly: both arms trade recall for precision against a base that over-speaks, so MIR *rises* under both; the pre-registered scoreboard direction was macro F1, and MIR is reported beside it rather than absorbed. The Stage-1 base-checkpoint probes (W2S_DIAG_RESULT) reproduce the wild battery's per-checkpoint sign pattern on a second external corpus: the trainee base carries no unsolicited-question excitation (-0.069), and the two text models carry a small positive one (Qwen2.5-7B $+0.037$, Qwen3-8B $+0.047$, the largest; W2S_DIAG_RESULT).

F PUBLIC BENCHMARKS, IN FULL

PREREG_PUB (two participation benchmarks) and PREREG_CAP (four capability benchmarks, with amendment A1) were frozen before any of their scoring ran; PUB_RESULT and CAP_RESULT hold every number below, and PUB_AUROC_RESULT the threshold-free readings. Tables 6, 7 and 8 give When2Speak, Speak or Stay Silent and the capability battery.

When2Speak, full test set. All 21,687 contexts of the token task's test split are read with the frozen subsample's readout (App. E): the renormalised two-way read over the benchmark's own label tokens at the assistant position, SPEAK when $P > 0.5$. On the 3,810 contexts shared with that subsample, re-scoring reproduces its macro F1 for all three models ($0.331 \to 0.332$, $0.784 \to 0.786$, $0.638 \to 0.642$); 48, 8, and 18 decisions flip under bf16 nondeterminism. Published numbers are quoted from Nama et al. (2026) and were not re-run; their models generate the label token after training on the full training split, whereas our arms read it and saw a 3,000-conversation subset. Both pre-registered readings of PREREG_W2S replicate on the full set: both arms improve macro F1 over the base (P1), and, with E_{QO} the pre-registered guard analogue, the mean speaking probability on SILENT-labelled unsolicited-question contexts minus that on SILENT-labelled statements, and PD/CS the per-decision and conversation-sum credit arms, $D = E_{QO}(\mathrm{CS}) - E_{QO}(\mathrm{PD}) = 0.068$ $[0.048,\ 0.088]$ over 1,064 test conversations (P2; subsample 0.072). D is a difference in differences: the raw unsolicited-question means are 0.105 (PD) and 0.286 (CS), a gap of 0.181 $[0.160,\ 0.201]$, and the statement means 0.079 and 0.191, a gap of 0.113 $[0.108,\ 0.118]$ (W2S_QO_RESULT; subsample 0.180 and 0.108). When2Speak ships no conversation identifiers; the 1,064 clusters are reconstructed from the overlap of consecutive 8-turn windows, and splitting instead wherever consecutive windows share no turn (1,735 clusters of at most 25 windows) leaves the interval at $[0.047,\ 0.090]$.

Threshold-free readings. Both When2Speak arms turn conservative, so a reading at $P > 0.5$ cannot separate selectivity from silence; we therefore also read the stored probabilities without a threshold (PUB_AUROC_RESULT; conversation-bootstrap 95% intervals, B=9999). Against the SPEAK label the per-decision arm reaches an AUROC of 0.879 $[0.873,\ 0.886]$ and an AUPRC of 0.580, the conversation-credit arm 0.688 $[0.679,\ 0.698]$ and 0.219, and the base 0.445 and 0.112; inside the unsolicited-question cell the AUROCs are 0.884 and 0.618. Held at the conversation-credit arm's own answering rate (0.123 of all contexts read as SPEAK), the per-decision arm recalls 0.528 $[0.509,\ 0.547]$ of the SPEAK-labelled contexts at a false-interruption rate of 0.063 $[0.059,\ 0.067]$, against 0.266 and 0.102 for conversation credit, and at conversation credit's recall its false-interruption rate is 0.011; per-decision supervision is therefore more selective, not only quieter.

Table 6: When2Speak, full test set (21,687 contexts, 12.9% SPEAK). Published rows from Nama et al. (2026), zero-shot → fine-tuned on the full training split (macro F1 ranges 0.206–0.569 and 0.697–0.747); ours with conversation-bootstrap 95% intervals. FIR: false-interruption rate, the share of SILENT-labelled contexts read as SPEAK; MIR: missed-intervention rate, the share of SPEAK-labelled contexts read as SILENT (1−recall).

model	macro F1 ↑	FIR ↓	MIR ↓	recall ↑
GPT-4o-mini (published)	0.569 → 0.697	0.257 → 0.163	—	—
GPT-4.1-mini (published)	0.499 → 0.729	0.435 → 0.105	—	—
Qwen3-4B (published)	0.206 → 0.737	0.897 → 0.037	—	—
Llama-3.1-8B (published)	0.337 → 0.740	0.731 → 0.044	—	—
Qwen3-8B (published)	0.466 → 0.739	0.000 → 0.036	—	—
Llama-3.3-70B (published)	0.325 → 0.747	0.749 → 0.050	—	—
Qwen3-VL-8B, zero-shot	0.299 [0.290, 0.307]	0.742 [0.730, 0.754]	0.324 [0.305, 0.343]	0.676 [0.657, 0.695]
+ per-decision labels	0.680 [0.669, 0.690]	0.013 [0.011, 0.015]	0.714 [0.698, 0.732]	0.286 [0.268, 0.302]
+ conversation credit	0.584 [0.575, 0.592]	0.102 [0.098, 0.106]	0.734 [0.718, 0.751]	0.266 [0.249, 0.282]

Speak or Stay Silent. Bhagtani et al. (2026) label, for every non-speaking participant at each utterance boundary of AMI meetings, *Friends* episodes, and SPGI earnings calls (Grossman et al., 2025), whether that participant takes the next turn; test labels are balanced by construction. The adapter speaks more on all three corpora, so a larger addressee gap does not by itself raise the benchmark's speak-or-silent accuracy. We render every test row into the trainee's own decision format (`build_soss.py`): the engine's transcript renderer and duration rule ($\max(2, \lceil \text{words}/2.5 \rceil)$ s per turn, 1 s gaps), the decision read 1 s after the current turn ends, names mapped as in our AMI battery (AMI letters to four names, *Friends* characters kept, SPGI identifiers as "Participant k"), and one setting line per corpus under the AMI battery's instruction paragraph; 81 contexts exceed 3,072 tokens and are scored whole. None of the four arms saw these corpora in training. The addressee gap G compares speaking probability when the target is addressed with when another participant is addressed; SPGI's addressees are heuristic, and the benchmark's AMI split comes from the corpus our AMI battery used, so it is not independent evidence about AMI. Zero-shot, the decision-credit adapter changes balanced accuracy by $+0.7$ and $+0.2$ points on AMI and *Friends*, intervals including zero, and by $+1.9$ $[+1.3, +2.6]$ on SPGI; its addressee gap moves by 0.103 $[0.038, 0.169]$ on *Friends*, where addresses are vocatives (from an inverted -0.107 to -0.004), by 0.053 $[0.044, 0.063]$ on SPGI, and by -0.027 $[-0.111, 0.053]$ on AMI. The adapter raises mean speaking probability on every corpus (AMI $0.313 \rightarrow 0.422$, *Friends* $0.625 \rightarrow 0.755$, SPGI $0.144 \rightarrow 0.233$), which is why its gains on addressed turns (I1) come with losses on the silent categories. The two episode-level arms move the other way: they answer only 1–16% of explicitly addressed turns on AMI and SPGI, which costs balanced accuracy on AMI (-4.5 and -4.0 points) and, for the critic, on *Friends* (-4.1); their small SPGI gains ($+0.9$, $+0.7$) come with a shrinking addressee gap. Episode return's *Friends* gap change ($+0.086$ $[+0.008, +0.164]$) falls in the TRANSFER branch, but its addressee AUROC there moves by -0.024 and it answers only 10.5% of explicitly addressed turns (Table 7), so its inverted gap closes because the arm goes quiet on addressed and other-addressed turns alike, not because it ranks addressed turns higher. Each is a single trained seed, and no training-method conclusion is drawn.

On Speak or Stay Silent the same rank-based reading asks whether an arm orders addressed above other-addressed turns regardless of how much it speaks: the addressee AUROC of $P(\text{speak})$ under the decision-credit adapter moves from 0.458 to 0.511 on *Friends* ($+0.053$ $[+0.012, +0.095]$), from 0.667 to 0.676 on SPGI ($+0.009$ $[+0.003, +0.015]$) and from 0.538 to 0.559 on AMI ($+0.020$ $[-0.019, +0.076]$; meeting-cluster bootstrap, B=9999; `PUB_AUROC_RESULT`), so the change in the addressee gap is not only the adapter's overall rise in speaking probability on *Friends* and, narrowly, on SPGI, while on AMI it is unresolved. The episode-return and critic arms raise this AUROC on no corpus (*Friends* -0.024 and -0.112, SPGI -0.018 and -0.006).

Table 7: Speak or Stay Silent, zero-shot. Balanced accuracy, per-category accuracy (I1 addressed and speaks, I2 speaks unaddressed, S1 silent and unreferenced, S2 silent though referenced) and the addressee gap G; contrasts against the base with meeting-cluster bootstrap 95% intervals and the pre-registered branch, fixed in `PREREG_PUB`: TRANSFER if the interval lies above 0, HARM if it lies below 0, NONE if it spans 0. The primaries are decision credit minus base in balanced accuracy and in G, per corpus; the episode-level rows are descriptive, and no familywise claim is made across them (App. M).

corpus	arm	bal. acc.	I1	I2	S1	S2	G
AMI	base	54.6	54.7	29.6	74.5	72.3	$+0.092$
	decision credit	55.4	66.7	41.9	64.0	60.8	$+0.065$
	episode return	50.2	8.2	2.2	96.7	97.0	$+0.013$
	critic	50.7	15.7	4.4	94.1	95.2	$+0.011$
Friends	base	48.6	65.5	58.9	60.7	27.1	-0.107
	decision credit	48.7	84.1	74.2	44.5	11.0	-0.004
	episode return	48.0	10.5	15.3	91.9	78.9	-0.021
	critic	44.5	12.7	21.4	89.0	64.0	-0.055
SPGI	base	49.1	23.2	5.5	91.9	81.0	$+0.126$
	decision credit	51.0	35.6	8.7	88.1	77.8	$+0.179$
	episode return	49.9	1.0	0.1	99.6	99.3	$+0.045$
	critic	49.8	3.2	0.3	99.2	97.4	$+0.058$

corpus	arm − base	Δ bal. acc. [95% CI]	branch	ΔG [95% CI]	branch
AMI	decision credit	$+0.7$ $[-1.0, +2.8]$	NONE	-0.027 $[-0.111, +0.053]$	NONE
	episode return	-4.5 $[-7.4, -2.0]$	HARM	-0.078 $[-0.239, +0.070]$	NONE
	critic	-4.0 $[-6.8, -1.7]$	HARM	-0.081 $[-0.241, +0.061]$	NONE
Friends	decision credit	$+0.2$ $[-2.0, +2.4]$	NONE	$+0.103$ $[+0.038, +0.169]$	TRANSFER
	episode return	-0.6 $[-3.3, +2.1]$	NONE	$+0.086$ $[+0.008, +0.164]$	TRANSFER
	critic	-4.1 $[-6.8, -1.4]$	HARM	$+0.052$ $[-0.025, +0.129]$	NONE
SPGI	decision credit	$+1.9$ $[+1.3, +2.6]$	TRANSFER	$+0.053$ $[+0.044, +0.063]$	TRANSFER
	episode return	$+0.9$ $[+0.2, +1.5]$	TRANSFER	-0.080 $[-0.091, -0.069]$	HARM
	critic	$+0.7$ $[+0.1, +1.4]$	TRANSFER	-0.067 $[-0.078, -0.057]$	HARM

General capability. MMLU (Hendrycks et al., 2021) (all 14,042 test questions, zero-shot), MME (Fu et al., 2023) (2,374 yes/no questions over 1,187 images), AI2D (Kembhavi et al., 2016) (3,088), and the letter-option subset of RealWorldQA (xAI, 2024) are read by one closed-set rule: the argmax over the option-letter tokens (MME: *Yes*/*No*) at the assistant position, with no generation or parsing, so absolute accuracies are not leaderboard numbers. Amendment A1, recorded before any outcome was read, computes the vision patch embedding as its exact linear equivalent after the stock bf16 Conv3d fell back to a host-bound kernel (four MME items agree to 10^{-4} in $P(\text{Yes})$). The decision rule was fixed in advance: an arm retains capability if the lower 95% bound of its accuracy change against the base exceeds -2 points on all four benchmarks. The RealWorldQA letter-option subset has 426 of its 765 items. Under MME's own scoring the base, decision credit, episode return, and critic reach perception 1741.8, 1744.2, 1733.6, and 1734.8 and cognition 669.3, 670.7, 669.3, and 669.3. All three trained arms meet the pre-registered margin on every benchmark.

Table 8: General capability. Base accuracy and each arm's change in points with paired-bootstrap 95% intervals (MME resampled by image pair) and the pre-registered verdict.

arm	MMLU	MME	AI2D	RealWorldQA	verdict
base (accuracy, %)	74.38	89.64	83.68	72.77	
decision credit − base (points)	$+0.01$ $[-0.15, +0.17]$	-0.08 $[-0.29, +0.13]$	$+0.16$ $[-0.16, +0.49]$	$+0.47$ $[-0.94, +1.88]$	RETAINED
episode return − base (points)	-0.07 $[-0.24, +0.09]$	-0.21 $[-0.46, +0.04]$	-0.06 $[-0.36, +0.23]$	-0.47 $[-1.88, +0.70]$	RETAINED
critic − base (points)	-0.08 $[-0.24, +0.08]$	-0.21 $[-0.46, +0.00]$	$+0.00$ $[-0.29, +0.29]$	$+0.00$ $[-1.41, +1.41]$	RETAINED

G CASE STUDIES, IN FULL

In the main-text figure's typical other-addressed case decision credit speaks with probability 0.002 against 0.285–0.349 for the other three arms; in the typical own-addressed case every arm answers (0.831 for decision credit); and in the boundary case (0.998) the question concerns a fact only the trainee holds, and every arm speaks. Every case in Figures 8 and 9 is chosen by `select_cases.py`

from the frozen shared-context battery (PREREG_FIX: every logged decision of the base arm's frozen lockbox-v9 run, session 0, read by the four sequence-trained Qwen3-VL-8B arms at the identical context), and CASES_RESULT lists the chosen contexts with all four arms' values. For each scored window type, the *typical* case is the context whose decision-credit-minus-base difference in $P(\text{speak})$ is closest to that window type's mean difference, with ties going to the earliest episode and tick; the window means are -0.353 for other address (115 contexts), -0.120 for own address (186), -0.129 for pause (60), and $+0.168$ for bad news (226). The *boundary* case is the other-addressed context where decision credit is most likely to speak, a worst case for our arm by construction; its question concerns the trainee's private fact, the temptation that App. Q varies experimentally. Transcript lines are the base arm's logged prompt with the engine's in-progress marker shown as "(still speaking)" and a marked ellipsis where a line is cut. The battery scored through bf16 servers, where the decision logit is quantized and arms can tie (its record of Figure 8a reads 0.321 for all three arms other than decision credit); the figures print fp32 re-reads of the same contexts (CASES_FP32_RESULT), which separate the arms and move no value by more than 0.08; the cases were selected on the battery's values. Figure 17 shows the complementary episode-level view: one live rollout of the base and decision-credit arms.

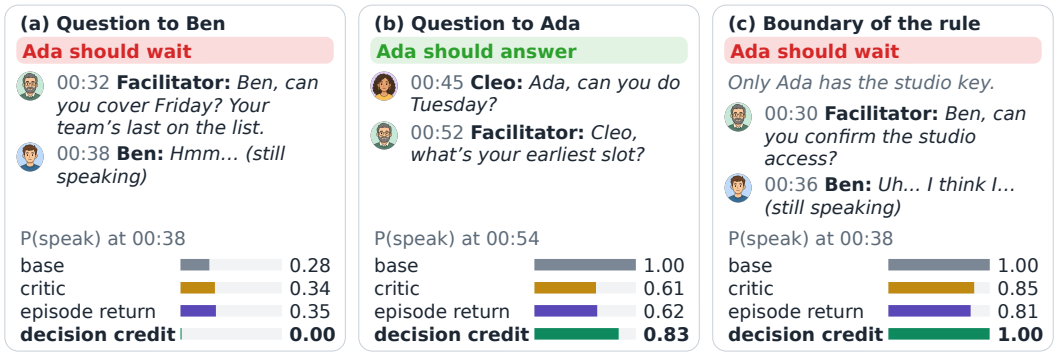

Figure 8: Three shared-context decisions, read by the base and the three sequence-trained arms. Chosen by rule: (a) and (b) are the typical other- and own-addressed contexts (decision-credit-minus-base difference closest to the window's mean); (c) is the other-addressed context where decision credit speaks most readily (values below; two more cases in Fig. 9). Lines are verbatim from the prompt; probabilities are fp32 re-reads; in (b) Ada's own window is still open when the facilitator turns to Cleo. In (c) only Ada holds the fact and every arm speaks, although in aggregate a private fact does not raise speaking (Appendix Q).

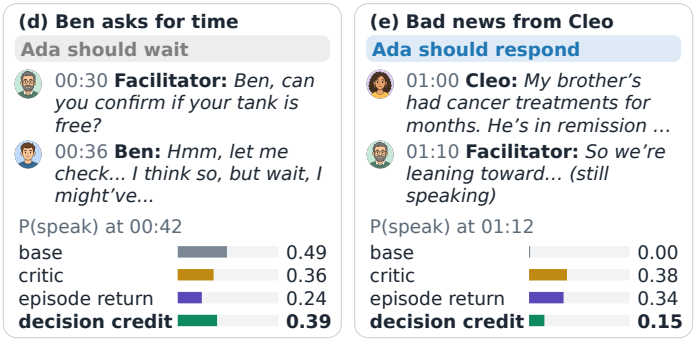

Figure 9: The two remaining typical cases, chosen by the same rule as Figure 8a,b. (d) During a requested pause every trained arm speaks less than the base (0.489): episode return least (0.240), decision credit most (0.391). (e) After bad news the base stays silent while the facilitator is still speaking; the trained arms respond with probability 0.153–0.377.

H THE RATING AUDIT

The audit behind the 0.36 figure of Section 3.1 (JUDGE_AUDIT_RESULT; warmth, competence, and engagement ratings of 420 episodes) uses Qwen3-VL-32B as the observer: given the whole transcript from a third-person view, it rates each participant on 1–9 scales for warmth, competence, and engagement (temperature 0), and Ada's mean rating is the dependent variable. Of 516 rated episodes from twelve experimental sources, 420 carry annotated moments and enter the regression (184 base-policy and 236 trained-adapter episodes over ten scenario families; between-episode SD of the rating 0.640). Three errors are detected mechanically against the annotated windows: answering for another participant (n=150), filling a requested pause (73), and ignoring bad news (215). One ordinary least-squares regression per error, with the error indicator, Ada's speaking rate, and scenario and source fixed effects, gives -0.228 rating points for answering for another participant (-0.355 SD; scenario-cluster wild-bootstrap interval $[-0.316, -0.138]$), -0.152 for filling a pause ($[-0.326, +0.023]$), and -0.131 for ignoring bad news ($[-0.309, +0.048]$); the pooled bad-news effect of -0.344 is an artifact of source differences. The penalty for answering for another participant is therefore real, the largest of the three, and small against the rating's spread: under a normal model of top-quartile selection, an episode with that violation is selected with probability 0.152 against 0.250 without it (empirically, 0.183 against 0.287), a selection pressure of about $1.6\times$. This is why rating-based selection alone slid toward indiscriminate silence in our early experiments (early runs, not reported here) and why the structural credit of Section 2 exists. Asked directly whether Ada answered for someone else (the DETECT run on the same 420 episodes), the observer returns a yes-biased flag, sensitivity 0.98 at specificity 0.25, so it fires on three quarters of clean episodes, whereas the same prompt separates ignored bad news at sensitivity 0.82 and specificity 0.87; the observer shares the trainee's model family (Qwen3-VL-32B), which any rating-based selection would inherit.

Peer ratings of the trained arms. At the end of every episode the two LLM peers rate Ada's warmth, competence, and engagement (1–9) and name the participant they preferred; these ratings are stored for every episode of sealed set 9 (PEER_RATING_RESULT, 240 episodes per arm, 24 scenarios; the original run, before the escape-decode fix of App. AB). They are the closest available measure of how the trained models' contributions land, and they fall: Ada's mean rating is 6.748 under the base and 6.384, 6.184, and 6.519 under decision credit, episode return, and the critic; the peers prefer Ada in 0.656 of base episodes and 0.583, 0.527, and 0.600 of the arms'. With scenario fixed effects the contrasts against the base are -0.364 $[-0.644, -0.083]$, -0.564 $[-0.913, -0.228]$, and -0.229 $[-0.521, +0.067]$ rating points, adjusting for Ada's speaking rate in the episode leaves them at -0.392 $[-0.666, -0.119]$, -0.605 $[-0.931, -0.293]$, and -0.244 $[-0.520, +0.012]$, and further holding fixed what Ada did at the four annotated moments (other-address violation, own-address answer, pause violation, bad-news response) and how often she interrupted leaves -0.346 $[-0.623, -0.074]$, -0.530 $[-0.819, -0.237]$, and -0.171 $[-0.446, +0.090]$. The arms therefore do not lose rating by speaking less or by the deference they were trained to show; the peers are reacting to something else in how the trained models take part, most plausibly the utterances themselves. PREREG_UTT then rated matched utterances blind: for the own-addressed and bad-news contexts of the test set (412 rated per arm), each arm's utterance and the base's were generated in the same context by the simulator's second call and shown in random order to Qwen3-VL-32B, which scored each from 1 to 9 as a contribution for Ada to make at that moment. Decision credit's utterances average 6.40 against the base's 6.85, a difference of -0.454 $[-0.630, -0.292]$ (higher in 0.330 of contexts; own address -0.527 $[-0.821, -0.261]$, bad news -0.394 $[-0.608, -0.174]$), which falls short of the pre-registered 0.50-point criterion for a quality cost although the interval excludes zero, so a smaller cost is present; episode return -0.701 $[-0.928, -0.501]$, a cost by the pre-registered rule (at least 0.50 points with the interval below zero), and the critic -0.461 $[-0.672, -0.262]$.

I THE WILD DIAGNOSIS, IN FULL

Per-checkpoint elevations on the 1909 deduplicated AMI contexts (meeting-clustered CR1 + Webb, G=22): base -0.090 $[-0.169, -0.012]$ (and -0.069 on When2Speak, App. E); Qwen2.5-7B $+0.053$ $[+0.03, +0.08]$ (the only CI excluding zero upward); Qwen3-8B $+0.015$ (own -0.095: the inverted ordering); e14_W -0.114; JoyAI-VL -0.018 at a 0.020 neutral floor; the adapter's suppression relative to its base, the frozen transfer branch, is -0.011 $[-0.053, +0.032]$; Mistral at

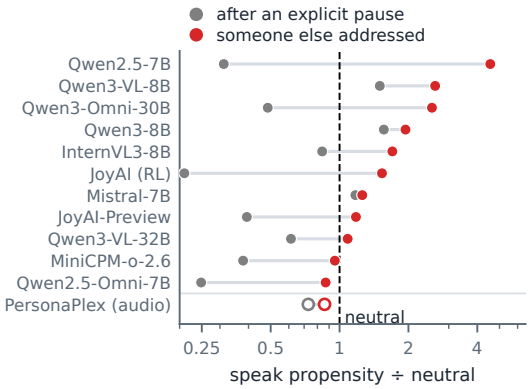

Figure 10: Eleven checkpoints, identical contexts. Speak propensity relative to neutral after an explicit pause (grey) and when someone else is addressed (red); the duplex speech model, probed through audio, is the hollow row. **All eleven rank the addressee guard above the explicit pause** (point estimates; the difference's interval excludes zero in nine).

floor ($P(\text{speak}) < 0.003$, uninformative); MiniCPM-o-2.6 $-0.054\,[-0.09,\ -0.01]$, probed via the vLLM route its simulator number used after the transformers loader's random-init guard correctly refused a partial load. The pre-registered audit fired before interpretation and caught 123 duplicate rows (119 within-class from mutual same-second addressing, 4 cross-class-ambiguous, dropped); render spot-checks passed $10/10$ and act-token ids match the simulator head per tokenizer. Floor-state sensitivity (someone is mid-turn at 66–77% of probe moments in every class): the base's negative elevation persists in both strata, so busy-floor aversion does not explain the non-reproduction. Artifacts: `WILD_RESULT`, `PREREG_WILD`, `ami_probe_contexts.jsonl`.

J THE ELEVEN-CHECKPOINT DIAGNOSIS, PER CHECKPOINT

The eleven checkpoints are Qwen3-VL-8B and 32B (Bai et al., 2025), two JoyAI-VL-Interaction checkpoints (Yao et al., 2026), InternVL3-8B (Zhu et al., 2025), Qwen3-8B (Yang et al., 2025), Mistral-7B-Instruct (Jiang et al., 2023), Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen3-Omni-30B-A3B (Xu et al., 2025b), Qwen2.5-Omni-7B (Xu et al., 2025a), and MiniCPM-o-2.6 (Yao et al., 2024). In its own rollouts on the development set (a behavioural ratio, a different instrument from the shared-context read tabulated below) the trainee base's propensity is uneven across the four moments: own address $4.63\times$ neutral, an explicit floor claim $0.77\times$, and a question addressed to *someone else* $3.82\times$. In the shared-context read (Table 9, plotted in Fig. 10) the same checkpoint reads $1.50\times$ (pause) and $2.62\times$ (other); the controlled pair Qwen3-8B / Qwen3-VL-8B behaves alike ($1.94\times$, $2.62\times$); MiniCPM-o-2.6 and Qwen2.5-Omni-7B are the two checkpoints whose addressee guard is neutral rather than elevated, which is why the main text claims elevation in most but not all checkpoints. Uncertainty (`PREREG_DIAGCI`; scenario-clustered percentile bootstrap, $B{=}2000$): every published ratio reproduced exactly under a hard gate ($11/11$; the Qwen3-Omni-30B number required the vLLM route its original probe used—the transformers Thinker path gives 1.57 where the published route gives 2.53, a route sensitivity specific to that MoE omni checkpoint, both row sets in the supplementary archive). Other-address ratio CIs: InternVL3 $[1.48, 1.88]$, Qwen2.5-7B $[2.65, 8.87]$, Qwen3-8B $[1.67, 2.31]$, Qwen3-VL-8B $[1.81, 3.66]$ exclude 1.0; JoyAI-VL $[0.55, 3.18]$, 30B $[0.82, 6.10]$, 32B $[0.34, 2.37]$, Preview $[0.69, 1.82]$, Mistral $[0.73, 1.83]$, Qwen2.5-Omni $[0.60, 1.17]$, MiniCPM $[0.65, 1.24]$ do not. The pause-below-guard ordering difference excludes 0 in nine of eleven (Mistral $[-0.49, +0.61]$ and 32B $[-0.37, +1.87]$ are the exceptions). Full table: `DIAGCI_RESULT`. In sum, other-addressed questions elicit more speaking than pause requests in all eleven checkpoints (nine intervals exclude zero) and more than neutral contexts in nine of eleven, with ratios from 1.08 to 4.56; the ratio's interval excludes 1 in four, including the trainee base.

Table 9: The eleven checkpoints on identical contexts: speak-propensity ratio to neutral after an explicit pause and when someone else is addressed (PersonaPlex: audio instrument, not counted).

checkpoint	kind	explicit pause	**other addressed**
Qwen2.5-7B-Instruct	text	0.31	**4.56**
Qwen3-8B	text	1.56	**1.94**
Mistral-7B-Instruct	text	1.17	**1.26**
Qwen3-VL-8B	VL	1.50	**2.62**
InternVL3-8B	VL	0.84	**1.70**
Qwen3-VL-32B	VL	0.61	**1.08**
Qwen3-Omni-30B-A3B	streaming/omni	0.49	**2.53**
JoyAI-VL-Interaction	proactive RL	0.21	**1.53**
JoyAI-VL-Preview	proactive RL	0.39	**1.18**
MiniCPM-o-2.6	streaming/omni	0.38	0.95
Qwen2.5-Omni-7B	streaming/omni	0.25	0.87
PersonaPlex-7B[†]	full-duplex speech	0.73	0.86

[†]Measured through its own audio interface, not the token read: behavioural onset within the window over TTS-rendered meetings (`PREREG_PPX` A1; windows capped at the 10 s tail where longer). It is not counted in any "eleven" claim. The ratios are compressed by a near-ceiling baseline—the model speaks in 0.917 of *neutral* windows (single seed, 10 s-capped windows; App. Y's two-seed read gives 0.958), filling any silence regardless of conversational structure (`PPX_A1_RESULT`); all four window CIs overlap, and the guard-above-pause ordering ($0.86 > 0.73$) matches the eleven text checkpoints' direction. The dyadic prior in its plainest behavioural form: any pause is its turn.

K CHECKPOINTS OUTSIDE THE PROBING PROTOCOL

MiniCPM-o-4.5 is unprobeable with our tooling for two recorded reasons: the `transformers` path requires the `minicpmo` package, whose `torchaudio` pin is ABI-incompatible with the training environment's torch, and vLLM 0.11.2's MiniCPM-O implementation asserts `version == (2,6)`. MiniCPM-o-2.6 was therefore recorded as the family representative *before* it was probed (`PREREG_DIAG` Amendment 3) rather than silently substituted. Moshi and PersonaPlex (Défossez et al., 2024; Roy et al., 2026) are audio-native speech-to-speech architectures whose speak/wait decision is not expressible as a completion over a text prompt, and DuplexPO (Li et al., 2026) has released no weights. Diagnosing audio-native duplex models needs a different instrument than ours—so we built one for PersonaPlex (`PREREG_PPX`, frozen before any inference): 24 minimal vocative pairs, TTS meeting audio byte-identical except whether the final question names Ada (the model's role-prompt name) or another party, run through NVIDIA's own offline pipeline at a fixed seed. Role-prompted as a named participant, the model answers questions addressed by name to someone *else* at 0.792 (own name 0.917; paired difference $+0.125$, CI [-0.042, +0.292], n=24; one sampling seed, where the two-seed means of App. Y are 0.771 and 0.833), and in 8 of 8 control items it speaks over the addressed party's own answer (`PPX_RESULT`). One model, one voice prompt, synthetic input off its training distribution. This probe tests addressed-question behavior in a speech interface. The same instrument then measured the model after training through its own text stream (App. Y).

L THE COMPONENT ABLATION

Table 10 lists the component checks of the historical configurations; its advantage-magnitude row carries the dose-preserving flat-versus-graded test (`PREREG_FLAT`, lockbox-14), and App. AE(iv) reconciles it with the other two trainers.

Table 10: Component checks in the historical training configurations. These observations concern the tested implementation and training regimes. They motivate explicit wait targets and neutral anchoring; the controlled study separately evaluates credit construction.

component or design choice	**evidence**
decision-focused training data	historical configurations using per-decision labels reduce guard elevation; the controlled study finds improvements under all four credit constructions on shared training contexts (Section 3.3)
sign-flip supervision	provides corrective targets for negatively credited actions, including waiting when a response is wanted; positive-credit waits remain supervised without this path
soft-KL anchor to base's own $P(\text{speak})$	without: silence collapse; hard BC anchors overshoot (0.065 vs base 0.126); per-example reweighting fails
Bernoulli-sampled decision head	greedy collapses to silence below $P{=}0.5$
conditional / removable	**evidence**
advantage magnitudes (*conditional; condition measured*)	under the shipped mass-normalizing trainer, the flat ±1 clone is indistinguishable from ours on both sealed sets at both levels (matched pools, App. AE(iv))—because that normalizer compresses *both* arms to near-flat; under a dose-preserving trainer the graded shape beats the flat clone on the guard (-0.171 $[-0.266, -0.077]$, $p{=}0.004$, `PREREG_FLAT`), paying on the pause window
advantage-mass normalization	same ablation; needed only under mass imbalance
corrective-row normalization	main effect -0.089 on DEV at 8B; justified only in the 32B regime (39:1653 corrective-row imbalance)

M Inference conventions

Rademacher weights are degenerate at small G (p-floor ≈ 0.125 at $G{=}4$), which is why the wild bootstrap uses Webb's six-point distribution throughout (Webb, 2023), and the cluster-robust variance is the CR1 form ($\frac{G}{G-1} \cdot \frac{n-1}{n-k}$ times the sandwich), not the bias-reduced CR2 of Bell & McCaffrey (2002); with few clusters its validity rests on the Webb wild bootstrap rather than on the variance estimator. Per-arm retention is printed wherever the per-episode ratio appears, and App. AE(i) quantifies how far it differs by arm—the residual endogeneity that §2.3 then removes by construction is a property of the environment, not of one estimator. Where a run or collection is the unit of inference we aggregate first and report n as runs. `speak_rate` is a mediator for behavioural outcomes and a confounder for rating outcomes, so the outcome type determines the covariate set. No p-value is read as evidence for a null; equivalence uses TOST with declared margins (Lakens, 2017; Schuirmann, 1987). **Multiplicity.** Each battery pre-registers its primaries (one or two co-primaries) and is adjudicated on them alone; we make no familywise claim across batteries, and the 26 contrasts of Table 1 are reported as 26 dependent tests, not as a corrected family. For a family-level bound: of the 24 contrasts recomputed at build time, all 24 survive Holm's step-down at $\alpha{=}0.05$ and 19 the Bonferroni threshold $0.05/24{=}0.0021$; the largest Webb p is 0.0073 (`MULTIPLICITY_RESULT`). Secondary outcomes are descriptive and carry no adjustment; self-rejections count against the study, not the family. **Training-run variance.** In this historical comparison, each primary arm is one training run and intervals are scenario-clustered, so seed-to-seed variation of training is absent from the intervals except where measured: `PREREG_SEEDS` evaluated the shipped recipe at seeds 1 and 2 on lockbox-v9, and `PREREG_SEEDS_A1` (frozen before its statistic was computed) resamples scenarios and the three seeds jointly: ours−base -0.345 $[-0.484, -0.230]$, ours−critic -0.190 $[-0.245, -0.132]$, ours−episode return -0.132 $[-0.185, -0.073]$, between-seed SD 0.036 on each (`SEEDS_A1_RESULT`); comparators are single runs whose run variance is unmeasured.

N Per-battery detail for the granularity comparison

The episode- and decision-level accounts behind Table 1 (§3.2) and Figure 3 (§3), in full; Fig. 11 plots the set-9 re-run and JoyAI-VL arms.

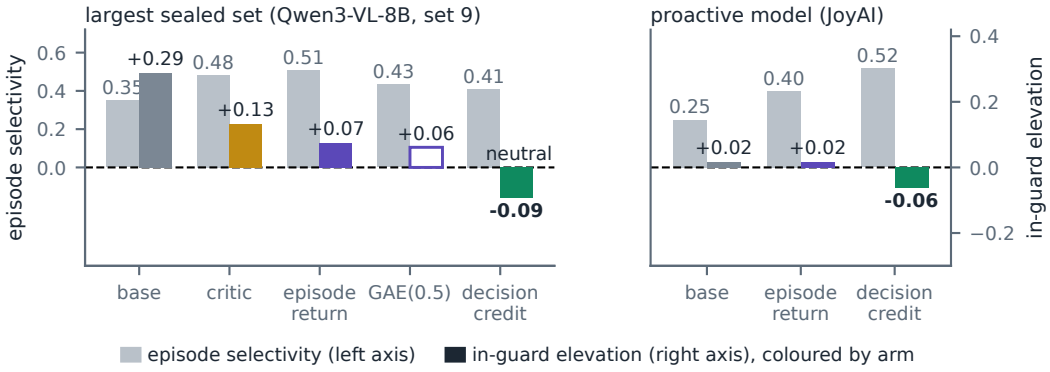

Figure 11: Scoreboards rise; the guard does not follow. In the historical configurations: episode selectivity (grey, left axis) and decision-level in-guard elevation (right axis, coloured by arm) for each arm on the largest sealed set (left; the escape-decode re-run of set 9, the "set 9, re-run" row of Table 1) and on the proactive model's set (right), from the logs Table 1 reads; the panels share both axes. **Every episode-level arm posts a higher scoreboard than the base without bringing the guard below neutral (on JoyAI-VL the episode return leaves it where the base left it,** $+0.02$**); decision credit is the only arm not shown above neutral.**

The earlier sealed sets. On lockbox-v5 ours is the only arm whose guard ratio is not above neutral (0.84 vs critic 1.39, episode return 1.23)—the episode ledger is mixed and v7's verdict was partial. Two later sets froze the selection-free pooled per-tick estimator as *the* primary before they existed: on lockbox-v8 both co-primaries confirm (ours−critic $-0.175\,[-0.250,\ -0.099]$, webb-$p$=0.0045; ours−episode return $-0.138\,[-0.221,\ -0.054]$, p=0.0073; ours−base -0.471) while the episode return posts the best episode aggregate of any arm (0.583 vs 0.483), and the per-episode ratio, a disclosed-bias secondary reported with retention, reads ours 1.03 ($38/60$) against critic 1.39, episode return 1.38 and base 5.20 (30–$33/60$)—ours at neutral, every baseline elevated; on lockbox-v9 (G=24, n=120 per arm) ours−critic $-0.228\,[-0.288,\ -0.168]$ and ours−episode return $-0.159\,[-0.218,\ -0.100]$ (both $p < 10^{-4}$; ours−base -0.367), and two further training seeds of the shipped recipe keep all nine contrasts negative with intervals excluding zero (`SEEDS_RESULT`). The information-symmetric arm (`PREREG_SYM`, `grpo_dense`) is the shipped episode-return construction with the group-normalised return replaced by the per-decision structural credit; both frozen co-primaries confirm along this sequence of configurations—labels $-0.083\,[-0.134,\ -0.032]$ (elevation $+0.050 \to -0.033$), delivery $-0.073\,[-0.126,\ -0.019]$—and it replicates the dissociation internally (better norm, worse scoreboard, 0.408 vs 0.508). The tuned baselines (`PREREG_TUNE`) received an LR sweep selected on DEV in-guard elevation; the sweep moved both off the shipped rate and improved the critic (-0.053, p=0.008), and both primaries survive (ours−tuned critic $-0.172\,[-0.237,\ -0.106]$, ours−tuned episode return $-0.179\,[-0.241,\ -0.117]$).

The GAE arm (`PREREG_GAE`, frozen before its rows existed). Episode-return and Monte-Carlo-critic credit constructions do not by themselves cover temporal credit assignment's standard machinery. Generalised advantage estimation with per-decision rewards r_t (the same structural credit content), the same five-fold cross-fitted ridge V (out-of-fold $R^2 = +0.349$), γ=1, $\delta_t = r_t + V(s_{t+1}) - V(s_t)$ and $A_t = \sum_k \lambda^k \delta_{t+k}$ at λ=0.5 interpolates between the critic arm (λ=1) and the one-step residual; rows pass the same scale and mass normalization and anchor rule as the episode-return and critic arms, with the per-arm row and update counts in Table 4, and the arm is evaluated on lockbox-v9 under the ESC convention against the existing runs (`GAE_RESULT`). In-guard elevation: base $+0.288$, critic $+0.133$, episode return $+0.075$, GAE $+0.061$, ours -0.094. Primary ours−GAE $-0.161\,[-0.215,\ -0.107]$ (webb-p=0.0002) holds; GAE−critic $-0.065\,[-0.115,\ -0.015]$ (p=0.015), GAE−episode return $-0.008\,[-0.051,\ +0.036]$ (p=0.72), GAE−base $-0.215\,[-0.339,\ -0.091]$. Episode selectivity $+0.433$ with violation 0.408 and own-answered 0.842. This offline GAE construction, run at one setting (λ=0.5, γ=1, one seed), improves over the critic but does not resolve an advantage over episode return. The historical configurations also differ in retained rows and training dose. Against the two sequence-level comparators—episode return and a critic-weighted supervised arm whose value function is ridge-regressed on the policy's own hidden states with 5-fold episode-grouped cross-fitting (out-of-fold $R^2 = +0.349$; the pre-registration required positivity for the comparison to count)—the episode-level ledger across four evaluations (lockbox-v4 and v5, each against both

comparators) is mixed: on lockbox-v4 ours beats episode return under the frozen rule (SURVIVES, $+0.200$, p=0.010) and the critic contrast is UNRESOLVED ($+0.183\,[-0.026,\ +0.393]$, p=0.065); on lockbox-v5 both baselines reach within the frozen 0.05 gap of ours ($+0.033$ against the critic, -0.017 against episode return)—the pre-registered verdict is REFUTED, reported in those words. These strong-baseline primaries are distinct from the two sets' author-shift primaries, which compare the same recipe with the base (App. P).

The pre-declared fixed-effect meta-analysis of the critic contrast is $+0.139\,[+0.016,\ +0.261]$; we label it a summary and disclose its fragility: without the development-set estimate (on which the shipped arm was selected, and which its own pre-registration marks non-confirmatory) the combination is $+0.11\,[-0.02,\ +0.24]$ and includes zero; with only three studies the declared z interval becomes $[-0.130,\ +0.407]$ under a $t(2)$ critical value, the convention every input to it uses; and the v4/v5 inputs evaluate the same checkpoints, sharing training-draw noise. The pooled episode-level critic contrast is therefore not established.

Is a linear value head the weak link? `PREREG_MLP`, frozen before the head was built, swaps the ridge for a 2-layer MLP (hidden 256, early stopping, inner-CV on training folds only) and changes nothing else—same features, same target, same episode-grouped cross-fitting, same downstream pipeline. Its own diagnostic gate settles what it is: out-of-fold $R^2 = +0.287$ against the ridge's $+0.349$, so by the pre-registration's words it is *not a more expressive critic in effect*, and we say so rather than claiming we tested one. Under the ledger's estimator it stays above neutral on both sets it was evaluated on (in-guard elevation $+0.038$ on lockbox-v5, $+0.061$ on lockbox-v7) and ours beats it on both ($-0.164\,[-0.218,\ -0.110]$, webb-$p$=0.0004; $-0.159\,[-0.237,\ -0.081]$, p=0.0047); its episode-level guard ratio is above neutral on v5 (1.11) and below on v7 (0.81), the set-dependence episode return also shows. Reported here rather than folded into Table 1, whose ledger is the four-arm one (`MLP_RESULT`).

A sixth sealed set (lockbox-v7, 12 fresh model-authored families, G=12) was authored under a pre-registration freezing the decision-level guard-ratio contrast as *the* primary, with a FAIL branch pre-committed as a seventh self-rejection. Verdict: PARTIAL. The vs-critic contrast resolves as a frozen primary ($-0.451\,[-0.850,\ -0.053]$, webb-$p$=0.027); the vs-episode-return contrast does not ($-0.379\,[-0.803,\ +0.044]$, p=0.066) and acquires its verdict only on lockbox-v8 (the lockbox-v8 account above). On this set ours is again the only arm below neutral (0.84; critic 1.36, episode return 1.19, base 2.70)—the same configuration as lockbox-v5, now on a second naive set, with the vs-critic half carrying a verdict. Figure 17 shows what this contrast is made of in one matched v7 episode: base answers a question directed at someone else at P=0.56 while ours holds the same guard at $P \leq 0.001$ and answers its own address at P=0.99. Episode selectivity on v7 is consistent with the REFUTED episode verdict: the episode return's aggregate (0.650) sits numerically above ours (0.483) while its guard ratio stays above neutral. Retention on this set is $48/38/33/30$ of 60 for ours / episode return / base / critic (App. AE(i)). (App. AE(iii) additionally finds the end-anchored episode scoreboard itself is tilted: the episode return's advantage on it depends on interruptions launched during the question, which that scoring cannot see.)

An eighth sealed set (lockbox-v9, 24 fresh families, G=24, n=120/arm, `PREREG_V9`) froze the same decision-level estimator a third time and, for the first time, the symmetric overlap-anchored selectivity of App. AE(iii) as an episode-level co-primary family. The end-anchored scoreboard meanwhile ranks episode return first a fourth time (0.508) and ours *below base* (0.317 vs 0.325 in the original run; the escape-decode re-run plotted in Figs. 3 and 11 reads 0.408 vs 0.350)—on a set where ours' decision-level contrast against base is -0.367, ours is the only arm below neutral (0.91), and ours retains the most guard-window episodes of any arm ($106/120$). The anchoring convention can invert even the ours-versus-base episode comparison.

O PRE-REGISTERED CONFIRMATIONS ON RESEARCHER-AUTHORED SEALED SETS

Six-batch replication (lockbox-v2), including collection and training variance. Six independent volume-matched collection→credit→training runs (disjoint collection seeds, training seed = batch index), blinded until all finished, rule and permitted wording frozen. Because these are separate *training* runs rather than one checkpoint scored repeatedly, the run can be the unit of inference: with n=6 runs, the decision-level contrast is $-0.438\,[-0.451,\ -0.425]$ and episode selectivity

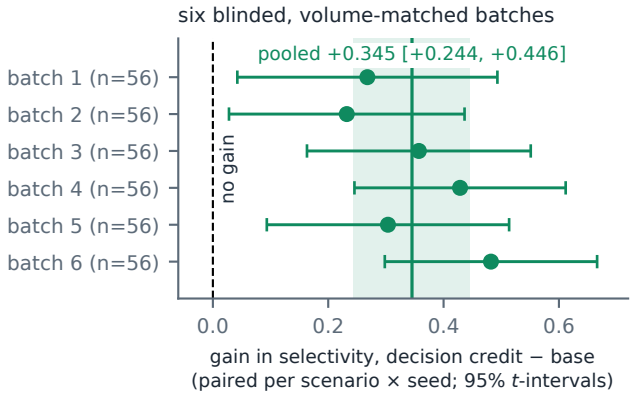

Figure 12: The effect replicates across six blinded batches. Independent volume-matched batches under a frozen rule, each batch's paired mean with its t-interval; the band is the pooled interval. **All six batches move in the same direction, and the pooled interval excludes zero.**

$+0.345$ [$+0.244$, $+0.446$], both $6/6$ runs in the expected direction (`RUNLEVEL_RESULT`); every run's in-guard elevation lands between -0.03 and -0.00 against a base at $+0.42$. That interval carries collection, credit-construction and training-seed variance; the other historical single-run comparisons condition on a trained checkpoint. The later controlled study also includes training-seed uncertainty.

Episode-level detail: Δ selectivity $+0.345$, t-CI [+0.244, +0.446], hierarchical bootstrap [+0.215, +0.463]; all six batch means positive; every leave-one-out interval excludes zero (Fig. 12). The gain comes from suppression: own-address answering does not clear both intervals ($+0.068$). speak_rate passes TOST at δ=0.15 in all six batches. **Coverage (lockbox-v3).** Twelve further families, three conditions fixed in advance, all hold: $0.150 \rightarrow 0.683$, $\beta = +0.533$ [+0.276, +0.791], Webb-p=0.0002; eleven of twelve families improve, none regress; three families with *negative* base selectivity all flip positive. Three disjoint collection seeds (post-hoc extension, fixed n=3) replicate at $\Delta = +0.383$ [+0.259, +0.508]. **Cross-scale (32B, n=3).** $\Delta = +0.643$; the t-interval (df=2) and the hierarchical bootstrap both exclude zero, with the caveats n=3 implies—but the *route* differs: the 8B gains by suppressing, the 32B by activating, at the cost of more pause interruptions. The credit regulates allocation, not volume; its direction is set by whichever way the base is deficient. Tracing the 32B pause regression identified corrective-row imbalance (39:1653) as the mechanism. **Dev ranking does not predict held-out ranking.** The best development cell (0.786) scored 0.583 on unseen families while the pre-registered policy (0.643) scored 0.683: swapping the primary after seeing DEV would have cost 0.10 of confirmed effect.

P THE AUTHOR-SHIFT PAIR: A CONTROLLED EXPERIMENT ABOUT EVALUATION

Are the results an artifact of our own scenario-writing idiom? On twelve families authored by a 32B model with structural validation only, the pre-registered author-shift primary (decision credit against the base, episode selectivity) **failed** ($+0.117$ [-0.172, +0.405]). The mandated post-failure trace found the cause was not that the model-authored moments were less tempting—the *base's* own in-guard propensity is nearly the same on the human-written set and on this one (0.512 vs 0.468; `AUTHORSHIFT_RESULT`)—but the *author's verbosity*: model authors wrote $2.05\times$ longer turns (18.9 vs 9.2 words), floor occupancy cut decidable ticks $30.2 \rightarrow 22.6$, and every episode-level contrast compressed mechanically. Lockbox-v5 enforces the human sets' verbosity envelope and temptation matching in the validator (calibrated on the 50 human-written families first); on it the author shift **passes both frozen primaries** against the base: selectivity $+0.300$ [+0.064, +0.536] (p=0.011) and guard ratio -2.60 [-3.82, -1.39] (p=0.0004), on the most tempting set yet (base guard ratio 3.40). The same two sets also carry the strong-baseline primaries, decision credit against episode return and against the critic; that ledger is in App. N.

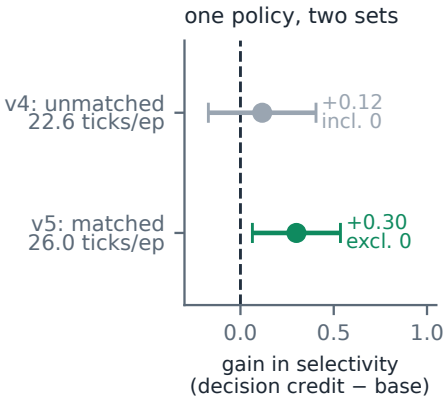

Figure 13: The author-shift pair. The same recipe, its decision-credit arm against the base, scored on two sealed sets: the frozen author-shift primary's interval includes zero on lockbox-v4 and excludes it on lockbox-v5. Row labels give the base's decidable ticks per episode (human-written sets: 30.2). **The sets differ in how they were written (v5 enforces the human verbosity envelope and temptation matching), not in the recipe.**

One regression on this set is reported rather than hidden: pause violations worsen in every trained arm ($0.367 \rightarrow 0.517$–0.600), ours included. The failed and passed tests are reported together (Fig. 13): same recipe—the pair shows that the same recipe can fail and pass an episode-level contrast when set construction (occupancy and temptation matching) differs—scoreboard masking's evaluation-side face: the aggregate can move, or refuse to move, for reasons that have nothing to do with the norm. This artifact entered two of our own pre-registrations before we recognized it as a family.

Q THE TEMPTATION PROBE

`PREREG_TEMPT` (frozen before any probe output) asks which half of temptation matching carries the elevation: the trainee's private fact, or the transcript around it (the peer-addressed question followed by the addressee's hesitation). On the logged decision contexts of the ESC-convention base re-run on lockbox-v9 (session 0 of the escape-decode re-run log of App. AB, a different rollout from the frozen run behind Table 21, so window counts differ by a few ticks; 957 ticks over 24 scenarios: every other-address guard tick, own-address and bad-news ticks capped at four per episode, every pause tick, and the first three neutral ticks from tick 4 on, where the base rarely speaks, so this probe's neutral level and elevation are internal to the T/U/N comparison and not comparable with Table 21's all-tick $+0.251$) the decision head is re-read under three system prompts identical except for one line: the private fact as logged (T), replaced by an irrelevant office fact (U), or removed (N). Positive control: the transformers decision head tracks the served one at Spearman ρ=0.999.

P(speak)	n	T (as logged)	U (irrelevant fact)	N (no fact)
other-address guard	115	0.562	0.665	0.677
own-address	187	0.599	0.724	0.699
pause	66	0.548	0.790	0.695
bad news	229	0.526	0.763	0.684
neutral	360	0.021	0.023	0.047
elevation (guard − neutral)		$+0.541$	$+0.642$	$+0.630$

Scenario-clustered bootstrap (B=9999): the elevation under U is $+0.642$ $[+0.524,\ +0.775]$ (primary, holds); T−U is -0.101 $[-0.151,\ -0.057]$ and T−N -0.089 $[-0.156,\ -0.033]$ (`TEMPT_RESULT`). Holding the answer does not create the violation; if anything the relevant fact makes the base slightly more reticent everywhere. What the base model's other-address elevation (Table 1: $+0.392$ and $+0.251$; $+0.54$ to $+0.64$ on this probe's tick sample under all three fact conditions above) measures is therefore the transcript half—a question to someone else that goes unanswered for a beat—which is the structure the guard is about.

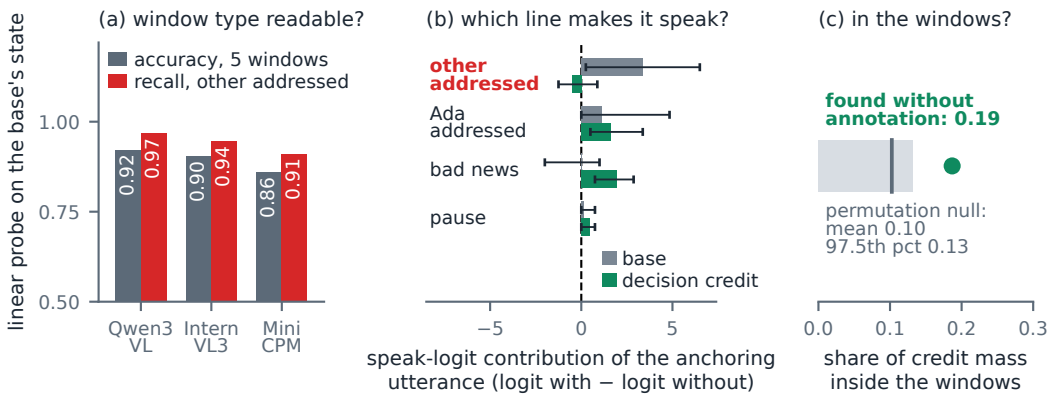

Figure 14: Three descriptive measurements on the base checkpoint. (a) A linear probe on the base's hidden state at the act position recovers the window type on three lineages. (b) Removing the anchoring utterance: the question addressed to someone else raises the base's speak logit and is inert after decision credit. (c) Credit localized by the divergence of two base rollouts, without annotation, lands inside the windows above its permutation null. These describe available information and input sensitivity; the state-transfer study of §4 provides the causal intervention.

R WHERE THE DECISION LIVES, IN FULL

Representation (`PREREG_PROBE3`). Figure 14 summarises the three measurements below. The exploratory 8B probe (`PROBE_REPR_RESULT`: balanced accuracy 0.932 on the E16 collection) was licensed for the paper only by replication. `PREREG_PROBE2` replicated it on InternVL3's training collection and *missed* its recall bar (0.711 on n=45 other-address decisions; accuracy 0.815 passed), so no sentence was made. `PREREG_PROBE3` re-ran the identical construction (five-fold episode-grouped, class-balanced logistic regression on final-layer decision-position states, window labels from the frozen definitions) on each lineage's sealed base logs, where the guard is dense:

lineage (base)	decisions	other n	bal. acc	other recall	guard recall	precision	specificity
Qwen3-VL-8B (v9, both base sessions)	5955	240	0.921	0.967	0.971	0.857	0.992
InternVL3-8B (two sets)	7424	275	0.903	0.945	0.949	0.850	0.992
MiniCPM-o-2.6	4882	201	0.859	0.910	0.960	0.798	0.986

Bars: balanced accuracy $>$ 0.80 (chance 0.20), other-address recall $>$ 0.90; all three pass (`PROBE3_RESULT`). Window type is linearly decodable from the state used for the action prediction. Folds separate episodes, not scenario families, so this result does not establish generalization to unseen scenario structures or knowledge of a social norm.

Attribution. Utterance-level occlusion of the decision head on identical lockbox-v7 contexts, 24 decisions per window per arm (exploratory; the verdict file required a log-odds re-analysis before use, `ATTRIB_LOGODDS_RESULT`). Median contribution of the window's anchoring utterance to the speak logit (logit with the full context minus logit with that utterance removed; positive means the line excites speaking), bootstrap over decisions:

arm	window	Δ logit (with − without)	anchoring line is the top-$\lvert\Delta\rvert$ line
base	other-address	$+3.37\,[+0.25, +6.50]$	0.17
base	own-address	$+1.11\,[+0.00, +4.84]$	0.21
base	pause	$+0.11\,[+0.00, +0.75]$	0.04
base	bad news	$+0.00\,[-2.00, +1.00]$	0.29
ours	other-address	$-0.50\,[-1.25, +0.88]$	0.12
ours	own-address	$+1.63\,[+0.50, +3.37]$	0.21
ours	pause	$+0.44\,[+0.00, +0.75]$	0.08
ours	bad news	$+1.94\,[+0.75, +2.87]$	0.62

Removing the peer-addressed question reduces the base model's speaking logit on average. That line has the largest absolute occlusion effect in only 17% of the sampled base contexts. After training, its interval includes zero. Whole-utterance removal changes several input properties, so this is evidence of input sensitivity rather than an isolated addressee mechanism. Appendix Q separately tests the supplied private fact.

Discoverability (`PREREG_DLA_S2`, amendment A1). The base policy evaluated twice on identical lockbox-v9 scripts (240 matched pairs; 83 with an outcome difference); per matched tick, weight $\propto \Delta R\,(1-\cos(h_t^A, h_t^B))$ with h the base's decision-position state and R the episode's own-answered minus other-violation; the fraction of credit mass inside any annotated window is 0.187 against a within-pair permutation null of 0.103 (97.5th percentile 0.132; excess $+0.084$; gate $+0.05$): CONFIRM (`DLA_S2_RESULT`). The $\Delta R{=}0$ pairs, whose weights come from divergence alone, overlap at 0.263: policies diverge where the floor is free, which is where the windows are, so Stage 3 (`PREREG_DLA_S3`, frozen before any pair existed) trains a divergence-localised arm on training-pool pairs, with no window annotation in its pipeline, against a free-floor occupancy control and against the episode return.

Stage 3 (`PREREG_DLA_S3`). The base policy twice on identical training scripts (240 pairs, 120 with an outcome difference, 2,986 matched ticks; 0.41 of the divergence mass falls inside the unused windows); per matched tick the two runs' decisions receive $\pm\Delta R\, v_t$ with v_t the normalised divergence (DLA) or uniform over free-floor ticks (FREE); the episode-return arm's normalisation, anchors, trainer and epoch count (one epoch; the smaller pool yields 279 and 374 updates, Table 4); lockbox-v9 under the ESC convention (`DLA_S3_RESULT`). In-guard elevation: base $+0.288$, FREE $+0.196$, critic $+0.133$, episode return $+0.075$, DLA $+0.072$, GAE $+0.061$, ours -0.094. P2, DLA$-$FREE $-0.128\,[-0.207,\ -0.049]$ (webb-p=0.004), holds: the localisation is not floor occupancy. P1, DLA$-$episode return $-0.007\,[-0.040,\ +0.026]$ (p=0.65), fails: discovered credit reaches the decision exactly as far as the information-symmetric episode baseline, and DLA$-$ours is $+0.158\,[+0.102,\ +0.214]$. At the episode level DLA ties the episode return for the best selectivity of any arm ($+0.508$ each; violation 0.375 vs 0.400) while its decision-level elevation sits at the episode return's—the dissociation once more. The pre-committed readings named neither this cell (P1 fails with P2 holding and the guard moved -0.215 against base); the discovered place is real and it trains, but not yet as far as naming it. Audit note: after the shared $|w| < 0.05$ drop, DLA keeps 2,968 credit rows of 5,972 candidate decisions (divergence spreads a pair's credit over ~ 25 ticks), FREE 3,988; the drop rule was frozen for all arms and is not revisited here.

S THE IDENTICAL-SUPPORT COMPARISON, IN FULL

Identical support. All arms train on the 1,564 credited decisions and 782 anchors of the decision-credit collection; prompts, recorded actions, and neutral anchor targets are identical across arms (manifest hashes in the supplementary archive), while the effective supervised target depends on the credit sign, and no arm drops or re-filters rows after its credit is assigned. Credits: *decision credit* r_t; *dense* $(r_t-\mu)/\sigma$ within scenario; *episode* $(R-\mu_R)/\sigma_R$ within scenario, assigned to the same rows; *flat* $\mathrm{sign}(r_t)$. Relative to decision credit, dense flips 77 targets (all pause waits), episode flips 607, flat and the sign-preserving variant none (Table 11). Each arm then balances speak-target and wait-target mass and rescales its mean absolute credit to one: a row with target y and credit A_t carries the weight $c_y\,|A_t|$, where $c_y = N/(2M_y)$ and M_y is the total absolute credit of the rows with target y, so the total weight on speak targets equals that on wait targets and the mean weight is one (`build_data.py`). The objective is binary cross-entropy on the renormalised speak/wait mass at the decision position, plus the anchor term of Eq. 1 with λ=0.3, with fixed whole-dataset denominators rather than per-micro-batch ones (the corrected normaliser of App. AE(iv)). LoRA rank 16, α=32, dropout 0.05 on all attention and MLP projections; AdamW, learning rate 10^{-4}, weight decay 0.01, 5% warm-up, gradient clipping 1, effective batch 16, three epochs $= 441$ updates for every arm; five seeds (0–4), with the same initialisation and row order per seed across arms. A fifth credit, *dense (sign-preserving)* r_t/σ, scales without centring and flips no target; it is reported in the tables as a comparator only. Its reported estimates, like the four primary arms', use all five seeds.

The other windows. Episode return keeps the sharpest bad-news selectivity, $3.4\times$ its neutral rate against the base's $1.9\times$; decision credit, dense, and flat answer bad news at about $1.7\times$ their neutral

rate (below the base's ratio) but also speak more in neutral contexts, and they move pause responses by about $+0.03$ (Table 12).

Where the targets flip. Table 11 places the flips (`FLIPS_RESULT`). Under decision credit the target is the window's prescribed action whatever the base did, so a flip trains the row toward the other action. Episode return's 607 flips fall in every window and on both recorded actions (0.388 of the credited rows; 220 targets become speak and 387 become wait), because a scenario-standardised return carries no information about the moment: 82 of the 183 own-address rows and 39 of the 153 other-address rows are trained toward the wrong action, and the arm still matches decision credit on the guard (Table 13). Dense credit's 77 flips are all pause waits (77 of 358): a positive wait credit smaller than its scenario's mean credit becomes negative after centring. Mixed credit flips 382 rows, 124 pause waits and 236 bad-news waits among them, which is where it loses bad-news response (Table 17). A flip-count-matched random-flip control is follow-up work.

Table 11: Where the supervised targets flip relative to decision credit, on the 1,564 credited rows shared by every arm, by window and by the action the base policy recorded on the row. Cells: flipped rows / rows in the cell. Flat and the sign-preserving dense variant flip none. From `FLIPS_RESULT`.

Window	Recorded action	Rows	Episode return	Dense	Mixed
Own address	speak	133	58	0	1
Own address	wait	50	24	0	12
Other address	speak	86	20	0	0
Other address	wait	67	19	0	8
Pause	speak	36	11	0	1
Pause	wait	358	170	77	124
Bad news	speak	108	37	0	0
Bad news	wait	726	268	0	236
Total		1,564	607	77	382

Scoring precision. Under bf16 the decision logit of a context moved by up to 0.25 depending on which other contexts shared its padded batch (a probability discrepancy of 0.057 on the preflight sample; 0.12 with eager attention); every score in this comparison is therefore computed in strict fp32, under which batched and single-context scoring agree to 5×10^{-6}, at 1.15 contexts per second. Training remains bf16. The matched-rate target of §3.4, 0.629, is the base model's own-address answering rate read in bf16; one context of 186 crosses the threshold under this fp32 rescoring, hence the 0.624 of §3.3 and Table 2.

Matched-rate readout. For each arm and seed the speak threshold on the frozen v9 contexts is swept until own-address answering equals the base model's exactly, the context at the threshold counting fractionally (a deterministic split of the tied mass, which is a randomised threshold and the counterpart of the interpolation on the envelope in Appendix AI), and the other-address violation is read at that threshold; the bootstrap resamples scenario families and training seeds together (1,999 draws) and recomputes the base model's target and every arm's threshold inside each replicate, so the matching itself is resampled. The deployment rows freeze the threshold on the development split (v8) at the base model's development answering rate and apply it unchanged to v9. Table 12 gives the per-window rates at the matched rate, Table 13 the pairwise contrasts, Table 15 the deployment rows, and Figure 15 the whole operating curve.

Table 12: Identical support, four credits: per-window speak rates at the base model's own-address answering rate on v9. Every arm, the additional sign-preserving dense arm included, averages five seeds.

Credit	Own ↑	Other ↓	Pause ↓	Bad news ↑	Neutral
Base	0.624	0.548	0.600	0.531	0.287
Episode return	0.624	0.205	0.510	0.669	0.195
Dense	0.624	0.249	0.633	0.868	0.522
Dense (sign-preserving)	0.624	0.245	0.657	0.835	0.512
Flat	0.624	0.228	0.627	0.849	0.495
Decision credit	0.624	0.249	0.627	0.850	0.511

Table 13: Pairwise contrasts against decision credit on other-address violation at the matched rate; scenario×seed bootstrap intervals.

Decision credit minus	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Pooled [95% CI]
Dense	-0.009	+0.043	-0.017	-0.026	+0.009	+0.000 [-0.050, +0.043]
Episode return	+0.000	+0.078	+0.043	+0.087	+0.009	+0.043 [-0.081, +0.160]
Flat	+0.009	+0.035	-0.009	+0.061	+0.009	+0.021 [-0.025, +0.073]

The planned contrasts of Table 13 (protocol fixed before training, after the historical set-9 results) compare the credit constructions with one another. Table 14 adds, as a secondary readout that was not pre-registered, each arm minus the base at the same matched rate under the same scenario×seed bootstrap (1,999 draws, the threshold re-matched inside every replicate; `ARM_BASE_RESULT`). Every construction lowers other-address violation by between 0.299 and 0.343, with intervals that exclude zero in every case (mixed credit by 0.421 $[-0.530, -0.306]$), and among the study's five arms the per-seed contrasts range from -0.252 to -0.417 (mixed credit's from -0.348 to -0.461). The constructions differ in what they do elsewhere: decision credit raises bad-news response by $+0.319$ $[+0.177, +0.437]$ and neutral speaking by $+0.223$ $[+0.089, +0.346]$, and dense and flat behave alike, whereas episode return leaves neutral speaking at the base level (-0.092 $[-0.185, +0.011]$) and raises bad-news response less ($+0.138$ $[-0.007, +0.264]$). Read directly, episode return minus decision credit is -0.316 $[-0.427, -0.184]$ on neutral speaking and -0.181 $[-0.322, -0.052]$ on bad-news response, against -0.043 $[-0.160, +0.081]$ on the guard, so on the guard the ordering is unresolved (four of five seeds favour episode return; no equivalence margin was declared, App. M), and elsewhere they differ.

Table 14: Arm minus base (secondary, not pre-registered) at the base model's own-address answering rate (0.624) on the v9 test contexts; scenario×seed bootstrap 95% intervals (1,999 draws), five seeds per arm, mixed credit included. Negative is less speaking than the base. Each estimate has its 95% interval beneath it; the last block is episode return minus decision credit under the same bootstrap. From `ARM_BASE_RESULT`.

Arm − base	Other address ↓	Pause ↓	Bad news ↑	Neutral
Decision credit	-0.299	$+0.027$	$+0.319$	$+0.223$
	$[-0.416, -0.183]$	$[-0.109, +0.140]$	$[+0.177, +0.437]$	$[+0.089, +0.346]$
Dense	-0.299	$+0.033$	$+0.337$	$+0.234$
	$[-0.413, -0.184]$	$[-0.108, +0.161]$	$[+0.195, +0.453]$	$[+0.092, +0.356]$
Dense (sign-preserving)	-0.303	$+0.057$	$+0.304$	$+0.224$
	$[-0.427, -0.185]$	$[-0.083, +0.182]$	$[+0.171, +0.425]$	$[+0.091, +0.347]$
Episode return	-0.343	-0.090	$+0.138$	-0.092
	$[-0.458, -0.219]$	$[-0.225, +0.039]$	$[-0.007, +0.264]$	$[-0.185, +0.011]$
Flat	-0.320	$+0.027$	$+0.318$	$+0.208$
	$[-0.443, -0.198]$	$[-0.150, +0.143]$	$[+0.180, +0.439]$	$[+0.059, +0.328]$
Mixed credit	-0.421	-0.120	$+0.175$	-0.040
	$[-0.530, -0.306]$	$[-0.262, +0.084]$	$[+0.065, +0.306]$	$[-0.123, +0.109]$
Episode return − decision credit				
	-0.043	-0.117	-0.181	-0.316
	$[-0.160, +0.081]$	$[-0.258, +0.037]$	$[-0.322, -0.052]$	$[-0.427, -0.184]$

Deployment rows. At thresholds frozen on development data, the four arms cut the violation rate from 0.539 to 0.144–0.165 (Table 15); own-address answering then moves from 0.597 to 0.452–0.581, episode return answering most at decision credit's violation rate.

Table 15: Deployment rows. The threshold is frozen on the development split and applied unchanged to v9.

Credit (dev-frozen threshold)	Own	Other	Pause	Bad news	Neutral
Base	0.597	0.539	0.583	0.522	0.271
Episode return	0.581	0.160	0.480	0.619	0.167
Dense	0.452	0.144	0.443	0.754	0.332
Dense (sign-preserving)	0.500	0.181	0.490	0.772	0.383
Flat	0.506	0.165	0.443	0.788	0.353
Decision credit	0.486	0.160	0.433	0.765	0.351

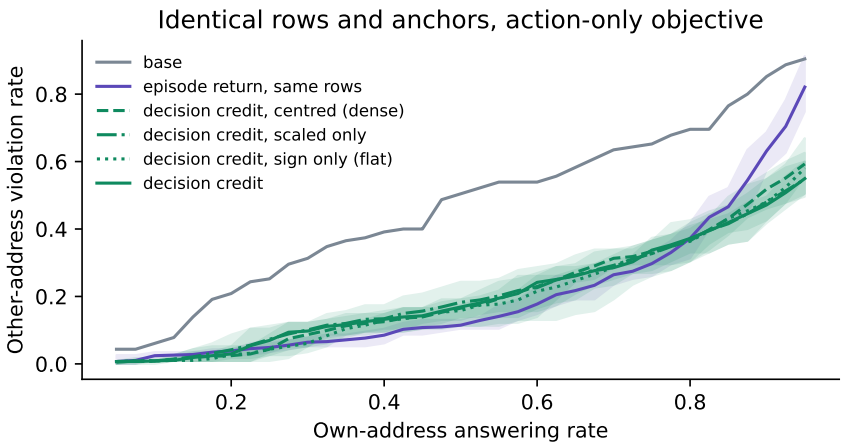

Figure 15: The whole operating curve, identical rows. Other-address violation against own-address answering as one speak threshold is swept per arm on v9; lines are seed means, bands the seed range. All trained arms improve the addressee trade-off over the base. The primary matched-rate contrasts do not resolve differences between decision credit and dense, flat or episode return (Table 13; the sign-preserving dense arm has no planned contrast). Every arm has five seeds.

Two threshold-free readings, not pre-registered, summarise Figure 15 beyond the single matched point (`CONTROLLED_AUROC_RESULT`; seed means, with the seed range in parentheses and no interval). The own-versus-other AUROC of the decision logit is 0.531 for the base and 0.786 (seed range 0.769 to 0.822) for decision credit, 0.782 for dense, 0.784 for the sign-preserving variant, 0.792 for flat and 0.777 (0.762 to 0.795) for episode return, so the four constructions rank own-addressed above other-addressed questions equally well and mixed credit ranks them better (0.863, seed range 0.814 to 0.893). The mean other-address violation over own-address answering rates from 0.4 to 0.8, the area under the matched-rate curve divided by the width of the band, is 0.553 for the base, 0.234 (0.185 to 0.254) for decision credit, 0.236 for dense, 0.239 for the sign-preserving variant, 0.226 for flat, 0.194 (0.175 to 0.221) for episode return and 0.127 for mixed credit; across the band episode return therefore sits lowest with a seed range that overlaps decision credit's, and the ordering that the matched point leaves unresolved is small and consistent rather than a property of one operating point. Above an own-address rate of about 0.8 episode return crosses above the decision-credit arms (Fig. 15); every arm's seed mean stays below the base at every rate.

T UNTRAINED BASELINES ON THE SAME CONTEXTS

Two policies need no training: the base model instructed with the norm, and a rule that answers only when last addressed by name. Both were frozen before scoring (`PREREG_INFER`) and are read on the 2,884 contexts of Table 2 with its fp32 readout (`INFER_RESULT`). The *prompted base* is the base checkpoint with one sentence appended to its behaviour instructions: "Only answer a question when it is addressed to you, Ada, by name. When a question is addressed to someone else, stay silent and let them answer, even if you know the answer." The *vocative rule* R speaks if and only if the most recent utterance by another speaker contains the vocative "Ada" and Ada has not spoken since; R2 also requires that utterance to end with a question mark. Table 16 gives every window.

The instruction makes the model quieter, not more selective. At $p \geq 0.5$ the appended sentence lowers speaking in every window, own-addressed contexts included (0.624 to 0.323). Matched to the base model's own-address answering rate, its other-address violation is 0.557 against 0.548 (difference $+0.009$, $[-0.064, 0.093]$, 1,999 scenario-family draws), and it speaks more than the base in neutral and bad-news contexts; its own-versus-other AUROC on these contexts is 0.530 against the base's 0.531. The pre-registered reading for this outcome is that the prompt does not move the guard; the four trained arms reach 0.205–0.249 at the same rate.

The rule is exact where it fires and silent elsewhere. R never speaks over an addressee but answers only 0.167 of own-addressed polls: in 155 of the 186 own-addressed contexts another participant has spoken after the question to Ada, often mid-sentence, so the most recent utterance carries no vocative

(RULE_DIAG_RESULT). R also speaks in 0.483 of pause polls; R2 answers 0.108 of own-addressed polls and nothing else. At the rules' answering rates the trained arms' other-address violation is 0.012–0.033 (at 0.167) and 0.009–0.024 (at 0.108), against the base's 0.165 and 0.070: training brings the model close to the rule's guard at the rule's operating point and, unlike the rule, can be operated at any answering rate (Fig. 4a).

Table 16: Untrained baselines: per-window speak rates on the 2,884 shared contexts. Upper block: the base and the prompted base at $p \geq 0.5$, and the two rules' own decisions. Lower block: the base and the prompted base, each with one speak threshold set so that own-address answering equals the base model's.

operating point	policy	own address	other address	pause	bad news	neutral
$p \geq 0.5$	base	0.624	0.548	0.600	0.531	0.289
	prompted base	0.323	0.296	0.233	0.354	0.170
rule	R (vocative)	0.167	0.000	0.483	0.040	0.088
	R2 (vocative + ?)	0.108	0.000	0.000	0.000	0.004
own address = 0.624	base	0.624	0.548	0.600	0.531	0.287
	prompted base	0.624	0.557	0.617	0.708	0.455

A four-times larger model with the same prompt. Amendment A1 repeats the prompted base on Qwen3-VL-32B-Instruct in bf16, with the 32B base as its own reference (INFER32_RESULT). The 32B base is quiet and, at $p \geq 0.5$, not tempted: it answers 0.317 of own-addressed questions and speaks in 0.070 of other-addressed against 0.110 of neutral contexts. The sentence does not move its guard: held at the 32B base's own-address rate of 0.317, the prompted model's other-address violation is 0.061 against 0.070 (-0.009 [-0.049, $+0.030$]), and at the 8B base's rate of 0.624 both read 0.139. The pre-committed threshold reading (P32 ≤ 0.249) is therefore met by the 32B base's own selectivity rather than by the prompt: at 0.624 the untrained 32B base (read in bf16, so its logits carry the bf16 quantisation that App. S measures, unlike the fp32 8B scores) violates the guard less often than the four primary 8B arms (0.205 to 0.249), though not than mixed credit (0.127). Model scale and decision-level training are two different levers on this simulator; an appended norm is neither. Its set-9 own-versus-other AUROC is 0.831 [0.750, 0.900] and the matched other-address rate 0.139 has interval [0.053, 0.252] (INFER32_AUROC_RESULT). On the 41 addressee-swap pairs (SWAP32_RESULT; bf16, contexts scored singly, exploratory) the untrained 32B shows a paired own − other gap of 0.570 [0.432, 0.707] (0.548 after renaming), ranks the own-addressed member first in 39 of 41 pairs, reaches an AUROC of 0.910 on the 82 contexts, and at its matched threshold answers the own-addressed member and waits on the other in 0.610 of pairs [0.463, 0.756], between the 8B base (0.366) and the trained decision-level arms (0.610 to 0.693); its mean own-addressed speaking probability is only 0.634 because it mostly waits when the original addressee is still replying (0.348 on the 23 continuing pairs against 0.999 on the 18 clean ones). **Two credit variants aimed at the remaining costs.** Decision credit holds the guard but speaks more than the base in neutral contexts (0.511 against 0.287) and does not improve pause handling; episode return keeps neutral speaking low but answers less after bad news. PREREG_MIX, frozen before either variant's dataset existed, tests two changes on exactly the controlled study's rows: a stronger neutral anchor ($\lambda = 1$ instead of 0.3, nothing else changed) and a *mixed* credit that adds decision credit and episode return at equal mean magnitude (which flips 382 of the 1,564 targets relative to decision credit). Its pre-committed reading asks whether a variant keeps the guard (other-address violation at the matched rate at most 0.03 above decision credit's 0.249) and removes the neutral cost (neutral speaking at least 0.10 below 0.511). The $\lambda = 1$ anchor (seeds 0, 1, 2) reaches 0.249 other-addressed, 0.600 pause, 0.847 bad-news, and 0.475 neutral speaking at the matched rate: it keeps the guard and gives no relief of the neutral cost. Mixed credit (seeds 0, 1, 2, 3, 4) reaches 0.127 other-addressed, 0.480 pause, 0.706 bad-news, and 0.248 neutral speaking at the matched rate: it keeps the guard and removes the neutral cost. With scenario families and the seeds of both arms resampled together (1,999 draws; five seeds of each arm; MIX_CONTRAST_RESULT), mixed credit minus decision credit is -0.122 [-0.218, -0.031] on other-address violation and -0.263 [-0.344, -0.144] on neutral speaking, at a cost of -0.144 [-0.209, -0.043] on bad-news response, whereas its differences from episode return are unresolved on every window (other address -0.078 [-0.184, $+0.008$], neutral $+0.053$ [-0.029, $+0.185$], bad news $+0.037$ [-0.055, $+0.170$]); this reading was not pre-registered, and it places mixed credit between the two constructions rather than above both, keeping episode return's low neutral rate while giving up part of decision credit's bad-news response. Table 17 gives every

window (VARIANTS_RESULT); the runs are single seeds where the amendments say so, and the intervals resample scenario families only.

Table 17: Two credit variants on the identical rows, at the base model's own-address answering rate on the v9 test contexts. Brackets: variant minus base, scenario-cluster bootstrap 95% intervals; the five-seed means of decision credit and episode return are repeated from Table 12.

arm (seed)	other addr. ↓	pause ↓	bad news ↑	neutral
base	0.548	0.600	0.531	0.287
decision credit, five-seed mean	0.249	0.627	0.850	0.511
episode return, five-seed mean	0.205	0.510	0.669	0.195
decision credit, $\lambda = 1$ (0)	0.209 $[-0.50, -0.25]$	0.600 $[-0.17, +0.14]$	0.827 $[+0.11, +0.38]$	0.458 $[-0.00, +0.24]$
decision credit, $\lambda = 1$ (1)	0.261 $[-0.43, -0.20]$	0.550 $[-0.20, +0.11]$	0.858 $[+0.16, +0.41]$	0.430 $[-0.01, +0.22]$
decision credit, $\lambda = 1$ (2)	0.278 $[-0.41, -0.18]$	0.650 $[-0.11, +0.17]$	0.854 $[+0.16, +0.42]$	0.537 $[+0.09, +0.32]$
mixed credit (0)	0.096 $[-0.60, -0.37]$	0.500 $[-0.27, +0.09]$	0.708 $[+0.01, +0.30]$	0.236 $[-0.20, +0.06]$
mixed credit (1)	0.139 $[-0.57, -0.33]$	0.550 $[-0.20, +0.10]$	0.730 $[+0.03, +0.31]$	0.279 $[-0.16, +0.08]$
mixed credit (2)	0.087 $[-0.58, -0.36]$	0.417 $[-0.43, +0.03]$	0.730 $[+0.03, +0.32]$	0.279 $[-0.19, +0.19]$
mixed credit (3)	0.200 $[-0.48, -0.26]$	0.500 $[-0.28, +0.10]$	0.699 $[+0.01, +0.29]$	0.263 $[-0.18, +0.08]$
mixed credit (4)	0.113 $[-0.56, -0.34]$	0.433 $[-0.31, +0.00]$	0.664 $[-0.02, +0.27]$	0.182 $[-0.26, -0.03]$

U BIDIRECTIONAL STATE PATCHING, IN FULL

The development selection is separate from the test comparison; the test contexts come from an existing evaluation set, and the full test-layer sweep is exploratory. **Protocol.** For each held-out context we run the base checkpoint (Qwen3-VL-8B) and the decision-credit adapter on identical inputs, capture the residual state at the decision position after decoder blocks $\{4, 8, \ldots, 32\}$ of 36 (never the final block, whose replacement would replace the prediction itself), and replace only that position in the other model, leaving every transcript token's state untouched. The development split (v8; 112 contexts, at most two per scenario and window by a fixed hash) selects the block by the gain in own-minus-other separation from trained→base patching, which is largest at block 32; the held-out split (v9; 227 contexts, 24 scenario families) is scored once at that block with five controls: the reverse patch, a self patch (numerical identity; error 0), the mean development-split state difference added as a fixed vector, a norm-matched random vector, and an output logit bias fitted on the development split to the patched other-address level (-7.73) and then frozen. Address-only pairs change the vocative of the triggering question from Ben or Cleo to Ada, leaving every other line unchanged (41 pairs whose question line is unique at its timestamp and opens with a single Ben or Cleo vocative; in 23 the original addressee's reply, marked `[still speaking]`, follows in both versions; in these continuing pairs the own-addressed version has the original addressee already replying, and we report the 18 clean pairs as the primary stratum), and a globally renamed copy (Ada/Ben/Cleo → Mira/Noah/Iris) repeats the comparison (Table 19). Scenario families are the bootstrap unit (B=9999). The patched base's separation (0.207) and the adapter's own (0.206) differ by $+0.000$ with a scenario-family bootstrap 95% interval of $[-0.040, +0.046]$ (B=1999, seed 2917; $[-0.043, +0.044]$ at B=9999), so the transferred state reproduces the adapter's separation to within 0.046 in either direction (EQUIV_RESULT); no equivalence margin was declared for this diagnostic, and the interval is reported as the bound rather than as a verdict. Every score is computed in strict fp32: under bf16 a context's decision logit moved by up to 0.25 depending on which other contexts shared its padded batch (a probability discrepancy of 0.057 on the preflight sample), whereas in fp32 batched and single-context scoring agree to 5×10^{-6}. The dev-selected block is the deepest candidate, so the layer curve is reported on v9 for every candidate block (Fig. 5a) and the result establishes an effect at late sampled blocks without identifying a unique critical layer. The patch replaces only the decision position, so earlier adapter changes elsewhere, read later through attention, are not excluded. Table 18 lists the per-window means of every intervention. This follow-up study uses existing v8/v9 contexts; it is not a newly collected evaluation set. A donor patch needs the trained model's state for each input; it is a causal diagnostic, not a deployable method.

Table 18: Mean $P(\text{speak})$ per window on held-out v9 at block 32 for every intervention. Separation (own − other) of the trained-state → base patch against the output bias $0.310\,[0.160,\ 0.466]$, the mean state difference $0.228\,[0.078,\ 0.373]$ and the random direction $0.203\,[0.039,\ 0.361]$ (scenario bootstrap, B=9999); a self patch reproduces the base exactly. Threshold-free, the own-vs-other AUROC of $P(\text{speak})$ is 0.472 for the base, 0.472 under the output bias, 0.475 under the mean shift, 0.472 under the random vector, 0.687 with the trained state ($+0.214$ over the base, $[0.124,\ 0.300]$) and 0.691 for the adapter itself; the full contrasts and intervals are in the analysis files of the supplementary archive.

Intervention	Own ↑	Other ↓	Pause ↓	Bad news ↑
Base	0.675	0.678	0.671	0.637
Full adapter	0.489	0.283	0.541	0.766
Output bias	0.267	0.371	0.254	0.268
Trained state → base	0.511	0.305	0.615	0.811
Base state → trained	0.608	0.620	0.605	0.586
Mean state difference	0.620	0.641	0.593	0.586
Random direction	0.699	0.694	0.707	0.664

Addressee-swap pairs. The 41 pairs come from 22 scenario families, and Table 19 gives every intervention on them: the paired own-minus-other gap and the share of pairs in which the own-addressed version receives the higher speaking probability (concordance), with scenario-family bootstrap intervals (B=9999). In probability space the bias's negative gap is a saturation effect of the constant shift. The base separates these minimal pairs (0.308) although its heterogeneous test-set gap is -0.003, so a bare vocative to Ada is a cue the base partly reads once wording and topic are held fixed; the output bias, which shifts every logit by the same constant, changes the gap only through saturation, and with the original names it stays at the base level (0.334, concordance 0.854). The trained state raises the gap to $0.633\ [0.506,\ 0.754]$ with perfect concordance, numerically above the adapter's own $0.571\ [0.462,\ 0.679]$ (overlapping intervals); the patch places the adapter's decision state into a model whose other positions are unadapted. Renaming every participant lowers each trained gap by about 0.06 and leaves the ordering and the concordance unchanged.

Table 19: Addressee-swap pairs (n=41 pairs, 22 scenario families). Paired own-minus-other gap in $P(\text{speak})$ and concordance (share of pairs ranked own first), with scenario-family bootstrap 95% intervals (B=9999); original names and every participant renamed Mira/Noah/Iris. From `AUROC_RESULT`.

	Original names		Renamed	
Intervention	gap [95% CI]	concordance [95% CI]	gap [95% CI]	concordance [95% CI]
Base	$0.308\ [0.145,\ 0.481]$	$0.854\ [0.756,\ 0.950]$	$0.309\ [0.155,\ 0.476]$	$0.854\ [0.756,\ 0.950]$
Base with output bias	$0.334\ [0.225,\ 0.459]$	$0.854\ [0.750,\ 0.950]$	$0.408\ [0.281,\ 0.537]$	$0.854\ [0.756,\ 0.950]$
Base carrying the trained state	$0.633\ [0.506,\ 0.754]$	$1.000\ [1.000,\ 1.000]$	$0.574\ [0.443,\ 0.700]$	$1.000\ [1.000,\ 1.000]$
Full adapter	$0.571\ [0.462,\ 0.679]$	$1.000\ [1.000,\ 1.000]$	$0.509\ [0.397,\ 0.627]$	$1.000\ [1.000,\ 1.000]$

Replication on the controlled adapters. Section 4 patched the seed-0 decision-credit adapter of the sequence-level protocol. Repeating both directions with the controlled study's adapters at blocks 20, 24, 28 and 32 on the same 227 contexts (no development selection; not pre-registered; `CONTROLLED_PATCH_RESULT`), the five decision-credit seeds give a separation gain over the base of $+0.027$ at block 20, $+0.162$ at 24, $+0.185$ at 28 and $+0.205\ [+0.015,\ +0.408]$ at 32 when the trained state enters the base, and a change of $-0.225\ [-0.380,\ -0.062]$ at block 32 when the base state enters the adapter (seed × scenario bootstrap, 5 seeds); the episode-return seed-0 adapter gives $+0.278\ [+0.121,\ +0.435]$ and $-0.229\ [-0.351,\ -0.101]$ at block 32. The late-block transfer of Section 4 therefore replicates across five decision-credit seeds, and an episode-return adapter's late state transfers a heterogeneous gap as well; what separates the constructions is whether the state follows the name, on pairs that hold everything else fixed: on the 41 addressee-swap pairs (`CONTROLLED_SWAP_RESULT`; seed × family bootstrap over each arm's five seeds), the paired own-minus-other gap with the original names is decision credit $+0.702\ [+0.585,\ +0.815]$, dense $+0.695\ [+0.578,\ +0.796]$, flat $+0.733\ [+0.615,\ +0.843]$, episode return $+0.216\ [+0.157,\ +0.280]$, against the base's $+0.308\ [+0.146,\ +0.486]$ (concordance: decision credit, dense and flat rank the own-addressed member first in all 205 seed×pairs, episode return in 197 of 205, the base in 35 of 41); with every participant renamed it is decision credit $+0.672\ [+0.552,\ +0.778]$, dense $+0.669$

[+0.546, +0.776], flat +0.689 [+0.567, +0.801], episode return +0.220 [+0.166, +0.283], base +0.309. Mixed credit, scored on the same pairs afterwards (MIXED_SWAP_RESULT), widens the gap to 0.737 [0.610, 0.854] (0.721 renamed) with concordance 0.995. At each arm's matched set-9 threshold (SWAP_MATCHED_RESULT; exploratory), the share of pairs on which an arm answers the own-addressed member and waits on the other is 0.649 for decision credit, 0.610 for dense, 0.693 for flat, 0.366 for the base and 0.351 for episode return (decision credit − episode return +0.298 [+0.100, +0.484]; episode return − base −0.015 [−0.195, +0.157]; on the clean pairs 0.889, 0.478 and 0.500); on the swap contexts the own-versus-other AUROC is 0.961 for decision credit and 0.766 for episode return (base 0.792), and with own answering held at 0.624 on the swap set alone decision credit answers the other-addressed member in 0.000 of pairs and episode return in 0.259. Stratified by whether anyone is speaking at the poll (CONTROLLED_SWAP_STRATA_RESULT; not pre-registered): on the 18 pairs where nobody is speaking, decision credit, dense and flat give the own-addressed version a mean speaking probability of 0.985 and episode return 0.488 (base 0.999), with $P(\text{other})$ 0.056, 0.083 and 0.048 against 0.215 (base 0.495); on the 23 continuing pairs the decision-level gaps stay large (0.525 to 0.573) while episode return's is 0.171 (base 0.155); over all 41 pairs the decision credit minus episode return paired gap is +0.487 [+0.374, +0.594], and on the 18 clean pairs, the primary stratum, +0.656 [+0.548, +0.756] (seed × family bootstrap). On pairs that hold wording, topic, position and time fixed, the three decision-level constructions more than double the base's separation, whereas episode return does not raise it (point estimate below the base's, unresolved): the addressee itself is what decision-level supervision teaches.

V On top of a proactive model trained with per-second RL

A demanding case: does *addressee-conditioned* credit add anything to a model that already learns when-to-speak per second with RL? Applying decision credit unchanged to JoyAI-VL: the first run's (lockbox-v3) frozen primary failed by 0.0161 at one collection batch while its guard-ratio secondary passed cleanly ($1.33 \to 0.47$ against a fixed 1.0 threshold)—a pattern consistent with low power, which the pre-registered replication then tested. The pre-registered replication (lockbox-v6; three independent collections, fresh model-authored sealed set, rule frozen before the set existed) confirms both primaries: pooled selectivity +0.206 [+0.045, +0.367] (p=0.017), guard ratio −0.400 [-0.670, -0.130] (p=0.002); every batch delta positive (Fig. 16). The batch-level t-interval alone (n=3) does not exclude zero and the claim rests on the pooled frozen primaries. On these terse families JoyAI-VL's base shows no elevated guard ratio, and the gains run through activation (own-address answering $0.433 \to 0.628$; bad-news response $0.500 \to 0.911$) while pause interruptions worsen—the base-deficit rule a third time. Does an *episode-level learner* on JoyAI-VL exhibit the same dissociation as on the 8B trainee? A pre-registered episode return baseline (PREREG_J3) answers: three episode return adapters trained on the *identical* collections with the 8B-identical construction—using episode-return credit and its associated row-selection pipeline—evaluated on a fresh sealed set (lockbox-10, G=12). Both frozen co-primaries confirm. Decision level: ours−episode return −0.080 [−0.122, −0.038] (webb-p=0.0026). Episode level, under the symmetric overlap-anchored instrument of App. AE(iii): +0.433 [+0.185, +0.681] (p=0.0020; ours 0.767 vs episode return 0.333)—on this trainee the arm separation is large enough for even the fair aggregate to resolve. It is also the cleanest case of an aggregate rising through answering alone: the episode return on JoyAI-VL leaves the decision-level guard *where base left it* (+0.004 [−0.038, +0.045], p=0.85) while improving the episode aggregates (own-address answering 0.650 vs base 0.400; end-anchored selectivity 0.400 vs 0.250)—in this configuration the episode-return arm improves the scoreboard with no detectable change in the decision-level guard, on the model class where timing is the native objective. Ours is again the only arm below neutral (guard ratio 0.63; base 1.71, episode return 1.05; retention $49/51/44$ of 60; the base's guard ratio is set-dependent, 1.33 on lockbox-v3, 0.81 on lockbox-v6 and 1.71 on lockbox-10, as it is for the 8B trainee across its sets, App. N), answers more (0.633), responds to every bad news (1.000), and worsens pause violations ($0.533 \to 0.667$)—the base-deficit rule, disclosed as elsewhere.

W The recipe on two further trainee lineages

Does any of this depend on the Qwen3-VL trainee? We applied the recipe *unchanged* to InternVL3 (Zhu et al., 2025) under a pre-registration frozen before any collection episode existed

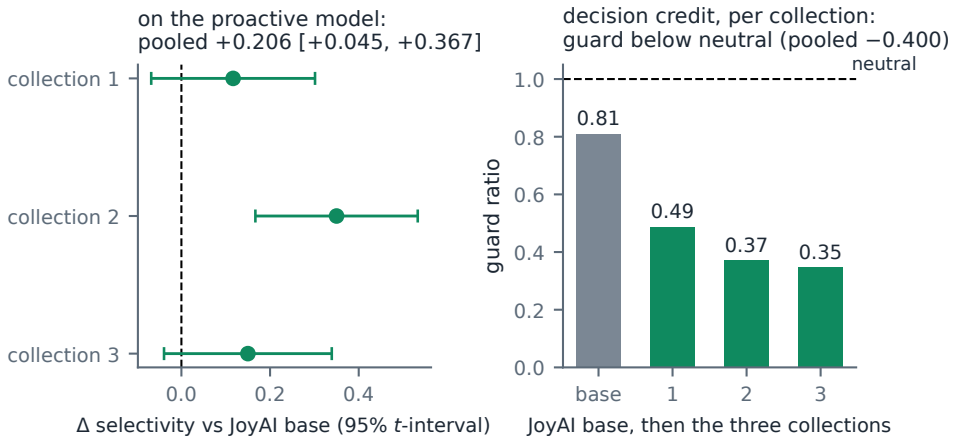

Figure 16: JoyAI-VL replication (lockbox-v6). Three independent collections; both pooled frozen primaries pass. Every collection's point estimate raises selectivity (the pooled interval $[+0.045,\ +0.367]$ excludes zero; two single-collection intervals do not), and decision credit lowers the guard ratio below the base's in each.

(PREREG_IVL); the trainee was selected by exploratory behavioural profiles run first (InternVL3 violates in 0.383 of episodes with only a weak in-rollout guard ratio of 1.07—a deficit of a different *shape*: talkativeness inside guards rather than addressee-specific elevation). Scope disclosed in advance: InternVL3's text backbone is Qwen2.5, so this tests a second independently post-trained *lineage*, not a second text family. Same collections feed a decision-credit arm and an episode-return arm using different credit and row-selection constructions; same trainer, same hyperparameters; fresh sealed set (G=12). **Three of four frozen primaries confirm**: decision-level ours−base $-0.224\,[-0.281,\ -0.167]$ and ours−episode return $-0.135\,[-0.172,\ -0.097]$ (both webb-p=0.0002), and overlap-anchored episode selectivity vs base $+0.267\,[+0.102,\ +0.432]$ (p=0.0012); the episode contrast vs episode return is directionally positive but unresolved ($+0.183\,[-0.041,\ +0.408]$, p=0.084) and per the pre-registration is claimed as nothing more. The signature repeats: violations halve ($0.350 \to 0.167$) while speak_rate *rises* ($0.261 \to 0.315$); guard ratio $1.33 \to 0.48$; pause violations worsen ($0.483 \to 0.617$)—the base-deficit rule a fourth time. One heterogeneity: episode return *partially* moves this trainee's guard (-0.078, p=0.025), unlike on JoyAI-VL (no detectable movement) or the Qwen3-VL trainee (above neutral)—how far episode-level credit moves the norm varies by base (it lowers elevation on some sets and leaves it at the base on others, Table 1), while moment-level credit moves it decisively on all three.

The same battery design then ran on **MiniCPM-o** (Yao et al., 2024), the streaming/omni lineage representative (text backbone Qwen2.5-7B (Yang et al., 2024), as for InternVL3) and, on paper, the case with the least to repair: its base is *propensity-light* (behavioural guard ratio 1.21; violations 0.167, the lowest of any base) and its episode aggregates suggest nothing needs fixing (overlap selectivity 0.600, numerically above every trained arm). The decision level disagrees, and the frozen primaries resolve it (PREREG_MCPM): ours−base $-0.237\,[-0.305,\ -0.169]$ (p=0.0006), ours−episode return $-0.173\,[-0.218,\ -0.128]$ ($p < 10^{-4}$); the episode-level B family fails on this base in *both* directions (-0.067 vs base; $+0.133$ vs episode return; neither resolves) and per the pre-registration changes no claim. Ours ends with the deepest guard suppression of any battery ($1.21 \to 0.31$) while speaking 58% *more* ($0.196 \to 0.309$) and lifting bad-news response $0.583 \to 0.950$—on the quietest base, the aggregate cannot even see what needs repair, and the decision level repairs it anyway. The vs-episode-return decision-level contrast then carried five pre-registered verdicts across four trainees, and carries six across five with the text-only trainee of App. X; the episode return's own reach remains heterogeneous (-0.065 here, p=0.004).

Late critic arms (PREREG_CRIT2). On MiniCPM-o the cross-fitted value head fits the policy's own hidden states at out-of-fold $R^2 = +0.012$, in effect a centred return-to-go comparator; that critic arm's decision-level contrast against base is $+0.000\,[-0.034,\ +0.035]$, and decision credit beats it by $-0.238\,[-0.290,\ -0.186]$ (Table 1). The InternVL3 critic arm was voided by its pre-registered gate; JoyAI-VL and Qwen3-8B have no critic arm; these three cells read — (no valid arm) in the ledger.

X A TEXT-ONLY LLM TRAINEE

Every other trainee in the main text is a vision-language or speech model; PREREG_LLM, frozen before the model scored a context, repeats the controlled study's key contrast on Qwen3-8B (Yang et al., 2025), the text-only model of the same family and size. Its tokenizer is identical to Qwen3-VL-8B's, so the decision readout, the prompts, the 1,564 credited rows and 782 anchors, the trainer and the fp32 scoring are unchanged; thinking is disabled in the chat template, and a preflight confirmed that batched and single scoring agree to 3×10^{-4} in the decision logit (LLM_RESULT).

The text-only base shows the same temptation. At $p \geq 0.5$ the text-only base speaks in 0.913 of other-addressed contexts against 0.533 of neutral ones (own address 0.968, pause 1.000, bad news 0.863), so the pre-registered diagnosis replicates; this base speaks far more often than the vision-language base in every window, which sets the matched operating point at an answering rate of 0.968.

The training result replicates on the text-only trainee. Table 20 holds every arm at this base's own-address answering rate (0.968). *Decision credit* (seeds 0, 2) moves other-addressed responses from 0.913 to 0.452 at that rate, a change of -0.461 (seed 0: -0.504 $[-0.653, -0.310]$; seed 2: -0.417 $[-0.605, -0.149]$), which the pre-committed reading calls "replicates"; at the same threshold it speaks in 0.779 of neutral, 0.942 of pause, and 1.000 of bad-news contexts (base 0.524, 1.000, 0.863). Two readouts that were not pre-registered put the same arm beside Table 2: at the vision-language base's own-address rate of 0.624 its other-address violation is 0.165 against this base's 0.374, and its own-versus-other AUROC is 0.847 against 0.655. *Episode return* (seeds 0, 1) moves other-addressed responses from 0.913 to 0.778 at that rate, a change of -0.135 (seed 0: -0.070 $[-0.196, +0.092]$; seed 1: -0.200 $[-0.333, +0.059]$), which under the decision-credit rule would count as "replicates in part"; at the same threshold it speaks in 0.531 of neutral, 0.908 of pause, and 0.991 of bad-news contexts (base 0.524, 1.000, 0.863). Two readouts that were not pre-registered put the same arm beside Table 2: at the vision-language base's own-address rate of 0.624 its other-address violation is 0.174 against this base's 0.374, and its own-versus-other AUROC is 0.790 against 0.655. Each arm is one or two training seeds (the run set was set by GPU capacity, as recorded in the pre-registration's amendment A4), so seed variance is not estimated and the intervals, which resample scenario families only, understate it.

Table 20: Text-only Qwen3-8B on the frozen test contexts. Speak rates per window; the base at $p \geq 0.5$ and, below, every arm with one speak threshold set so that own-address answering equals the base's rate. Brackets: arm minus base, scenario-cluster bootstrap 95% intervals.

Qwen3-8B (text)	own addr.	other addr. ↓	pause ↓	bad news ↑	neutral
base, $p \geq 0.5$	0.968	0.913	1.000	0.863	0.533
base, matched	0.968	0.913	1.000	0.863	0.524
episode return (0), matched	0.968	0.843 $[-0.20, +0.09]$	0.900 $[-0.19, -0.01]$	0.996 $[+0.02, +0.27]$	0.611 $[-0.13, +0.25]$
episode return (1), matched	0.968	0.713 $[-0.33, +0.06]$	0.917 $[-0.15, +0.00]$	0.987 $[+0.01, +0.25]$	0.451 $[-0.22, +0.23]$
decision credit (0), matched	0.968	0.409 $[-0.65, -0.31]$	0.933 $[-0.14, +0.00]$	1.000 $[+0.02, +0.28]$	0.782 $[+0.03, +0.43]$
decision credit (2), matched	0.968	0.496 $[-0.60, -0.15]$	0.950 $[-0.13, +0.00]$	1.000 $[+0.02, +0.28]$	0.776 $[+0.00, +0.42]$

The full recipe on the text-only trainee. PREREG_LLMSEQ, frozen before any of its rows existed, ran the unchanged Table 1 pipeline on Qwen3-8B: own rollouts against the JoyAI-VL peers, the two credit constructions, the sequence-level trainer, and 120 live episodes per condition on lockbox-v9, one seed per arm and no critic arm. The pre-registered decision-level contrasts (selection-free FWL, CR1 + Webb) are -0.549 $[-0.675, -0.423]$ against the base (confirmed, Webb $p < 0.001$) and -0.222 $[-0.316, -0.128]$ against episode return (confirmed, $p < 0.001$); episode return against the base is -0.326 $[-0.389, -0.262]$. The episode-level co-primaries with symmetric overlap anchoring read 0.217 $[+0.091, +0.342]$ against the base and 0.117 $[-0.038, +0.272]$ against episode return (confirmed, not confirmed; the pre-registration's power caveat applies). The Table 1 result therefore holds on a trainee with no vision or speech component. The episode metrics are in LLMSEQ_BATTERY_RESULT. Four further training seeds per arm (amendments A1 to A3, each frozen before any replication seed's evaluation was read (A2 and A3 after seed 0's verdict); same rows, trainer and evaluation) give seed 1: -0.510 $[-0.628, -0.393]$ against the

base (confirmed) and -0.167 $[-0.239, -0.095]$ against episode return (confirmed) at the decision level, 0.167 $[+0.050, +0.283]$ and 0.025 $[-0.120, +0.170]$ at the episode level; seed 2: -0.601 $[-0.703, -0.500]$ against the base (confirmed) and -0.278 $[-0.349, -0.207]$ against episode return (confirmed) at the decision level, 0.225 $[+0.107, +0.343]$ and 0.108 $[+0.014, +0.203]$ at the episode level; seed 3: -0.499 $[-0.602, -0.397]$ against the base (confirmed) and -0.133 $[-0.219, -0.048]$ against episode return (confirmed) at the decision level, 0.075 $[-0.011, +0.161]$ and -0.117 $[-0.225, -0.009]$ at the episode level; seed 4: -0.690 $[-0.784, -0.596]$ against the base (confirmed) and -0.382 $[-0.467, -0.297]$ against episode return (confirmed) at the decision level, 0.325 $[+0.192, +0.458]$ and 0.150 $[+0.006, +0.294]$ at the episode level (`LLMSEQ_S1_RESULT`, `LLMSEQ_S2_RESULT`, `LLMSEQ_S3_RESULT`, `LLMSEQ_S4_RESULT`); Table 1 keeps seed 0. Across the 5 training seeds the decision-level contrast averages -0.570 (seed SD 0.078) against the base and -0.237 (SD 0.098) against episode return, every seed confirming both; the across-seed summary is exploratory.

Y PersonaPlex as a trainee

Instrument and trainee. PersonaPlex-7B is Moshi's architecture: a 7B temporal transformer over seventeen 12.5 Hz streams (its own text stream, eight of its own audio codebooks, eight of the user's), a depth transformer for the audio codebooks, and a voice prompt replayed as stored embeddings. The decision we train is the text stream's choice between the padding token and a word. A persistent runner reproduces NVIDIA's offline pipeline exactly (gate G1: the four archived items of App. K re-generated bit-for-bit; two RNG-consumption differences had to be found first—the offline script seeds *before* building its two audio codecs on the GPU, and a CUDA-graph capture consumes the generator differently from a replay, so the runner rebuilds that sequence per item). A teacher-forced trainer rebuilds the delayed stream layout from the recorded rollout; the recorded per-step $\log p(\mathrm{PAD})$ pins the alignment (gate G2: median $|\Delta|$ of 0.001 nats at the derived layout against about 1.5 at either one-frame shift, and every one-frame displacement of the delayed streams worse in the decision range), and a freshly attached zero adapter is the base model exactly (gate G3). LoRA rank 16 on the fused q/k/v, output and gating projections of all 32 layers; AdamW 10^{-4}, three epochs, eight rollouts per step, seed 0; a 0.1-weighted cross-entropy on the base's own tokens anchors both arms. Evaluation merges the adapter into the weights and runs the untouched compiled path, so base and arms share one sampling code path.

Pool, rollouts, credit. 130 fresh scenarios from the sealed sets' own generator and validator (topics checked disjoint from all 168 used before; 118 train, 12 held out as DEV), rendered with the frozen edge-tts voices (facilitator en-US-JennyNeural, Ben en-US-GuyNeural, Cleo en-US-AriaNeural; PersonaPlex speaks through its NATF0 voice prompt), Ada never in the input; after each anchored utterance the gap equals the frozen window length so Ada's decision window is silent. $K{=}4$ rollouts per render at the frozen sampling settings. Decisions: speak for the question that names Ada, wait for the question naming a peer, the peer's thinking pause and the personal bad news (the frozen plan scores bad news as a wait for this trainee, where the text trainees have a response window), read by the frozen rule (≥ 3 word tokens in the window); $c_d{=}\pm 1$; return $R{=}\sum_d c_d$. Frame losses: positive weight, cross-entropy of the taken tokens; negative weight, $-\log p(\mathrm{PAD})$ if the model spoke and $-\log(1-p(\mathrm{PAD}))$ (unlikelihood of silence) if it stayed silent; each decision contributes $|w_d|$ times its window mean. MOMENT uses $w_d{=}c_d$; EPISODE uses the GRPO-normalised R of the rollout within its scenario group, on the same decision frames.

The first battery and its audit. Under `PREREG_PPX_TRAIN` the primary failed: MOMENT$-$BASE on the sealed other-addressed items -0.104 $[-0.250, +0.042]$, while the own-name floor held (0.885), pause and bad-news speaking fell to 0.000, neutral speaking from 0.958 to 0.271, and the DEV return reached $+3.83$ of 4 (`PPX_TRAIN_RESULT`). The pre-committed rule was to audit before reading. G1–G3 re-ran clean on the pool; the failure was in the pool: the generator opens 129 of 130 scripts with the peer-addressed question and places the Ada-addressed one third in 125, whereas the sealed items ask at utterance 3–4 and none of the 24 is a script's other-address moment. Position separated every credited window without the name; on DEV the decision-credit arm read other 0.00 at utterance 0 and own 0.92 at utterance 2, and on the sealed items it answered mid-conversation questions whichever name they carried.

Amendment A1: the name as the only cue. Frozen after that analysis and before any new render (PREREG_PPX_TRAIN_A1): every scenario rendered twice, as written and with the two vocatives swapped, and every mid-conversation leading-vocative question to a peer added as a wait decision—own-addressed questions at utterances 0 and 2 (129 and 125), peer-addressed at 0, 2 and 3 (129, 125, 175). Nothing else changed: 944 rollouts, 354 optimiser steps per arm, gates re-run (same numbers), the base's sealed runs reused (the runner is deterministic per item and seed).

	other	own (pairs)	own (A1)	pause	bad news	neutral	barge-in
n items	24	24	24	24	24	72	8
BASE	0.771	0.833	0.812	0.771	0.792	0.958	1.000
EPISODE	0.771	0.792	0.875	0.333	0.646	0.896	1.000
MOMENT	**0.542**	0.979	1.000	0.062	0.062	0.542	0.750

$P(\text{speak})$ per window on the sealed instruments, item means over two sampling seeds (42424242, 42424243) of the one trained adapter per arm; seed 42424242 alone gives App. K's single-seed base readings (other 0.792, own 0.917) (PPX_TRAIN_A1_RESULT). Other minus neutral: BASE -0.187, EPISODE -0.125, MOMENT 0.000. P1: MOMENT$-$BASE on other $-0.229\,[-0.396,\,-0.062]$ (bootstrap over items, B=9999); P2: own pooled $0.990\,[0.969,\,1.000]$; S1: EPISODE$-$BASE $+0.000\,[-0.167,\,+0.167]$; S2: MOMENT$-$EPISODE $-0.229\,[-0.396,\,-0.062]$. DEV return: base -2.42, episode -1.00, moment $+4.58$ of a maximum 4.72; on DEV the moment arm waits at every peer-addressed question (0.00 at utterances 0, 2 and 3) and answers every Ada-addressed one (1.00 at utterances 0 and 2)—the separation follows the name, not the position. Read with its limits: one model, one voice prompt, synthetic TTS input; open-loop evaluation, with decisions read from the text stream (≥ 3 word tokens, so short backchannels do not count); duplex ability, latency and general capability not measured; the sealed items come from different scenario families with two-to-four-turn prefixes, and there the violation falls by 0.229 (about 30%) rather than to zero (0.542, equal to the arm's neutral speaking rate, against $1.85\times$ neutral for its own name, own (A1) 1.000 over neutral 0.542); the barge-in controls move from $8/8$ to $6/8$. The pre-committed reading is the one the file carries: the ladder separates on a full-duplex speech trainee, and the episode arm reproduces, in behavioural form, the historical Qwen3-VL pattern of Table 1; the identical-support comparison (§3.3) does not attribute that pattern to summation.

Z THE THIRD NORM, IN FULL

The third norm is audience-conditioned disclosure: hold a private fact while a non-entitled party is present and share it once the audience changes, which requires remembering conversational state. Per-arm decomposition on the sealed C set (G=12, n=60/arm): $P(\text{speak} \mid \text{hold})$ base 0.791, critic 0.345, episode return 0.327, ours 0.384; neutral $0.226/0.333/0.170/0.173$; share $0.493/0.443/0.439/0.567$. Hold elevations $+0.565/+0.012/+0.157/+0.212$—the critic is the best arm on this norm's guard. The three frozen primaries all fail: C1, hold elevation ours$-$episode return $+0.038\,[-0.049,\,+0.124]$ and ours$-$critic $+0.204\,[+0.107,\,+0.300]$ (the critic is better); C2, share elevation ours$-$base $+0.128\,[-0.054,\,+0.310]$. Against the base, ours lowers hold elevation by 0.349 ($[-0.499,\,-0.199]$), a secondary. The neutral-pool hygiene sensitivity (dropping 6 s after each audience change) moves nothing by more than 0.004. When arms do speak inside the hold window, mean lexical overlap with the private fact is 0.21–0.25 for every arm including base, so the violations are floor-taking rather than targeted disclosure, and the norm's failure is not a disclosure-rate artifact. Episode level: hold violation $0.317/0.217/0.350/0.400$, share response $0.733/0.833/0.783/0.733$—the critic wins both. All numbers from C_RESULT; design and frozen readings in PREREG_C.

AA THE SECOND NORM, IN FULL

Per-arm outcomes on the sealed U set (mean $\pm$ scenario-clustered bootstrap SE, computed from the evaluation logs), followed by the pre-committed cross-norm transfer measurement. Verdict: PARTIAL. The three decision-level primaries resolve (ours$-$episode return $-0.072\,[-0.137,\,-0.007]$, webb-$p$=0.030; ours$-$critic $-0.148\,[-0.216,\,-0.080]$, p=0.0006; ours$-$base $-0.362\,[-0.584,\,-0.140]$,

p=0.0007), while the episode co-primary, selectivity against base, misses (+0.267 [−0.015, +0.549], p=0.054; U_RESULT).

	base	**ours**	episode return	critic
speaks when it should (critical) ↑	0.73 ±0.08	**0.85** ±0.05	0.82 ±0.06	0.77 ±0.03
speaks when it should not (low-stakes) ↓	0.33 ±0.09	0.18 ±0.05	**0.12** ±0.04	0.38 ±0.05
episode selectivity ↑	0.40 ±0.12	0.67 ±0.08	**0.70** ±0.08	0.38 ±0.07
speak rate	0.20 ±0.02	0.29 ±0.02	0.21 ±0.01	0.30 ±0.01
in-guard elevation (decision level)	+0.206	**−0.181**	−0.114	−0.038

Cross-norm transfer, measured as pre-committed. PREREG_U declared one further measurement in advance: the U-trained policies evaluated on the *addressee* sealed set, reported whichever way it landed—the worry being interference, since fixing one non-action norm had repeatedly perturbed another (App. AC). The result is the opposite of interference. All three U-trained arms are *more* addressee-protective than base at the decision level (ours −0.292, p=0.004; episode return −0.261, p=0.0008; critic −0.234, p=0.012), despite never seeing an addressee guard in training—and the ours arm does it while talking *more* than base (0.433 vs 0.395), so for that arm the transfer is conditional, not global suppression (the episode return and critic arms speak less, so volume accounts for part of theirs). This refines, rather than contradicts, the window-specificity result of App. AH: within the addressee family, removing a window leaves its behaviour untrained; but *u_wait* and *other_address* share an abstract structure—defer when another party is the appropriate handler—and credit placed on one instance of that structure reaches the other. What transfers appears to be the deference relation, not the window.

AB UTTERANCE RENDERING: THE ESCAPE-DECODE FIX AND ITS RE-RUN

Generated utterances were stored with non-ASCII characters as literal \uXXXX escapes that the reader never decoded, so they surfaced verbatim in transcripts (the first symptom). This is one bug with two disagreeing halves—an ASCII-escaping producer and a non-decoding consumer—and no assertion in this repository would have caught it. It has a second symptom: because an escape costs six characters where the glyph costs one, the utterance call's token budget can be exhausted *inside* one, leaving a dangling fragment ("…that\u20") that no decoder can resolve—4.9% of ours' utterances, 2.9% critic's, 2.3% the episode return's, none of base's. Both decoders now strip a trailing incomplete escape.

We had in fact *seen* the first symptom: one figure script carried a local regex that decoded the escapes before plotting a quote. The symptom was patched where it was visible and the cause was never traced, which is the failure mode worth naming—a workaround at the display layer removes the evidence that anything is wrong upstream. An audit of the rest of the train/eval interface found no second instance of the same class: no training target contains a quote character that `stop=['"']` could truncate on, and the act-token priming matches the target prefixes exactly.

We also checked what the defect can and cannot have reached among the results. The rating audit of §3.1 (the 516-episode measurement, together with the three counterfactual distillation protocols kept in the accompanying artifacts) draws 236 of its 420 regression episodes from early trained adapters whose logged utterances do carry literal escapes (the 184 base episodes carry none); its source fixed effects absorb a source-wide shift (JUDGE_AUDIT_RESULT). What the defect can reach is anything a fine-tuned arm's transcript feeds: Ada's own later decisions within a rollout, the peers, and the peer ratings of App. H, which fall under every trained arm—which is why PREREG_ESC makes those ratings a named secondary rather than leaving the question open. Under the fix they do not move (ESC_RESULT, after minus before, 120 episodes per arm): base +0.136 [−0.117, +0.389], decision credit −0.001 [−0.458, +0.455], episode return −0.068 [−0.502, +0.367], critic −0.053 [−0.489, +0.384]; the re-run detects no change in the rating drop under the trained arms (decision credit −0.001 [−0.458, +0.455]), a non-detection with wide intervals rather than an exclusion.

The defect also reached the human-validation package (Table 3; instrument and analysis frozen, collection not yet run), and there it was not cosmetic but a break in the blinding: on the pairwise form the fine-tuned transcript carried visible escapes and the base transcript did not, in 30/30 of Study B's

original items (amendment A1 has since extended it to 60 pairs), so an annotator could have sorted the arms without reading a word. The generator now decodes every transcript before it reaches a human; the unblinding key is byte-identical to the pre-fix package, so the items and their assignment are unchanged and only the rendering differs. The study had not yet been run.

AC PAUSE-GUARD RESCALING: TWO VARIANTS AND THEIR OUTCOMES

We tested the naive form under a pre-registration (`PREREG_PFIX`: W_{pause} doubled to -1.5, everything else frozen at e14's values, fresh sealed set). **It failed its primary** (pause violations $+0.150\,[-0.065,\ +0.365]$ against e14), and the mandated code audit locates the miss in the pipeline rather than the mechanism: the row-level mass balancing reshaped the intended $2.0\times$ pause dose into $1.45\times$ while cutting the addressee guard's corrective mass to $0.73\times$ (anti-window mass is conserved, so boosting one guard deflates the others), and the micro-batch normalizer of App. AE(iv) compresses realized ratios further. Decision-level pause propensity confirms the dose never arrived (-0.266 vs e14's -0.263 against base—indistinguishable), while the episode-level worsening tracks the arm's higher speak rate—the volume mechanism this section describes, acting on its own would-be fix. The addressee primary survived the 27% dose cut (-0.293, p=0.0004), consistent with the pool-scoped magnitude ablation.

The table below gives, for the eight historical configurations (settings) behind the pause regression noted in §5, the mean change in pause violation, split by whether the method's net effect on that configuration was to activate or to suppress speaking; the two pre-registered fixes follow.

the method's net effect	settings	mean Δ pause violation
activates (base under-answers; Δspeak > 0)	6	$+0.114$
suppresses (base over-speaks; Δspeak ≤ 0)	2	-0.125

The second test (`PREREG_DOSE`) corrected both pipeline layers that the first test's audit identified—the per-micro-batch CE denominator became a dataset-global constant, and the $2\times$ pause dose was applied after the build so the conserved anti-mass rebalancing could not deflate the addressee guard. The build gate verified the delivered ratio at exactly $\times 2.000$ with addressee mass untouched, and the two arms' training-time global scales (wbar 0.4570 vs 0.5437) sit in exactly the dosed-to-shipped total-mass ratio, reconfirming delivery from inside the trainer. The dose arrived; the pause propensity did not move ($+0.089$, p=0.13), and episode pause violations worsened (0.517 vs 0.333). The normalizer-only control arm sharpens the mechanism: with magnitudes delivered at shipped values, the addressee guard *strengthens* (-0.119, p=0.0045 vs e14) while pause *worsens* ($+0.102$, p=0.041)—delivered credit redistributes the same speech volume toward the strongly-guarded window's complement. Full numbers in `DOSE_RESULT`. Rebalancing guard magnitudes therefore requires the corrected normalizer first; App. AE(iv) now has two independent demonstrations.

AD THE EFFECT IS NOT AN ARTIFACT OF THE DECISION RATE

The engine polls the trainee every 2 s. Re-evaluating base and ours at 1 Hz and 0.25 Hz keeps the sign of the decision-level contrast at every rate, at about half its training-rate size at 1 Hz ($-0.423\,[-0.567,\ -0.279]$ at 0.25 Hz, $-0.418\,[-0.550,\ -0.286]$ at the 0.5 Hz training rate, $-0.209\,[-0.397,\ -0.021]$ at 1 Hz; `TICK_RESULT`). Level shifts across rates are mechanical occupancy effects—decidable ticks per episode rise $16.1 \to 26.0 \to 37.3$—so only the within-rate contrast is interpreted.

AE ROBUSTNESS OF THE DECISION-LEVEL ANALYSES

A post-hoc audit of the analysis code surfaced three properties of the frozen estimators; (iv) adds a caveat on the magnitude ablation and (v) a pre-registered window-length check. All frozen verdicts stand as adjudicated under the estimators their pre-registrations named; what follows is disclosure plus sensitivity, never substitution.

(i) The guard-ratio sample is policy-selected. An episode enters the per-episode estimator only if some poll fell inside the 6 s guard window, and a talkative policy suppresses its own polls while speaking. Retention differs by arm: on lockbox-v7, $48/60$ (ours), $38/60$ (episode return), $33/60$ (base), $30/60$ (critic); on lockbox-v5, $47/38/44/38$; on lockbox-v8, $38/30/33/30$; on lockbox-v9 the direction inverts ($106/68/76/82$ of 120—that set's base is the chattiest arm), so the selection is policy-dependent in sign, not merely in size. This is the same floor-occupancy mechanism the author-shift pair exposes (App. P), now inside our own primary. A fixed-state re-scoring settles what it costs: with every arm scored on the same $2{,}884$ logged contexts (the shared-context scoring of §2.3; the verdict file of `PREREG_FIX`), ours$-$critic is $-0.230\,[-0.291,\ -0.170]$ and ours$-$episode return $-0.190\,[-0.254,\ -0.126]$, within 0.03 of the rollout-scored lockbox-v9 re-run contrasts (-0.226, -0.164; Table 1), so retention differences are a real defect of the estimator that turns out not to carry the result. **(ii) A selection-free re-estimate agrees, and more sharply.** Regressing $P(\text{speak})$ on in-guard versus neutral over *all* episodes with scenario-clustered CR1 gives, for the difference in in-guard elevation, $-0.198\,[-0.283,\ -0.112]$ (ours vs critic) and $-0.157\,[-0.240,\ -0.074]$ (ours vs episode return) on lockbox-v7, and $-0.198\,[-0.252,\ -0.144]$ and $-0.159\,[-0.217,\ -0.102]$ on lockbox-v5—four intervals excluding zero, with effect sizes replicating to two decimals across two independently authored sets (the critic pair to three). We deliberately do *not* promote this to the primary: it was computed after seeing that the frozen estimator left the vs-episode-return contrast unresolved, and swapping estimators at that point is the failure mode this paper's whole protocol exists to prevent. It is reported as a robustness check that agrees in direction, and a pre-registered replication with this estimator as the frozen primary is the correct way to convert it into a verdict—lockbox-v8 (`PREREG_V8`, App. N) has since run exactly that replication, and both co-primaries confirmed. **(iii) End-anchored windows have an unscored zone—on both sides, and it is not used equally.** The frozen episode scoring counts only utterances that *start after* the trigger utterance ends: speech launched *during* the other-addressed question is an uncounted violation, and an answer launched during one's *own* question is an uncounted response. The zone is asymmetrically occupied: scoring it raises the episode return's guard violations most ($0.317 \to 0.650$ on v8, the largest jump of any arm; base $0.633 \to 0.750$ on v5), while ours has the largest share of uncounted own-address answers (0.217 of v8 episodes). Re-anchoring *both* window types symmetrically, ours' overlap selectivity leads on v5 (0.517 vs episode return 0.383, critic 0.350) and v8 (0.417 vs 0.300, 0.333), trails episode return on v7 (0.400 vs 0.500), and leads pooled over all three naive sets (n=180/arm, G=37): ours 0.444, episode return 0.394, critic 0.344, base 0.278—the end-anchored pooled ranking, which places episode return first and ours third, *inverts* under the symmetric instrument, though the pooled ours$-$episode return contrast ($+0.050\,[-0.076,\ +0.176]$) includes zero and is reported as such (ours$-$critic $+0.100\,[-0.019,\ +0.219]$, p=0.085). The REFUTED end-anchored episode verdicts stand as adjudicated; the sharpened reading is that the aggregate's *sign* depends on an anchoring convention, and the convention that favours the scoreboard leader is the one with the blind zone—scoreboard masking extends to the scoreboard's own definition. Lockbox-v9 subsequently froze the symmetric instrument as an episode-level primary family on a fresh set (App. N); it did not resolve there, and per its pre-registration that outcome changes no claim—the corrected aggregate is a fairer instrument, not a sharper one. Where the arm separation is larger it does resolve: the JoyAI-VL battery confirms it as a frozen co-primary (App. V).

(iv) A caveat on the magnitude ablation. The training loss normalizes by weight mass within each two-example micro-batch rather than across the optimizer step, which compresses the intended $8{:}1$ advantage-magnitude range to roughly $1.4{:}1$ in realized gradient share. Table 10's mass-normalised-trainer finding that a flat ± 1 clone matches the full machinery is therefore partly mechanical: the implementation was already close to flat. We report the ablation as run and flag that a corrected normalizer would need it re-run before "magnitudes do not matter" could be asserted of the method rather than of this implementation.

A contrast of $+0.017\,[-0.168,\ +0.201]$ once supported that row and was described as a second pool; it is not used, because it compared a clone built from one collection (164 episodes) against an arm trained on another (180 episodes), so it varied the training data as well as the weighting and was never a test of magnitudes; it was also on lockbox-v3, the same set as the figure it was contrasted with, not a second one. The matched rebuild exists precisely because of this—`L2_ruleBC` is the flat clone of *method-v2's own* pool—and only that pair was evaluated on two sealed sets, so it is the only pair we report. On held-out lockbox-v4 the clone and ours are indistinguishable at the decision level ($-0.022\,[-0.123,\ +0.080]$, webb-$p$=0.66; negative would favour ours) and at the episode level ($+0.100\,[-0.118,\ +0.318]$, p=0.32; positive would favour ours); on lockbox-v3,

the declared secondary whose families the policy had seen, likewise ($+0.028\,[-0.000,\ +0.055]$, p=0.051, and $-0.067\,[-0.250,\ +0.117]$, p=0.41). Both arms move well away from base (on v4, ruleBC−base $-0.148\,[-0.291,\ -0.005]$; ours−base $-0.163\,[-0.274,\ -0.051]$). The reading we can defend is therefore narrow and stated as such: *within* moment-level credit, under the mass-normalising trainer, this implementation's graded magnitudes buy nothing detectable over a flat ±1 rule, on either set, at either level—the contribution is the moment specification. The re-run this caveat called for was made: once the trainer preserves the dose, the graded shape beats the flat clone on the guard ($-0.171\,[-0.266,\ -0.077]$, p=0.004, `PREREG_FLAT`, Table 10), paying on the pause window; under the action-only objective with the corrected normaliser (App. S), sign-only and graded credit are not resolved at the matched answering rate ($+0.021\,[-0.025,\ +0.073]$, decision credit minus flat). Three trainers, three readings, one statement: magnitudes matter to the extent the trainer delivers them and the outcome is the guard window, and the ingredient that holds under every trainer is the moment-level sign. That is a claim about magnitudes, not about the paper's thesis: `ruleBC` is itself a moment-level learner; the historical comparison against *episode*-level learners given the same reward content is App. N, and under identical support the five-seed study leaves that ordering unresolved (§3.3). `RULEBC_MATCHED` records the pairing rules and names the comparison we could not make—`e14_W` versus `L_ruleBC`, the other matched pair, has no sealed set in common. The PFIX battery (App. AC) independently confirms the mechanism: an intended $2\times$ guard rescale was reshaped by these same layers before reaching the optimizer.

(v) Guard-window duration (`PREREG_WINDUR`, frozen before any re-scored number existed). The frozen other-address guard is 6 s. Re-scoring the lockbox-v9 logs under the ESC convention with the guard set to 2, 3 and 4 s (polls are 2 s apart, so 2 s admits one poll per window; the other windows are unchanged) gives, per arm, the in-guard elevation and the contrasts against decision credit (scenario-clustered CR1, Webb wild bootstrap, B=9999; `WINDUR_RESULT`):

	in-guard elevation					ours − comparator		
guard	base	critic	episode return	GAE	ours	base	critic	episode return
2 s	+0.210	+0.077	+0.054	+0.008	−0.212	−0.415 [−0.571, −0.259]	−0.280 [−0.320, −0.239]	−0.256 [−0.303, −0.209]
3 s	+0.225	+0.082	+0.062	+0.025	−0.178	−0.399 [−0.545, −0.254]	−0.254 [−0.311, −0.197]	−0.232 [−0.273, −0.191]
4 s	+0.256	+0.099	+0.062	+0.036	−0.168	−0.414 [−0.546, −0.281]	−0.262 [−0.320, −0.203]	−0.225 [−0.273, −0.178]
6 s	+0.288	+0.133	+0.075	+0.061	−0.094	−0.371 [−0.502, −0.241]	−0.226 [−0.279, −0.174]	−0.164 [−0.216, −0.113]

Every interval excludes zero at every length (ours−GAE likewise: -0.223, -0.206, -0.209, -0.161); the contrasts are largest at 2 s and smallest at 6 s (strictly monotone only against episode return), and the 6 s row reproduces the published numbers exactly. The pre-committed reading: the shortest length at which all three primary contrasts hold is the shortest length tested.

AF AMENDMENT REGISTRY

The following amendments were recorded: thirteen entries during the historical campaign (follow-on pre-registrations such as `PREREG_M2`/`M3` and amendments, the three `PREREG_DIAG` amendments counted separately), one in the capability battery, and fourteen during the follow-up studies (`PREREG_PPX` A1, `PREREG_INFER` A1, `PREREG_MIX` A1–A5, `PREREG_LLM` A1–A4, `PREREG_LLMSEQ` A1–A3). The rule they all obey: an amendment is made *before* the data it governs exists, states what moves and what may not, and never deletes the original. The registry, dated: `PREREG_lockbox` (2026-08-09, before any lockbox episode existed; the file records the legitimacy check); `PREREG_M2` and `PREREG_M3` (2026-08-11, follow-on designs with the M1/M2 verdicts standing and reported regardless); `PREREG_V5` (validator v2, before any policy run); `PREREG_DIAG` amendments 1–3 (2026-08-13: the vision–language confound control, the newest streaming model plus two launch fixes, and the MiniCPM family representative recorded before probing); `PREREG_human_A1` (2026-08-27, power: the original Study B had four scenario clusters and could not have rejected its own null; no threshold moved) and `PREREG_human_A2` (2026-08-30, the designed-item extension, before any item was authored); `PREREG_W2S_A1` (2026-08-31, before any training step: the conversation-sum arm as first frozen was label SFT with a per-conversation learning rate, which would have made the co-primary uninterpretable); and `PREREG_DLA_S2_A1` (2026-08-31, after a pairing run matched zero pairs and before any statistic existed on a valid pair); and `PREREG_PPX_TRAIN_A1` (2026-09-02, after the first battery's sealed evaluation was complete and analysed and before any render of the rebalanced pool existed: the training pool's positional confound, located by the pre-committed audit, with the first battery reported alongside);

and PREREG_SEEDS_A1 (2026-09-02, before its statistic was computed: the seed-inclusive interval on the existing runs; the frozen PREREG_SEEDS is verified byte-identical to its committed version); and PREREG_CAP amendment A1 (2026-09-22, before any outcome was read: the vision patch embedding computed as its exact linear equivalent after the stock bf16 Conv3d fell back to a host-bound kernel; App. F); PREREG_PPX_A1 (2026-09-01, before any new render: the four-window diagnosis row for PersonaPlex through the audio instrument, added beside the untouched primary battery); PREREG_INFER A1 (2026-09-25, before any 32B context was scored: the same frozen norm sentence on Qwen3-VL-32B with the 32B base as its own reference, both read in bf16; App. T); PREREG_MIX A1–A4 and PREREG_LLM A1–A4 (2026-09-25, all before any variant or LLM score existed: GPU scheduling, micro-batch sizes and which seeds could be trained and scored under the memory available, with the pre-committed readings applied to the mean over the scored seeds); PREREG_MIX A5 (2026-09-25, after the seed-0/1 results were read: further seeds under unchanged rules, the mean and the seed count reported whatever they show); and PREREG_LLMSEQ A1–A3 (2026-09-25, each frozen before any replication seed's evaluation was read: training seeds 1–4 as disclosed replications, the Table 1 row and the pre-registered verdicts staying seed 0's). Three of these amendments repair designs that would otherwise have been unrunnable, degenerate or confounded. Amendments on three studies were written after an earlier outcome on the same study was read—PREREG_PPX_TRAIN A1 (rebalanced pool, first battery kept), PREREG_LLMSEQ A2–A3 and PREREG_MIX A5 (further seeds)—each under unchanged rules; PREREG_DLA_S2 A1 followed a zero-pair pipeline check. None changes a rule or moves a threshold after data.

AG Pre-registered stopping rules and the nine experiments they stopped

Improvements we developed and rejected by rules fixed before their results existed, reported as failures rather than reworded: a post-hoc extension (missed by 0.0026); an iteration-ratchet fix (by 0.001; the residual exposed credit-composition drift); 32B adaptive windows (by 0.060); the first JoyAI-VL primary (by 0.0161; later confirmed when powered); the v4 author shift (traced to the occupancy confound and re-run under control); and a "streaming objective actively worsens guards" prediction, refuted at the episode level while the decision-level in-guard propensity rose $0.512 \to 0.803$ exactly as built—our own pre-registration fooled by the aggregation it criticizes.

A seventh joined during the follow-up studies: the naive fix for the pause regression noted in §5 (double the pause-guard weight; App. AC) failed its pre-registered primary, with the audit tracing the miss to the normalization layers of App. AE(iv), which reshaped the dose before it reached the optimizer.

An eighth joined when the third norm's battery failed all three frozen primaries (App. Z): on the audience-conditioned disclosure norm, decision credit shows no advantage over episode-level learners and trails the critic.

A ninth: the corrected-pipeline pause dose (PREREG_DOSE) delivered a build-gate-verified $2\times$ dose and the pause propensity still did not move—the volume budget, not dosing (App. AC). Three of the misses were narrow, which is when frozen rules matter most; the eighth is the reason the central claim carries a boundary, and the ninth is why its one regression is reported as a characterised mechanism rather than an open tuning problem. Frozen primaries that failed but were not candidate improvements are reported where they arose and indexed in Table 3: the lockbox-v5 episode-level verdict (REFUTED), PROBE2's recall bar, PPX-TRAIN's first battery (confounded pool), DLA-S3's P1, and ISO's ours-below-episode-return test.

AH Boundary map

Credit teaches the credited moments: removing a window from credit leaves its behaviour at base (the held-out-window runs on the development scenarios, REPORT_E8), with no detectable transfer between windows of the addressee family; across norms, credit placed on the second norm's deference window does reach the addressee guard (-0.292 at the decision level, App. AA). Data-hungry: $60 \to 180$ episodes moves selectivity $0.571 \to 0.643$ with diminishing returns. Windows must be annotated but need not be hand-annotated *here*: a keyword regex over the transcript retains 84% of the oracle-window gain and is statistically indistinguishable from it (regexwin_result); under

synthetic corruption it retains 57% (e18_doseresponse); two LLM detectors were poor at the same task, and the regex is template-matched to our generator—detection in natural conversation is unsolved. Iterated self-training ratchets talkativeness down (35%; ITER2_RESULT) via re-anchoring and credit-composition drift; not recommended. In the historical campaign the observer's rating of the window-trained arm did not move detectably ($7.14 \rightarrow 6.79$, p=.18); on sealed set 9 the LLM peers rate every trained arm lower (decision credit -0.364 $[-0.644, -0.083]$; App. H), and the blind utterance judge marks decision credit's matched utterances down by 0.454 (App. H). Structural credit enforces the specified policy, which a general rating under-prices; whether people endorse it is what the frozen human study will answer (instrument and analysis frozen, collection pending; Table 3).

AI HISTORICAL TRAINING COMPARISONS AND ARTIFACT INVENTORY

Table 1 legend: arms are whole pipelines, one checkpoint each (Table 4), scored on their own rollouts; bold* marks a confirmed planned primary; † a post-hoc selection-free re-estimate (App. AE); — no valid arm (App. W); shading follows App. B; the set-9 rows share contexts.

The following results use the sequence-level trainer and the per-arm datasets in Table 4. They establish behavioral changes under the recorded configurations, rather than isolating credit arithmetic. The controlled comparison of Section 3.3 fixes training support, anchors, and update count. For the earlier campaign, every frozen primary confirmed on four of the fifteen sealed test sets and no frozen decision-level other-address contrast against the base missed on any (Table 5 names, set by set, the primary or co-primary that missed elsewhere; the researcher-authored sets are in App. O); partial and negative results are retained there. **Reading Table 1.** Its intervals are scenario-clustered CR1/t intervals conditional on one trained checkpoint per arm; the three-seed check of Appendix M adds seed variance. The three set-9 rows share contexts, and MiniCPM-o's critic fits its value head at out-of-fold $R^2 = +0.012$ (Appendix W).

At the base model's own-address answering rate (0.629), decision credit reduces other-address violation relative to the base by 0.252 (signed contrast $[-0.383, -0.165]$) and relative to the critic by 0.165 ($[-0.407, -0.054]$). Its difference from episode return is $+0.043$ $[-0.053, +0.111]$, so the pre-registered ours-below-episode-return contrast fails and the battery's verdict is PARTIAL: two of its three frozen contrasts confirmed. The episode-return and decision-credit arms differ in bad-news responsiveness (0.504 versus 0.805) and neutral speaking (0.147 versus 0.462).

How the matched-rate readout is computed. The rules were frozen in PREREG_ISO before any curve existed. Every arm is read on the same 2,884 frozen lockbox-v9 contexts (24 scenario families), so nothing but the policy differs, and a threshold on $P(\text{speak})$ is swept over every observed probability; where several thresholds give the same own-address answering rate, the lowest other-address violation reachable there is kept, the same rule for every arm. The target rate is the base model's own-address answering rate at τ=0.5 (0.629 in the bf16 serving read), and each arm's violation at that rate is read by linear interpolation on its envelope. The scenario-clustered bootstrap (B=9,999) recomputes the target rate, every envelope and every threshold inside each replicate, so the matching itself is resampled; 24 clusters is small, and the intervals are approximate. A tie-rule audit moves every arm by at most 0.01. Read instead at the controlled study's fp32 target of 0.624 (116 of 186 own-address contexts), the same envelopes give -0.261 against the base, -0.165 against the critic and $+0.035$ against episode return, so no contrast moves by more than 0.009 and no verdict depends on the target's precision (the arm scores are the bf16 serving read; ISO_FP32_RESULT). Decision credit sits below the base at every answering rate in $[0.05, 1.00]$: at answering rates 0.60, 0.629 and 0.75 the other-address violation reads base 0.548/0.548/0.661, critic 0.410/0.461/0.643, episode return 0.217/0.252/0.348 and decision credit 0.252/0.296/0.391. Matched instead on the base's neutral speaking rate (0.301), decision credit reads 0.157 against base 0.548, critic 0.574 and episode return 0.461.

The base model's own-address values, reconciled. Five numbers describe the same Qwen3-VL-8B base on lockbox-v9 and differ only in instrument or pool: (1) 0.629 is the share of the 186 own-address contexts with $P(\text{speak}) > 0.5$ in the bf16 serving read, the matched-rate target above and in §3.4; (2) 0.624 is the same share under the strict-fp32 rescoring of the controlled study (§3.3, Table 2), where one context crosses the threshold; (3) 0.603 is the mean probability over those 186

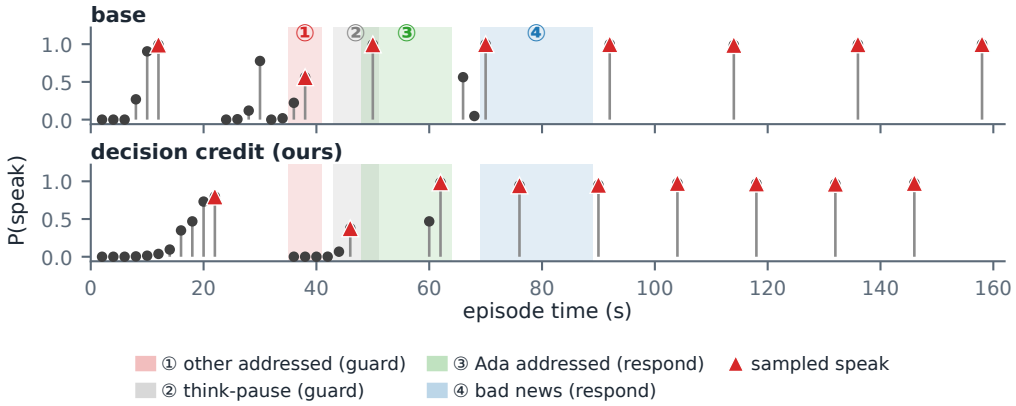

Figure 17: The decision-level contrast in a single episode (base is Qwen3-VL-8B; per-tick $P(\text{speak})$ read from the evaluation logs; shaded spans are the scored windows; triangles are sampled speaks). At ① the facilitator asks *Ben*, not Ada—the base rises to $P{=}0.56$ and answers for him; decision credit stays at the floor. At ③ Ada is addressed and both speak. At ② decision credit speaks once inside the pause guard ($P{=}0.38$, before Ada's own window opens): the pause regression of §5 in one episode. **The scoreboard nets the base to 0 and decision credit to 1 and never sees the pause speak at ②; the decision level sees each.**

contexts (Table 21); (4) 0.597 is the deployment row (Table 15), read at a threshold frozen on the v8 development split rather than matched on v9; (5) the three context pools are three logs: 2,884 contexts are session 0 of the frozen v9 base run (Tables 21 and 2, this readout); 5,955 decisions pool both base sessions of that set (the probe of App. R); and 957 ticks are a capped sample from session 0 of the escape-decode base re-run (App. Q).

Artifact inventory. The historical campaign contains **64 pre-registrations**, **over 110 evaluation logs** covering more than 14,900 episodes, **over 20 collections**, **over 70 training sets**, and **over 40 verdict files**. These are repository inventory counts, not counts of independent replications. The ledger generator checks 24 of its 26 contrasts against stored verdict or sensitivity files; the two it cannot check are the vs-base cells of sets 5 and 7, whose rows (†) re-estimate post hoc, under the selection-free estimator, batteries whose frozen primaries used the per-episode estimator (App. AE(ii)). Nine candidate improvements were rejected by their own frozen rules and are listed in Appendix AG. Follow-up studies on existing contexts are described in their own protocols rather than counted as fresh sealed sets.

AJ Additional figures and tables

This appendix collects the historical-campaign floats: Figure 17 (one matched lockbox-v7 episode), Figure 18 (scoreboard masking on lockbox-v5), Figure 19 (the cells of Table 1 as a forest plot), Figure 20 (the early four-scenario granularity comparison) and Table 21 (the fixed-state decomposition behind the matched-rate readout of App. AI).

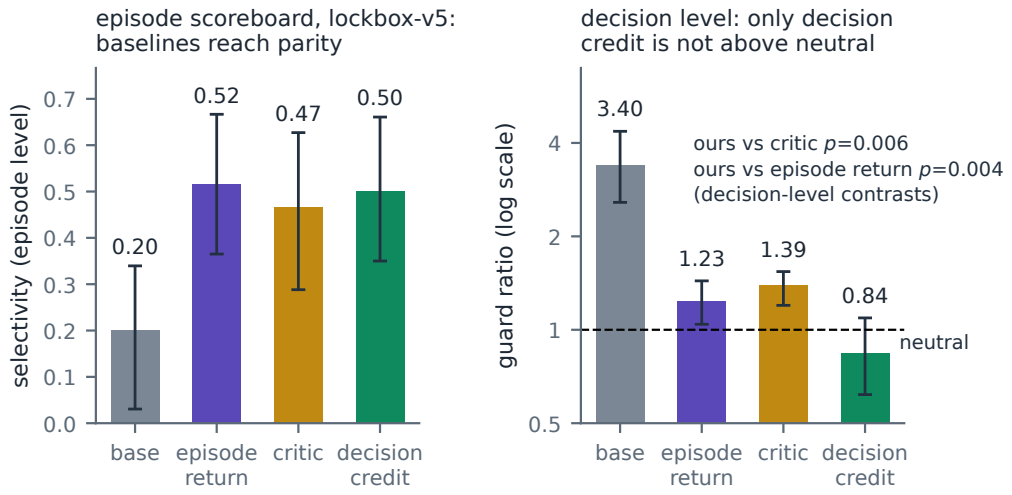

Figure 18: Scoreboard masking on lockbox-v5. Learned baselines reach episode parity (left) while only decision credit's decision-level guard ratio is not shown above neutral (right). Annotated p-values are contrasts against ours; each baseline's own interval excludes 1.0 from above (episode return $[1.04, 1.44]$, critic $[1.20, 1.54]$) whereas ours' includes it ($0.84\,[0.62,\ 1.09]$)—ours is the only arm not shown to be above neutral, a weaker statement than being shown below it.

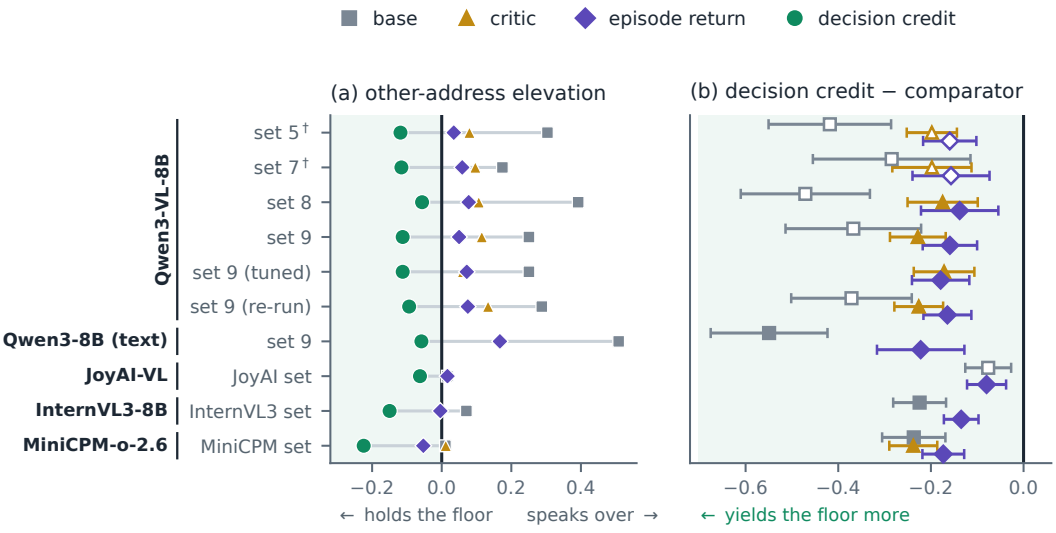

Figure 19: The Table 1 pipeline contrasts as a forest plot. The cells of Table 1, as a picture; every contrast is negative. *(a)* Raw in-guard elevation of each arm: every base and every critic is *above* zero, and ours is below zero in all ten rows—episode return crosses below only on the two non-Qwen3-VL lineages, InternVL3 and MiniCPM-o (-0.004, -0.053), where ours reaches -0.150 and -0.224. *(b)* The 26 contrasts with scenario-clustered intervals (dependent, not 26 replications: the three set-9 rows share contexts, and the set-5 and set-7 rows are post-hoc re-estimates whose frozen verdicts were v5 episode-level REFUTED and v7 PARTIAL, App. N, App. AE(ii)); filled markers are pre-registered frozen primaries confirmed under their own rule, hollow markers are post-hoc sensitivity or declared secondaries. No interval touches zero. These compare whole pipelines (Table 4); on identical rows (Table 2) the matched-rate contrasts resolve no ordering among the constructions on aggregate violation, and the swap pairs of App. U separate them.

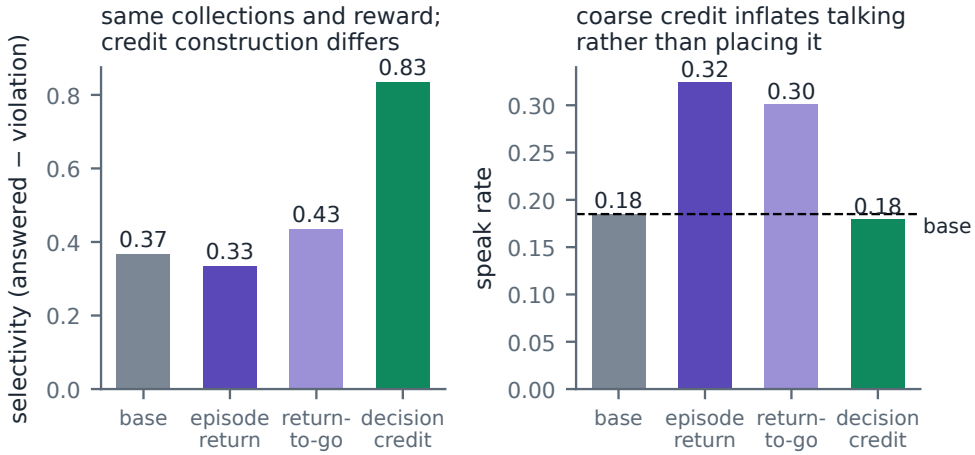

Figure 20: Early historical configurations: selectivity and speak rate. These E16 configurations share source collections and a reward definition; credit construction also changes row retention and anchors. Return-to-go credits each decision with the z-scored sum of its own and subsequent rewards; the episode return here is the globally z-scored variant. Four scenarios; each arm's log holds 60 rollouts, of which the 30 session-0 episodes are plotted. With $G{=}4$ no contrast against the base resolves (decision credit vs base selectivity $+0.467\,[-0.244,\ +1.177]$; return-to-go $+0.067\,[-0.594,\ +0.728]$; episode return $-0.033\,[-0.411,\ +0.344]$; decision credit vs episode return $+0.500\,[+0.051,\ +0.949]$, webb-$p$=0.052). With $G{=}4$ the CR1 interval and the Webb p need not agree, as that last contrast shows. Decision credit raises selectivity without raising the speak rate above the base's; under identical rows (Table 12) the episode-return arm speaks less at neutral than decision credit (0.195 vs 0.511), so the speak-rate difference here reflects the pipeline, not credit granularity.

Table 21: The decomposition the pooled contrast hides, at fixed states. All four arms scored on one frozen context set (session 0 of the base arm's frozen lockbox-v9 run), so the columns are directly comparable. Ours is the only arm that speaks *less* over an addressee than in neutral context, and it reaches that while holding the *highest* neutral propensity of any arm—including base. Bold marks the best arm per row under that row's stated direction; it is not always ours.

$P(\text{speak} \mid \cdot)$ at the *same* 2884 contexts	base	critic	episode return	ours	n
someone else addressed (GUARD: LOWER BETTER)	0.556	0.387	0.338	**0.203**	115
pause (GUARD: LOWER BETTER)	0.601	0.419	**0.402**	0.472	60
Ada addressed (RESPOND: HIGHER BETTER)	**0.603**	0.437	0.460	0.483	186
bad news (RESPOND: HIGHER BETTER)	0.534	0.460	0.447	**0.702**	226
neutral (NO NORM APPLIES)	0.305	0.285	0.277	0.332	2297
in-guard *elevation* (addressee − neutral)	+0.251	+0.102	+0.061	**−0.129**	

AK EXTENDED RELATED WORK

Turn-taking and participation. Conversation analysis describes how a current speaker selects the next speaker and how participants self-select (Duncan, 1972; Sacks et al., 1974; Schegloff, 2000). Computational systems model timing cues, turn completion, and future voice activity (Ekstedt & Skantze, 2020; Elmers et al., 2025; Gravano & Hirschberg, 2011; Raux & Eskenazi, 2009), and addressee identification and response selection are established multi-party tasks (Gu et al., 2021; Jovanović et al., 2006; op den Akker & Traum, 2009; Ouchi & Tsuboi, 2016; Zhang et al., 2018). Our focus is the participation policy of a generative agent: how training changes its decision to answer, and whether the change can be traced through an internal intervention.

Streaming and socially interactive agents. Moshi, PersonaPlex, and JoyAI-VL make listening and speaking explicit parts of interaction (Défossez et al., 2024; Roy et al., 2026; Yao et al., 2026); recent work trains conversational dynamics and evaluates proactive behavior (Li et al., 2026; Lu et al., 2026). When2Speak, Speak or Stay Silent, and ModeratorLM bring participation and turn roles into multi-party language-model evaluation (Bhagtani et al., 2026; Mitra et al., 2026; Nama et al., 2026), and social-agent benchmarks and training methods evaluate broader conversational behavior (Wang et al., 2024; Zhou et al., 2024), as does long-horizon agent-society simulation (Wang et al., 2026). We complement these efforts with matched-responsiveness evaluation and causal tests of a learned addressee distinction.

Supervision and internal interventions. Process supervision, reward decomposition, and fine-grained feedback study how learning signals reach individual decisions (Arjona-Medina et al., 2019; Chan et al., 2024; Lightman et al., 2024; Uesato et al., 2022; Wu et al., 2023); our controlled study compares alternative credit constructions on a fixed set of participation decisions. Activation patching tests causal contributions of internal states (Zhang & Nanda, 2024), representation interventions can adapt model behavior (Wu et al., 2024), and recent speech-model work also steers the transition between listening and generation (Chang et al., 2026). We use bidirectional donor-state transfer to study a different distinction: whether the current question is addressed to the agent or to another participant.