BEYOND TURN TIMING: AUDITING MULTI-AXIS REWARDS FOR FULL-DUPLEX SPEECH REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A full-duplex speech agent must respond appropriately while deciding when to speak, wait, or yield. We study reinforcement learning with three sources of feedback: deterministic timing rules, semantic judgments of response appropriateness, and perceptual judgments of speech quality. A generative reward model supplies the two quality scores for joint text–audio rollouts. We normalize each reward component within a group of continuations and use the resulting advantages to update the policy through its text-token probabilities. We compare sequence-level optimization with component-wise optimization over event-specific positions. On Full-Duplex-Bench and SID-Bench, the trained policies produce fewer unwanted responses but also answer fewer inputs that require participation. Lower aggregate latency on Full-Duplex-Bench comes from a changed set of answered inputs; SID shows longer delays even on shared responding inputs. Adding learned rewards does not yield a consistent quality advantage over timing rules alone. The study provides a working method for combining timing and quality feedback in full-duplex post-training and identifies where its behavioral outcomes fall short: improved restraint has yet to translate into more useful responses.

1 INTRODUCTION

A full-duplex speech agent must decide when to speak, wait, or yield while producing an appropriate response (Levinson & Torreira, 2015; Skantze, 2021). A hesitation may call for silence, a completed question for an answer, and an interruption for a timely handover. These events require feedback about both interaction timing and response quality. Post-training must also connect that feedback to the decisions made by a streaming policy.

Moshi supports concurrent listening and speaking through parallel speech streams and time-aligned text (Défossez et al., 2024). In this architecture, the reward can assess generated audio while reinforcement learning (RL) updates the probabilities of generated text tokens. The choice of reward matters because timing and quality can disagree. A prompt reply may be irrelevant, while an appropriate reply may arrive too late. A silent agent avoids interruptions but also misses questions. Timing rules can evaluate annotated boundaries; semantic and perceptual judgments assess what the agent says and how it sounds.

Recent methods combine interaction and response-quality rewards (Ohashi et al., 2026) or optimize conversational dynamics over selected temporal windows (Li et al., 2026). Dual-Axis GRM assesses semantic quality and interaction timing (Chen et al., 2026). We study a related design in which timing remains rule-based and a generative reward model (GRM) supplies separate semantic and perceptual judgments. Separating the components allows us to inspect their variation among alternative responses and to assign them to different generated positions.

For each conversational context, we sample a group of joint text–audio continuations and standardize each available reward component within that group. We then compare two ways of applying the advantages. Seq combines them before optimizing the generated sequence; Event assigns each component to relevant positions and clips its objective separately. A timing judgment can then supervise turn-boundary decisions, while a semantic judgment supervises the lexical response. This

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.

comparison tests the full optimization schemes, including their different reward-combination and normalization choices.

We train these policies from the same Moshi backbone and evaluate them on Full-Duplex-Bench (FDB) (Lin et al., 2025) and SID-Bench (Xia et al., 2026). The comparisons separate two questions: how behavior changes when learned quality rewards are added to timing rules, and how it changes when the optimization scheme changes under the same reward source. For each comparison, we report which inputs receive a response before examining the timing and quality of those responses. Otherwise, a policy that stops answering difficult inputs can appear faster on the answers it still produces.

The experiments expose a recurring evaluation problem. Across both benchmarks, reduced unwanted speech accompanies fewer responses on inputs that call for participation. Yet the timing interpretation differs: response selection explains the aggregate FDB speed advantage, while SID shows longer delays on common responding inputs. These observations motivate analyzing reward design through the behaviors it separates. An agent that waits more successfully during pauses must still recognize when waiting is no longer appropriate.

We make three contributions:

- **Multi-axis speech post-training.** We combine rule-based timing with separate GRM semantic and perceptual rewards, using joint text–audio rollouts to optimize a streaming policy through text-token probabilities.
- **Reward and optimization comparisons.** We evaluate rule-only and two learned-reward variants, and compare sequence-level and event-level optimization while holding the reward source fixed.
- **A paired account of behavioral changes.** Across FDB and SID, we identify fewer unwanted responses alongside reduced response coverage. Shared-response analysis distinguishes changes in response timing from changes in which inputs are answered.

2 Related Work

Full-duplex modeling and evaluation. dGSLM learns two-channel conversational dynamics from audio (Nguyen et al., 2023); Moshi adds time-aligned text to parallel speech generation (Défossez et al., 2024). Turn-taking research and Voice Activity Projection distinguish yielding, holding, responding, and backchanneling (Skantze, 2021; Ekstedt & Skantze, 2022). FDB evaluates canonical interaction behaviors and, in v1.5, overlapping speech (Lin et al., 2025; 2026). SID-Bench measures semantic-aware interruption detection through false-alarm and response-delay penalties (Xia et al., 2026). Our evaluation keeps response coverage separate from conditional timing and quality.

Reward learning and full-duplex RL. Learned feedback supports behavior alignment and instruction following (Christiano et al., 2017; Ouyang et al., 2022). Full-duplex methods optimize event-specific interaction and response quality (Ohashi et al., 2026) or separate conversational dynamics from semantic generation (Li et al., 2026). Dual-Axis GRM assesses semantic quality and turn-taking (Chen et al., 2026). Our focus is the integration of separate quality axes with deterministic timing and text-stream credit in a streaming policy. The evaluation compares configurations within this pipeline, rather than ranking the published systems.

Relative advantages and scoring interfaces. GRPO uses within-group comparisons with a clipped surrogate (Shao et al., 2024; Schulman et al., 2017). GDPO separates component normalization, DAPO filters zero-contrast groups, and SA-MRPO reweights saturated objectives (Liu et al., 2026; Yu et al., 2026; Wang et al., 2026). These methods address how reward variation produces a learning signal. Scorer discrimination and proxy overoptimization concern the reliability of the scores themselves (Lambert et al., 2025; Gao et al., 2023). Our learned-reward variants use a deployed Qwen3-Omni scorer (Xu et al., 2025); downstream evaluation uses a fixed judging protocol (Zheng et al., 2023).

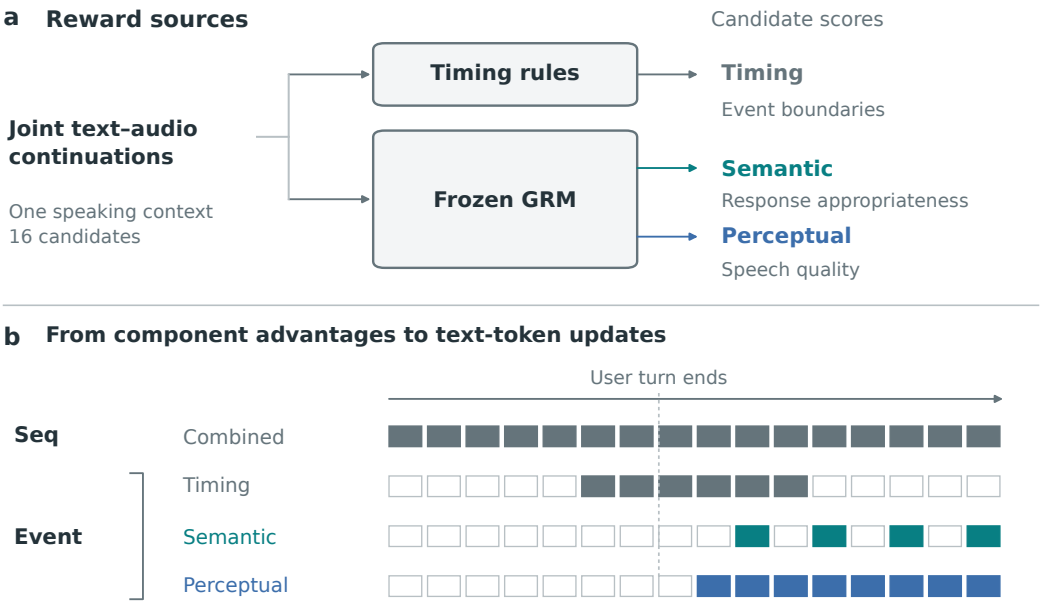


Figure 1: Reward sources and temporal credit. (a) For speaking events, timing rules and a frozen GRM assess each joint text–audio continuation; hold events use timing rules alone. Retained reward components are standardized within each group. (b) Illustrative generated text positions for a completed turn; filled cells receive credit. Seq combines advantages before clipping over the sequence. Event clips components separately over timing, lexical-response, and speech-aligned positions. Both use sequence-wide KL regularization. Positions are schematic, without a measured time scale.

3 GRM-Guided Full-Duplex Post-Training

The method combines GRM training, online reward computation, and policy optimization. The policies share a streaming backbone and differ in reward source or optimization scheme. Figure 1 shows the reward sources and the positions they supervise; Appendix Figure 5 gives the full training flow.

3.1 Streaming Policy and Interaction Events

Let $u_{1:t}$ denote the user audio observed by time t, and let h_t contain the generated assistant history. A full-duplex policy produces a text token z_t and audio codes a_t conditioned on this streaming state. We write the rollout as $y = (z_{1:T}, a_{1:T})$. Training examples specify an interaction event e and its temporal landmarks. The event determines the desired behavior: maintain the user's floor during a pause, initiate a response after a turn boundary, produce a brief backchannel at an opportunity, or yield and resume after an interruption.

For each context and event, the rollout policy samples $G = 16$ continuations. The group shares the same user input, so its rewards compare alternative actions in the same conversational situation. We train on event segments from Seamless Interaction (Agrawal et al., 2025). The schedule samples all four event families, with up to 30 seconds of context and a context inclusion probability of 0.5. The policy starts from the pretrained Moshika checkpoint; the reference policy is an independent frozen copy.

3.2 Generative Reward Modeling

The GRM uses the Thinker component of Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) to supply semantic and perceptual judgments through structured, fixed prompts. Semantic assessment targets response appropriateness; perceptual assessment targets speech quality. Keeping these out-

puts separate permits each component to be normalized and assigned credit independently. Timing remains an external deterministic reward, rather than a learned GRM output.

GRM training and fixed-scorer evaluation. GRM training comprises grounding supervision, reward-score supervision, and same-context ranking. Grounding addresses the interpretation of interaction inputs; score supervision supplies structured quality judgments; ranking compares alternative candidates for a common context. We evaluate the original scorer (GRM) and a scorer from the rubric-revision training line (GRM-Rubric), which targets more explicit score-level descriptions and perceptual criteria. Their use across policy comparisons is specified in Appendix B. Both remain fixed during policy optimization. This comparison evaluates the resulting scorers; it does not isolate the contribution of each training stage.

3.3 Event Rewards and Component Aggregation

The rule reward r_i^{rule} measures event timing from generated speech activity and event annotations. During a hold event, it penalizes premature speech onset, speech duration, and sustained takeovers. For a completed turn, it rewards prompt speech within 0.8 seconds, penalizes early speech, and assigns a penalty to never responding. Interruption rewards combine yielding with resuming after the user's interrupting utterance. Backchannel rewards match short utterances to annotated opportunities and penalize missed opportunities and takeovers. Appendix C records the numerical settings.

Learned components apply to speaking events; hold events use the rule component to evaluate appropriate silence. Each sampled continuation therefore receives the reward components applicable to its event. A reward component contributes to an update only if all candidates in the group receive a valid score. This complete-group requirement discards otherwise available scores when any candidate lacks one; Appendix E quantifies the resulting loss of semantic feedback.

3.4 Policy Optimization and Temporal Credit

For a retained reward component k, we compute population-standardized advantages

$$A_i^k = \frac{r_i^k - \bar{r}^k}{\max(\sigma_k, \epsilon)}, \qquad \bar{r}^k = \frac{1}{G}\sum_{i=1}^{G} r_i^k, \tag{1}$$

with $A_i^k = 0$ when $\sigma_k < \epsilon$ and equal component weights. We compare *sequence-level joint optimization* (Seq) with *event-level component-wise optimization* (Event). Seq combines standardized components before applying the clipped surrogate over the generated text sequence. Event instead clips each component separately and assigns credit through event-relevant masks. Its timing, semantic, and perceptual masks select boundary, lexical-response, and speech-aligned positions, respectively. Both retain sequence-wide KL regularization. Optimization uses text-token probability ratios with full-model adaptation; audio-code likelihoods are not part of the ratio. Equation 2 specifies the Event update; Appendix C gives mask definitions and numerical settings.

Let M_i^k select generated text positions receiving credit from component k, and let ρ_{it} be the current-to-rollout text-token probability ratio. Writing the population objective averaged over sampled candidates, the component-masked Event objective is

$$\begin{aligned}\mathcal{L}(\theta) = & -\frac{1}{G}\sum_{i=1}^{G}\sum_{k}\frac{\sum_t M_{it}^k \min\big(\rho_{it}A_i^k, \text{clip}(\rho_{it}, 1-\eta, 1+\eta)A_i^k\big)}{\sum_t M_{it}^k} \\ & + \beta\, \mathbb{E}_{i,t\in M_i^{\text{KL}}} D_{\text{KL}}(\pi_\theta(\cdot \mid h_{it}) \| \pi_{\text{ref}}(\cdot \mid h_{it}))\,. \end{aligned} \tag{2}$$

Here η is the clipping range, β weights KL regularization, and M_i^{KL} includes all generated text positions. Empty component masks contribute zero. Seq and Event differ in both temporal masks and the order of clipping and combination.

The two optimization schemes differ in how they combine reward components and distribute their advantages over time. Seq lets every retained reward assess the same generated text sequence, including the decisions leading up to the response. Event associates timing with boundary decisions and quality with response positions, reducing the temporal span over which each component contributes. Its normalization by selected mask length also changes the weight assigned to an individual

Policy	Rewards	Optimization	FDB inputs	SID inputs
Base	No RL	None	727	2,100
Rule/Seq	Timing	Seq	727	2,100
GRM/Seq	Timing + GRM	Seq	727	2,100
GRM-Rubric/Seq	Timing + GRM-Rubric	Seq	727	2,100
Rule/Event	Timing	Event	727	Not tested
GRM/Event	Timing + GRM	Event	727	Not tested

Table 1: Policy variants and evaluation coverage. Rewards lists the training signals: Timing denotes the rule reward; each GRM adds semantic and perceptual scores. Seq and Event are the optimization schemes defined in Section 3. FDB inputs and SID inputs are sample counts per policy, shared within each benchmark. Not tested denotes an evaluation that was not run. Each RL policy uses one training seed.

token. Moreover, combining advantages before clipping can allow opposing component preferences to cancel before the surrogate is evaluated; separately clipped components need not produce the same update. We therefore compare Seq and Event as complete optimization schemes, accounting for both reward combination and temporal credit.

4 Evaluation Design

We evaluate response decisions on common benchmark inputs and conditional performance on eligible outputs. This separates whether a policy responds from the timing and quality of its responses.

4.1 Policies and Benchmarks

Policies and training scope. Table 1 organizes policies by reward source and optimization scheme. Policy names specify both choices: Rule/Seq uses timing rules with sequence-wide optimization; GRM/Event adds the original scorer with event-local optimization. The RL runs share a data schedule, training seed, and rollout budget. Inputs and decoding settings are fixed within each benchmark; Appendix B records the training settings.

FDB scoring and decoding. FDB evaluates pause handling, turn-taking, backchannels, and interruptions. Lower takeover rates are desirable when the agent should wait; higher response rates are desirable when it should answer. Quality is rated among eligible interruption responses. Eligibility depends on automatic speech recognition (ASR), so short or poorly transcribed speech may count as no response. Decoding and judging are fixed across policies, and FDB was used during development. Appendix D specifies the input composition, eligibility rules, and scoring implementation.

SID scoring. SID complements event-specific interaction evaluation with English and noise/silence inputs. We map output-audio activity to detection time and apply the released notebook formulas for average penalty time (APT), false interruption rate (FIR), and interruption response latency (IRL). Coverage additionally records whether positive inputs receive a response. Appendix I details the scoring adaptation and source grouping.

4.2 Paired Comparisons and Uncertainty

We pair outputs by benchmark input and report paired bootstrap intervals (Efron, 1992). FDB binary comparisons use exact McNemar tests (McNemar, 1947) with Holm correction (Holm, 1979); this comparison family was specified retrospectively. FDB intervals resample clips, while SID intervals resample source recordings. All intervals condition on fixed checkpoints and decoding; continuous-outcome and SID intervals are pointwise. Paired timing and quality use jointly eligible inputs, excluding gained or lost responses. Appendices D and I specify the estimands and resampling settings.

Policy	Synth. ↓	Candor ↓	Turn ↑	BC ↓	Lat. ↓	Intr. ↑	Quality ↑
Base	38.7	50.5	73.1	32.7	0.199	92.0	3.500
Rule / Seq	37.2	48.1	68.9	36.4	0.177	92.0	3.435
GRM / Seq	32.8	47.2	68.9	40.0	0.272	90.0	3.339
GRM-Rubric / Seq	37.2	49.1	68.9	40.0	0.241	91.5	3.492

Table 2: Main Full-Duplex-Bench v1 results. Synth. and Candor are pause takeover rates; BC is backchannel takeover; Turn and Intr. are response rates, all in percent. Lat. is turn latency in seconds. Quality is the interruption rating (0–5) among rated responses. Every policy is evaluated on all 727 inputs.

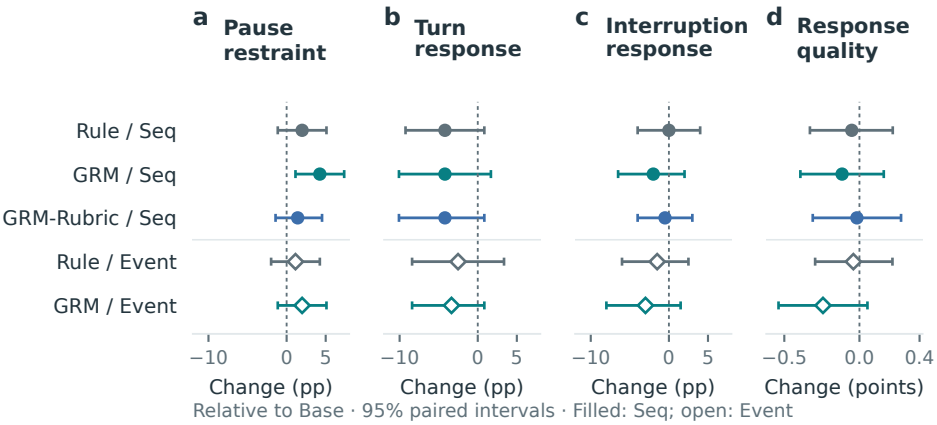


Figure 2: Behavioral effects of RL relative to Base. The same five policies are compared on (a) pause restraint (one minus takeover rate), (b) completed-turn response, (c) interruption response, and (d) interruption quality. Positive differences favor the RL policy. Panels a–c use a common percentage-point scale; panel d uses rating points on jointly rated responses. Bars show pointwise 95% paired bootstrap intervals on 353, 119, 200, and 169–177 clips, respectively.

5 Results

We first examine response decisions and conditional quality on FDB, then analyze its aggregate latency comparison. SID tests whether restraint and responsiveness also differ under a separate speech-activity-based scoring protocol.

5.1 FDB Separates Pause Restraint from Response Coverage

Lower pause takeover rates accompany lower turn coverage. All three Seq policies reduce observed pause takeovers on both pause subsets (Table 2). GRM/Seq has the largest pooled reduction, from 162 to 147 takeovers across 353 inputs: 24 inputs improve and nine regress. At the same time, its completed-turn responses fall from 87 to 82 out of 119, as they do for the other Seq policies. The changes therefore combine better restraint inside a user's turn with more missed opportunities after a completed turn. Figure 2 makes this distinction visible. The pause result is a descriptive effect: none of the 25 FDB binary comparisons survives Holm correction (Appendix A).

These outcomes distinguish two behaviors that a single measure of speaking frequency would conflate. A pause requires the agent to preserve the user's floor, whereas a completed turn requires it to initiate speech. Suppressing both responses improves one outcome and worsens the other. The shared direction across rule-only and learned-reward Seq policies suggests that increased restraint is a broader effect of the evaluated training configurations; it does not identify an additional benefit of learned quality feedback.

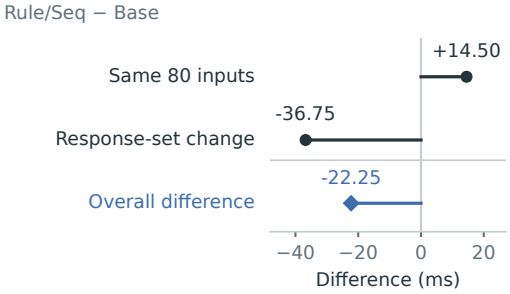
Conditional quality and event-specific behavior. On 177 jointly rated interruption inputs, GRM-Rubric/Seq exceeds Rule/Seq by 0.056 rating points, with a 95% interval of $[-0.209, 0.322]$. The

a Paired response counts

Response set	Inputs
Both policies	80
Base only	7
Rule/Seq only	2
Neither policy	30

119 completed-turn inputs

b Sources of the latency difference



+14.50 − 36.75 = −22.25 ms

Figure 3: Lower aggregate latency reflects a different response set. (a) Response counts on 119 FDB completed-turn inputs, using the benchmark's ASR eligibility criterion. (b) Rule/Seq is 22.25 ms faster in aggregate, but 14.50 ms slower on the 80 inputs answered by both policies (95% paired interval: $[-98.5, 132.5]$ ms). The response-set change contributes -36.75 ms to this descriptive decomposition; negative values favor Rule/Seq.

	APT (s)			FIR	IRL	Positive response
Policy	All	English	Noise/silence	(%)	(s)	(%)
Base	3.472	2.679	6.010	3.81	4.884	28.82
Rule/Seq	3.478	2.732	5.867	3.38	5.134	26.00
GRM/Seq	3.444	2.754	5.650	3.48	5.024	25.27
GRM-Rubric/Seq	3.436	2.733	5.687	3.33	5.088	25.64

Table 3: SID-Bench results on 2,100 shared inputs. APT, FIR, and IRL follow the released notebook formulas; lower is better. FIR divides false triggers by all inputs; IRL averages strictly positive delays. Positive response is the fraction of 1,100 positive inputs with detected output speech, including early responses.

comparison does not establish a quality advantage. Backchannel behavior also differs from pause restraint: Seq takeover rates rise from Base's 32.7% to 36.4–40.0%, while backchannel frequency falls from 0.123 to 0.111–0.118 per second. Fewer acknowledgments can therefore coexist with more floor-taking responses. Successful backchanneling requires producing a brief response at the right opportunity, so silence alone cannot satisfy this event's objective.

5.2 Response Selection Reverses an Aggregate Latency Comparison

Rule/Seq lowers aggregate turn latency by 22.25 ms but responds to fewer inputs. Both policies respond to 80 inputs; Base alone responds to seven and Rule/Seq alone to two. The seven Base-only responses have a mean latency of 568.57 ms, so excluding these relatively slow responses changes the aggregate comparison. On the 80 shared responses, Rule/Seq is 14.50 ms slower on average. Response-set composition contributes -36.75 ms, yielding the observed -22.25 ms aggregate difference (Figure 3).

This decomposition locates the apparent speed advantage in which inputs receive answers. It does not establish slowing on common inputs: the shared-response interval spans zero. It does show that an aggregate latency reduction can occur without a corresponding reduction on shared responses. Coverage, shared-response latency, and the response-set contribution answer different questions and should be read together. Appendix D gives the exact identity.

5.3 SID Records Fewer Triggers and Reduced Responsiveness

GRM-Rubric/Seq has the lowest observed APT and FIR (Table 3). Its false triggers fall from 80 to 70 relative to Base, a change of -0.476 percentage points with pointwise interval $[-0.927, -0.047]$.

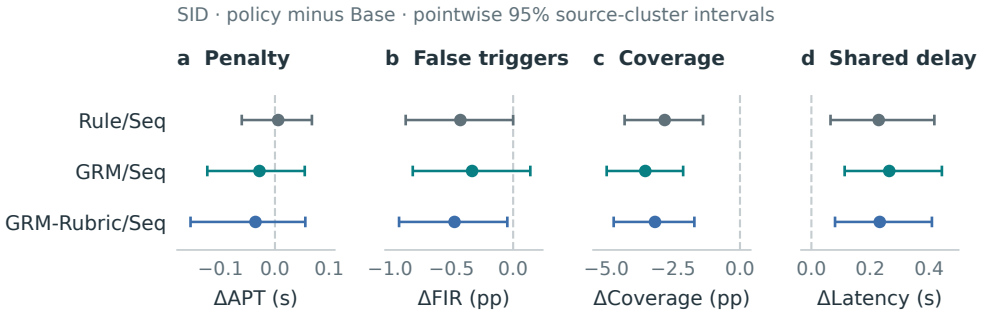


Figure 4: SID paired effects relative to Base. Points show differences; bars are pointwise 95% intervals from 20,000 paired source-recording bootstrap resamples. APT and FIR use 2,100 inputs from 903 sources; coverage uses 1,100 positive inputs. Shared positive-delay sets contain 262, 255, and 257 inputs for Rule/Seq, GRM/Seq, and GRM-Rubric/Seq, respectively. Lower APT, FIR, and latency are desirable; higher coverage is desirable. Intervals condition on fixed checkpoints and decoding, without multiplicity adjustment.

The APT reduction is 0.036 s, with an interval spanning zero. These two measures weight errors differently: FIR counts triggers, whereas APT charges a false trigger the full input duration. Eliminating a trigger on one 91.263-s noise input contributes more than the policy's net 75.271-s total penalty reduction. The small aggregate APT change consequently summarizes substantial offsetting changes across inputs.

The scene breakdown identifies where these changes occur. All three RL policies lower noise/silence APT but raise English APT. GRM/Seq has the lowest noise/silence penalty, whereas GRM-Rubric/Seq has the lowest overall penalty. For GRM-Rubric/Seq, the noise/silence change is -0.323 s and the English change is $+0.054$ s; weighting them by 500 and 1,600 inputs gives the -0.036 s overall change. Its aggregate ranking thus depends on the benchmark's scene mixture.

All three RL policies reduce positive response coverage from 28.82% to 25.27–26.00%. They also add 0.229–0.264 s on shared positive-delay responses (Figure 4); all three pointwise intervals exclude zero for coverage and shared latency. SID therefore provides evidence of later responses on common inputs, beyond the response-set effect seen on FDB. Its activity-based criterion also finds reduced coverage, so the repeated direction across benchmarks is not confined to FDB's ASR eligibility rule. Appendix I reports the full contrasts.

5.4 Event-Local Optimization Changes the Behavioral Profile

Holding the rule reward fixed, Event raises completed-turn coverage from 68.9% to 70.6% relative to Seq, while increasing turn latency from 0.177 to 0.258 s and backchannel takeover from 36.4% to 47.3%. Using the same GRM scorer in both schemes, Event raises turn coverage to 69.7% and lowers aggregate turn latency to 0.217 s, but interruption coverage falls from 90.0% to 89.0%. Event-local optimization therefore changes different behaviors in different directions. The configuration changes temporal masks, normalization denominators, and clipping order together; the comparison evaluates that combined choice. Paired interruption-quality intervals span zero for both reward sources (Appendix A).

What changes when learned rewards are added? The Seq comparisons separate the reward-source question from the Seq/Event question. On FDB, both learned-reward policies retain the rule-only policy's 68.9% completed-turn coverage, while their other outcomes differ. GRM gives the strongest pause restraint but lower interruption coverage and conditional quality; GRM-Rubric remains closer to the rule-only interruption profile. SID provides a complementary comparison: GRM-Rubric/Seq lowers APT by 0.042 s relative to Rule/Seq, but the paired interval is $[-0.160, 0.049]$ s. Its positive response coverage is 25.64%, compared with 26.00% for Rule/Seq. The learned-reward variants therefore change some event outcomes while leaving others unchanged or worse; neither scorer dominates the rule-only policy on the reported measures.

6 Discussion and Limitations

The strongest common pattern is reduced participation: the trained Seq policies interrupt less during pauses and trigger less often on SID, but also respond to fewer completed turns and positive SID inputs. Because the rule-only policy shows this pattern too, the change cannot be attributed specifically to learned quality feedback. The reward-source comparisons are needed to determine what that feedback adds beyond the timing objective.

Two reasons for a latency change. FDB and SID distinguish two effects of post-training. On FDB, dropping relatively slow responses changes the mean, even though responses on common inputs are not faster on average. On SID, the shared responses themselves are later. The latter result cannot be explained solely by which responses enter the mean. Together, the comparisons show why response coverage is needed to interpret latency, and why conditioning on shared responses supplies information that coverage alone cannot provide.

Scene mixture and aggregate performance. The SID scene breakdown illustrates why a benchmark average is only one view of a policy. In a setting with frequent background audio, suppressing unintended triggers may matter more; in an exchange that demands prompt answers, missed and delayed responses carry a greater cost. The English and noise/silence penalties move in opposite directions here, so changing their relative frequency changes the aggregate assessment. The same issue appears within FDB: preserving the user's floor during a pause and acknowledging a backchannel opportunity require different actions. Reporting the outcome vector allows these priorities to be assessed explicitly, without assuming that the benchmark's mixture defines every use case.

GRM and optimization limits. The quality results need to be read alongside the rewards that reached optimization. In the GRM-Rubric/Seq audit, complete-group deletion removes 82.9% of available semantic scores before advantages are computed (Appendix E). A separate perceptual probe returns the maximum score for every GRM-Rubric output (Appendix G). These observations give specific reasons to examine score availability and discrimination before interpreting a small policy-quality difference as a verdict on learned rewards. They do not establish the cause of the benchmark outcomes. The experiments also compare complete scorers and optimization schemes, without isolating individual GRM training stages or temporal masks.

Statistical and evaluation scope. One backbone, one training seed per configuration, and development use of FDB limit generalization. No FDB binary comparison survives Holm correction, and the shared-response FDB latency interval includes zero. FDB clip intervals omit source dependence and judge variability; SID source-cluster intervals remain pointwise and checkpoint-conditional. FDB eligibility depends on ASR, whereas SID uses output-audio activity. Short or poorly transcribed speech can be excluded by one criterion without being silent.

Implications for reward design. The results favor evaluating a timing reward by the decisions it distinguishes. Pause suppression, turn initiation, and brief acknowledgment demand different actions, even when all affect total speaking time. Learned quality feedback must then discriminate among responses in contexts where these decisions matter. The recorded deletion rule removes semantic comparisons before either Seq or Event can assign them credit. Preserving available scores and validating their ordering are therefore concrete next steps for this pipeline. A matched intervention should test whether these changes improve response quality while retaining coverage, rather than treating increased reward availability as the endpoint (Appendix H).

7 Conclusion

We studied full-duplex speech RL with rule-based timing, learned semantic and perceptual rewards, and two optimization schemes. The policies reduce unwanted speech on FDB and SID, but also lose response coverage. Lower aggregate FDB latency comes from response selection; SID shows later responses even on common inputs. Adding learned rewards does not produce a consistent quality gain over rules alone. Improving this method requires both reliable quality feedback during training and evidence that better restraint preserves the responses users need.

Reproducibility Statement

The evidence package contains per-example outcomes, paired statistics, missingness histograms, generated tables, and source-artifact hashes. Its analysis scripts recompute FDB comparisons, the response-selection decomposition, supplementary reward diagnostics, and SID scores and tables. The reproducibility manifest supplies artifact identities and the package README lists the corresponding commands. This supports downstream recomputation; it does not recover missing partial-group scores or independently reproduce scorer training. Appendix B records policy identities and optimization settings; Appendix D defines the FDB estimands, and Appendix I specifies SID scoring.

Ethics Statement

This work analyzes recorded model outputs and existing conversational datasets. It does not include a new human-subject experiment or a human preference study. Dataset access and redistribution remain subject to the original providers' terms; raw conversational audio is not part of the manuscript evidence package. Automated judgments can misrepresent accents, disfluencies, and conversational styles, so the reported scores should not be interpreted as universal measures of interaction quality.

AI Use Statement

An AI coding assistant was used to draft and edit text, inspect experiment artifacts, implement statistical analyses, and prepare LaTeX tables and figures. Quantitative results are derived from recorded experiment outputs. AI assistance also supported literature retrieval and bibliography preparation.

References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*, 2025.

Yifu Chen, Shengpeng Ji, Zhengqing Liu, Qian Chen, Wen Wang, Ziqing Wang, Yangzhuo Li, Tianle Liang, and Zhou Zhao. Dual-axis generative reward model toward semantic and turn-taking robustness in interactive spoken dialogue models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 173–208, 2026.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. *arXiv preprint arXiv:2205.09812*, 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731, 2015.

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.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. Full-Duplex-Bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19447–19451. IEEE, 2026.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11: 250–266, 2023.

NVIDIA. Parakeet TDT 0.6B V2. Hugging Face model card, 2025. URL `https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2`.

Atsumoto Ohashi, Neil Zeghidour, Alexandre Défossez, and Eugene Kharitonov. Multi-faceted interactivity alignment in full-duplex speech models. *arXiv preprint arXiv:2606.11167*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 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*, 67:101178, 2021.

Yixuan Wang, Yifei Chen, Haichao Zhang, Haozheng Luo, Xander Wu, Jie Ni, Yun Fu, Nuno Vasconcelos, and Yijiang Li. Learn what's left, not what's mastered: Saturation aware advantage reweighting for multi-reward policy optimization. *arXiv preprint arXiv:2608.16072*, 2026.

Kangxiang Xia, Bingshen Mu, Xian Shi, Jin Xu, and Lei Xie. Semantic-aware interruption detection in spoken dialogue systems: Benchmark, metric, and model. *arXiv preprint arXiv:2603.24144*, 2026.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Complete Benchmark Results

Policy	Synth. ↓	Candor ↓	Turn ↑	BC ↓	Lat. ↓	Intr. ↑	Quality ↑
Base	38.7	50.5	73.1	32.7	0.199	92.0	3.500
Rule / Seq	37.2	48.1	68.9	36.4	0.177	92.0	3.435
GRM / Seq	32.8	47.2	68.9	40.0	0.272	90.0	3.339
GRM-Rubric / Seq	37.2	49.1	68.9	40.0	0.241	91.5	3.492
Rule / Event	36.5	50.0	70.6	47.3	0.258	90.5	3.508
GRM / Event	37.2	48.1	69.7	41.8	0.217	89.0	3.287

Table 4: All six policies. Definitions and units match Table 2.

Policy	BC takeover ↓	BC freq.	BC dist. ↓	Intr. latency ↓	Rated %
Base	32.7	0.123	0.731	1.234	92.0
Rule / Seq	36.4	0.113	0.740	1.282	92.0
GRM / Seq	40.0	0.111	0.742	1.351	90.0
GRM-Rubric / Seq	40.0	0.118	0.738	1.464	91.5
Rule / Event	47.3	0.104	0.733	1.232	90.5
GRM / Event	41.8	0.114	0.725	1.511	89.0

Table 5: Backchannel (BC) and interruption metrics. Takeover and rated coverage are percentages; frequency is per second and latency is in seconds. The evaluator's timing distance is the SciPy Jensen–Shannon distance, although the original metric key uses "JSD." Rated coverage is reported separately from response coverage.

Policy vs. Base	Outcome	n	Δ (pp)	95% CI (pp)	p_{Holm}
Rule / Seq	Pause takeover	353	-1.98	[-5.10, +1.13]	1.000
GRM / Seq	Pause takeover	353	-4.25	[-7.37, -1.13]	0.338
GRM-Rubric / Seq	Pause takeover	353	-1.42	[-4.53, +1.42]	1.000
Rule / Event	Pause takeover	353	-1.13	[-4.25, +1.98]	1.000
GRM / Event	Pause takeover	353	-1.98	[-5.10, +1.13]	1.000

Table 6: Paired pooled-pause effects. Negative takeover differences indicate improvement. The 95% intervals are pointwise; p-values are adjusted over the 25 binary comparisons defined in the main text.

All interruption responses receive one unique sample-level outcome. Rule/Event has 182 nonempty outputs, of which 181 satisfy the benchmark's duration-or-chunk-count turn criterion. All 181 eligible responses have ratings; its response rate and rated coverage are both 90.5%. The additional short output is excluded by the same criterion used for every policy. Response and rated counts also coincide for the other five policies. Historical retry-inflated rate fields are retained in the machine-readable evidence beside the recomputed rates.

Seq/Event comparisons. The rule-only change from Seq to Event raises turn response from 68.9% to 70.6% and turn latency from 0.177 to 0.258 seconds. It also raises backchannel takeover from 36.4% to 47.3%. Paired interruption-quality intervals span zero for both Rule/Event minus Rule/Seq and GRM/Event minus GRM/Seq (Table 7). These configurations change masks, normalization denominators, and clipping order together, so they do not isolate temporal credit assignment.

Selected FDB paired contrasts. All three Seq policies have lower pause takeover rates than Base on both subsets (Table 2). GRM/Seq has the largest observed reduction, from 162 to 147 takeovers across 353 pause clips. Twenty-four inputs improve and nine regress, giving a difference of -4.25 percentage points with a paired 95% interval of $[-7.37, -1.13]$. The exact unadjusted p-value is 0.0135, but its Holm-adjusted value is 0.338. None of the 25 binary comparisons remains significant at 0.05 after correction.

At the same time, completed-turn coverage falls from Base's 87/119 responses to 82/119 for each Seq policy. For GRM/Seq, the paired difference is -4.20 percentage points with a 95% interval of

Comparison	Paired n	Difference	95% pointwise CI
Rule/Seq − Base	176	-0.051	$[-0.330, +0.222]$
GRM/Seq − Base	173	-0.116	$[-0.393, +0.162]$
GRM-Rubric/Seq − Base	177	-0.017	$[-0.311, +0.277]$
Rule/Event − Base	173	-0.040	$[-0.295, +0.220]$
GRM/Event − Base	169	-0.243	$[-0.538, +0.053]$
GRM-Rubric/Seq − Rule/Seq	177	$+0.056$	$[-0.209, +0.322]$
Rule/Event − Rule/Seq	174	$+0.063$	$[-0.236, +0.351]$
GRM/Event − GRM/Seq	167	-0.132	$[-0.407, +0.138]$

Table 7: All archived paired interruption-quality comparisons. Differences use jointly rated examples and a 0–5 scale. Intervals are descriptive and pointwise; all include zero. They do not establish equivalence.

$[-10.08, 1.68]$. Figure 2 presents restraint and response coverage together, preventing lower pause takeover from standing in for overall interaction improvement.

Learned rewards do not establish better interruption quality. Rule/Seq retains Base's 92.0% interruption response rate; GRM/Seq and GRM-Rubric/Seq reach 90.0% and 91.5%, respectively. On 177 jointly rated inputs, GRM-Rubric/Seq minus Rule/Seq has a mean quality difference of $+0.056$ with a 95% interval of $[-0.209, 0.322]$. This comparison establishes neither a quality gain nor equivalence between the policies. Backchannel behavior also differs from pause restraint. Seq backchannel takeover rates rise from Base's 32.7% to 36.4–40.0%, while frequency falls from 0.123 to 0.111–0.118 per second (Appendix A).

B Training and Policy Identity

The training configurations use full-model FSDP adaptation, bfloat16, gradient checkpointing, and 16 policy ranks. The five archived completion summaries report 3,200 global rollout groups and 3,200 optimizer updates per policy rank, with 16 candidates per group. Across 16 ranks, this corresponds to 200 local groups per rank and 16 completion-level optimizer steps per local group; synchronized optimizer steps are not multiplied by the rank count. AdamW (Loshchilov & Hutter, 2019) processes completion microbatches of size one, with 16 optimizer steps per local group. The five policies share the same training schedule, initialization seed, and trainable parameter count (7,687,729,152). Table 8 lists the GRM-Rubric/Seq optimization settings; the accompanying machine-readable training contracts record each policy separately.

Parameter	Value
Base policy	Moshika, bfloat16
Reference	Independent frozen pre-RL policy
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$
Learning rate / schedule	2×10^{-7} / cosine
Weight decay / gradient clip	0.1 / 2.0
PPO clip / exact KL coefficient	0.2 / 0.01
Group size / training seed	16 / 799
Maximum context / inclusion probability	30 seconds / 0.5
Rollout audio temperature / top-k	0.8 / 250
Rollout text temperature / top-k	0.7 / 0 (unrestricted)
Policy loss	Generated text-token probabilities

Table 8: Frozen optimization settings for the GRM-Rubric/Seq run. Evaluation text top-k is 25, separately fixed across all policies.

GRM/Seq and GRM/Event use the same reward model, so their comparison holds scorer identity constant while changing the optimization scheme. GRM/Seq and GRM-Rubric/Seq instead compare two reward models under the same Seq scheme. GRM-Rubric is evaluated only with Seq; the study therefore does not estimate its interaction with Event optimization. The accompanying reproducibility manifest links each named policy to its training and evaluation artifacts.

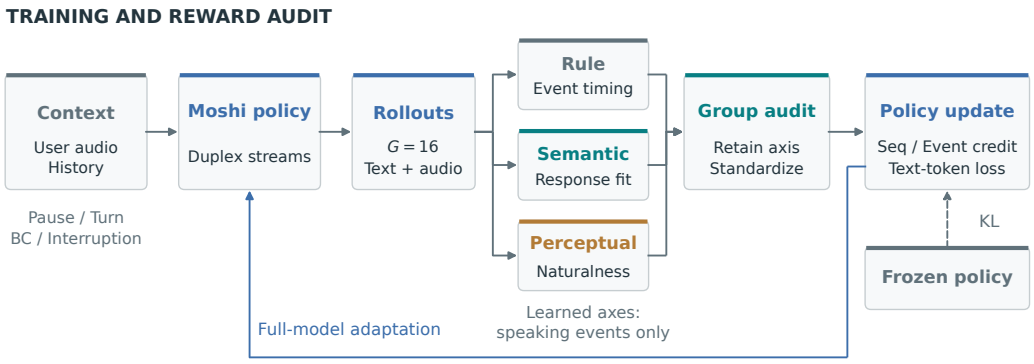


Figure 5: Training and reward audit. Each context produces 16 joint text–audio rollouts. Event annotations define scoring targets and credit windows. A rule scorer measures timing; a fixed learned scorer supplies semantic and perceptual components on speaking events. Retained components are standardized within groups before Seq/Event credit assignment. Optimization uses text-token probabilities, full-model adaptation, and frozen-reference KL. BC denotes backchannel.

C Reward and Credit Settings

Optimization objective. The main-text objective (Equation 2) defines the Event update. The settings below specify its event windows and the Seq comparison.

Seq assigns credit over the generated text sequence. **Event** routes components to event-relevant windows: timing near turn or interruption boundaries, semantic feedback to lexical response positions, and perceptual feedback to speech-aligned text positions, including padding decisions. The component masks are illustrated in Figure 1b; the KL term remains sequence-wide. The policy objective uses text-token probabilities with full-model parameter adaptation. Updating the text timing stream can change when synthesized speech starts and stops. This is a text-stream surrogate driven by joint text–audio outcomes; the probability ratio does not include audio-code likelihoods. Full-model adaptation specifies the trainable scope, not a measured nonzero gradient for every parameter.

The deterministic rule uses speech-activity intervals and event geometry. In a hold segment, speech onset costs 0.5, speech duration costs 0.5 per second, and a takeover longer than 0.8 seconds incurs an additional 0.5 penalty. The turn reward assigns 1.0 to initiation within $[0, 0.8]$ seconds, decays toward zero by 1.5 seconds, and penalizes never speaking by 2.0. Early onset costs 1.5 plus 0.5 per second. Generic responses and audio artifacts have configured costs of 0.5 and 1.0.

For backchannels, the maximum short-utterance duration is 0.8 seconds and the timing tolerance is 1.0 second. A matched event receives 1.0; unmatched events cost 0.5 and takeovers cost 1.0. Interruption scoring uses a 0.16-second yielding grace period, a 1.0-per-second yielding cost thereafter, and a resumption reward of 1.0 in $[0, 0.8]$ seconds after the user's end. Late resumption decays to zero at 1.5 seconds; never resuming costs 2.0. Early resumption costs 1.0 plus 0.5 per second.

The configured event-local windows begin 0.4 seconds before ordinary turn boundaries and extend through the late-response region at 1.5 seconds; the semantic response cap is 2.4 seconds. Backchannel opportunity windows extend 0.4 seconds on either side of the reference, with a 1.6-second takeover endpoint and 0.8-second semantic cap. Interruption windows include 0.24 seconds before yielding, the 0.16-second grace period, and the post-interruption response region. Speech-based perceptual masks include a 0.16-second margin. All five policies use a zero-second offset when mapping decision times to the audio timeline. This shared setting does not establish the accuracy of text-to-audio temporal alignment; the available evidence contains no independent alignment measurement.

Input set	Inputs	Response or event criterion	Uncertainty unit
FDB pauses	353	ASR duration ≥ 1 s or > 3 chunks	Clip
FDB completed turns	119	Same ASR criterion; latency on eligible responses	Clip
FDB backchannels	55	Benchmark voice-activity and ASR rule	Descriptive summary
FDB interruptions	200	Same ASR criterion; quality on rated responses	Clip
SID all inputs	2,100	First audio activity above -45 dBFS for two 0.08-s frames	Source recording
SID positive inputs	1,100	Activity detection, including early responses; IRL uses positive delays	Source recording

Table 9: Evaluation protocols and denominators. SID positive inputs are a subset of all SID inputs, not an additional dataset. Shared-response comparisons further restrict eligibility to a policy pair's intersection. FDB uncertainty uses clip resampling; SID uses source-recording resampling. Neither captures training-seed variation.

D Metric Reconstruction and Statistical Estimands

Table 9 summarizes the input sets, eligibility rules, and resampling units used in the two benchmarks.

Benchmark composition and implementation. FDB v1 contains 137 synthetic pause, 216 Candor pause, 119 Candor turn-taking, 55 ICC backchannel, and 200 synthetic interruption clips. These total 727 inputs per policy. Lower takeover rates are desirable during pauses and backchannels; higher response rates are desirable for completed turns and interruptions. Interruption quality is rated on a 0–5 scale among eligible responses. For pauses, completed turns, and interruptions, eligibility requires automatic speech recognition (ASR) duration of at least one second or more than three chunks. Short or poorly transcribed speech may therefore count as no response (Appendix D).

Decoding and judging are fixed across policies (Appendix D). Transcription uses NeMo Parakeet-TDT-0.6B-v2 (NVIDIA, 2025); the recorded semantic judge is `gpt-5-mini-0807`. FDB v1 was used during development.

SID input composition. Base and the three Seq policies share 1,600 English inputs (1,100 positive, 500 negative) and 500 noise/silence inputs. Decoding settings match FDB. We map first output-audio activity to a detection time and apply the released notebook's scoring formulas. Metrics are average penalty time (APT), false interruption rate (FIR), and interruption response latency (IRL). Appendix I specifies this adaptation and its 903 source recordings.

Decoding and scoring. We prepend three seconds of streaming silence and remove that prefix from the output; annotations retain their original clock. Audio and text temperatures are 0.8 and 0.7, top-k values are 250 and 25, and the decoding seed is 20260729. No trailing silence is appended. Transcription uses NeMo Parakeet-TDT-0.6B-v2 (NVIDIA, 2025), and the recorded semantic judge is `gpt-5-mini-0807`. This scoring pipeline is fixed across policies; FDB v1 was used during development.

Take-turn decisions. For pause, turn, and interruption subsets, the v1 evaluator treats an empty ASR chunk list as no turn. Otherwise, it measures duration between the first chunk start and the final chunk end. A response counts as a turn when duration is at least one second or there are more than three chunks. For a null final end timestamp, the pause evaluator uses the final chunk start. Backchannel takeover uses a separate voice-activity and ASR rule; its aggregate values are taken from the recorded evaluator output, not from the nonempty-transcript indicator.

Latency. Turn latency subtracts the turn annotation's first timestamp from the first output chunk start. Interruption latency subtracts the interrupting utterance's end. Negative differences are clipped to zero for the reported benchmark means, and raw differences are retained in the evidence. Nonre-

sponses have undefined latency. Recomputed aggregate turn latencies and rated interruption latencies agree with the stored evaluator values for all six policies.

Coverage and quality. The interruption rate is the number of unique examples satisfying the turn criterion divided by 200. Semantic quality is the mean available judge rating among eligible responses. A missing rating is not a zero-quality judgment. Paired quality uses only examples with ratings for both compared policies. Paired latency uses only examples eligible for both policies. Consequently, a paired latency difference need not equal the difference of the two aggregate means. The jointly eligible set is specific to each policy pair; it excludes responses gained or lost by either policy. These conditional comparisons should not be interpreted as an overall quality ranking across all six policies.

Response-selection decomposition. To separate timing from response membership, let A and B be two policies' response sets, with shared set $C = A \cap B$. Define $U_A = A \setminus B$ and $U_B = B \setminus A$, and let $\bar{\ell}_{A,C}$ denote policy A's mean latency on C. For nonempty C, the aggregate difference decomposes exactly as

$$\bar{\ell}_A - \bar{\ell}_B = (\bar{\ell}_{A,C} - \bar{\ell}_{B,C}) + \frac{|U_A|}{|A|}(\bar{\ell}_{A,U_A} - \bar{\ell}_{A,C}) - \frac{|U_B|}{|B|}(\bar{\ell}_{B,U_B} - \bar{\ell}_{B,C}). \tag{3}$$

The first term measures the shared-response difference; the remaining terms account for response-set composition. An empty exclusive set contributes zero; its mean need not be defined.

Equation 3 follows by writing each policy's aggregate mean as a response-count-weighted average over its shared and exclusive sets. The shared set must be nonempty; empty exclusive sets contribute zero. For Rule/Seq versus Base on completed turns, $|A| = 82$, $|B| = 87$, $|C| = 80$, $|U_A| = 2$, and $|U_B| = 7$. Shared-set means are 181.25 and 166.75 milliseconds. The exclusive-set means are 0 and 568.57 milliseconds, respectively, under the evaluator's zero-clipped latency convention. The shared-set difference is $+14.50$ milliseconds and the composition contribution is -36.75 milliseconds. Their sum, -22.25 milliseconds, equals the aggregate difference. Exclusive-set means are descriptive summaries of small selected sets, not estimates of a population effect.

Uncertainty. For paired vectors x and y, we form $d_i = x_i - y_i$, sample n indices with replacement 20,000 times using NumPy's default generator initialized with 20260916, and take the 2.5th and 97.5th percentiles of the resampled means. For binary outcomes, the exact McNemar test is the two-sided binomial test on discordant pairs under success probability $1/2$. Holm adjustment uses a family of 25 comparisons: pooled pauses, synthetic pauses, Candor pauses, completed turns, and interruptions for each RL policy against Base. Continuous-outcome intervals are descriptive pointwise intervals. All calculations condition on the existing generated outputs. Endpoints are pointwise, and the family of comparisons was specified retrospectively. Clip-level resampling treats clips as exchangeable; without conversation-level source keys, repeated-source dependence remains unaccounted for.

E Implementation Diagnostics for Learned Rewards

These diagnostics document an implementation limitation of the recorded learned-reward pipeline. They are not a proposed method or evidence that aggregation caused the benchmark behaviors.

E.1 Audit Scope

Here, "abstention" denotes recorded score absence, including potentially inapplicability, explicit abstention, and parsing or service failure. The archive does not distinguish these causes. We inspect 1,600 rollout groups from eight of the GRM-Rubric/Seq run's 16 policy ranks. These contain 25,600 candidates across 404 pause, 399 turn, 396 backchannel, and 401 interruption groups. The audit follows three quantities: candidate scores supplied, scores retained after aggregation, and groups retaining nonconstant scores. These measurements describe the archived ranks and do not estimate score correctness or gradient contribution.

Event	Groups	Sem. retained (%)	Sem. varying (%)	Perc. varying (%)
Turn	399	9.3	9.3	82.7
Backchannel	396	12.4	9.8	18.4
Interruption	401	16.2	16.2	77.3
All speaking	1196	12.6	11.8	59.6

Table 10: Reward availability in GRM-Rubric/Seq training. Sem. and Perc. denote semantic and perceptual components. Retention requires a complete scored group; varying additionally requires nonconstant scores. Perceptual retention is 100% for these events. Denominators are all groups of the corresponding event type.

E.2 Aggregation Deletes Available Scores

The scorer abstains on 25.3% of turn candidates, 38.8% of backchannel candidates, and 14.2% of interruption candidates. Because any abstention deletes the group's semantic component, only 37/399, 49/396, and 65/401 groups retain semantic rewards, respectively (Table 10). Across speaking events, 4,981 of 19,136 candidates abstain, while 1,045 of 1,196 groups lose the component. Thus a 26.0% candidate-level abstention rate translates into 87.4% group-level removal. More directly, the scorer supplies $19{,}136 - 4{,}981 = 14{,}155$ semantic scores, but only $151 \times 16 = 2{,}416$ survive in complete groups: aggregation discards 11,739 available scores (82.9%). The denominator excludes the 404 pause groups, where semantic feedback is intentionally inapplicable.

Retention and contrast are separate checks. Ten of the 49 complete backchannel groups have constant semantic scores, leaving 39/396 with semantic contrast. Across speaking events, $37 + 39 + 65 = 141$ of 1,196 groups (11.8%) retain nonconstant semantic scores. Perceptual scores are retained in every speaking-event group, yet only $330 + 73 + 310 = 713$ groups (59.6%) have nonconstant perceptual scores. In particular, 323/396 backchannel groups have constant perceptual scores (Table 10; Appendix Figure 6). The components therefore differ in both retention and contrast despite equal configured weights. A separate 24-source audio probe also produces saturated perceptual scores (Appendix G). Nonconstant scores are necessary for nonzero component advantages, but scoring validity, component cancellation, clipping, and token masks still affect the update. The 82.9% deletion rate measures score entries, not independent information or lost gradient magnitude.

E.3 What Frozen Candidates Reveal About Alternative Aggregation

Retention depends on the group requirement. For an archived group g with k_g scored candidates, uniformly selecting m of its 16 candidates retains the component with probability

$$q_g(m) = \frac{\binom{k_g}{m}}{\binom{16}{m}}, \qquad \binom{k_g}{m} = 0 \text{ when } k_g < m. \tag{4}$$

Averaging over archived groups preserves their observed within-group dependence, without assuming independent abstentions. Reducing the subgroup size from 16 to two increases expected retention from 9.3% to 58.5% for turns. The corresponding changes are 12.4% to 45.6% for backchannels and 16.2% to 74.4% for interruptions. These are exact retention probabilities on frozen candidates, not results from training with smaller groups (Appendix F).

Mean imputation preserves ordering but changes scale. Mean imputation keeps the original group size and fills missing scores with the mean of the k scored candidates. Missing candidates receive zero semantic advantage; scored candidates retain their ordering. Above the standard-deviation threshold, their advantage magnitudes scale by $\sqrt{G/k}$ relative to normalization over the scored subset. However, averaging over all G candidates reduces the unclipped mean gradient contribution by $\sqrt{k/G}$ relative to averaging over that subset. This identity assumes unit probability ratios and identical token normalization (Appendix F). Thus, retaining scores and retaining update magnitude are distinct questions even on fixed candidates.

Event	$m = 2$	$m = 4$	$m = 8$	$m = 16$	At least one score
Turn	58.5	38.8	20.7	9.3	100.0
Backchannel	45.6	32.0	21.1	12.4	98.2
Interruption	74.4	57.1	35.7	16.2	100.0

Table 11: Frozen-candidate subgroup analysis: expected percentage of groups retaining semantic feedback under complete-group deletion. The last column gives the percentage of original 16-candidate groups with at least one semantic score. These are reward-availability measurements.

Availability counts do not identify restored contrast. The archive retains missingness indicators but omits numeric scores from partial groups. Consequently, imputation could yield between 141 and 1,165 nonconstant semantic groups, compared with 141 under deletion (Appendix Table 12). The number of additional groups could therefore be zero. The interval separates recoverable score availability from unobserved score variation; it is not a confidence interval or a prediction of improvement. Mean imputation also leaves quality-dependent missingness unresolved. A matched deletion-versus-imputation training experiment remains outstanding, with its required validation specified in Appendix H.

F Subgroup Retention and Imputation Derivation

Exact subgroup retention. Equation 4 counts uniformly selected subgroups of size m from each original group of 16 candidates. Of the $\binom{16}{m}$ possible subgroups, exactly $\binom{k_g}{m}$ contain only scored candidates. Averaging over archived groups preserves their empirical within-group dependence of abstention. The calculation keeps generated candidates and missingness indicators fixed.

For turn events, retention falls from 58.5% at $m = 2$ to 9.3% at $m = 16$. Interruption retention falls from 74.4% to 16.2%; backchannel retention falls from 45.6% to 12.4%. For $1 \le m \le k_g < G$, $q_g(m+1)/q_g(m) = (k_g - m)/(G - m) \le 1$, so retention cannot increase with subgroup size under this rule. These probabilities describe aggregation over the fixed rollouts.

Mean imputation preserves available comparisons and changes their scale. An alternative aggregation fills abstaining candidates with the mean of the k scored candidates. If their mean and population variance are μ_k and σ_k^2, the completed group has mean μ_k and variance $(k/G)\sigma_k^2$. For nonzero variance, its normalized advantage is

$$A_i^{\text{impute}} = \begin{cases} \sqrt{G/k}\,(r_i - \mu_k)/\sigma_k, & i \text{ scored}, \\ 0, & i \text{ abstained}. \end{cases} \tag{5}$$

The formula applies above the standard-deviation threshold. It preserves the order of scored candidates but increases their individual advantage magnitude relative to normalization over the scored subset alone. This is not a guarantee of a larger aggregate gradient: the optimizer still averages over all G candidates, and abstentions may depend on response quality. All-abstaining groups remain absent; a single scored candidate produces a constant completed group and hence zero advantage.

For mean imputation, the completed sum is $k\mu_k + (G - k)\mu_k = G\mu_k$. The unscored candidates add zero squared deviation, so the population variance is

$$\frac{1}{G}\sum_{i \in S}(r_i - \mu_k)^2 = \frac{k}{G}\sigma_k^2, \tag{6}$$

which yields Equation 5. In our implementation, standard deviations below 10^{-6} produce zero advantage. A group with one scored candidate, or multiple equal scores, therefore supplies zero advantage after imputation.

The archived rollouts retain candidate abstention indicators and numeric scores for complete groups. Partial groups retain abstention indicators; their pre-aggregation numeric scores were omitted from the training logs. The subgroup analysis and the count of imputation-eligible groups use these recorded indicators. Advantage scaling follows analytically from Equation 5.

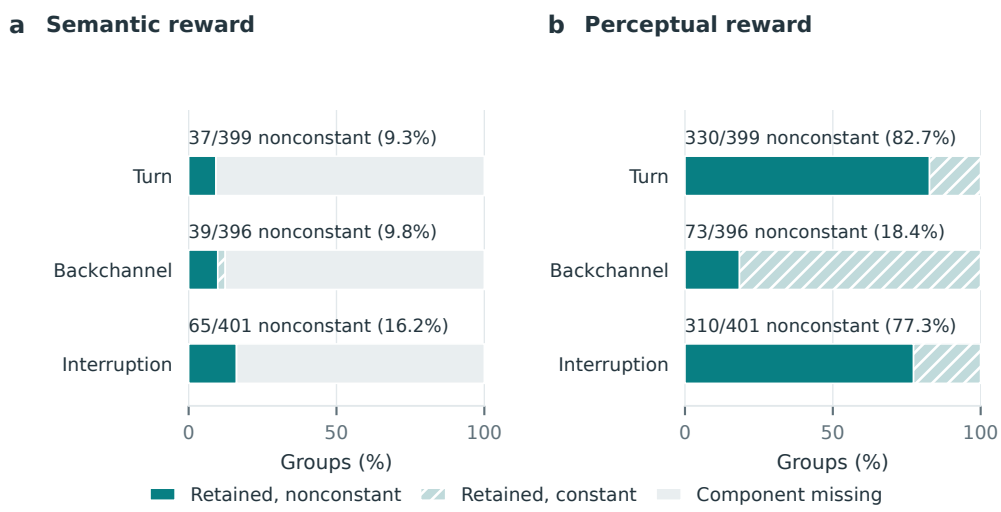


Figure 6: Measured component availability and within-group variation, using the same event-wise denominators as Table 10. Each bar partitions all groups into missing, retained constant, and retained nonconstant components; hatching marks constant scores.

Identifiable bounds on restored contrast. Let H_j count groups with j scored candidates and let C count complete groups known to have nonzero score variance. Under the same numerical standard-deviation threshold, the recorded complete groups remain unchanged by imputation. Groups with zero or one score cannot become nonconstant. A partial group with at least two scores may or may not have nonzero variance, because its scores were not logged. Consequently,

$$C \le N_{\text{nonzero}}^{\text{impute}} \le C + \sum_{j=2}^{G-1} H_j. \tag{7}$$

Here $C = 141$, $H_0 = 7$, $H_1 = 14$, and ten complete groups have constant semantic scores. The resulting range is $[141, 1165]$, or 11.8–97.4% of the 1,196 speaking-event groups. The number of additional nonzero groups lies in $[0, 1024]$. These are deterministic identification bounds conditional on the archived counts and original scores, not confidence intervals or predicted training gains.

Event	Groups	Complete, varying	At least two scored	Imputation range
Turn	399	37	397	[37, 397]
Backchannel	396	39	377	[39, 367]
Interruption	401	65	401	[65, 401]
Total	1196	141	1175	[141, 1165]

Table 12: Semantic signal under mean imputation. "Complete, varying" is observed under deletion; "at least two scored" is a necessary but insufficient condition for contrast. The range bounds the total nonzero groups after imputation and excludes known complete-group ties.

Normalization and update scale. When the standard-deviation threshold is inactive, imputed advantages satisfy $\sum_i A_i = 0$ and $G^{-1}\sum_i A_i^2 = 1$. Individual scored advantages grow by $\sqrt{G/k}$, but averaging their contributions over G rather than k changes the scale in the opposite direction. At unit probability ratios, with identical token normalization and no clipping, the mean reward-gradient contribution equals $\sqrt{k/G}$ times the mean over the k scored candidates normalized within that subset. This algebra concerns a fixed group. It does not remove bias from quality-dependent abstention, model shared parameters, or subsequent policy changes.

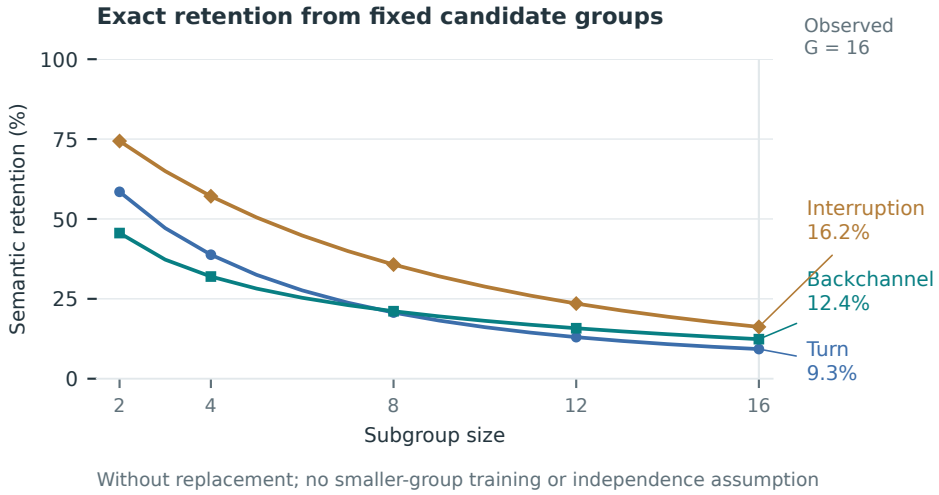


Figure 7: Exact reward-retention probability as a function of subgroup size, using the same candidates and abstention decisions. Increasing subgroup size decreases complete-group retention under the complete-group deletion rule.

G Perceptual Score Saturation Diagnostic

The archived diagnostic compares GRM and GRM-Rubric on the same 24 source items with eight audio variants per item: clean audio, a duplicate, micro-dropout, six-bit mu-law, one-bit quantization, polarity inversion, rate scaling, and soft clipping. All 192 items receive a valid perceptual score in each run. GRM assigns 5/5 to 191 items and 1/5 to one clean item. GRM-Rubric assigns 5/5 to all 192. Its seven within-source clean-versus-variant contrasts are consequently all ties. These counts are recomputed from the archived paired-score CSV.

Condition	Valid pairs	GRM at 5/5	GRM-Rubric at 5/5
Clean	24	23	24
Duplicate	24	24	24
Micro-dropout	24	24	24
Six-bit mu-law	24	24	24
One-bit quantization	24	24	24
Polarity inversion	24	24	24
Rate scaling	24	24	24
Soft clipping	24	24	24
Total	192	191	192

Table 13: Maximum perceptual scores in the paired probe. Every condition reuses the same 24 source items. Counts describe scorer outputs; no human reference rating or perceptual ordering is assumed.

This establishes score saturation on the recorded probe. It does not establish that every transformation worsens perceived quality: polarity inversion can preserve perception, and rate changes need not be objectionable. It also does not distinguish insufficient perturbation, scorer insensitivity, or audio-path failure in those historical runs. The probe was used for development, has only 24 source items, and has no independent listening validation or certified held-out split. Its 192 conditions must not be treated as 192 independent sources. Stage-specific degradation or recovery would require matched intermediate checkpoints and verified audio consumption in addition to this output comparison.

H Requirements for a Matched Intervention

The following is a proposed experiment, not a completed result. First, verify audio consumption and same-context semantic ranking, and distinguish inapplicability, absent speech, explicit abstention, and parsing or service failure where observable. Archive every pre-aggregation score and compare deletion with semantic-only mean imputation on the same candidates. Record restored contrast, advantages, clipping, component cancellation, and mask sizes.

A fresh deletion control and imputation arm should share the base policy, GRM-Rubric scorer, data schedule, group size, Seq objective, and training budget. Repeated paired seeds and a fixed checkpoint-selection rule are needed beyond a historical control. Evaluation should include independent inputs, response coverage, content quality, timing, and sensitivity to speech-activity and ASR thresholds. Human checks should include gained and lost responses. Increased contrast without improved behavior would show that score deletion alone is insufficient to explain the policy's limitations.

I SID-Bench Scoring and Paired Results

Aggregate and coverage contrasts. GRM-Rubric/Seq has the lowest observed overall APT and FIR (Table 3). Relative to Base, false triggers fall from 80 to 70, a -0.476-point change with a pointwise source-cluster 95% interval of $[-0.927, -0.047]$. Its APT change is -0.036 seconds, with interval $[-0.155, 0.056]$. All three RL policies have lower noise/silence APT but higher English APT. The aggregate therefore combines different changes across scenes.

Positive response coverage falls from 28.82% for Base to 25.27–26.00% for the RL policies. On shared positive-delay responses, their latency is 0.229–0.264 seconds higher than Base's. Each coverage-change interval lies below zero, and each shared-latency interval lies above zero (Figure 4; Appendix I). Thus, the lower false-trigger counts accompany fewer and later positive responses in these fixed checkpoints. Neither this co-occurrence nor the aggregate penalty establishes an unavoidable trade-off or a benefit from learned rewards over rules alone.

Inputs and scoring. The 2,100 inputs represent 903 source recordings: 403 English recordings and 500 noise/silence recordings. Output activity uses a -45 dBFS threshold with at least two consecutive 0.08-s active frames. The first activity onset, after removing the three-second prefix, is the predicted detection time. This adapts a speech-generating policy to SID's detection task; it does not use a separate interruption classifier.

We use the released notebook formulas. A prediction within 50 ms of the annotated time has zero penalty. Otherwise, an early positive trigger or a trigger on a negative input incurs the full annotated duration; a missed positive incurs the remaining duration after the event; a late positive incurs its delay. A silent negative has zero penalty. FIR counts false triggers over all inputs, and IRL averages strictly positive delays. Positive response coverage counts all positive inputs with detected speech, including early speech. The released README's prose differs from the notebook for FIR and IRL; our results use the latter consistently.

All intervals use 20,000 paired source-recording bootstrap resamples (seed 20260917), preserving each recording's clips across policies. They are pointwise percentile intervals for fixed checkpoints and decoding, without multiplicity adjustment. Table 14 reports all six overall comparisons. Learned-reward policies have lower APT point estimates than Rule/Seq, but both APT and FIR intervals include zero in these comparisons.

For Rule/Seq, GRM/Seq, and GRM-Rubric/Seq versus Base, respectively, shared positive-delay sets contain 262, 255, and 257 inputs. Their latency changes are $+0.229$ s $[0.065, 0.418]$, $+0.264$ s $[0.113, 0.443]$, and $+0.232$ s $[0.080, 0.409]$. Positive response coverage changes by -2.818 pp $[-4.320, -1.383]$, -3.545 pp $[-4.992, -2.128]$, and -3.182 pp $[-4.729, -1.708]$.

English APT increases by 0.053, 0.075, and 0.054 s, respectively; the corresponding intervals are $[0.015, 0.094]$, $[0.041, 0.114]$, and $[0.021, 0.090]$. Noise/silence APT changes by -0.143, -0.360, and -0.323 s, with intervals $[-0.397, 0.083]$, $[-0.741, -0.032]$, and $[-0.823, 0.044]$. Full-duration penalties make APT sensitive to long clips: removing one false trigger on a 91.263-s noise input contributes more than GRM-Rubric/Seq's net 75.271-s total penalty reduction. Counts and scene-

Comparison	ΔAPT (s), 95% CI	ΔFIR (pp), 95% CI
Rule/Seq − Base	$+0.006$ $[-0.061, +0.068]$	-0.429 $[-0.873, +0.000]$
GRM/Seq − Base	-0.028 $[-0.125, +0.055]$	-0.333 $[-0.815, +0.140]$
GRM-Rubric/Seq − Base	-0.036 $[-0.155, +0.056]$	-0.476 $[-0.927, -0.047]$
GRM/Seq − Rule/Seq	-0.034 $[-0.115, +0.031]$	$+0.095$ $[-0.234, +0.420]$
GRM-Rubric/Seq − Rule/Seq	-0.042 $[-0.160, +0.049]$	-0.048 $[-0.333, +0.234]$
GRM-Rubric/Seq − GRM/Seq	-0.007 $[-0.127, +0.091]$	-0.143 $[-0.488, +0.189]$

Table 14: SID paired differences (first policy minus second). Intervals resample source recordings. pp: percentage points.

specific results therefore help interpret the aggregate. The evidence package contains all 8,400 policy–input scores and the archived audit, including the evaluation protocol and artifact provenance.