FORECASTING BEFORE THE OUTCOME: CATALOGUE-GATED PREQUENTIAL COGNITIVE PERFORMANCE UNDER PARTICIPANT AND DOMAIN SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Online assistance needs a performance forecast before the response and a principled way to decide whether source knowledge transfers to the current participant. We introduce the Support-Aware Prequential Forecaster (SAPF), which uses a fixed catalogue of supported experiment identifiers to retain source predictions when transfer is expected and switches to hierarchical Beta adaptation for unseen tasks. Every route decision and probability is committed before the current response; feedback updates only subsequent trials. On an external auditory cohort with 13 participants and 28,452 trials, SAPF achieves NLL 0.066 and AUROC 0.929, while immediate source exclusion increases first-five-trial Brier by 0.039 relative to retaining the source. A feedback-weighted SafeMix baseline is useful when catalogue membership is misleading, reducing macro Brier from 0.113 to 0.071 under identifier conflict. Two-expert and temporal ablations show that the main gain comes from the route decision and from recovering source weight after transient failures. In a prefix-causal replay with a 10% budget cap, SAPF covers 42.2% of errors at a 5.8% call rate, compared with 9.0% coverage at a 9.5% call rate for random allocation. These results position pre-response forecasting as an explicit transfer-control layer: retain source knowledge when catalogue support holds, fall back at unsupported cold starts, and update only when later feedback justifies a change. SAPF thus provides a simple way to decide, before each response, whether source knowledge should be used or replaced by online evidence.

1 INTRODUCTION

Cognitive monitoring combines task behavior with signals such as pulse-wave morphology, electrodermal activity, and movement (Fekri Azgomi et al., 2023; Khazaei & Faghih, 2024; Abbaspourazad et al., 2023; Narayanswamy et al., 2025; Pillai et al., 2025). An online assistant faces a more specific decision: whether to intervene before the learner responds. It needs a probability computed from evidence available at the intervention time. A posterior that uses the current response, its reaction time, or a later physiological window can describe a completed trial, but cannot guide an action on that trial. Prospective risk and retrospective state reconstruction are different estimands and require different comparison protocols.

Knowledge-tracing models learn from response histories (Piech et al., 2015; Pandey & Karypis, 2019; Ghosh et al., 2020; Liu et al., 2022; 2023a; Yang et al., 2023), whereas physiological state-space models reconstruct evolving latent states (Khazaei & Faghih, 2024). Comparing them requires the same pre-response information set. Transfer adds a second uncertainty: target participants may encounter tasks or stimuli absent from source training. The first prediction must decide whether to reuse the source model while balancing rapid local adaptation against population-level pooling. Test-time adaptation can update from the arriving stream (Wang et al., 2020; 2022a; Niu et al., 2023; Zhao et al., 2023), but it does not remove this initial transfer decision.

The central question is when source knowledge should give way to online feedback. SAPF uses task metadata to make that decision before the first target response. Represented experiment identifiers retain the source predictor; unseen identifiers activate hierarchical Beta adaptation. SafeMix offers a complementary route when membership does not predict transferability: past predictive loss

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

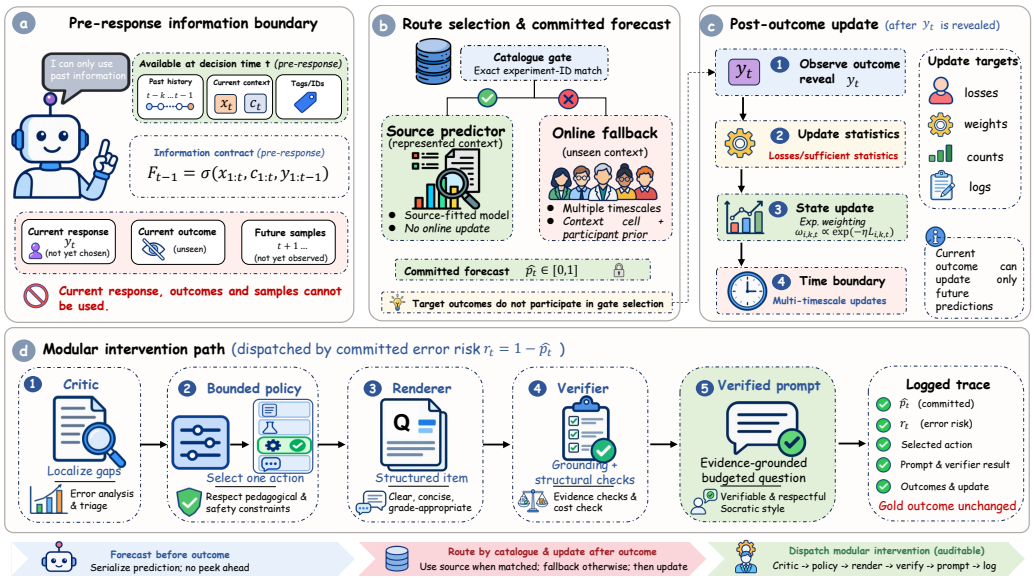


Figure 1: Source-catalogue routing, post-response updating, and the modular intervention path.

determines the source weight. This distinction separates immediate source exclusion from learned source reuse. We examine both decisions through cold-start tests, route interventions, and source–Context Beta ablations. Within-stream perturbations then test how quickly source weight recovers when reliability returns.

A forecast determines when assistance may be useful; a Socratic Agent determines what question to ask. The Agent localizes a reasoning span, selects one bounded pedagogical move, and checks the question and trace. Figure 1 illustrates the resulting evaluation path. Forecasting, localization, and human response are evaluated on their respective task streams. We make three contributions:

- A pre-response forecasting formulation with testable temporal invariance and a regret decomposition separating fallback adaptation from route disagreement.
- Complementary catalogue and feedback-based routes that protect unseen-task cold starts and retain transferable source knowledge under metadata mismatch.
- A comparative evaluation of source retention, suppression, and recovery, with two-expert baselines, temporal perturbations, and prefix-causal budgeted allocation.

2 Related Work

Cognitive-state and performance modeling. Wearable studies infer cognitive state from PPG, EDA, motion, and task behavior through physiological state-space models and learned representations (Fekri Azgomi et al., 2023; Khazaei & Faghih, 2024; Spathis et al., 2021). Recent foundation models scale this idea across participants, modalities, optical physiology, and motion, with growing attention to personalization and missing sensors (Abbaspourazad et al., 2023; Narayanswamy et al., 2025; Pillai et al., 2025; Chen & Yang, 2025; Xu et al., 2025; Wu et al., 2025). A parallel literature predicts performance from response histories: DKT and its attentive or context-aware successors model longer learning trajectories (Piech et al., 2015; Pandey & Karypis, 2019; Ghosh et al., 2020; Wong et al., 2022), while standardized benchmarks and controlled baselines reveal sensitivity to preprocessing, leakage, and architecture choice (Liu et al., 2022; 2023a;b; Yang et al., 2023). Large time-series models pursue broader transfer through pretraining (Goswami et al., 2024; Ansari et al., 2024). These lines improve representation and state reconstruction; our concern is the narrower prediction contract required when a probability must be available before the response that an intervention could change.

Prequential adaptation under shift. Test-time adaptation learns from an incoming stream through entropy objectives, augmentation, teacher updates, anti-forgetting, reliability weighting, or a limited

label budget (Wang et al., 2020; Zhang et al., 2022; Wang et al., 2022a; Niu et al., 2022; 2023; Lee et al., 2024; Yang et al., 2024; Gui et al., 2024). Source-free time-series methods further reduce target-side optimization and storage (Patel et al., 2025), whereas dynamic-stream studies show that temporal ordering, hyperparameters, and collapse can dominate the outcome (Yuan et al., 2023; Zhao et al., 2023; Press et al., 2023). Label-shift methods estimate target priors from calibrated predictions (Lipton et al., 2018; Wen et al., 2024). Our setting differs because labels do arrive, but only after their forecasts have been committed. This is the prequential object studied by Dawid (1984): proper scores assess the probabilities that were actually issued (Glenn et al., 1950), calibration complements ranking (Guo et al., 2017; Marx et al., 2023), and deployment loss determines which correction matters (Sun et al., 2023; Amoukou et al., 2024). Exponential weighting supplies the fixed-expert comparator used here (Freund & Schapire, 1997; Cesa-Bianchi et al., 2006); specialist and tracking formulations instead accommodate expert availability or a changing best expert (Freund et al., 1997; Herbster & Warmuth, 1998). SAPF makes source availability explicit through a fixed metadata catalogue, while SafeMix addresses the different case in which source reliability must be learned from delayed outcomes.

Socratic question generation. Socratic question generation turns an identified reasoning gap into a prompt that preserves the learner's responsibility for the answer (Ang et al., 2023; Pothong et al., 2025); recursive Socratic decomposition has also been used to organize LLM reasoning (Qi et al., 2023). Sampling agreement (Wang et al., 2022b) and process verification (Lightman et al., 2024) improve reasoning selection, while self-feedback (Madaan et al., 2023; Shinn et al., 2023), consistency checks (Manakul et al., 2023), tool-mediated critique (Gou et al., 2024), and explicit verification chains (Dhuliawala et al., 2024) inspect generated answers. External grounding remains important because self-correction alone is unreliable (Huang et al., 2024). ProofWriter and EntailmentBank expose evidence membership and proof order for direct verification (Tafjord et al., 2021; Dalvi et al., 2021). The Agent studied here binds these ideas to a committed forecast: forecasting determines when assistance is offered, while grounding and structural checks determine what the intervention may say.

3 Catalogue-Gated Prequential Forecasting

3.1 Prediction contract

For participant i, context c_t, pre-response covariates x_t, and correctness $y_t \in \{0, 1\}$, define

$$\mathcal{F}_{t-1} = \sigma(x_{1:t}, c_{1:t}, y_{1:t-1}), \qquad \hat{p}_t = f(\mathcal{F}_{t-1}). \tag{1}$$

Current-response reaction time, y_t, and future samples are excluded from the information set. Each forecast is recorded and scored before the state update. With T_i trials for participant i, the primary risk is participant-balanced Brier,

$$\widehat{R}_{\mathrm{PB}}(f) = \frac{1}{P} \sum_{i=1}^{P} \frac{1}{T_i} \sum_{t=1}^{T_i} (\hat{p}_{i,t} - y_{i,t})^2, \tag{2}$$

so session length does not reweight the participant population.

Let $S(c) \in \{0, 1\}$ denote membership in the fixed source catalogue. The catalogue contains the `experiment` identifiers observed during source training; $S(c_t) = 1$ when the current identifier matches an entry. Missing or unseen identifiers take the fallback route. This deterministic rule establishes metadata support. SafeMix estimates source reliability from delayed outcomes when transferability is uncertain.

3.2 Two causal routes

If $S(c_t) = 1$, SAPF returns the source-fitted prediction p_t^{src}. If $S(c_t) = 0$, expert $k \in \mathcal{K}$ predicts

$$p_{t,k} = \frac{s_{i,c,t-1} + \kappa_k \bar{p}_{i,t-1}}{n_{i,c,t-1} + \kappa_k}, \quad \mathcal{K} = \{0.25, 1, 5, 20, 100\}, \tag{3}$$

where $s_{i,c,t-1}$ and $n_{i,c,t-1}$ summarize past outcomes in the participant–context cell, and $\bar{p}_{i,t-1}$ is a participant posterior mean. Small κ adapts rapidly; large κ pools aggressively.

The participant mean uses a $\mathrm{Beta}(1,1)$ prior,

$$\bar{p}_{i,t-1} = \frac{s_{i,t-1}+1}{n_{i,t-1}+2}. \tag{4}$$

This prior initializes every expert at probability $1/2$. Context keys follow the recorded task schema: n-back level and block in the source task, the corresponding task-level field in development cohorts, and stimulus class in the auditory cohort. Weights begin uniformly and states reset between participants; dataset-specific mappings and numerical clipping are detailed in the supplement.

Before y_t is observed,

$$\hat{p}_t = \sum_k w_{i,k,t-1} p_{t,k}, \quad w_{i,k,t-1} = \frac{\exp(-\eta L_{i,k,t-1})}{\sum_j \exp(-\eta L_{i,j,t-1})}, \tag{5}$$

with cumulative past Brier loss $L_{i,k,t-1}$ and $\eta = 0.25$, selected on the development cohorts. After y_t arrives, losses, weights, and counts update. Weights are shared across a participant's unseen contexts, but the current outcome cannot affect its own prediction.

Each step follows route selection, forecast issuance, scoring, and state update. Supported trials leave fallback weights unchanged, while the prediction log retains the selected route, expert probabilities, and weights. The source predictor is a logistic–HGB ensemble selected on participant-disjoint development data. It uses the current task context, trial index, lagged correctness, and response-time summaries at participant, session, block, n-back, and action scales. Current responses, future observations, and test-participant labels are excluded. Complete optimization settings appear in the supplement.

3.3 Relationship to SafeMix

SafeMix retains the source expert alongside participant Beta, hierarchical context Beta, EMA, and rolling-context experts in every context. If $q = (\rho, (1-\rho)/(K-1), \ldots)$ is the fixed initial distribution, then

$$w_{i,k,t} \propto q_k \exp(-\eta L_{i,k,t}), \qquad \hat{p}_t^{\mathrm{mix}} = \sum_k w_{i,k,t-1} p_{t,k}. \tag{6}$$

SAPF fixes source eligibility from metadata before the first target outcome, whereas SafeMix learns source reliability from delayed feedback. The former switches immediately at a catalogue boundary; the latter can retain a transferable source estimator under metadata mismatch.

The two-expert ablation retains source and Context Beta, with $q = (1/2, 1/2)$ and $\eta = 0.5$. A constant 50:50 mixture isolates the value of weight adaptation. A standard fixed-share comparator applies $w_{t+1} \leftarrow 0.99\,\widetilde{w}_{t+1} + 0.01q$ after each exponential update, preserving weight for subsequent source recovery. These supplementary settings are fixed across evaluated streams and separate expert composition, loss-based weighting, and weight renewal.

3.4 Evidence-grounded intervention agent

Benchmark annotations provide the reference error type and evidence span used only for evaluation. FOCUS and DeepSeek-generated outputs supply candidate predictions. The Critic selects an error type and span, and a bounded policy then requests evidence, exposes an assumption, asks for a counterexample, or abstains. The selected action is rendered as one Chinese-language Socratic item. Verification compares the selected span and action with the item evidence and proof structure; items that fail are regenerated or withheld. Pre-response risk controls dispatch but cannot alter the Critic's evidence or revise the committed forecast.

Online allocation uses risk $r_t = 1 - \hat{p}_t$ and a prefix budget: select trial t only if $r_t \geq \tau_t$ and the previous call count is below $\lfloor bt \rfloor$. One rule fixes τ_t from source-development risks; another uses the $(1-b)$ quantile of risks strictly before t, initialized by that threshold. Neither uses the final horizon, future scores, or current outcome. Full-stream top-b selection is retained only as an offline reference. Historical labels measure error coverage, not the effect that an actual intervention would have on the response.

3.5 GUARANTEES AND COMPLEXITY

The catalogue decision depends only on source metadata and c_t; the experts and their weights depend on statistics through $t-1$. With participant states reset between streams, every stored SAPF forecast is measurable with respect to $\mathcal{F}_{t-1}$.

For participant i, let $U_{i,T}$ be the fallback trials and let P_U count participants with a nonempty fallback route. The exp-concavity of Brier loss gives, for $0 < \eta \leq 1/2$,

$$\sum_{i:U_{i,T}\neq\varnothing} \left[\sum_{t\in U_{i,T}} \ell_t(\hat{p}_t) - \min_k \sum_{t\in U_{i,T}} \ell_t(p_{t,k}) \right] \leq \frac{P_U \log K}{\eta}. \tag{7}$$

This standard exponential-aggregation bound controls excess loss relative to one fixed shrinkage expert per participant. The experiments use $\eta = 0.25$ selected on development data; the SafeMix extension appears in the supplement.

If I_T and U_T denote supported and unseen trials, cumulative loss decomposes as $L_T^{SAPF} = L_{I_T}^{\mathrm{src}} + L_{U_T}^{\mathrm{EWA}}$. For each participant, let k_i^* be the best fixed shrinkage expert on SAPF's realized fallback set, using a prespecified grid member if that set is empty. Comparator $Q = (S^*, k^*)$ uses the source predictor when $S^*(c_t) = 1$ and p_{t,k_i^*} otherwise. With $M_{01} = \{t : S = 0, S^* = 1\}$ and $M_{10} = \{t : S = 1, S^* = 0\}$, define

$$D_T(Q) = \sum_{t\in M_{01}} [\ell_t(p_{t,k_i^*}) - \ell_t(p_t^{\mathrm{src}})]_+ + \sum_{t\in M_{10}} [\ell_t(p_t^{\mathrm{src}}) - \ell_t(p_{t,k_i^*})]_+. \tag{8}$$

Against this comparator,

$$L_T^{SAPF} - L_T^Q \leq D_T(Q) + \frac{P_U \log K}{\eta} \leq |M_{01} \cup M_{10}| + \frac{P_U \log K}{\eta}. \tag{9}$$

$D_T(Q)$ prices both directions of catalogue disagreement, while the second term controls aggregation on the realized fallback route. The disagreement cost can grow linearly; neither bound ensures correct routing or recovery after the best expert changes. Fixed-share is an empirical tracking comparator here (Herbster & Warmuth, 1998). Prediction and update require $O(K)$ time and $O(K + C_i)$ state. The supplement gives the derivation.

4 EXPERIMENTAL PROTOCOL

Evaluation is participant-disjoint and prequential. The Regulation of Brain Cognitive States cohort supplies source fitting and validation, three public n-back cohorts guide development, and an auditory-oddball cohort provides external evaluation with 13 participants and 28,452 trials. Transfer, temporal, verifier, and allocation analyses replay the recorded forecasts in their original order without refitting the source models. The public cohorts are hosted by OpenNeuro (Markiewicz et al., 2021).

Every method receives the same pre-response history. Baselines cover participant and context Beta estimators, recency rules, causal state models, GRU/DKT, and a causal Transformer. The sequence baselines are source-fitted and frozen at evaluation, so this comparison diagnoses transfer under the shared information contract rather than universal architecture superiority. Participant-balanced Brier is the primary metric; NLL and ECE measure probability quality, while AUROC, Error AP, and risk–coverage measure ranking. Participant bootstrap intervals preserve method pairing. For the external cohort, all intervals resample the 13 participants; the 28,452 trials are not treated as independent units. The external comparison family uses Holm–Bonferroni correction (Neuhauser & Ruxton, 2024); the supplement reports dataset mappings, model settings, and additional analyses.

5 RESULTS

5.1 INFORMATION BOUNDARY

Before comparing models, we isolate the effect of the information boundary. Observing the current outcome lowers Brier from 0.149 ± 0.046 to 0.137 ± 0.051 and raises AUROC from 0.851 ± 0.070

Table 1: Forecasting before versus after observing the current response.

Target	P-Brier ↓	NLL ↓	ECE ↓	AUROC ↑	Uses y_t?
Prequential forecast ($\mathcal{F}_{t-1}$)	0.149 ± 0.046	0.453	0.051	0.851 ± 0.070	no
Outcome posterior ($\mathcal{F}_{t-1}, y_t$)	0.137 ± 0.051	0.416	0.046	0.873 ± 0.077	yes
Posterior minus forecast	-0.012	-0.037	-0.005	$+0.022$	—

Table 2: External forecasting performance; Error AP ranks incorrect responses.

Method	Brier ↓	P-Brier ↓	NLL ↓	ECE ↓	Pooled AUROC ↑	Error AP ↑
Catalogue-gated and online estimators						
SafeMix (soft)	**0.016**	0.016 ± 0.016	0.072	0.016	0.915	<u>0.454</u>
Context Beta ($\kappa = 1$)	0.017	0.017 ± 0.017	<u>0.067</u>	**0.002**	<u>0.929</u>	0.390
Rolling-20 + Platt	0.017	0.018	0.094	0.044	0.904	**0.478**
Rolling-20	0.018	0.018 ± 0.015	0.095	0.046	0.904	**0.478**
Hierarchical logistic	0.017	0.017 ± 0.018	0.070	0.012	0.904	0.438
Participant Beta	0.023	0.024 ± 0.028	0.100	0.003	0.794	0.067
State-space and sequence estimators						
Bayesian/ARCH causal filter	0.036	0.037 ± 0.035	0.157	0.062	0.672	0.049
Source-fitted state ensemble	0.084	0.084 ± 0.011	0.335	0.240	0.511	0.027
Causal Transformer	0.132	0.132 ± 0.015	0.447	0.325	0.569	0.033
GRU/DKT	0.162	0.162 ± 0.010	0.514	0.371	0.680	0.053
SAPF	**0.014**	0.017 ± 0.017	**0.066**	**0.002**	**0.929**	**0.490**

to 0.873 ± 0.077; Table 1 shows corresponding gains in NLL and ECE. The difference reflects posterior reconstruction rather than prospective prediction. Every subsequent comparison evaluates probabilities committed before the response.

An outcome-permutation check verifies this ordering at trial level. Once the available history is fixed, permuting y_t leaves the stored prediction for trial t unchanged and affects only later forecasts after the response enters the online state. The reported forecasts are measurable with respect to $\mathcal{F}_{t-1}$.

5.2 External fallback and cold-start transfer

The external cohort combines participant and context shift, neither represented in source training, and is absent from the source catalogue; it tests fallback generalization, with gate selection assessed in the cross-environment experiments below. SAPF attains the lowest pooled Brier (0.014), NLL (0.066), tied-highest AUROC (0.929), and highest Error AP (0.490). Table 2 places both adaptive methods ahead of source-fitted models on proper scores; SAPF improves participant-balanced Brier over Participant Beta by 0.006 (Holm-adjusted $p = 0.005$). Figure 2 shows the paired effects and calibration profile.

Before feedback accumulates, SAPF lowers first-five-trial Brier by 0.039 relative to SafeMix (95% interval 0.018–0.054) by activating fallback immediately on unsupported trials; SafeMix starts with source weight and learns from losses. The gap narrows as feedback arrives, with hard and soft adaptation converging after thousands of trials. This cold-start separation is obscured by pooled full-stream averages. SAPF also leads Error AP (0.490), ahead of Rolling-20 (0.478), SafeMix (0.454), and Context Beta (0.390). Proper scores measure probability quality; Section 5.6 tests error retrieval under budget.

5.3 Routing under transfer shift

Catalogue routing depends on membership being informative about source transfer. Across represented holdouts and three shifted development cohorts, Figure 3 shows that the source route leads on represented tasks and the fallback leads on every unseen catalogue. Selecting between them gives macro Brier 0.167, while the regret diagnostics remain within the theoretical envelope across regimes.

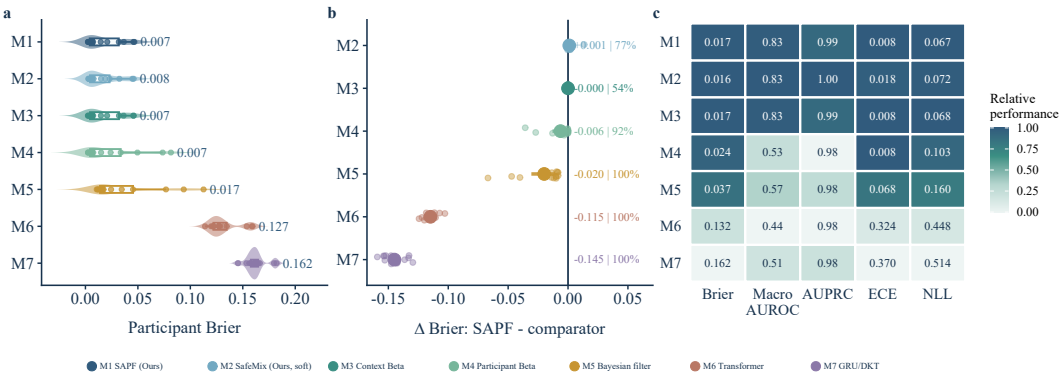


Figure 2: External validation. (a) Participant Brier; (b) paired differences; (c) participant-level metrics.

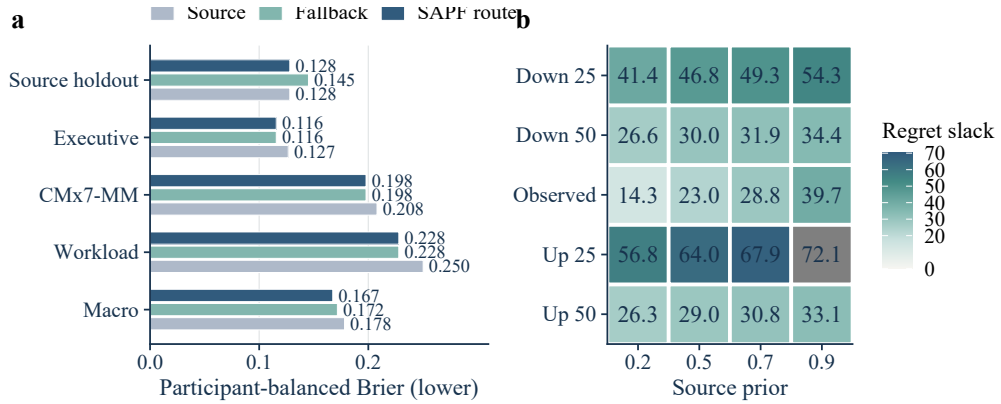


Figure 3: Routing performance and regret slack. (a) Brier for source, fallback, and SAPF routes; (b) regret slack across source priors and regimes.

When identifiers and transferability disagree, SafeMix reaches macro Brier 0.071 versus 0.113 for exact-ID routing, a paired reduction of 0.042 with a 95% interval of 0.038–0.046. Its loss-based weights preserve useful source predictions even when the identifier points to the wrong route. On the mixed-context stream, SafeMix likewise matches direct source reuse at Brier 0.124 and improves on learned and entropy gates by 0.003–0.005. Hard routing is best suited to a trustworthy support boundary, whereas feedback weighting handles ambiguous metadata.

The route intervention separates the two terms in the analysis. The bound controls aggregation after fallback selection, while the measured route gap captures the cost of choosing the wrong branch. Online expert weighting addresses the first term; the catalogue decision addresses the second.

5.4 Robustness and Efficiency

The ablations locate the gain in the routing mechanism rather than one particular hyperparameter. Compact and single-scale expert grids remain within 0.002 Brier of the full grid, while removing task context increases Brier by up to 0.011 (supplement). Under stream perturbations, context noise, delayed feedback, and missing history produce gradual degradation; past-label corruption is more damaging because each error enters subsequent updates (Table 3 and Figure 4). The robustness audit is separate from the 13-participant auditory cohort: it contains 14 held-out experiment streams, and $n = 14$ below counts those streams.

SAPF introduces no learned parameters in its fallback layer, whereas SafeMix stores expert weights. The supplement reports the corresponding per-trial computation alongside held-out-stream variation and stress behavior.

Table 3: Routing gains and stress responses. Routing gains are Brier reductions; stress ΔBrier is degradation from the clean stream.

Block	Condition / method	Primary value	SD	Reference	Gain / Δ	Evaluation scale
Catalogue routing: adaptive selection lowers Brier by 0.010–0.022						
SAPF route (Ours)	Source hold-out (represented)	**0.128**	0.037	0.145	+0.017	5 splits
SAPF route (Ours)	Executive-function n-back (unseen)	**0.116**	–	0.127	+0.011	40,009 trials
SAPF route (Ours)	CMx7-MM n-back (unseen)	**0.198**	–	0.208	+0.010	5,640 trials
SAPF route (Ours)	Multimodal workload n-back (unseen)	**0.228**	–	0.250	+0.022	7,200 trials
Stream stress at severity 0.8: gradual costs, with past-label flips as the failure boundary						
SAPF (Ours)	Context noise	0.012	0.017	clean	+0.012	$n = 14$ streams
SafeMix (Ours, soft)	Context noise	0.012	0.017	clean	+0.012	$n = 14$ streams
SAPF (Ours)	Feedback delay	0.018	0.019	clean	+0.018	$n = 14$ streams
SafeMix (Ours, soft)	Feedback delay	0.022	0.020	clean	+0.022	$n = 14$ streams
SAPF (Ours)	History missing	0.016	0.009	clean	+0.016	$n = 14$ streams
SafeMix (Ours, soft)	History missing	0.015	0.010	clean	+0.015	$n = 14$ streams
SAPF (Ours)	Past-label flips	0.236	0.181	clean	+0.236	$n = 14$ streams
SafeMix (Ours, soft)	Past-label flips	0.260	0.187	clean	+0.260	$n = 14$ streams
Per-trial computation (reported for completeness)						
Cost	Task history	2.58	0.720	–	–	73 values
Cost	Source ensemble	7.64	1.17	–	–	2,699 values
Cost	Adaptive state	18.6	2.13	–	–	2,728 values
Cost	SAPF fallback (Ours)	21.7	1.14	–	–	0 values
Cost	SafeMix (Ours, soft)	26.0	3.08	–	–	2,699 values

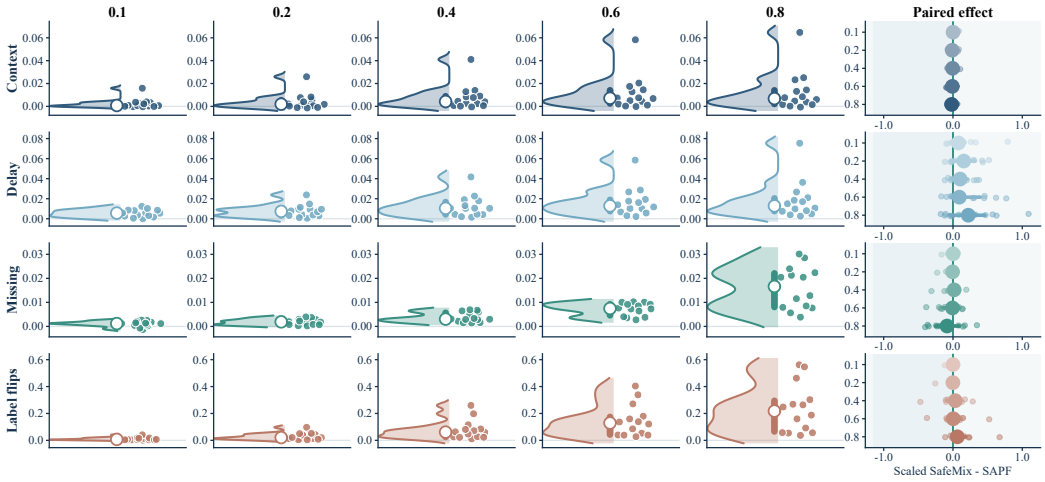


Figure 4: Held-out-stream ΔBrier for SAPF and SafeMix across four corruption families and five severities.

5.5 Source retention and recovery

A two-expert mixture captures nearly all of SafeMix's stationary-stream performance. Two-expert EWA and five-expert SafeMix score 0.1240 and 0.1239 on 20 source LOPO streams, and 0.0176 and 0.0164 externally. Their paired differences are small, whereas route-specific behavior remains pronounced: Context Beta leads during unseen-task cold start and SAPF retains source-level performance on represented streams. The supplementary comparison attributes the gain to adaptive source retention rather than expert count.

Reliability may also change after the initial route decision. The source-minus-context contrast changes sign at 285 of 940 source-task boundaries, motivating a controlled recovery test. We invert source probabilities during the middle third of each stream and restore them in the final third. Fixed-share then lowers Brier from 0.1446 for EWA to 0.1290, a paired reduction of 0.01558 with a 95% interval of 0.01074–0.02031. Its source weight recovers in 18/20 streams, compared with 2/20 for EWA and 1/20 for SafeMix. Sensitivity tests identify weight renewal as the driver of recovery under moderate and strong corruption.

5.6 Online allocation and intervention targeting

The online replay converts forecast risk into intervention decisions through a prefix-constrained threshold estimated from each participant's risk history. Under a 10% cap, SAPF covers 42.2% of errors at a 5.8% call rate, whereas random allocation covers 9.0% while calling on 9.5% of trials. SAPF retrieves about 4.7 times as many errors with fewer calls. Context Beta reaches similar coverage, and SafeMix covers 29.3% at a more selective 3.1% call rate. The supplement separates this prefix-causal replay from the retrospective full-stream risk-ranking reference.

Once a trial is selected, the Agent determines what to present. Across five benchmark runs, candidate selection raises span F1 from 0.438 for public vote to 0.513 ± 0.004, a paired gain of 0.075 with a 95% interval of 0.013–0.136; type F1 is 0.110 ± 0.005. Graph perturbations show that its structural checks respond to broken evidence links and proof order. The forecaster and Agent address complementary parts of the intervention: when to act and which reasoning span to target.

In the seven-participant pilot, item-specific prompts reduce confidence Brier by 0.033 relative to generic prompts, with a 95% interval of 0.004–0.065, and receive higher helpfulness and trust ratings. Accuracy is similar between conditions, but the calibration and preference effects show that targeted evidence changes how participants assess the assistance they receive.

6 Discussion

Transfer uncertainty comes from metadata and time. When catalogue membership marks support, hard routing acts before feedback and helps SAPF at cold start; when identifiers mislead, loss-based SafeMix retains useful source predictions. The ablations indicate that the gains come from matching support decisions and renewal to environmental change, rather than enlarging the ensemble; fixed-share can restore source weight after a transient failure. Under a budget constraint, historical quantiles retrieve more errors than random with fewer calls, showing that prequential forecasts retain actionable ranking information. Agent selection chooses the reasoning span, while the pilot supports calibration and perceived usefulness although accuracy remains similar. Forecasting, dispatch, and intervention content are distinct decisions that should be evaluated jointly in one prequential stream. SAPF is strongest when catalogue support is reliable at cold start, whereas SafeMix is useful when feedback must reveal transferability: metadata and feedback provide different evidence and are not interchangeable.

7 Conclusion

SAPF treats pre-response forecasting as a decision to retain, suppress, or relearn source knowledge. Catalogue routing protects unsupported cold starts, while SafeMix preserves useful source predictions when metadata is misleading. External transfer, identifier conflict, and controlled recovery experiments show that route and renewal decisions shape the observed transfer differences. Online replay converts calibrated risk into 4.7-fold greater error coverage than random allocation at a lower call rate. The method is strongest when catalogue support is reliable at cold start; feedback weighting remains important when identifiers do not predict transferability. Each forecast is committed before the current response, and later feedback updates only subsequent trials, linking calibration, transfer selection, and intervention targeting in one prequential evaluation.

AI USE STATEMENT

Generative AI tools assisted with code development. The authors verified the analyses and final manuscript.

ETHICS STATEMENT

This study uses de-identified data and an anonymous, minimal-risk interface evaluation without direct identifiers or sensitive information. Institutional ethics approval was not required.

REPRODUCIBILITY STATEMENT

The supplement specifies the causal information boundary, dataset mapping, participant-level splits, feature construction, model selection, baseline configurations, robustness protocols, statistical analyses, and computational resources used in the experiments. An anonymous code release is available at `https://anonymous.4open.science/r/SAPF-215C/`; it includes the core forecasting implementation, a data-free demo, tests, and environment specifications.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. *arXiv preprint arXiv:2312.05409*, 2023. URL https://arxiv.org/abs/2312.05409.

Salim I Amoukou, Tom Bewley, Saumitra Mishra, Freddy Lecue, Daniele Magazzeni, and Manuela Veloso. Sequential harmful shift detection without labels. *Advances in Neural Information Processing Systems*, 37:129279–129302, 2024.

Beng Heng Ang, Sujatha Das Gollapalli, and See Kiong Ng. Socratic question generation: A novel dataset, models, and evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 147–165, 2023.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024. URL https://arxiv.org/abs/2403.07815.

Nicolo Cesa-Bianchi, Gábor Lugosi, et al. *Prediction, learning, and games*, volume 1. Cambridge university press Cambridge, 2006.

Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, volume 2025, pp. 97378–97397, 2025. URL https://openreview.net/forum?id=b42wmsdwmB.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7358–7370, 2021.

A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2): 278–290, 1984.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Hamid Fekri Azgomi, Luciano R F. Branco, Md Rafiul Amin, Saman Khazaei, and Rose T Faghih. Regulation of brain cognitive states through auditory, gustatory, and olfactory stimulation with wearable monitoring. *Scientific reports*, 13(1):12399, 2023.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Yoav Freund, Robert E. Schapire, Yoram Singer, and Manfred K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pp. 334–343, 1997. doi: 10.1145/258533.258621.

Aritra Ghosh, Neil Heffernan, and Andrew S Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2330–2339, 2020.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024. URL https://arxiv.org/abs/2402.03885.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. *arXiv preprint arXiv:2404.05094*, 2024. URL https://arxiv.org/abs/2404.05094.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2): 151–178, 1998. doi: 10.1023/A:1007424614876.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Saman Khazaei and Rose T Faghih. Decoding a cognitive performance state from behavioral data in the presence of auditory stimuli. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:4270–4283, 2024.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, volume 2024, pp. 46977–47002, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pp. 3122–3130. PMLR, 2018.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, Jiliang Tang, and Weiqi Luo. pykt: a python library to benchmark deep learning based knowledge tracing models. *Advances in Neural Information Processing Systems*, 35:18542–18555, 2022.

Zitao Liu, Qiongqiong Liu, Jiahao Chen, Shuyan Huang, and Weiqi Luo. simplekt: a simple but tough-to-beat baseline for knowledge tracing. *arXiv preprint arXiv:2302.06881*, 2023a. URL https://arxiv.org/abs/2302.06881.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36:32958–32970, 2023b.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Christopher J Markiewicz, Krzysztof J Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncalves, et al. The openneuro resource for sharing of neuroscience data. *Elife*, 10:e71774, 2021.

Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36:25910–25928, 2023.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jacob Sunshine, Yun Liu, et al. Scaling wearable foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 59382–59413, 2025. URL https://openreview.net/forum?id=yb4QE6b22f. arXiv:2410.13638.

Markus Neuhauser and Graeme D Ruxton. Multiple testing. In *The Statistical Analysis of Small Data Sets*, pp. 122–126. Oxford University Press, 2024. doi: 10.1093/oso/9780198872979.003.0010.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. URL https://arxiv.org/abs/2302.12400.

Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837*, 2019. URL https://arxiv.org/abs/1907.06837.

Gaurav Patel, Christopher M Sandino, Behrooz Mahasseni, Ellen Zippi, Erdrin Azemi, Ali Moin, and Juri Minxha. Efficient source-free time-series adaptation via parameter subspace disentanglement. In *International Conference on Learning Representations*, volume 2025, pp. 90723–90764, 2025.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, volume 2025, pp. 48230–48261, 2025. URL https://openreview.net/forum?id=kYwTmlq6Vn.

Surawat Pothong, Machi Shimmei, Naoya Inoue, Paul Reisert, Ana Brassard, Wenzhi Wang, Shoichi Naito, Jungmin Choi, and Kentaro Inui. Focus: A benchmark for targeted socratic question generation via source-span grounding. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2938–2958, 2025.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. Rdumb: A simple approach that questions our progress in continual test-time adaptation. *Advances in Neural Information Processing Systems*, 36:39915–39935, 2023.

Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. The art of socratic questioning: Recursive thinking with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4177–4199, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Dimitris Spathis, Ignacio Perez-Pozuelo, Soren Brage, Nicholas J Wareham, and Cecilia Mascolo. Self-supervised transfer learning of physiological representations from free-living wearable data. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 69–78, 2021.

Zeyu Sun, Dogyoon Song, and Alfred Hero. Minimum-risk recalibration of classifiers. *Advances in Neural Information Processing Systems*, 36:69505–69531, 2023.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, 2021.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. URL https://arxiv.org/abs/2006.10726.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7191–7201. IEEE, 2022a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022b. URL https://arxiv.org/abs/2203.11171.

Hongwei Wen, Annika Betken, and Hanyuan Hang. Class probability matching with calibrated networks for label shift adaption. In *International Conference on Learning Representations*, volume 2024, pp. 29337–29367, 2024.

Cheryl Sze Yin Wong, Guo Yang, Nancy F Chen, and Ramasamy Savitha. Incremental context aware attentive knowledge tracing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3993–3997. IEEE, 2022.

Chenrui Wu, Haishuai Wang, Xiang Zhang, Chengqi Zhang, and Jiajun Bu. Efficient personalized adaptation for physiological signal foundation model. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 67833–67851. PMLR, 2025. URL https://proceedings.mlr.press/v267/wu25ah.html.

Maxwell Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik Raghuram, James Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data. In *International Conference on Learning Representations*, volume 2025, pp. 53188–53206, 2025. URL https://openreview.net/forum?id=k2uUeLCrQq. arXiv:2411.18822.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations*, volume 2024, pp. 55036–55055, 2024.

Shangshang Yang, Xiaoshan Yu, Ye Tian, Xueming Yan, Haiping Ma, and Xingyi Zhang. Evolutionary neural architecture search for transformer in knowledge tracing. *Advances in Neural Information Processing Systems*, 36:19520–19539, 2023.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15922–15932. IEEE, 2023.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. *arXiv preprint arXiv:2306.03536*, 2023. URL https://arxiv.org/abs/2306.03536.

SUPPLEMENTARY MATERIAL: FORECASTING BEFORE THE OUTCOME

Anonymous authors
Paper under double-blind review

CONTENTS

A Abbreviations and Statistical Conventions **14**

B Full Theoretical Development **14**
B.1 Prequential measurability . . . 15
B.2 Participant-wise regret on the fallback route . . . 15
B.3 Catalogue routing and route error . . . 16
B.4 Prediction, estimation, and intervention targets . . . 17

C Algorithm and Evaluation Protocol **17**

D Complete External Unseen-Context Evaluation **17**

E Support Routing and Regret **21**
E.1 Identifier–transferability conflict . . . 21
E.2 Predictive gates on the same-ID mixed stream . . . 22
E.3 Calibrated and hierarchical baselines . . . 22
E.4 Online-layer sensitivity . . . 23
E.5 Expert grid and context taxonomy . . . 23
E.6 Catalogue error sensitivity . . . 24
E.7 Cross-environment route selection . . . 24

F Transfer Decisions and Prefix-Causal Allocation **25**

G Information-Set Alignment **28**

H State Components and History Length **28**

I Sequence Baseline Capacity **30**

J Validation-Only Sensitivity **30**

K Auxiliary Sensor-Fusion Analysis **31**

L Modular Agent Evaluation **31**
L.1 Assisted-reasoning pilot . . . 33
L.2 Pilot item prompt cards . . . 33

A ABBREVIATIONS AND STATISTICAL CONVENTIONS

Because all forecasting datasets score trial correctness, Brier, NLL, ECE, and AUROC share a common probability scale despite differences in task context. The auditory analysis additionally reports Error AP with incorrect trials as the positive class. Agent benchmarks use span/type F1 and verifier accuracy, while the interface pilot uses within-participant behavioral and rating endpoints. Table S2 specifies the recorded context and fallback key for each dataset.

B FULL THEORETICAL DEVELOPMENT

For participant i, trial t presents covariates and context $z_t = (x_t, c_t)$ before correctness $y_t \in \{0, 1\}$ is observed. The filtration $\mathcal{F}_{t-1} = \sigma(z_{1:t}, y_{1:t-1})$ includes the current presented context but excludes

Table S1: Abbreviations and metric directions.

Abbreviation	Expansion	Use in this work
AUROC ↑	area under the ROC curve	threshold-free ranking of correctness or error risk
AUPRC/AP ↑	area under the precision–recall curve	positive class is named explicitly; Error AP treats incorrect responses as positive
CI	confidence interval	participant bootstrap or hierarchical replicate/item bootstrap as specified
ECE ↓	expected calibration error	binned calibration diagnostic
EWA	exponentially weighted aggregation	online combination of hierarchical Beta experts
HGB	histogram gradient boosting	source-fitted state predictor and sensitivity axis
NLL ↓	negative log likelihood	strictly proper probabilistic score
P-Brier	participant-balanced Brier score	averages trial Brier within participant before averaging participants
RT	response time	lagged behavioral state; current-response RT is excluded
SAPF	Support-Aware Prequential Forecaster	"support" denotes exact source-catalogue membership; prefix-causal online fallback
SafeMix	shift-aware soft mixture	soft-routing variant used as a principal comparator
SD	standard deviation	across seeds, participant-disjoint splits, or participants as captioned

Table S2: Context and metric definitions.

Dataset	Recorded task context	SAPF context key	Metric scale
Regulation of Brain Cognitive States	n-back level, block, session, intervention	n-back level × block	Trial correctness; probabilities score correctness
Executive Functioning Neurofeedback	n-back level and session	n-back level (block is the level label)	Trial correctness; probabilities score correctness
CMx7-MM	fixed 2-back task across sessions	participant history (task and block are constant)	Trial correctness; probabilities score correctness
Multimodal Cognitive Workload	n-back level	n-back level (block duplicates level)	Trial correctness; probabilities score correctness
Auditory Oddball	stimulus class and session	stimulus class (stored in `block`)	Trial correctness; Error AP reverses the positive class to error
ProofWriter / EntailmentBank / SciFact	item evidence and proof structure	not used by the forecaster	Span/type F1, coverage, and verifier accuracy
Human interface pilot	item and prompt condition	not used by the forecaster	Within-participant accuracy, confidence Brier, time, helpfulness, trust

the current outcome, current-response reaction time, and future signals. The support catalogue and source model are fixed before the target stream.

B.1 Prequential measurability

The catalogue indicator applies deterministic exact matching between c_t and the string-valued `experiment` identifiers observed during source fitting. Exact membership selects the source route; missing or unseen identifiers select the fallback. This fixed metadata rule is distinct from learned support estimation or OOD detection, and target outcomes update the online experts without changing catalogue membership. Each expert uses sufficient statistics through trial $t-1$, and each weight uses cumulative losses through $t-1$. Their weighted prediction is $\mathcal{F}_{t-1}$-measurable and is recorded before y_t is revealed.

B.2 Participant-wise regret on the fallback route

Let $U_{i,T}$ be participant i's fallback trials, let P_U count participants with $U_{i,T} \neq \varnothing$, and let K denote the fixed expert set. Participant states reset, so the potential argument is applied within each stream.

For $0 < \eta \le 1/2$, squared loss on $[0,1]$ is η-exp-concave:

$$\exp\{-\eta \ell_t(\textstyle\sum_k w_k p_k)\} \ge \sum_k w_k \exp\{-\eta \ell_t(p_k)\}.$$

Let $W_{i,t} = \sum_k \exp(-\eta L_{i,k,t-1})$. Taking logarithms of the exp-concavity inequality and telescoping $W_{i,t}$ over participant i's fallback trials gives

$$\sum_{i:U_{i,T}\neq\varnothing} \left[\sum_{t\in U_{i,T}} \ell_t(\hat{p}_t) - \min_k \sum_{t\in U_{i,T}} \ell_t(p_{t,k}) \right] \le \frac{P_U \log K}{\eta}. \tag{1}$$

The bound is pathwise and horizon-free, with each participant compared against the best fixed member of the prespecified shrinkage grid. The comparison class is fixed over time. The learning rate $\eta = 0.25$ is selected on development cohorts, and the corresponding certificate is reported for every stream.

For SafeMix, let q be the nonuniform initial expert distribution, with source mass ρ and $(1-\rho)/(K-1)$ on each online expert. Repeating the argument gives participant-wise regret at most $-\log q_{k_i^*}/\eta$ against participant i's best fixed expert. The forced-route and catalogue-error analyses below provide empirical route-cost checks for this decomposition.

B.3 Catalogue routing and route error

Let I_T and U_T be the catalogue and fallback routes. SAPF's loss decomposes exactly as

$$L_T^{SAPF} = \sum_{t\in I_T} \ell_t(p_t^{\mathrm{src}}) + \sum_{t\in U_T} \ell_t(p_t^{\mathrm{EWA}}). \tag{2}$$

For participant i, let k_i^* minimize cumulative expert loss on SAPF's realized fallback set $U_{i,T}$; if $U_{i,T}$ is empty, choose any prespecified member of the fixed grid. Fix any analysis route S^* and define comparator $Q = (S^*, k^*)$ by

$$q_t^Q = \begin{cases} p_t^{\mathrm{src}}, & S^*(c_t) = 1, \\ p_{t,k_i^*}, & S^*(c_t) = 0. \end{cases}$$

The comparator is fixed only for analysis; it is not queried by the algorithm. Partition route disagreements by direction,

$$M_{01} = \{t : S(c_t) = 0, S^*(c_t) = 1\}, \qquad M_{10} = \{t : S(c_t) = 1, S^*(c_t) = 0\},$$

and define

$$\begin{aligned} D_T(Q) = &\sum_{t\in M_{01}} [\ell_t(p_{t,k_i^*}) - \ell_t(p_t^{\mathrm{src}})]_+ \\ &+ \sum_{t\in M_{10}} [\ell_t(p_t^{\mathrm{src}}) - \ell_t(p_{t,k_i^*})]_+. \end{aligned} \tag{3}$$

To prove the bound, expand the difference between SAPF and Q. Trials with $S = S^* = 1$ cancel. On every realized fallback trial, add and subtract the fixed expert p_{t,k_i^*}:

$$\begin{aligned} L_T^{SAPF} - L_T^Q = &\sum_i \sum_{t\in U_{i,T}} [\ell_t(p_t^{\mathrm{EWA}}) - \ell_t(p_{t,k_i^*})] \\ &+ \sum_{t\in M_{01}} [\ell_t(p_{t,k_i^*}) - \ell_t(p_t^{\mathrm{src}})] + \sum_{t\in M_{10}} [\ell_t(p_t^{\mathrm{src}}) - \ell_t(p_{t,k_i^*})]. \end{aligned} \tag{4}$$

The first line is at most $P_U \log K/\eta$ by Eq. (1); replacing each signed disagreement term by its positive part gives $D_T(Q)$. The bound is

$$L_T^{SAPF} - L_T^Q \le D_T(Q) + \frac{P_U \log K}{\eta} \le |M_{01} \cup M_{10}| + \frac{P_U \log K}{\eta}. \tag{5}$$

The last inequality follows because Brier loss lies in $[0,1]$ and each mismatch appears once. The quantity $D_T(Q)$ is the realized price of selecting a catalogue route different from S^*, evaluated against the same participant-wise fixed expert as the regret comparator. The theorem controls aggregation conditional on the realized metadata decisions, with route identification entering through $D_T(Q)$. Forced-route experiments estimate the corresponding route-loss contrasts, while SafeMix evaluates nonuniform-prior aggregation under uncertain transfer.

B.4 PREDICTION, ESTIMATION, AND INTERVENTION TARGETS

A posterior state estimator may condition on y_t and optimize $\Pr(s_t \mid \mathcal{F}_{t-1}, y_t)$. SAPF targets $\Pr(y_t = 1 \mid \mathcal{F}_{t-1})$. These quantities answer different questions even when they share a latent-state model. The information-set experiment in Section G measures the optimism created by allowing the current outcome into a posterior estimate.

C ALGORITHM AND EVALUATION PROTOCOL

For each participant, SAPF initializes a $\mathrm{Beta}(1, 1)$ participant prior, empty participant–context counts, uniform expert weights, and zero cumulative expert losses. On trial t, the support catalogue selects either the source probability or the hierarchical fallback. Both route probabilities, all expert probabilities, and the current weights are retained in the prediction log. The selected probability is clipped to $[10^{-6}, 1 - 10^{-6}]$, recorded, and scored; only then is y_t used to update the participant and context counts. Expert losses update only on fallback trials. All online state resets between participants. The fallback requires $O(K)$ arithmetic per trial and $O(K + C_i)$ state for K experts and C_i observed contexts.

The source probability is $0.2p_t^{\mathrm{logit}} + 0.8p_t^{\mathrm{HGB}}$. The inputs are current trial index and task context plus lagged correctness and response-time summaries at participant, session, block, n-back, and action scales. Logistic regularization is $C = 0.01$. HGB uses seven leaves, minimum leaf size 100, ℓ_2 regularization 1, and 200 iterations. The support catalogue contains exactly the experiment identifiers present in source fitting; it is neither expanded nor inferred from target outcomes. The implementation reads the current identifier, tests exact catalogue membership, computes both route probabilities for route analysis, records the selected probability, reveals and scores y_t, and updates only the eligible online state.

All splits are participant-disjoint, and trials retain their recorded order. Neural baselines use five fixed seeds; deterministic online estimators have a single trajectory for each fixed stream. Participant-balanced Brier is the primary endpoint. NLL is a second proper score, ECE measures calibration, and AUROC measures ranking; precision–recall results name correctness or error as the positive class. Confidence intervals resample participants, paired tests retain participant pairing, and Holm correction is applied to the external SAPF-versus-baseline Brier family.

The calibrated baselines use only the three development cohorts for fitting. Platt scaling takes the logit of the causal Rolling-20 probability. The hierarchical logistic baseline fits regularized fixed effects for Rolling-20, history length, and task context on development data. A Gaussian participant random intercept starts at zero on the external stream and is updated by posterior-mode steps after each observed outcome. External participant identifiers do not enter the fixed effects, and the current outcome cannot affect its own forecast.

Table S3: Dataset roles and sample sizes.

Dataset	Participants	Trials	Outcome mapping	Role	Main use
Regulation of Brain Cognitive States	20	21,120	direct correctness	source/dev.	source fitting and represented holdouts
Executive Functioning Neurofeedback	24	40,009	direct event field	development	support selection
CMx7-MM	17	5,640	button reconstruction	diagnostic	parser and route validation
Multimodal Cognitive Workload	18	7,200	response accuracy	development	support selection
Auditory Oddball	13	28,452	official event field	external	unseen-context fallback

D COMPLETE EXTERNAL UNSEEN-CONTEXT EVALUATION

The external experiment is absent from the source catalogue, so its evaluation isolates hierarchical fallback adaptation. The analysis covers prevalence and cold-start behavior (Figure S1), four

external metrics (Figure S2), and participant-paired effects (Figure S3). Tables S4–S8 provide the corresponding estimates and subgroup values.

Table S4: Participant-paired external Brier effects ($n = 13$).

Comparator	ΔBrier ↓	95% CI	Holm p	CI < 0
Participant Beta	-0.006	[-0.013, -0.001]	0.005	yes
Participant EMA	-0.007	[-0.014, -0.002]	0.003	yes
Adaptive state	-0.046	[-0.049, -0.043]	0.003	yes
Source ensemble	-0.067	[-0.071, -0.062]	0.003	yes
Task history	-0.607	[-0.624, -0.587]	0.003	yes
Bayesian filter	-0.020	[-0.031, -0.011]	0.003	yes
GRU/DKT	-0.145	[-0.150, -0.140]	0.003	yes
Causal Transformer	-0.115	[-0.118, -0.112]	0.003	yes

Table S5: Participant-paired external cold start.

Trials	**SAPF**	**SafeMix**	Context Beta	Participant Beta	Δ vs SafeMix [95% CI]
1–5	**0.099 ± 0.073**	**0.139 ± 0.037**	0.092 ± 0.083	0.121 ± 0.054	-0.039 [-0.054, -0.018]
1–10	**0.059 ± 0.056**	**0.084 ± 0.037**	0.053 ± 0.060	0.075 ± 0.048	-0.025 [-0.034, -0.013]
1–20	**0.050 ± 0.039**	**0.063 ± 0.028**	0.049 ± 0.043	0.059 ± 0.035	-0.013 [-0.019, -0.007]
1–50	**0.034 ± 0.022**	**0.042 ± 0.018**	0.035 ± 0.025	0.039 ± 0.023	-0.008 [-0.010, -0.005]

Values are participant means ± SD. The final column is paired within participant; negative values favor SAPF. The first-five comparison is relative to SafeMix and other source-retaining baselines: SAPF is 0.099 versus SafeMix at 0.139, while Context Beta is lower at 0.092. The cold-start result supports an advantage over source retention, not uniform superiority over every fallback estimator.

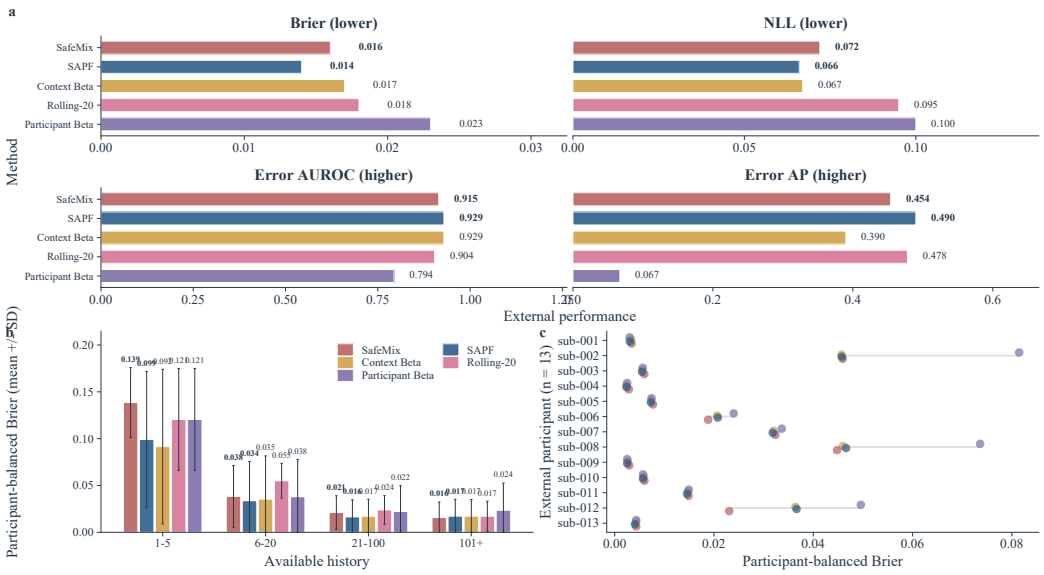


Figure S1: External performance and cold-start Brier across subgroups.

Errors comprise 2.4% of trials, so Error AP complements rather than replaces the proper scores. We report error-retrieval lift alongside proper-score skill and the evolution of participant-balanced Brier as feedback accumulates (Figure S1).

The session rows above are session-level summaries. The prevalence table below recomputes each metric on the pooled trial stream; because Error AP is nonlinear, its pooled value need not equal a trial-count-weighted average of the three session AP values. Brier, which is additive over trials, agrees after rounding.

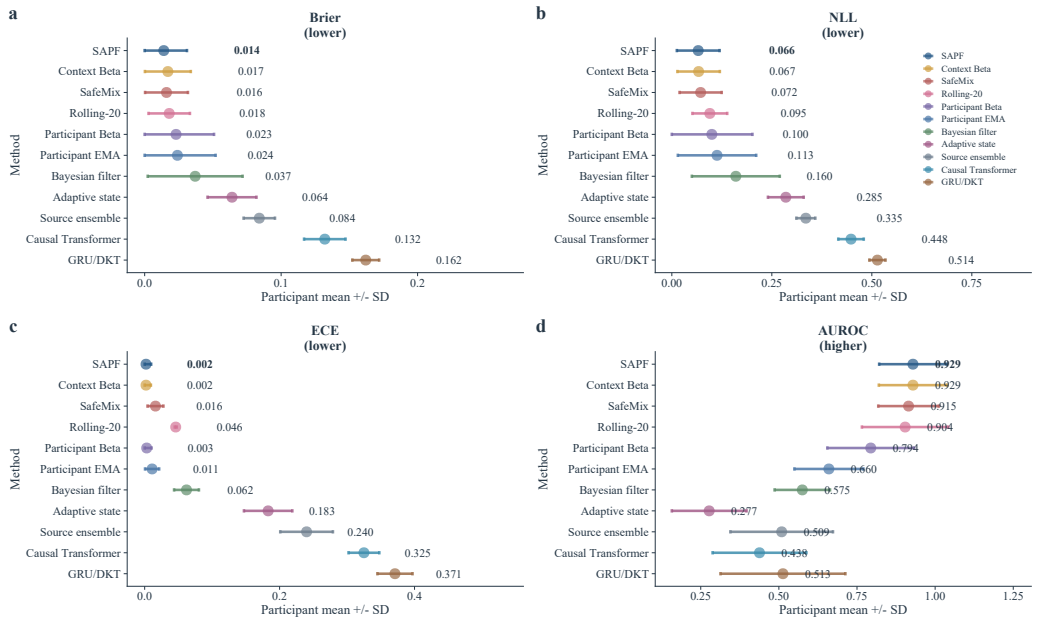


Figure S2: External method landscape across four metrics; lower is better for Brier, NLL, and ECE, higher for AUROC.



Figure S3: Paired external metric effects for SAPF across 13 participants. Numbers retain signed paired differences; colour shows scaled effect magnitude.

Table S6: Rare-error and proper-score metrics.

Method	Prev.	Brier ↓	NLL ↓	Err. AUC ↑	Err. AP ↑	AP lift ↑	Br. skill ↑	NLL skill ↑
SafeMix (Ours, soft)	0.024	0.016	0.072	0.915	0.454	18.8x	29.8%	28.7%
Context Beta	0.024	0.017	0.067	0.929	0.390	16.1x	26.4%	33.7%
Rolling-20	0.024	0.018	0.095	0.904	0.478	19.8x	22.9%	5.4%
Participant Beta	0.024	0.023	0.100	0.794	0.067	2.8x	0.0%	0.0%
SAPF (Ours)	**0.024**	**0.014**	**0.066**	**0.929**	**0.490**	**19.2x**	**30.5%**	**34.4%**

Table S7: Cold-start performance by history length.

Method	History	Trials	Brier ↓	Error AP ↑	Participant Brier ↓
Rolling-20	1-5	65	0.121	0.118	0.121 ± 0.054
	6-20	195	0.055	0.214	0.055 ± 0.019
	21-100	1,040	0.024	0.478	0.024 ± 0.015
	101+	27,152	0.017	0.491	0.017 ± 0.016
Context Beta	1-5	65	0.092	0.118	0.092 ± 0.083
	6-20	195	0.035	0.117	0.035 ± 0.046
	21-100	1,040	0.017	0.452	0.017 ± 0.018
	101+	27,152	0.017	0.401	0.017 ± 0.018
Participant Beta	1-5	65	0.121	0.118	0.121 ± 0.054
	6-20	195	0.038	0.039	0.038 ± 0.040
	21-100	1,040	0.022	0.039	0.022 ± 0.028
	101+	27,152	0.023	0.074	0.024 ± 0.029
SafeMix (Ours, soft)	1-5	65	0.139	0.133	0.139 ± 0.037
	6-20	195	0.038	0.139	0.038 ± 0.033
	21-100	1,040	0.021	0.465	0.021 ± 0.018
	101+	27,152	0.016	0.477	0.016 ± 0.017
SAPF (Ours)	**1-5**	**65**	**0.099**	**0.158**	**0.099 ± 0.073**
	6-20	**195**	**0.034**	**0.199**	**0.034 ± 0.042**
	21-100	**1,040**	**0.017**	**0.480**	**0.017 ± 0.018**
	101+	**27,152**	**0.014**	**0.503**	**0.014 ± 0.018**

Table S8: Stimulus and session subgroups.

Method	Stimulus	Session	Trials	Error prev.	Brier ↓	Error AP ↑
Context Beta	noise	1	1,352	0.009	0.012	0.060
	noise	2	1,457	0.005	0.005	0.015
	noise	3	1,455	0.001	0.001	0.002
	oddball	1	1,353	0.120	0.083	0.430
	oddball	2	1,449	0.171	0.097	0.578
	oddball	3	1,451	0.121	0.084	0.368
	standard	1	6,358	0.004	0.005	0.020
	standard	2	6,772	0.005	0.005	0.012
	standard	3	6,805	0.003	0.003	0.010
SafeMix (Ours, soft)	noise	1	1,352	0.009	0.010	0.122
	noise	2	1,457	0.005	0.005	0.016
	noise	3	1,455	0.001	0.002	0.004
	oddball	1	1,353	0.120	0.056	0.482
	oddball	2	1,449	0.171	0.065	0.591
	oddball	3	1,451	0.121	0.066	0.551
	standard	1	6,358	0.004	0.007	0.017
	standard	2	6,772	0.005	0.005	0.012
	standard	3	6,805	0.003	0.003	0.008
SAPF (Ours)	noise	1	1,352	0.009	0.010	0.069
	noise	2	1,457	0.005	0.005	0.016
	noise	3	1,455	0.001	0.001	0.002
	oddball	1	1,353	0.120	0.063	0.627
	oddball	2	1,449	0.171	0.070	0.674
	oddball	3	1,451	0.121	0.073	0.703
	standard	1	6,358	0.004	0.005	0.017
	standard	2	6,772	0.005	0.005	0.012
	standard	3	6,805	0.003	0.003	0.010

Pooling can conceal whether a rare outcome drives model ordering. Table S9 recomputes the external metrics within each stimulus class. Oddball trials have a 13.8% error rate, compared with less than 1% for noise and standard trials. The sequence baselines remain substantially worse on Brier and NLL in the oddball stratum. Among the adaptive methods, SAPF has lower NLL (0.208 versus 0.264 for SafeMix) and higher Error AP (0.669 versus 0.555), whereas SafeMix has lower Brier (0.062 versus 0.069). The metric-specific ordering is explicit rather than collapsed into a claim of uniform superiority.

Table S9: External performance by stimulus prevalence.

Stimulus	Method	Trials	Error rate	Brier ↓	NLL ↓	Error AP ↑
Noise	**SafeMix (Ours, soft)**	4,264	0.5%	0.005	0.039	0.075
	SAPF (Ours)	4,264	0.5%	0.005	0.031	0.046
	Context Beta	4,264	0.5%	0.006	0.033	0.040
	Rolling-20	4,264	0.5%	0.009	0.070	0.105
	Causal Transformer	4,264	0.5%	0.126	0.435	0.003
	GRU/DKT	4,264	0.5%	0.159	0.508	0.004
Oddball	**SafeMix (Ours, soft)**	4,253	13.8%	0.062	0.264	0.555
	SAPF (Ours)	4,253	13.8%	0.069	0.208	0.669
	Context Beta	4,253	13.8%	0.088	0.294	0.464
	Rolling-20	4,253	13.8%	0.080	0.270	0.570
	Causal Transformer	4,253	13.8%	0.162	0.511	0.173
	GRU/DKT	4,253	13.8%	0.182	0.555	0.333
Standard	**SafeMix (Ours, soft)**	19,935	0.4%	0.005	0.037	0.011
	SAPF (Ours)	19,935	0.4%	0.004	0.025	0.013
	Context Beta	19,935	0.4%	0.004	0.025	0.014
	Rolling-20	19,935	0.4%	0.007	0.063	0.011
	Causal Transformer	19,935	0.4%	0.126	0.435	0.029
	GRU/DKT	19,935	0.4%	0.159	0.506	0.024

E Support Routing and Regret

The forced-route ablation reverses the catalogue decision: represented source trials use the fallback, and unseen datasets use the source predictor. Catalogue routing lowers participant-balanced Brier in all four settings, with environments weighted equally in the macro comparison (Table S10).

The same-ID experiment removes one wearable-intervention condition per participant from source fitting while retaining the remaining conditions from the experiment. Its mixed test stream interleaves represented and withheld trials. SafeMix preserves source information on the withheld condition and matches source reuse overall (Table S11); in the history-stratified table, the refreshed protocol-catalogue SAPF results are lower than SafeMix at 1–5 and 6–20 prior trials (0.092 versus 0.109; 0.108 versus 0.137) and are comparable at 21+ (0.115 versus 0.120) (Table S12). The cross-environment route comparison below focuses on the prespecified represented-versus-unseen decision and reports the complete cohort-level values in Table S10.

Table S10: Cross-environment routing.

Method	Source hold-out	Executive	CMx7-MM	Workload	Macro	Worst	Oracle gap
Always source	0.128	0.127	0.208	0.250	0.178	0.250	0.011
Always fallback	0.145	0.116	0.198	0.228	0.172	0.228	0.004
SAPF route (Ours)	**0.128**	**0.116**	**0.198**	**0.228**	**0.167**	**0.228**	**0.000**

E.1 Identifier–transferability conflict

The metadata-alias experiment decouples experiment identifier from empirical transferability without refitting either model. Represented source trials receive an unseen identifier, while precomputed auditory predictions receive a source-catalogue identifier. The gate sees aliased metadata but not the

Table S11: Hard–soft routing.

Method	Rep.	Same-ID	Mixed	NLL	ECE	ΔBrier [95% CI]
SAPF hard (ID)	0.126	0.120	0.124	0.389	0.042	reference
SAPF hard (condition)	0.126	0.138	0.131	0.407	0.038	+0.007 [+0.004, +0.010]
SafeMix (Ours, soft)	0.126	0.120	0.124	0.389	0.037	+0.000 [-0.001, +0.001]
Always source	0.126	0.120	0.124	0.389	0.042	+0.000 [+0.000, +0.000]
Always fallback	0.142	0.138	0.141	0.432	0.034	+0.017 [+0.011, +0.024]
Context Beta	0.147	0.143	0.146	0.453	0.066	+0.022 [+0.015, +0.029]

Table S12: Same-ID cold start.

Method	Prior trials	Brier ↓ ± SD	NLL ↓ ± SD	ECE ↓ ± SD
SafeMix (Ours, soft)	1–5	0.109 ± 0.107	0.364 ± 0.281	0.198 ± 0.142
	6–20	0.137 ± 0.081	0.430 ± 0.212	0.166 ± 0.087
	21+	0.120 ± 0.070	0.379 ± 0.190	0.046 ± 0.025
Context Beta	1–5	0.192 ± 0.177	0.550 ± 0.451	0.284 ± 0.218
	6–20	0.210 ± 0.132	0.609 ± 0.332	0.257 ± 0.178
	21+	0.140 ± 0.081	0.439 ± 0.209	0.075 ± 0.033
Protocol-catalogue SAPF (Ours, hard)	**1–5**	$\mathbf{0.092 \pm 0.177}$	$\mathbf{0.360 \pm 0.451}$	$\mathbf{0.184 \pm 0.218}$
	6–20	$\mathbf{0.108 \pm 0.128}$	$\mathbf{0.405 \pm 0.321}$	$\mathbf{0.154 \pm 0.166}$
	21+	$\mathbf{0.115 \pm 0.079}$	$\mathbf{0.320 \pm 0.207}$	$\mathbf{0.040 \pm 0.026}$

route-loss contrast. After equal weighting of participant-level condition means, SafeMix reaches macro participant-balanced Brier 0.071, compared with 0.113 for exact-ID routing; the paired bootstrap difference is -0.042 (95% CI -0.046 to -0.038). SafeMix also matches the compatibility route selected from the two forced predictions, demonstrating adaptation when identifier and transferability disagree.

Table S13: Identifier–transferability conflict.

Policy	Transferable alias	Shifted alias	Macro P-Brier [95% CI]	Δ vs ID [95% CI]
SafeMix (ours, soft)	0.126	0.016	0.071 [0.055, 0.087]	-0.042 [-0.046, -0.038]
Compatibility route	0.126	0.017	0.072 [0.055, 0.087]	-0.041 [-0.046, -0.037]
Always fallback	0.142	0.017	0.080 [0.061, 0.097]	-0.033 [-0.036, -0.031]
Always source	0.126	0.084	0.105 [0.089, 0.120]	-0.008 [-0.011, -0.005]
Exact ID catalogue	0.142	0.084	0.113 [0.095, 0.131]	+0.000 [+0.000, +0.000]

E.2 Predictive gates on the same-ID mixed stream

The mixed stream also tests whether predictive confidence can replace or refine catalogue membership. For each held-out participant, logistic gates are fitted only to route predictions from the other 19 participants; their inputs comprise the two committed probabilities, disagreement, entropy, trial index, task level, wearable action, and the catalogue indicator. No target-participant outcome is used to train a gate. Using the refreshed aggregate in Table S14, SAPF catalogue has the lowest P-Brier (0.098 ± 0.073), NLL (0.307), ECE (0.030), and highest AUROC (0.898) among the deployable gates; its paired difference versus the SafeMix reference is -0.024 (95% CI $[-0.032, -0.019]$). The entropy rule agrees with the per-trial oracle more often but incurs greater probabilistic loss, showing that route hit rate does not account for the magnitude of a wrong forecast. The outcome-dependent oracle is a diagnostic replay from the route predictions and is not treated as a numerical lower bound because it uses a different aggregation.

E.3 Calibrated and hierarchical baselines

Calibration and hierarchical adaptation produce distinct effects (Table S15). Platt scaling improves the proper scores of Rolling-20 without changing its ranking. Hierarchical logistic regression adds a prequential participant random intercept to development-fitted context effects, reducing Brier to

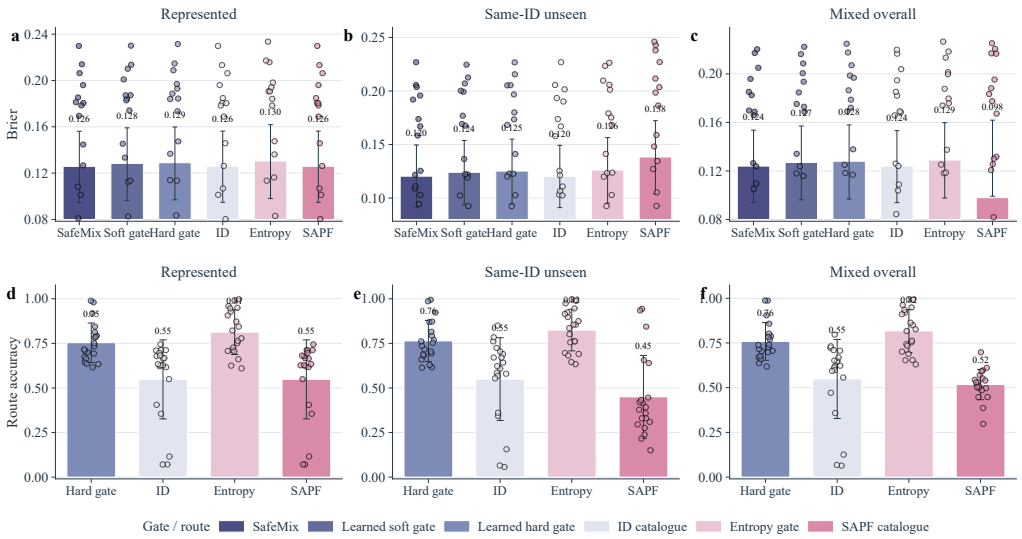


Figure S4: Gate and route comparison across represented, same-ID unseen, and mixed-support regimes.

Table S14: Mixed-support gate comparison.

Method	P-Brier mean ± SD	NLL	ECE	AUROC	Route agree.	Paired ΔBrier 95% CI
SafeMix (Ours, soft)	0.124±0.070	0.389	0.037	0.735	–	reference
ID catalogue	0.124±0.070	0.389	0.042	0.746	0.549	-0.000 [-0.001, +0.001]
Learned soft gate	0.127±0.072	0.397	0.032	0.707	–	+0.003 [+0.001, +0.005]
Learned hard gate	0.128±0.072	0.399	0.036	0.701	0.758	+0.004 [+0.002, +0.006]
Entropy gate	0.129±0.073	0.402	0.037	0.701	0.817	+0.005 [+0.003, +0.007]
SAPF catalogue (Ours, hard)	0.098±0.073	0.307	0.030	0.898	0.518	-0.024 [-0.032, -0.019]
Always fallback	0.141±0.081	0.432	0.034	0.638	–	+0.017 [+0.011, +0.023]
Trial oracle (diagnostic)	0.102±0.057	0.329	0.075	0.886	–	-0.022 [-0.028, -0.016]

0.017 and ECE to 0.012. SAPF retains lower NLL and ECE and higher discrimination, yielding different model orderings for probability quality and ranking.

Table S15: Development-fitted causal baselines on the external cohort.

Method	Brier ↓	P-Brier ↓	NLL ↓	ECE ↓	Pooled AUROC ↑	Correct AP ↑
Rolling-20	0.018	0.018 ± 0.015	0.095	0.046	0.904	0.995
Rolling-20 + Platt	0.017	0.018 ± 0.015	0.094	0.044	0.904	0.995
Hierarchical logistic	0.017	0.017 ± 0.018	0.070	0.012	0.904	0.996

E.4 Online-layer sensitivity

The development sweeps place the selected values inside broad performance plateaus (Table S16). SAPF changes by less than 0.001 in mean and worst-cohort participant-balanced Brier between $\eta = 0.25$ and 0.5. SafeMix is similarly stable for source prior $\rho \in \{0.2, 0.5, 0.7\}$, while $\rho = 0.9$ raises the worst-regime value to 0.192.

E.5 Expert grid and context taxonomy

The five-expert grid is compared with compact, fast-adaptation, pooling, and single-expert alternatives at a common learning rate. Across the four fallback streams, the largest absolute change from the

Table S16: Online-layer development sensitivity.

Component	Value	Mean P-Brier ↓	Worst-setting P-Brier ↓
SAPF η	0.25	0.180	0.228
	0.5	0.181	0.228
SafeMix ρ	0.2	0.163	0.190
	0.5	0.162	0.190
	0.7	0.163	0.190
	0.9	0.164	0.192

full grid is 0.002, indicating that performance is not tied to an isolated κ value. The context ablation separates the recorded task key from task-only, session-augmented, and participant-only state. Removing task context raises participant-balanced Brier by 0.011 on Executive Function and 0.010 on Cognitive Workload; session augmentation changes the auditory value by -0.001. These values support the dataset-specific mapping in Table S2.

Table S17: Shrinkage expert-grid ablation.

Dataset	Expert grid κ	K	P-Brier ± SD	95% CI	Δ from full
Executive Function	0.25,1,5,20,100	5	0.116 ± 0.032	[0.103, 0.128]	+0.000
	0.25,5,100	3	0.116 ± 0.032	[0.103, 0.127]	+0.000
	0.25,1,5	3	0.116 ± 0.031	[0.103, 0.127]	-0.000
	5,20,100	3	0.116 ± 0.032	[0.103, 0.128]	+0.000
	1	1	0.116 ± 0.031	[0.103, 0.128]	-0.000
CMx7-MM	0.25,1,5,20,100	5	0.198 ± 0.022	[0.189, 0.209]	+0.000
	0.25,5,100	3	0.198 ± 0.022	[0.189, 0.209]	-0.000
	0.25,1,5	3	0.198 ± 0.022	[0.189, 0.209]	+0.000
	5,20,100	3	0.198 ± 0.022	[0.189, 0.209]	+0.000
	1	1	0.198 ± 0.022	[0.189, 0.209]	+0.000
Cognitive Workload	0.25,1,5,20,100	5	0.228 ± 0.011	[0.223, 0.233]	+0.000
	0.25,5,100	3	0.228 ± 0.011	[0.223, 0.233]	-0.000
	0.25,1,5	3	0.229 ± 0.012	[0.224, 0.234]	+0.001
	5,20,100	3	0.229 ± 0.010	[0.224, 0.233]	+0.001
	1	1	0.230 ± 0.012	[0.224, 0.235]	+0.002
Auditory Oddball	0.25,1,5,20,100	5	0.017 ± 0.017	[0.009, 0.027]	+0.000
	0.25,5,100	3	0.017 ± 0.017	[0.009, 0.027]	+0.000
	0.25,1,5	3	0.017 ± 0.017	[0.009, 0.027]	-0.000
	5,20,100	3	0.018 ± 0.017	[0.009, 0.027]	+0.000
	1	1	0.017 ± 0.017	[0.009, 0.027]	+0.000

E.6 Catalogue error sensitivity

The controlled catalogue-error perturbation is retained in the diagnostic appendix because it is a failure-boundary analysis rather than core evidence. At 40% route error, participant-balanced Brier increases by 0.0040–0.0089 on unseen cohorts and by 0.0067 on the represented holdout; the complete heatmaps and route-loss table are provided there.

E.7 Cross-environment route selection

The replacement analysis tests the contribution that SAPF is designed to make: retain the source route on a represented hold-out and select the fallback on unseen cohorts. Every cohort is scored with the same participant-balanced Brier endpoint under three fixed policies, and the route policy is evaluated before target outcomes are observed. SAPF reaches the lower of the two fixed routes in each cohort, reducing the macro Brier from 0.178 for always-source and 0.172 for always-fallback to 0.167.

Table S18: Context-taxonomy ablation.

Dataset	Taxonomy	Cells	P-Brier ± SD	Δ from recorded
Executive Function	Recorded fields	3	0.116 ± 0.032	+0.000
	Task context	3	0.116 ± 0.032	+0.000
	Session augmented	6	0.115 ± 0.032	-0.000
	Participant only	1	0.127 ± 0.035	+0.011
CMx7-MM	Recorded fields	1	0.198 ± 0.022	+0.000
	Task context	1	0.198 ± 0.022	+0.000
	Session augmented	3	0.199 ± 0.022	+0.001
	Participant only	1	0.198 ± 0.022	+0.000
Cognitive Workload	Recorded fields	4	0.228 ± 0.011	+0.000
	Task context	4	0.228 ± 0.011	+0.000
	Session augmented	4	0.228 ± 0.011	+0.000
	Participant only	1	0.238 ± 0.006	+0.010
Auditory Oddball	Recorded fields	3	0.017 ± 0.017	+0.000
	Task context	3	0.017 ± 0.017	+0.000
	Session augmented	9	0.016 ± 0.016	-0.001
	Participant only	1	0.024 ± 0.028	+0.006

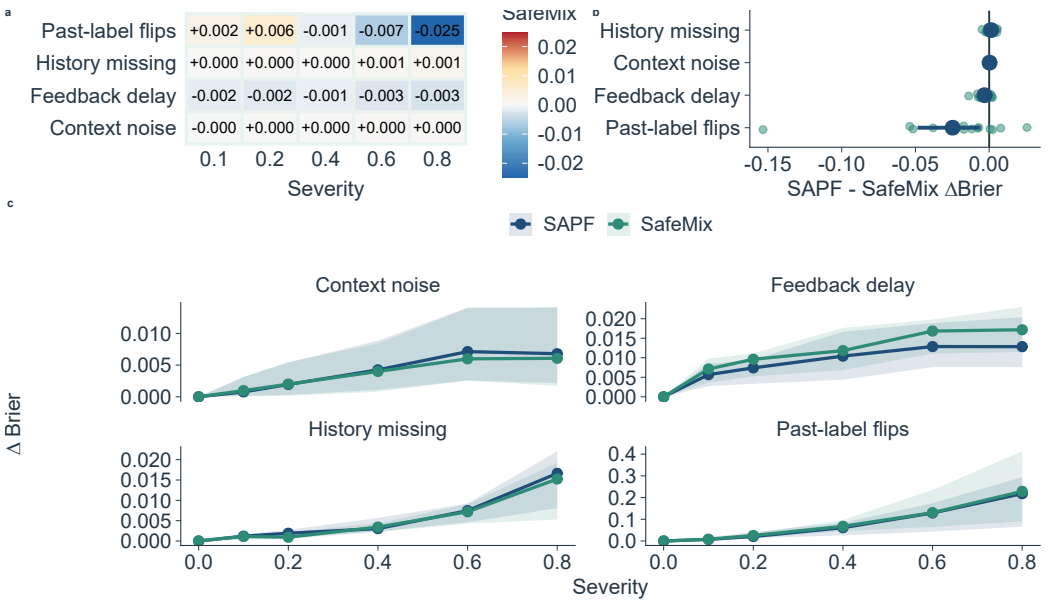


Figure S5: Paired corruption effects across methods and severities.

F Transfer Decisions and Prefix-Causal Allocation

We analyze the recorded predictions post hoc to examine expert composition and prefix-causal allocation. The analysis can be reproduced from the supplementary code; source models remain unchanged. Reported intervals use 10,000 paired participant-bootstrap resamples and are exploratory, unadjusted 95% intervals.

Data and state. The source comparison contains 20 participant-disjoint LOPO streams and 21,120 trials. The same-ID-unseen slice retains the full-stream adaptation state. The external comparison contains 13 participants and 28,452 trials, evaluated with five fixed source-model seeds. Each seed is replayed independently before averaging issued predictions. Replaying the original Context Beta and full SafeMix reproduces their stored probabilities to below 10^{-8} maximum absolute error. Stream order is the recorded outcome-time order; participant state resets only between participants. Current

Table S19: Complete 4-by-6 corruption matrix (14 participants per condition).

Corruption family	Method	Severity	ΔBrier $\uparrow$ (degradation) mean $\pm$ SD	Median
Context noise	SafeMix (Ours, soft)	0.000	$+0.000 \pm 0.000$	+0.000
	SAPF (Ours)	0.000	$+0.000 \pm 0.000$	+0.000
	SafeMix (Ours, soft)	0.100	$+0.002 \pm 0.004$	+0.001
	SAPF (Ours)	0.100	$+0.002 \pm 0.004$	+0.001
	SafeMix (Ours, soft)	0.200	$+0.004 \pm 0.007$	+0.002
	SAPF (Ours)	0.200	$+0.004 \pm 0.007$	+0.002
	SafeMix (Ours, soft)	0.400	$+0.008 \pm 0.011$	+0.004
	SAPF (Ours)	0.400	$+0.008 \pm 0.011$	+0.004
	SafeMix (Ours, soft)	0.600	$+0.011 \pm 0.015$	+0.006
	SAPF (Ours)	0.600	$+0.011 \pm 0.015$	+0.007
	SafeMix (Ours, soft)	0.800	$+0.012 \pm 0.017$	+0.006
	SAPF (Ours)	0.800	$+0.012 \pm 0.017$	+0.007
Feedback delay	SafeMix (Ours, soft)	0.000	$+0.000 \pm 0.000$	+0.000
	SAPF (Ours)	0.000	$+0.000 \pm 0.000$	+0.000
	SafeMix (Ours, soft)	0.100	$+0.007 \pm 0.004$	+0.007
	SAPF (Ours)	0.100	$+0.006 \pm 0.004$	+0.006
	SafeMix (Ours, soft)	0.200	$+0.010 \pm 0.007$	+0.010
	SAPF (Ours)	0.200	$+0.008 \pm 0.006$	+0.007
	SafeMix (Ours, soft)	0.400	$+0.014 \pm 0.011$	+0.012
	SAPF (Ours)	0.400	$+0.012 \pm 0.011$	+0.010
	SafeMix (Ours, soft)	0.600	$+0.019 \pm 0.015$	+0.017
	SAPF (Ours)	0.600	$+0.017 \pm 0.015$	+0.013
	SafeMix (Ours, soft)	0.800	$+0.022 \pm 0.020$	+0.017
	SAPF (Ours)	0.800	$+0.018 \pm 0.019$	+0.013
History missing	SafeMix (Ours, soft)	0.000	$+0.000 \pm 0.000$	+0.000
	SAPF (Ours)	0.000	$+0.000 \pm 0.000$	+0.000
	SafeMix (Ours, soft)	0.100	$+0.001 \pm 0.001$	+0.001
	SAPF (Ours)	0.100	$+0.001 \pm 0.001$	+0.001
	SafeMix (Ours, soft)	0.200	$+0.001 \pm 0.001$	+0.001
	SAPF (Ours)	0.200	$+0.002 \pm 0.001$	+0.002
	SafeMix (Ours, soft)	0.400	$+0.004 \pm 0.002$	+0.003
	SAPF (Ours)	0.400	$+0.004 \pm 0.002$	+0.003
	SafeMix (Ours, soft)	0.600	$+0.007 \pm 0.003$	+0.007
	SAPF (Ours)	0.600	$+0.007 \pm 0.003$	+0.008
	SafeMix (Ours, soft)	0.800	$+0.015 \pm 0.010$	+0.015
	SAPF (Ours)	0.800	$+0.016 \pm 0.009$	+0.017
Past-label flips	SafeMix (Ours, soft)	0.000	$+0.000 \pm 0.000$	+0.000
	SAPF (Ours)	0.000	$+0.000 \pm 0.000$	+0.000
	SafeMix (Ours, soft)	0.100	$+0.010 \pm 0.006$	+0.008
	SAPF (Ours)	0.100	$+0.012 \pm 0.011$	+0.007
	SafeMix (Ours, soft)	0.200	$+0.024 \pm 0.014$	+0.025
	SAPF (Ours)	0.200	$+0.029 \pm 0.025$	+0.021
	SafeMix (Ours, soft)	0.400	$+0.079 \pm 0.047$	+0.068
	SAPF (Ours)	0.400	$+0.078 \pm 0.072$	+0.061
	SafeMix (Ours, soft)	0.600	$+0.152 \pm 0.105$	+0.130
	SAPF (Ours)	0.600	$+0.145 \pm 0.118$	+0.129
	SafeMix (Ours, soft)	0.800	$+0.260 \pm 0.187$	+0.229
	SAPF (Ours)	0.800	$+0.236 \pm 0.181$	+0.218

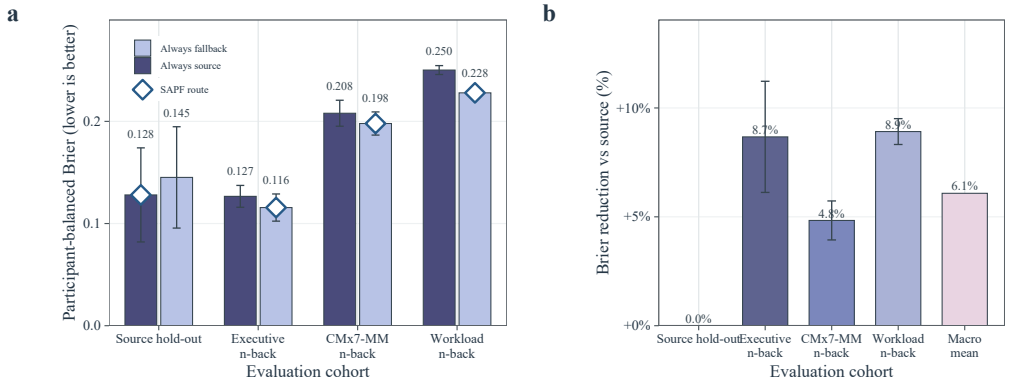


Figure S6: Cross-environment catalogue routing. Bars compare fixed routes; diamonds mark SAPF's selected route.

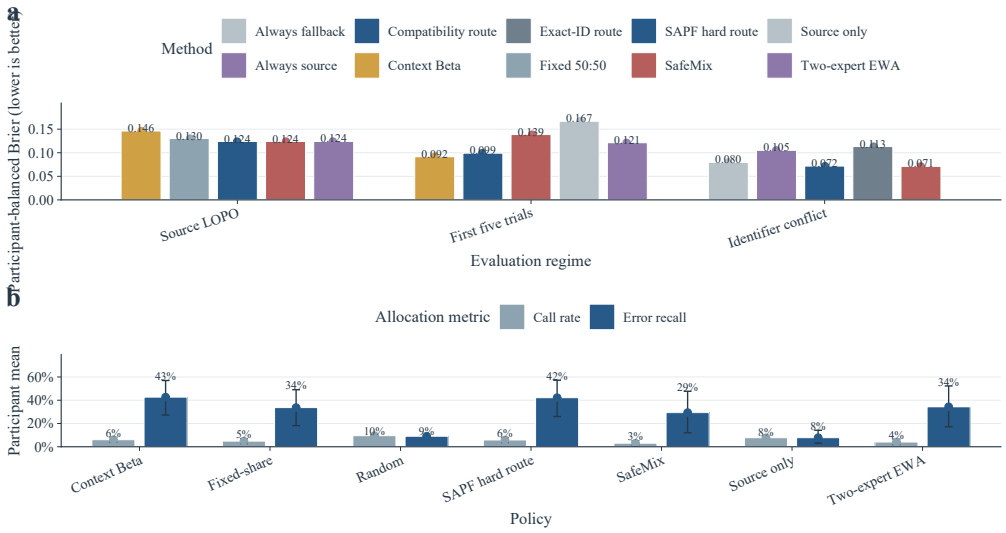


Figure S7: Transfer decisions and prefix-causal error coverage across evaluation regimes.

outcomes enter after prediction. Temporal checks perturb future labels and future risks to verify prefix invariance and the prefix budget at every step.

Comparators. Source only and Context Beta are the endpoints of the transfer decision. Fixed 50:50 tests whether averaging is enough. Two-expert EWA uses source and Context Beta with $\eta = 0.5$ and equal prior weights. Full SafeMix retains the original five experts and source prior one half. Two-expert fixed-share uses the same exponential update followed by a 0.01 share toward the equal prior, fixed across cohorts. Exact-ID uses experiment membership; the protocol gate additionally requires represented action metadata. Figure S7 summarizes the decision contrasts; Table S20 gives the complete transfer ablation.

Strictly online allocation. For each budget $b \in \{0.05, 0.10, 0.20\}$, a call is allowed only if past calls are fewer than $\lfloor bt \rfloor$. The fixed threshold is the source-LOPO $(1-b)$ risk quantile and never uses external outcomes or future external scores. The adaptive threshold is the empirical $(1-b)$ quantile of the participant's past risks; the source threshold initializes the first trial. Equal-risk ties pass the threshold, subject to the same prefix cap. Random priorities are Bernoulli(b), replayed 100 times with the same cap. They match capacity, not realized expenditure. No policy is forced to spend its remaining budget at stream end, and no future horizon is needed.

Table S20: Transfer ablations using fixed predictions; lower participant-balanced Brier is better.

Method	Source LOPO	Same-ID unseen	External	External first 5
Source only	0.1238	0.1203	0.0843	0.1668
Context Beta	0.1461	0.1433	0.0174	0.0915
Fixed 50:50	0.1301	0.1269	0.0340	0.1225
Two-expert EWA	0.1240	0.1206	0.0176	0.1212
Two-expert fixed-share	0.1257	0.1224	0.0185	0.1212
Full SafeMix	0.1239	0.1204	0.0164	0.1385
Exact-ID SAPF	0.1238	0.1203	0.0174	0.0991
Protocol gate	0.1308	0.1384	0.0174	0.0991

Table S21: Prefix-causal external allocation under a 10% budget cap; participant means with 13-participant intervals.

Score	Actual call rate	Error recall (95% CI)	Precision (95% CI)
Source only	0.078	0.078 [0.031, 0.142]	0.020 [0.007, 0.034]
Context Beta	0.061	0.426 [0.272, 0.569]	0.215 [0.103, 0.333]
Two-expert EWA	0.041	0.344 [0.172, 0.523]	0.218 [0.103, 0.337]
Two-expert fixed-share	0.049	0.337 [0.181, 0.490]	0.203 [0.096, 0.317]
Full SafeMix	0.031	0.293 [0.120, 0.477]	0.250 [0.111, 0.402]
Exact-ID SAPF	0.058	0.422 [0.260, 0.572]	0.201 [0.085, 0.325]
Random	0.095	0.090 [0.085, 0.095]	0.025 [0.010, 0.041]

At $b = 0.10$, source-calibrated thresholds yield no external calls for any forecast score, leaving precision undefined. The lower panel of Figure S7 shows the call-rate–recall trade-off, and Table S21 reports the historical-quantile values and intervals. Call rate and recall average all participants; precision averages participants with at least one call. Random repetitions are averaged within participant first. The supplementary tables report participant-paired differences against Context Beta and random, all three budgets, and the full-stream offline references. These prefix-causal replays measure selection of observed errors. They retain the recorded responses rather than estimating how an intervention would change them.

Offline full-stream reference. For comparison with the prefix-causal rules, the offline reference ranks external forecasts within participant using the complete stream and selects the highest-risk 5, 10, 20, or 30%. A budget-matched random reference uses 100 repetitions. At 10%, pooled-count error recall is 62.4% for SAPF, 58.7% for SafeMix, and 10.2% for random. These values are retrospective ranking results rather than deployable allocation estimates. In the budget table, SAPF also has higher precision and recall than SafeMix at 5% (0.185 and 0.384 versus 0.159 and 0.332) and 10% (0.150 and 0.624 versus 0.141 and 0.587). At 20%, SAPF retains higher recall (0.857 versus 0.827) while SafeMix has slightly higher precision (0.103 versus 0.100), so the budget comparison is metric-specific.

G Information-Set Alignment

Admitting the current outcome changes the target from causal prediction under $\mathcal{F}_{t-1}$ to posterior estimation. The posterior obtains lower apparent Brier in every seed, producing an optimism gap from -0.002 to -0.027 (Figure S8). This intervention quantifies the estimand change induced by current-outcome access.

NLL, ECE, and AUROC show the same information advantage as the seed-level Brier comparison. The improvement attributed to posterior estimation arises from its access to the current outcome rather than from the choice of score.

H State Components and History Length

This is a state-model diagnostic rather than a direct SAPF-versus-SafeMix comparison. Context state carries the largest ablation effect: removing it increases mean Brier from 0.1306 to 0.1362.

Table S22: Offline full-stream risk-ranking reference.

Dispatch score	Budget	Error precision ↑	Error recall ↑	Precision lift ↑	Recall lift ↑	Interventions per error ↓
Agent intervention budget: 5% of trials						
SafeMix (Ours, soft)	5%	0.159	0.332	6.600×	6.600×	6.400
Context Beta	5%	0.158	0.329	6.550×	6.550×	6.300
Participant Beta	5%	0.029	0.060	1.190×	1.190×	34.900
Random	5%	0.024	0.049	0.980×	0.980×	43.100
SAPF (Ours)	**5%**	**0.185**	**0.384**	**7.650×**	**7.650×**	**5.400**
Agent intervention budget: 10% of trials						
SafeMix (Ours, soft)	10%	0.141	0.587	5.850×	5.850×	7.100
Context Beta	10%	0.143	0.595	5.940×	5.940×	7.000
Participant Beta	10%	0.028	0.118	1.180×	1.180×	35.200
Random	10%	0.025	0.102	1.020×	1.020×	41.100
SAPF (Ours)	**10%**	**0.150**	**0.624**	**6.230×**	**6.230×**	**6.600**
Agent intervention budget: 20% of trials						
SafeMix (Ours, soft)	20%	0.103	0.827	4.130×	4.130×	10.000
Context Beta	20%	0.103	0.856	4.270×	4.270×	9.700
Participant Beta	20%	0.027	0.224	1.120×	1.120×	37.000
Random	20%	0.024	0.197	0.980×	0.980×	42.300
SAPF (Ours)	**20%**	**0.100**	**0.857**	**4.280×**	**4.280×**	**9.700**

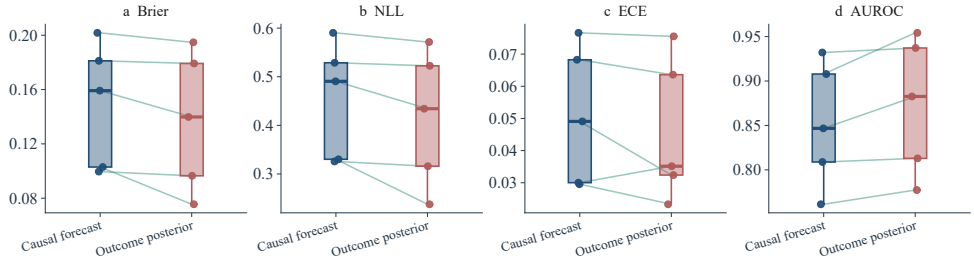


Figure S8: Information-set sanity check. Posterior scores use the current outcome and are not causal forecast evidence.

Table S23: Forecasting versus posterior estimation.

Information set	Brier ↓	NLL ↓	ECE ↓	AUROC ↑
Causal forecast $\mathcal{F}_{t-1}$	0.149 ± 0.046	0.453	0.051	0.851 ± 0.070
Outcome posterior $\mathcal{F}_t$	0.137 ± 0.051	0.416	0.046	0.873 ± 0.077

Short-term, long-term, and reaction-time history contribute smaller and partly redundant changes under the tested source representation. The history-stratified panel shows that state models retain their advantage over task history across all six bins.

Table S24: State-component ablation.

Configuration	Brier ↓	AUROC ↑	ECE ↓
Without response-time state	0.131 ± 0.042	0.878 ± 0.059	0.032 ± 0.017
Full state	0.131 ± 0.042	0.878 ± 0.058	0.035 ± 0.015
Without long state	0.131 ± 0.042	0.877 ± 0.060	0.039 ± 0.014
Without short state	0.131 ± 0.041	0.878 ± 0.058	0.040 ± 0.017
Without context state	0.136 ± 0.044	0.871 ± 0.066	0.052 ± 0.016
EMA online rule	0.151 ± 0.047	0.843 ± 0.072	0.048 ± 0.014
Beta online rule	0.163 ± 0.049	0.815 ± 0.090	0.038 ± 0.014

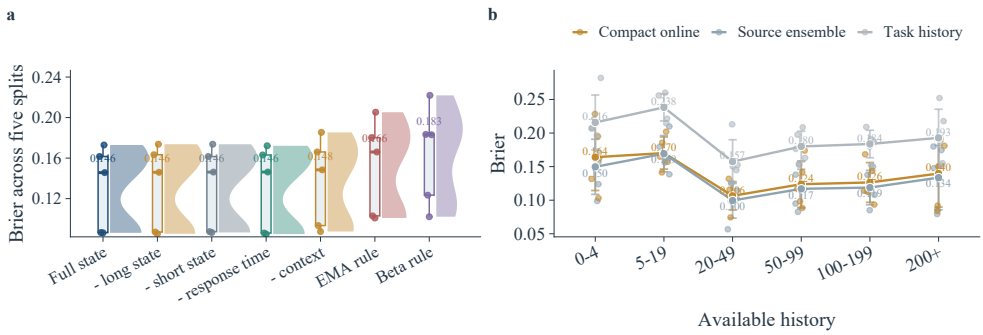


Figure S9: State-component and history-length diagnostic. Full state is the complete state estimator; lower Brier is better.

I Sequence Baseline Capacity

GRU and causal Transformer capacity is selected on validation data. Table S25 lists every evaluated architecture, and Table S26 separates architectural capacity from five-seed variation at the selected configurations.

Table S25: Sequence-model capacity sweep.

Family	Configuration	Val. Brier ↓	Parameters	Epoch	Seconds
GRU	h=64, lr=0.003	0.191 ± 0.091	15,809	16.600	3.300
	h=64, lr=0.001	0.192 ± 0.088	15,809	22.800	4.200
	h=16, lr=0.003	0.194 ± 0.095	1,649	13.400	2.700
	h=32, lr=0.003	0.207 ± 0.099	4,833	10.200	2.300
	h=32, lr=0.001	0.212 ± 0.088	4,833	22.000	3.600
TRANSFORMER	h=64, a=4, L=1, lr=0.001	0.187 ± 0.101	124,865	22.000	17.800
	h=64, a=4, L=2, lr=0.001	0.190 ± 0.104	158,337	23.200	38.100
	h=32, a=2, L=2, lr=0.001	0.201 ± 0.087	62,785	9.200	9.900
	h=32, a=2, L=1, lr=0.001	0.214 ± 0.067	54,241	10.800	5.100
	h=16, a=2, L=1, lr=0.001	0.261 ± 0.021	25,073	11.200	5.100

Table S26: Sequence-model performance across random seeds.

Model	Brier ↓	NLL ↓	ECE ↓
Causal GRU	0.193 ± 0.063	0.562 ± 0.151	0.089 ± 0.093
Causal Transformer	0.179 ± 0.074	0.532 ± 0.173	0.079 ± 0.037

Model	AUROC ↑	Correct AP ↑
Causal GRU	0.764 ± 0.131	0.680 ± 0.165
Causal Transformer	0.758 ± 0.185	0.688 ± 0.184

The selected architectures are then evaluated with five independent seeds while capacity remains fixed (Table S26).

J Validation-Only Sensitivity

Development-only sensitivity varies tree leaves, logistic regularization, and ensemble weight. Mean Brier changes by less than 5.4×10^{-4} across the evaluated range, yielding a broad validation plateau rather than a sharp optimum.

Leaf count, logistic regularization, and mixture weight all exhibit a shallow validation plateau. Table S27 reports the range of mean scores across each factor rather than repeating nearly identical grid rows; it is a stability check, not an additional performance claim. The complete evaluated grid is provided with the supplementary code.

Table S27: Validation-only sensitivity; ranges cover the three tested values per factor.

Factor	Values tested	Brier mean range ↓	AUROC mean range ↑	ECE mean range ↓
hgb_leaves	5, 7, 11	0.136–0.136	0.875–0.875	0.027–0.031
logistic_C	0.005, 0.010, 0.020	0.136–0.136	0.875–0.875	0.028–0.028
logistic_weight	0.100, 0.200, 0.300	0.136–0.136	0.874–0.875	0.028–0.029

K Auxiliary Sensor-Fusion Analysis

An auxiliary sensor-fusion analysis uses the source cohort, where PPG, EDA, and ACC share an aligned acquisition protocol. Single-channel, pairwise, and all-core missingness are evaluated across five participant-disjoint splits. The primary SAPF forecaster is behavioral and does not consume these channels; the auxiliary dropout-gated fusion model quantifies how a sensor-aware extension returns to the behavioral anchor when all three channels are unavailable.

Table S28: Auxiliary sensor-fusion proper scores.

Missing channels	Variant	Brier	Delta Brier	NLL
complete	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.130 ± 0.037	+0.000 ± 0.000	0.413 ± 0.104
	Full sensor fusion	0.130 ± 0.038	+0.000 ± 0.000	0.414 ± 0.107
ACC	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.129 ± 0.037	-0.001 ± 0.001	0.410 ± 0.104
	Full sensor fusion	0.129 ± 0.037	-0.001 ± 0.001	0.411 ± 0.104
PPG+EDA+ACC	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.128 ± 0.037	-0.002 ± 0.001	0.408 ± 0.106
	Full sensor fusion	0.129 ± 0.037	-0.001 ± 0.002	0.411 ± 0.104
EDA	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.129 ± 0.037	-0.001 ± 0.001	0.411 ± 0.105
	Full sensor fusion	0.130 ± 0.037	0.000 ± 0.001	0.415 ± 0.107
EDA+ACC	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.128 ± 0.037	-0.001 ± 0.001	0.409 ± 0.105
	Full sensor fusion	0.129 ± 0.037	-0.001 ± 0.002	0.411 ± 0.104
PPG	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.129 ± 0.037	-0.001 ± 0.001	0.411 ± 0.105
	Full sensor fusion	0.130 ± 0.037	0.000 ± 0.000	0.414 ± 0.107
PPG+ACC	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.128 ± 0.037	-0.001 ± 0.001	0.409 ± 0.105
	Full sensor fusion	0.129 ± 0.037	-0.001 ± 0.002	0.411 ± 0.104
PPG+EDA	Behavioral anchor	0.128 ± 0.037	+0.000 ± 0.000	0.408 ± 0.106
	Dropout-gated fusion	0.128 ± 0.037	-0.001 ± 0.001	0.409 ± 0.105
	Full sensor fusion	0.130 ± 0.037	0.000 ± 0.001	0.414 ± 0.107

We report proper-score, calibration, and ranking results separately so that every single, paired, and joint channel removal has a consistent metric-specific interpretation.

L Modular Agent Evaluation

The Agent evaluation separates localization, verification, allocation replay, and the exploratory assistance pilot. These use different task streams and do not test a prediction-driven closed loop. Benchmark error labels and spans are evaluation annotations. FOCUS and DeepSeek-generated outputs provide candidate predictions; the Critic selects a label and span, and the policy maps that prediction to one Socratic action. The verifier checks span existence, grounding, action count, answer withholding, and trace completeness against item evidence and proof structure. Localization, structural validity, and human assistance are distinct endpoints.

Table S30 separates verifier outcomes from policy actions. Of 2,947 intervention traces, 2,886 are accepted, six are rejected for an ungrounded span, and 55 abstain. The policy selects escalation 456 times as the stronger action in the bounded intervention inventory.

Table S29: Auxiliary sensor-fusion calibration and ranking.

Missing channels	Variant	ECE	AUROC
complete	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.044 ± 0.018	0.877 ± 0.057
	Full sensor fusion	0.044 ± 0.018	0.876 ± 0.057
ACC	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.040 ± 0.017	0.878 ± 0.058
	Full sensor fusion	0.039 ± 0.016	0.876 ± 0.056
PPG+EDA+ACC	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.035 ± 0.015	0.878 ± 0.058
	Full sensor fusion	0.040 ± 0.020	0.877 ± 0.058
EDA	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.041 ± 0.018	0.878 ± 0.057
	Full sensor fusion	0.045 ± 0.020	0.876 ± 0.059
EDA+ACC	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.038 ± 0.016	0.879 ± 0.058
	Full sensor fusion	0.040 ± 0.020	0.877 ± 0.058
PPG	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.041 ± 0.018	0.878 ± 0.058
	Full sensor fusion	0.044 ± 0.018	0.876 ± 0.057
PPG+ACC	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.038 ± 0.016	0.879 ± 0.058
	Full sensor fusion	0.038 ± 0.017	0.877 ± 0.056
PPG+EDA	Behavioral anchor	0.035 ± 0.015	0.878 ± 0.058
	Dropout-gated fusion	0.039 ± 0.016	0.879 ± 0.058
	Full sensor fusion	0.043 ± 0.020	0.876 ± 0.059

Table S30: Verifier decisions and policy actions over evaluated traces.

Stage	Outcome	Count	Rate
Verifier Decision	accept	2,886	97.93%
	reject	6	0.20%
	abstain	55	1.87%
Policy Action	assumption	801	27.18%
	counterexample	1,369	46.45%
	evidence	266	9.03%
	escalate	456	15.47%
	stop	55	1.87%

An accepted trace localizes the span "Feminism is often bad and unfair to men" and asks which assumption makes that claim follow. A rejected trace contains an empty predicted span and produces an incomplete assumption question; the verifier marks `ungrounded_span` before the prompt reaches the user. These cases make the operating rates traceable to specific grounding decisions.

Table S31: Matched five-run Agent comparisons. Mean $\pm$ run SD; 10,000 paired run/item bootstrap resamples estimate differences of the five-run means.

Endpoint ↑	Comparator	Verifier on	Comparator	Δ ↑	95% CI
Type macro F1	Verifier off	0.110 ± 0.005	0.110	+0.000	[0.000, 0.000]
Span token F1	Verifier off	0.513 ± 0.004	0.513	+0.000	[0.000, 0.000]
Type macro F1	Public vote	0.110 ± 0.005	0.135	-0.025	[-0.061, 0.011]
Span token F1	Public vote	0.513 ± 0.004	0.438	+0.075	[0.013, 0.136]

The repeated table and Figure S10 use five optimized generation runs (`run_01`–`run_05`) on 300 official-test items. Type macro F1 averages the fixed ten-label inventory. Span token F1 is the maximum overlap over accepted predicted and reference spans, averaged on gold-positive items. Missing public predictions on 30 items count as abstentions. Each bootstrap draws five paired runs with replacement and one shared item sample, then compares run-mean scores. Verifier-on and

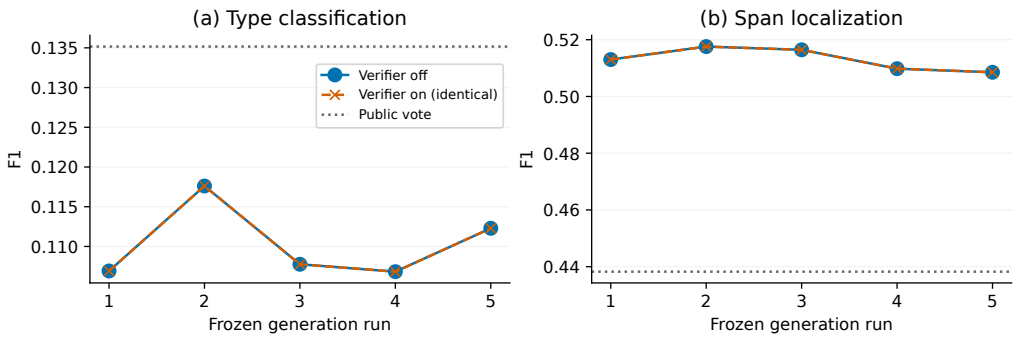


Figure S10: Five matched frozen runs on the common 300-item test set; verifier-on and verifier-off curves coincide.

verifier-off conditions use identical candidates and rendering. Their labels and span scores coincide in all five runs, locating the observed localization improvement in candidate selection rather than verification.

L.1 Assisted-reasoning pilot

The seven-participant assisted-reasoning pilot is exploratory and is not used to support the forecasting or verifier claims. Full participant-level results, including the intervals that cross zero for accuracy and response time, are retained in the diagnostic supplement alongside the consent and data-handling record.

L.2 Pilot item prompt cards

The pilot used a frozen 20-item bank. Each item below is shown as a separate interface card; the cards are English-gloss renderings of the prespecified Chinese-language contexts, claims, and prompts. Gold answers and evidence graphs supported scoring but were hidden from participants. The complete UTF-8 item bank remains in the diagnostic appendix and code artifact.

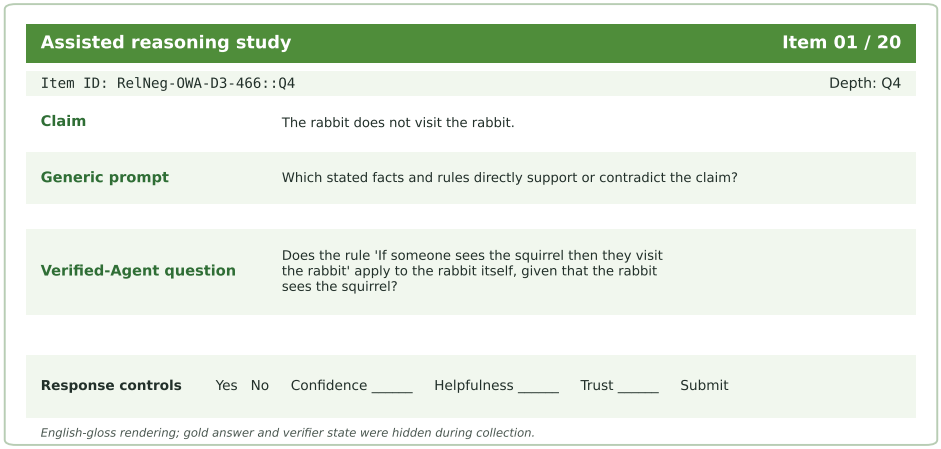


Figure S36. Pilot item 01.

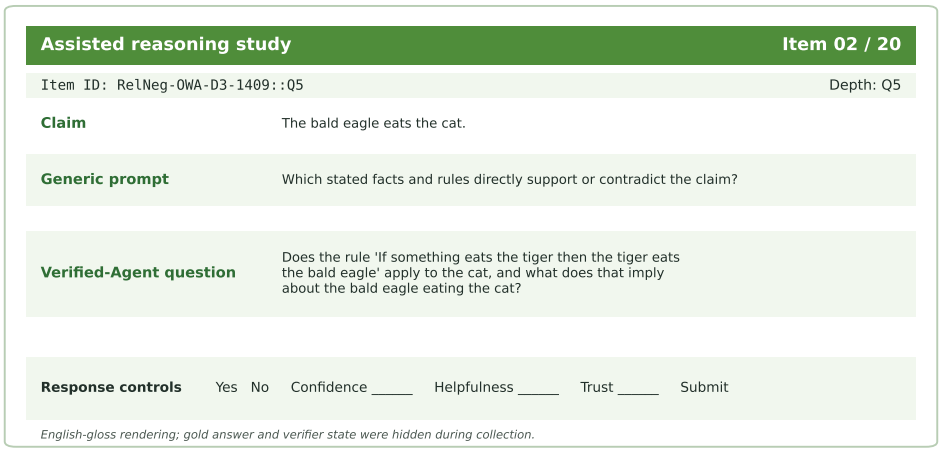


Figure S37. Pilot item 02.

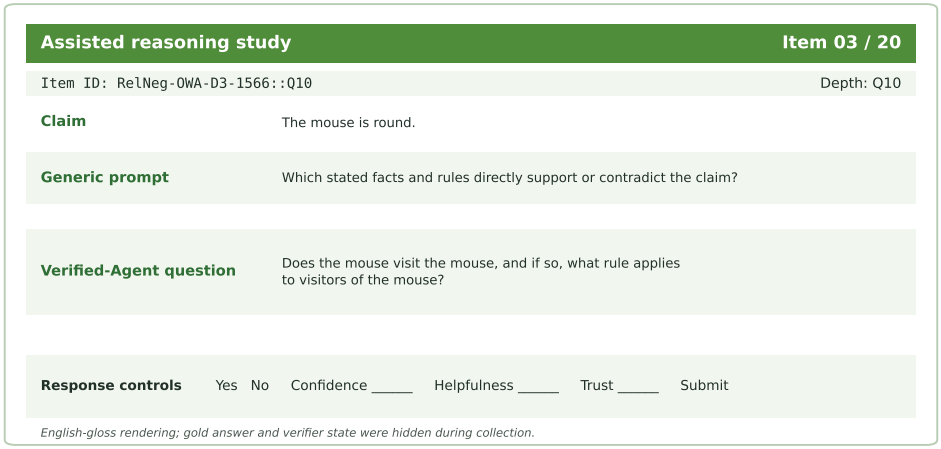


Figure S38. Pilot item 03.

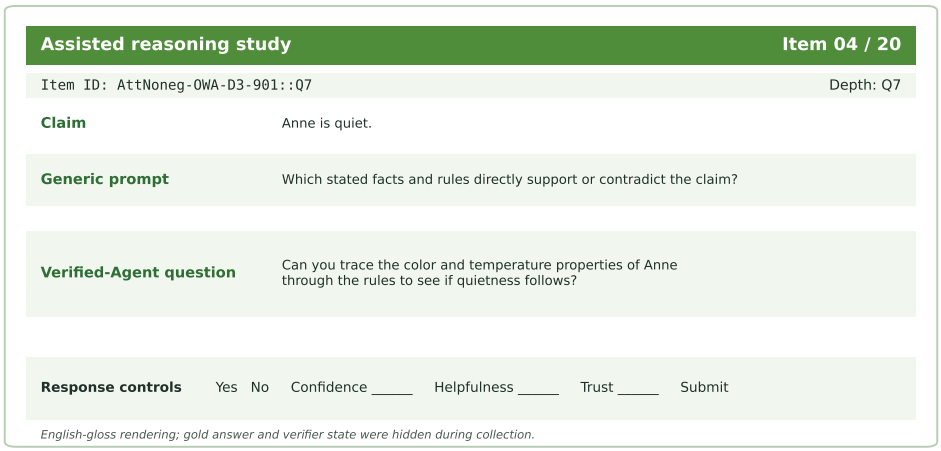


Figure S39. Pilot item 04.

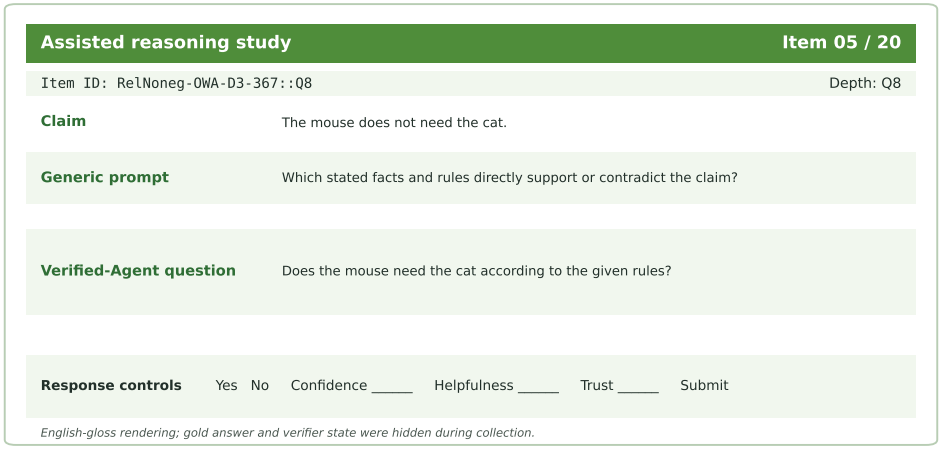


Figure S40. Pilot item 05.

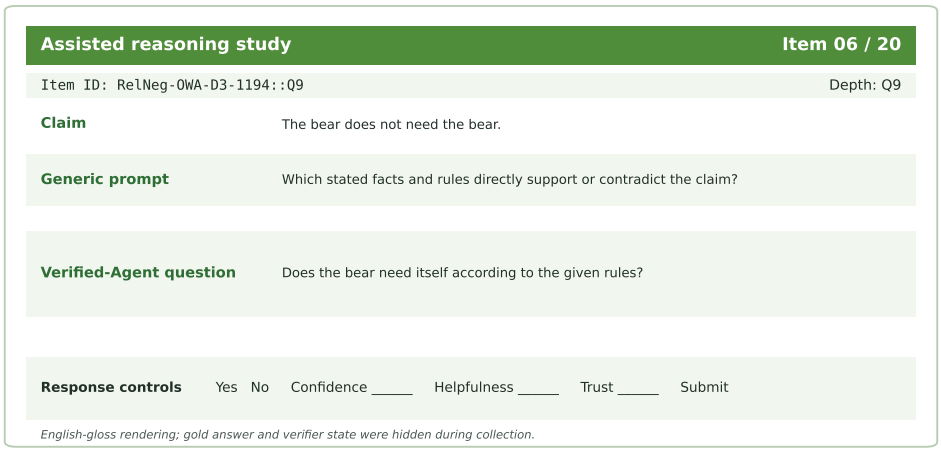


Figure S41. Pilot item 06.

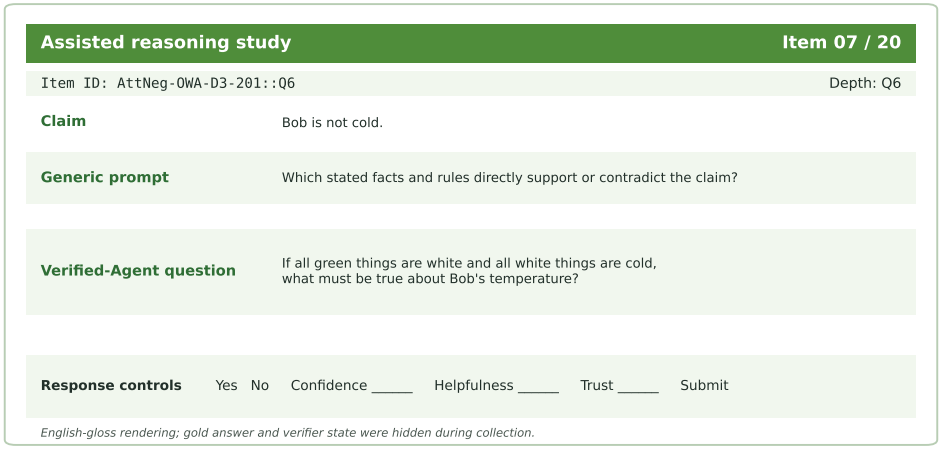


Figure S42. Pilot item 07.

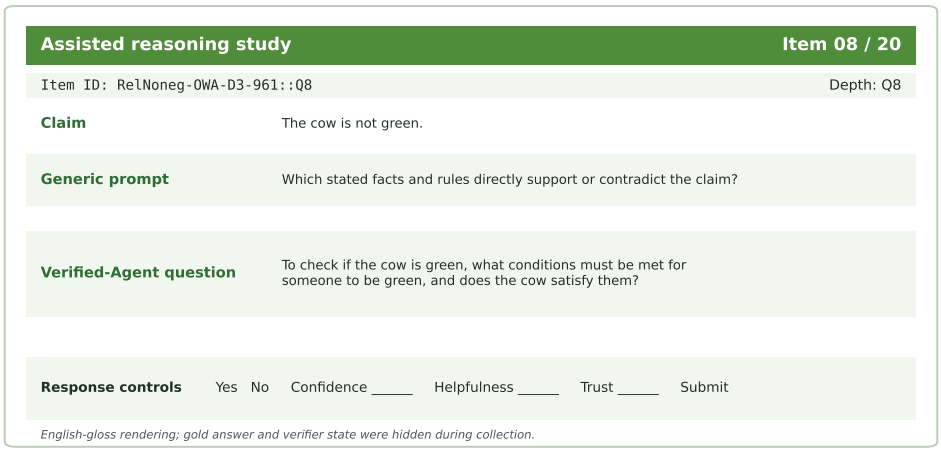


Figure S43. Pilot item 08.

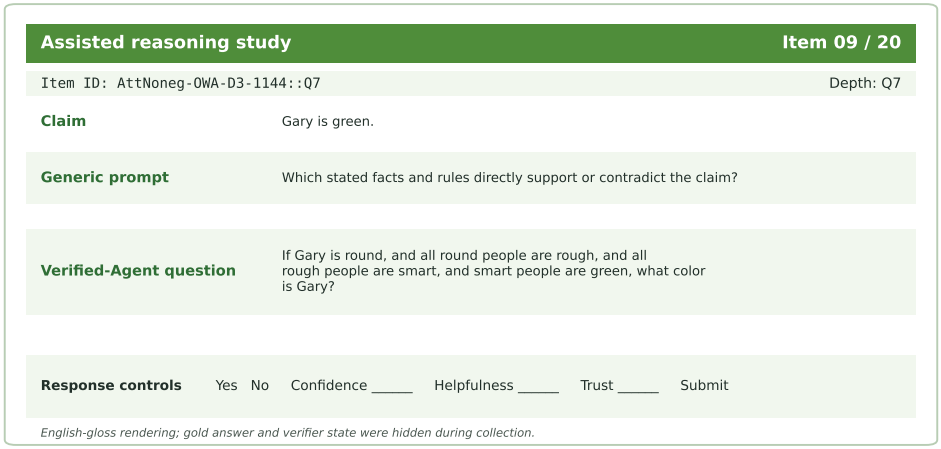


Figure S44. Pilot item 09.

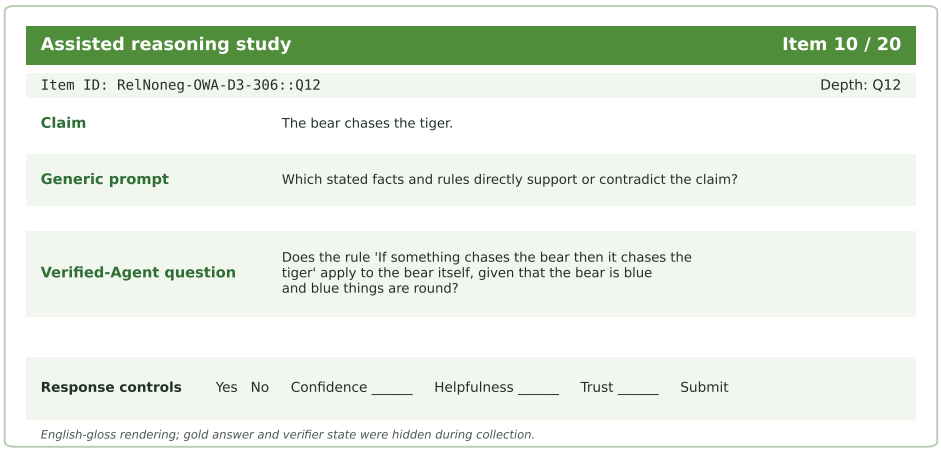


Figure S45. Pilot item 10.

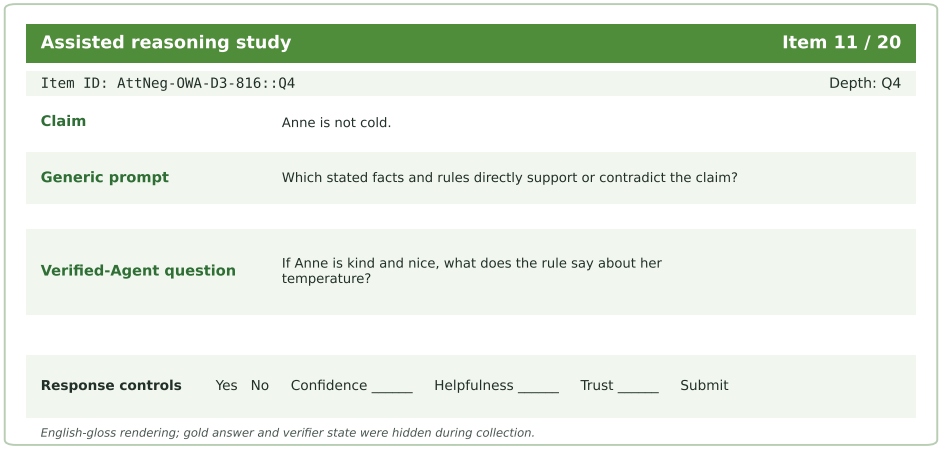


Figure S46. Pilot item 11.

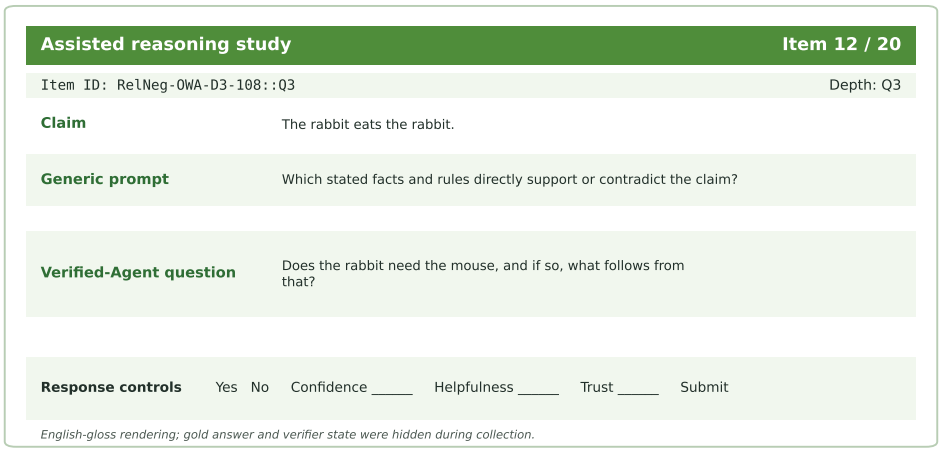


Figure S47. Pilot item 12.

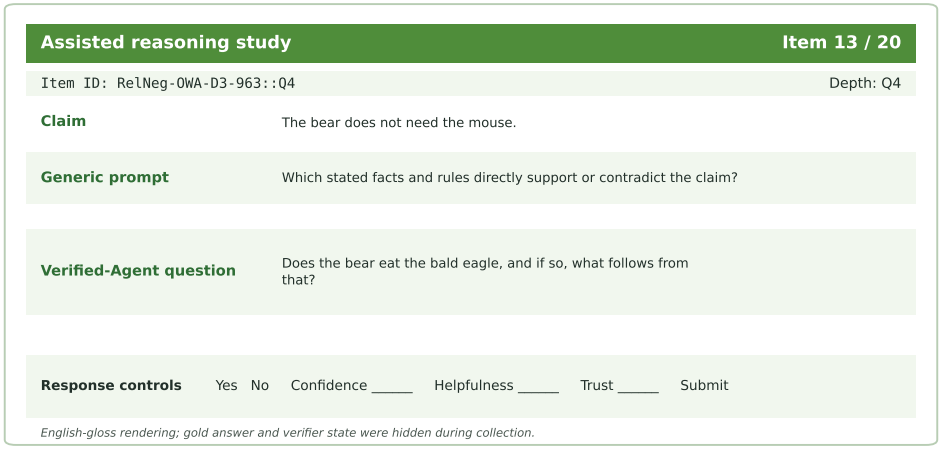


Figure S48. Pilot item 13.

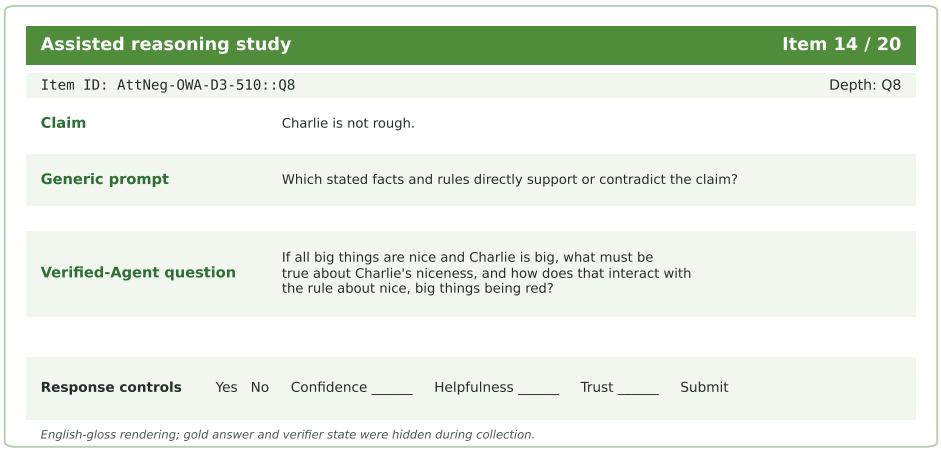


Figure S49. Pilot item 14.



Figure S50. Pilot item 15.

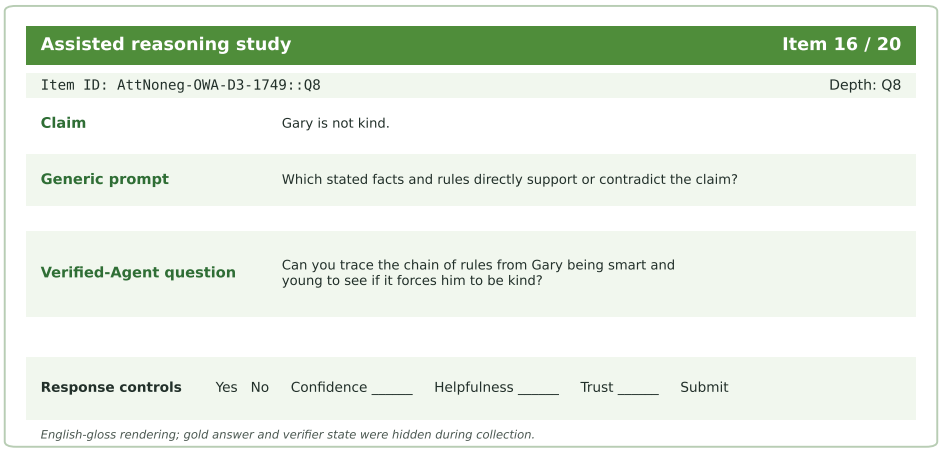


Figure S51. Pilot item 16.

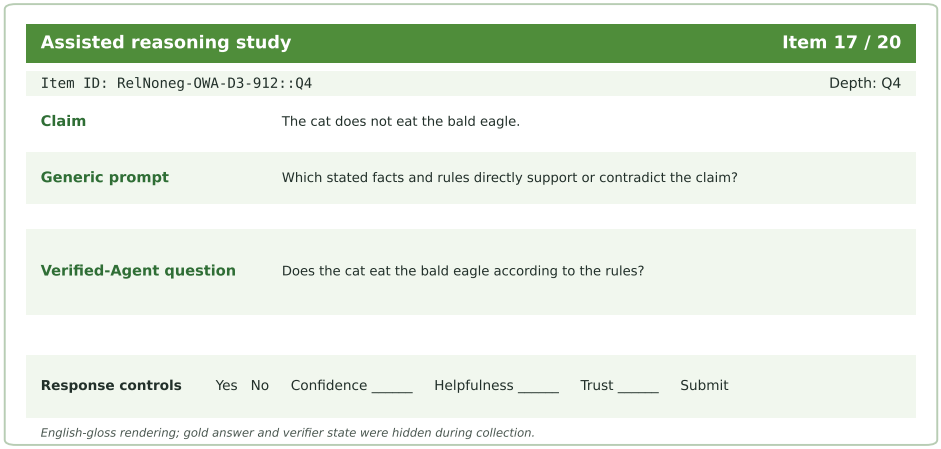


Figure S52. Pilot item 17.

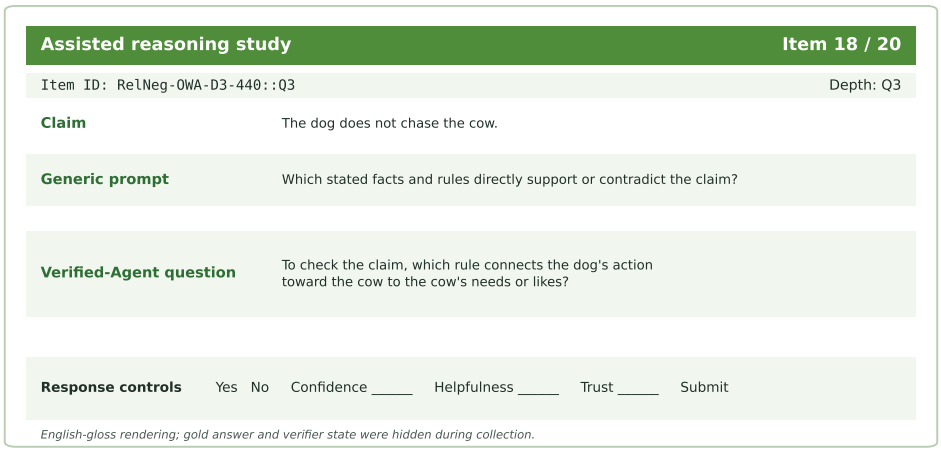


Figure S53. Pilot item 18.

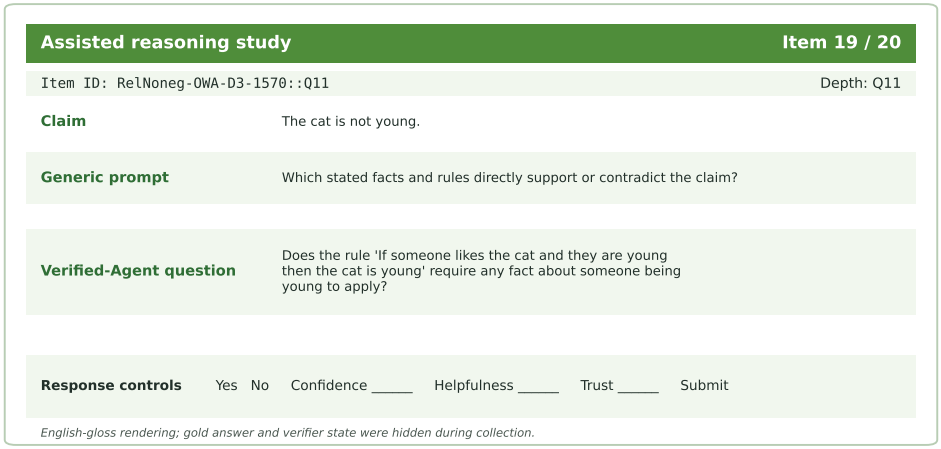


Figure S54. Pilot item 19.

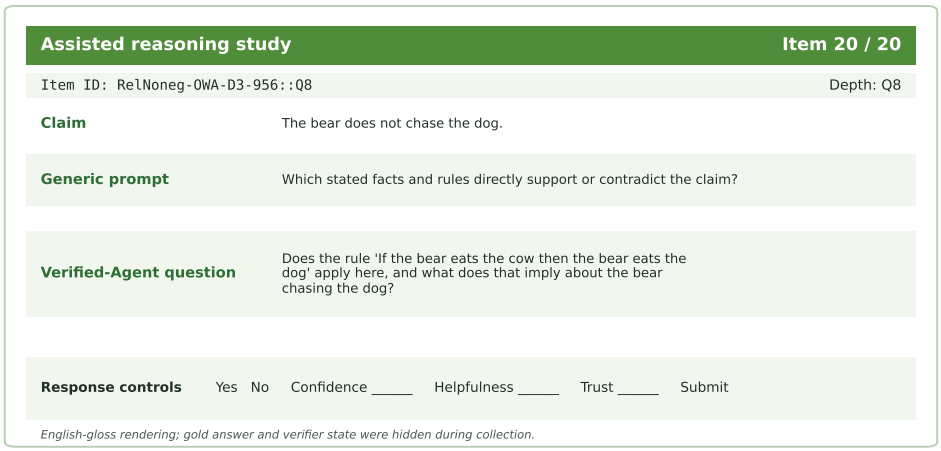


Figure S55. Pilot item 20.